Molecular dynamics of MHC genesis unraveled by sequence analysis of the 1,796,938-bp HLA class I region

Takashi Shiina*, Gen Tamiya*, Akira Oka*, Nobusada Takishima*, Tetsushi Yamagata*, Eri Kikkawa*, Kyoko Iwata*, Maiko Tomizawa*, Noriko Okuaki*, Yuko Kuwano*, Koji Watanabe†, Yasuhito Fukuzumi†, Shoko Itakura†, Chiyo Sugawara†, Ayako Ono†, Masaaki Yamazaki†, Hiroyuki Tashiro†, Asako Ando*, Toshimichi Ikemura‡, Eiichi Soeda§, Minoru Kimura*, Seiamak Bahram¶, and Hidetoshi Inoko*‖

*Department of Genetic Information, Division of Molecular Life Science, Tokai University School of Medicine, Bohseidai, Isehara, Kanagawa 259-1193, Japan; †Bioscience Research Laboratory, Fujiya Co., Ltd., Soya, Hadano, Kanagawa 257-0031, Japan; ‡Department of Evolutionary Genetics, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-0801, Japan; §Tukuba, Life Science Center, The Institute of Physical and Chemical Research (RIKEN), 3–1-1 Koyadai, Yatabe-choh, Tsukuba, Ibaraki 305-0861, Japan; and ¶Centre de Recherche d'Immunologie et d'Hématologie, 4 Rue Kirschleger, 67085 Strasbourg, France

Edited by Johannes van Rood, Leiden University, Leiden, The Netherlands, and approved August 25, 1999 (received for review July 6, 1999)

The intensely studied MHC has become the paradigm for understanding the architectural evolution of vertebrate multigene families. The 4-Mb human MHC (also known as the HLA complex) encodes genes critically involved in the immune response, graft rejection, and disease susceptibility. Here we report the continuous 1,796,938-bp genomic sequence of the HLA class I region, linking genes between *MICB* and *HLA-F*. A total of 127 genes or potentially coding sequences were recognized within the analyzed sequence, establishing a high gene density of one per every 14.1 kb. The identification of 758 microsatellite provides tools for high-resolution mapping of HLA class I-associated disease genes. Most importantly, we establish that the repeated duplication and subsequent diversification of a minimal building block, *MIC-HCGIX-3.8–1-P5-HCGIV-HLA class I-HCGII*, engendered the present-day MHC. That the currently nonessential *HLA-F* and *MICE* genes have acted as progenitors to today's immune-competent *HLA-ABC* and *MICA*/*B* genes provides experimental evidence for evolution by "birth and death," which has general relevance to our understanding of the evolutionary forces driving vertebrate multigene families.

The chromosome 6p21.3-located human MHC is densely packed with genes functioning at key checkpoints in the adaptive immune system (1). Preeminent among these are the antigen-presenting HLA class I and II molecules, which initiate the cell-mediated immune response by displaying antigenic oligopeptides to the $\alpha\beta$ T cell receptor (2). This interaction is central for restraining microbiological invasions, controlling malignant cell proliferation, and governing transplant success. The 4-Mb HLA segment is divided into three regions (from centromere to telomere): class II (1 Mb), class III (1 Mb), and class I (2 Mb) (1). They, by several criteria, occupy a unique position within the human genome, most notably an unusually high gene density of more than 180 genes per 4 Mb, the highest degree of genetic polymorphism ever encountered within the genome (with close to 900 alleles at the eight classical class I and II loci; see http://www.anthonynolan.com/HIG/nomenc.html for regular updates), and allelic and haplotypic association to more than 100 diseases (3).

The telomeric class I region spans 2 Mb from *MICB* to *HLA-F* and is known to contain six expressed HLA class I genes: the three classical (*HLA-A*, *HLA-B*, and *HLA-C*), the three nonclassical (*HLA-E*, *HLA-F*, and *HLA-G*) (1); and the two class I chain-related (*MICA* and *MICB*) (4–6) genes. This region also encompasses 12 HLA class I pseudogenes, truncated, or fragmental genes, (*HLA-X*, *-17*, *-30*, *-L/92*, *-J/59*, *-80*, *-21*, *-K/70*, *-16*, *-H/54*, *-90*, and *-75*) (7) and three class I chain-related pseudogenes (*MICC*, *MICD*, and *MICE*) (4).

Here we present the complete 1,796,938-bp HLA class I sequence solved by shotgun strategy. This continuum links the centromeric *MICB* gene to the telomeric *HLA-F* gene, allows a direct and in-depth analysis of the region with respect to overall structure, gene content, and microsatellite density, and permits us to grasp the complex development of this plastic segment of the vertebrate genome at the molecular level.

Materials and Methods

Yeast Artificial Chromosome (YAC), Bacterial Artificial Chromosome (BAC), P1-derived artificial chromosome (PAC), and Cosmid Clones. Large-insert bacterial clones were identified by PCR-based screening of a human BAC library (Research Genetics, Huntsville, AL) constructed from the B cell line, 978SK (Fig. 3*A*, which is provided on the PNAS web site, www.pnas.org.), and two PAC libraries derived from human lymphocyte DNA (Genome Systems, St. Louis) (Fig. 3*A*), and human male lymphocyte DNA (supplied by Pieter J. de Jong, Roswell Park Cancer Institute, Buffalo, NY) (Fig. 3*A*) (8). PCR screening and physical mapping followed the protocol provided by Research Genetics and Osoegawa *et al.* (8). Two YAC clones, 745D12 with a 590-kb insert and 960H11 with a 1,600-kb insert (Fig. 3*A*) (6, 9), were obtained from the Centre d'Étude du Polymorphisme Humain (Paris) YAC library constructed from the HLA-homozygous B cell line BOLETH (HLA-A2, -B62, -Cw10, -DR4, -DQ8, and -DR53) (10), and one YAC clone (Y109) with a 240-kb insert (Fig. 3*A*) (5) was obtained from a YAC library constructed from the B cell line CGM1 harboring the following HLA haplotypes (HLA-A3, -B8, -Cw-, -DR3, -DQ2, -DR52, -A29, -B14, -Cw-, -DR7, -DQ2, and -DR53) (11). Construction, handling, screening, and mapping of cosmid libraries derived from YAC clones 745D12 and 960H11 in the Super Cos1 cosmid vector (Stratagene) and from YAC Y109 in the pWE15 cosmid vector (Stratagene) have been described (5, 6, 9). Chromosomal mapping and chimerism analysis of these BAC, PAC, and cosmid clones were analyzed by fluorescent *in situ* hybridization as described (9).

Sequencing Strategy, Assembly, and Analyses. Three BACs, eight PACs, and 24 cosmid clones covering the 1.8-Mb segment from the *MICB* to *HLA-F* genes were subjected to nucleotide sequence determination by the shotgun strategy, as originally reported (12). Assembly and database analyses were performed following previously established procedures (6).

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: YAC, yeast artificial chromosome; BAC, bacterial artificial chromosome; PAC, P1-derived artificial chromosome; LTR, long terminal repeat; EST, expressed sequence tag.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. AB014077–AB014088, D84394, AB000882, and AB023048–AB023060.

‖To whom reprint requests should be addressed. E-mail: hinoko@is.icc.u-tokai.ac.jp.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Results

Contig Construction and Sequencing. To establish the genomic sequence of the entire HLA class I region between the *MICB* and *HLA-F* genes, a sequence-ready contig was first constructed by using one BAC-, two PAC-, and three YAC-derived cosmid libraries (5, 6, 9). By screening these libraries with human *Alu*-repeat probes, and gene-specific and sequence-tagged site-specific probes or PCR primers, seven BACs, 39 PACs, and 199 cosmid clones were isolated and assembled into a single contig after Southern hybridizations with clone-derived PCR products and *Eco*RI fragments (ref. 9 and unpublished work). In this manner, 245 densely overlapping clones spanning the HLA class I region between *MICB* and *HLA-F* were identified. Of these, three BACs, eight PACs, and 24 cosmids were selected for sequencing (Fig. 3*A*). Fluorescent *in situ* hybridization confirmed that all clone inserts were derived from chromosome 6, band p21.3 (data not shown). Furthermore, the physical map obtained by this BAC, PAC, and cosmid contig was consistent with that previously constructed on the basis of pulsed field gel electrophoresis analysis using independently isolated YAC clones (13). Altogether, these results suggested that these BAC, PAC, and cosmid clones used for sequencing were devoid of gross deletions, rearrangements, or chimerisms.

After shotgun sequencing, the total length of the contig linking *MICB* to *HLA-F* was established to be 1,796,938 bp (Fig. 3*B*). This result was obtained with a high redundancy of 7.14. All clone overlaps were ascertained at the nucleotide level. This defined length of roughly 1.8 Mb is slightly shorter than the previously predicted 2.0 Mb. The sequence contained precisely 483,365 A (adenine), 410,963 C (cytosine), 411,876 G (guanine), and 490,734 T (thymine), yielding an overall 45.8% G+C content, classifying this DNA segment within the relatively G+C-rich isochore H1 (14), putting it above the class II region that belongs to the G+C-poor isochore L (40%), but below the central class III segment, member of the most G+C-rich isochore H3 (around 53%) (15).

Gene Identification. Homology searches with the entire sequence were carried out against the latest updates of DNA databases by using FASTA, BLASTN, and BLASTX (http://www.ncbi.nlm.nih.gov/BLAST/). Searches for coding regions used the CRM/Grail Grail I, Ia, and II gene finding programs (genscan@gnomic.stanford.edu), the HEXON exon finding program, the GENSCAN gene prediction program (genscan@gnomic.stanford.edu), along with the SwissProt database and Smith-Waterman algorithm. As a result, the HLA class I region was found to include 23 known expressed genes, 12 new expressed genes (previously reported cDNA clones of unknown or ambiguous locations), three possibly expressed sequences (Fig. 3*B* and Table 1), and 22 potentially coding sequences (nearly 100% expressed sequence tag (EST)-matched sequences with exon-intron organization) (Table 2). Thus, a total of 37 new expressed genes or possibly coding sequences were identified. Our analysis also determined the precise location and structure of 30 known pseudogenes. Moreover, 37 new pseudogenes (Table 1) also were revealed during this study. These include one *MIC* (*MICF*), seven *P5*, six *HCGII*, seven *HCGIV*, three *HCGIX*, and 12 other pseudogenes. It must be noted that most of these pseudogenes are members of multigene families restricted to the HLA class I region. In sum, 127 genes were identified within this region, which corresponds to one gene per 14.1 kb. Among these, 60 are expressed genes or potentially coding sequences, corresponding to one expressed gene for every 29.9 kb.

Zooming In on the Segment Between the *S* and *HLA-E* Genes. The 625-kb segment between the *S* and *HLA-E* genes is the least-characterized segment of the class I region because of previous difficulties in clone coverage (9); so far only three genes (*TUBB*, *CAT56*, and *HSR1*) have been mapped in this segment. By genomic sequencing, nine new expressed genes and seven EST-matched sequences were recognized in this segment. This makes the 250 kb between the *PRG1* and *HLA-E* loci a gene-rich segment with 17 genes, including 15 expressed or potentially coding sequences (one expressed gene every 16.7 kb) (Fig. 3*B* and Tables 1 and 2). Novel genes noted in this segment are as follows (from centromere to telomere) (because of space restraints primary references for all these loci could be best accessed through their respective GenBank accession numbers). Transcription factor II H (*TFIIH*) is part of a protein complex involved in both transcription and DNA repair (Y07595); *DDR* encodes a receptor-tyrosine kinase up-regulated by the p53 (U48705) tumor suppressor gene; *PRG1* (*IEX-1*) translates into an early response protein carrying functional binding sites for p53 and NF-κB (X96438); *DBP2* gives rise to a putative nuclear ATP-dependent RNA helicase carrying a DEAH (Asp-Glu-Ala-His) box (AB001601); and *ABC50* encodes an ATP binding cassette protein stimulated by tumor necrosis factor α (AF027302). Finally, the *TC4* gene encoding a ras-like protein is juxtaposed (2.5 kb telomeric) to the *HLA-E* gene. It is noteworthy that all of these (six) newly mapped genes, packed around the *PRG1–HLA-E* segment, are likely to function in the process of DNA repair or cell proliferation, hence they may be possibly involved in the development of some cancers. The *CAT56* (U63336) gene located just telomeric of *ABC50* specifies a proline-rich protein homologous to the Wiskott-Aldrich syndrome gene located on Xp11.23-Xp11.22, mutation of which causes a rare immunodeficiency disorder affecting mainly platelets and lymphocytes. The *HSR1* gene codes for a putative GTP-binding protein and is located 52 kb centromeric to *HLA-E,* not 2 kb as previously reported (indeed, a number of inaccuracies in previously published mapping as well as sequencing data have been discerned; these are available on request from the authors). Finally, no function has yet been ascribed to *KIAA0170* (D79992), *FB19* (Y13247), or *GT260* and *GT478* (X90535 and X90538, respectively).

Repetitive Elements. Analysis of the complete sequence with the REPEATMASKER2 program unveiled the following numbers of repeats: 1,001 *Alus*, 105 *MIRs*, 411 *LINEs* (L1+L2), 290 *LTRs*, and 100 *MERs*. These collectively occupy 43.7% of the class I region, with *Alus* and *LINEs* representing 14.8% and 16.0% or one repeat per 1.8 kb and 3.6 kb, respectively. Although in case of *Alus*, this finding closely matches the theoretical value of 17.9% obtained for a genomic segment harboring a high G+C content of 45.8%, the *LINE* content is remarkably higher than the calculated value of 6.1% (16, 17).

Microsatellites. A total of 758 microsatellite repeats were identified in the 1,796,938-bp genomic sequence. These consist of 203 di-, 139 tri-, 273 tetra-, and 143 penta-nucleotide repeats (Fig. 3*C*), yielding an overall density of one microsatellite per 2.3 kb, significantly higher than the one per 6 kb previously predicted by Beckman and Weber (18). Among the 758 microsatellite identified here, 70 already have been subjected to polymorphism analysis within the Japanese population. As expected, 38 of these 70 microsatellites are quite polymorphic with an average of 8.9 alleles and a 0.66 polymorphism content value (19). As these polymorphic microsatellites are evenly dispersed throughout the class I region, they should serve as much needed genetic markers in linkage and association analysis, enabling investigators to precisely map class I-associated disease susceptibility loci (3).

From the Original Building Block to Today's MHC. Dot matrix analysis using the entire 1,796,938-bp class I sequence versus itself revealed numerous large scale duplications (Fig. 1 *A–C*). These include several prominent homology sections: (*a*) 35-kb downstream segments of both *HLA-B* and *HLA-C* genes as well as

Table 1. Genes identified in the HLA class I region

Location	Name	Orientation	Exons	Homology or prominent features
1-12930	MICB	C	6	MIC family gene
5800-16793	HCGIX-1	+	1	HCGIX family pseudogene
38694-39873	3.8-1.1	C	1	anonymous, 97.8% identity with 3.8-1 mRNA (L29376)
45291-47909	P5-1	C	2	P5 family gene,99.3% identity with P5-1 sequence (Z31714)
48610-49255	HLA-X	+	1	HLA class I fragment
95951-107670	MICA	C	6	MIC family gene
101016-111141	HCGIX-2	+	1	HCGIX family pseudogene
109250-110947	NOB1	?	?	anonymous
116371-163572	NOB3	?	?	anonymous
123014-124273	NOB2	?	?	anonymous
125044-127489	P5-8	C	1	P5 family pseudogene
128797-129729	HLA-17	+	1	HLA class I fragment
144295-144409	DHFRP	+	1	dihydrofolate reductase pseudogene
153008-153388	HCGIV-1	C	1	HCGIV family pseudogene
154115-157398	HLA-B	+	8	HLA class I gene
154602-165657	HCGII-1	+	1	HCGII family pseudogene
229156-230422	RPL3-hom	C	1	ribosomal protein L3 pseudogene
231286-235152	KIAA0055-hom	C	1	KIAA0055 pseudogene, 92.6% identity with KIAA0055 mRNA (D29956)
237231-237465	HCGIV-2	C	1	HCGIV family pseudogene
238645-241966	HLA-C	+	8	HLA class I gene
241967-249784	NOB5	?	?	anonymous
254997-311255	HCGIX-3	+	1	HCGIX family pseudogene
256575-257489	HCGII-2	+	1	HCGII family pseudogene
272884-276165	NOB4	?	?	anonymous
338533-344874	OTF3	+	5	octamer transcription factor 3
345434-350088	SC1	C	3	trans-acting factor
388788-394124	S	+	2	S protein
595113-600934	TFIIH	C	14	transcription/DNA repair factor
609914-620823	DDR	C	17	Receptor tyrosine kinase up-regulated by p53
764754-766101	PRG1	+	2	early response gene
784912-788802	TUBB	C	4	tublin beta protein
794128-809484	KIAA0170	+	14	anonymous, perfect match with KIAA0170 mRNA
811754-812545	RPL7A	+	1	ribosomal protein L7 pseudogene
836314-856186	DBP2	+	20	DEAH box RNA helicase
874002-876075	PROA-hom	C	1	Prothymosin A pseudogene
892085-908911	FB19	+	20	anonymous, perfect match with FB19 mRNA (Y13247)
904870-906088	GT26	C	5	anonymous
907759-908419	0GT478	+	1	anonymous
918004-937913	ABC50	C	24	TNF-alpha stimulated ABC protein
945583-952420	CAT56	C	4	anonymous, proline rich sequence
953147-963430	HSR1	+	12	GTP binding protein
1015100-1019761	HLA-E	C	8	HLA class I gene
1022700-1023350	TC4	C	1	ras-like protein
1052471-1052795	HCGII-3	C	1	HCGII family pseudogene
1089977-1094599	MICC	C	5	MIC family pseudogene
1104371-1105269	HCGII-4	+	1	HCGII family pseudogene
1137844-1138182	HCGII-5	+	1	HCGII family pseudogene
1157277-1158241	HLA-30	C	1	HLA class I fragment
1191469-1191866	GT257	+	1	anonymous
1198912-1199017	CAT75X	C	1	anonymous
1246433-1249666	HLA-L/92	C	8	HLA class I pseudogene
1304526-1324823	ZNF173	+	9	acid finger protein
1348395-1355279	RFB30	+	5	RING finger protein
1403613-1406366	HCGI	+	4	anonymous, 99.2% identity with HCGI mRNA (X81006)
1438955-1442076	HCGV	C	3	anonymous, 99.9% identity with HCGV mRNA (X81003)
1448094-1473892	HTEX4	+	4	anonymous, 99.9% identity with HTEX mRNA (AF032110)
1475444-1476873	HCGVII	C	1	anonymous, 94.5% identity with HTEVII mRNA (X80916)
1496241-1497513	HCGVIII-1	+	1	anonymous, 99.9% identity with HCGVIII mRNA (X92110)
1499717-1503078	HLA-J/59	C	8	HLA class I pseudogene
1503845-1504830	HCGIV-3	+	1	HCGIV family pseudogene
1506404-1508672	P5-2	+	1	P5 family pseudogene
1513265-1513720	3.8-1.2	+	1	3.8-1 pseudogene
1532752-1536915	HCGIX-4	C	1	anonymous, 97.3% identity with HCGIX gene (X92109)
1537217-1539144	MICD	+	4	MIC family pseudogene
1542262-1549351	HCGII-6	C	2	HCGII family pseudogene
1550743-1553088	HLA-80	C	6	HLA class I pseudogene
1554047-1554475	HCGIV-4	+	1	HCGIV family pseudogene
1560012-1561509	P5-3	+	1	P5 family pseudogene
1563922-1567123	HLA-A	C	8	HLA class I gene
1567027-1567711	P5-4	C	1	P5 family pseudogene
1567875-1568764	HCGIV-5	+	1	HCGIV family pseudogene
1575430-1575615	HLA-21	C	1	HLA class I fragment
1579883-1583258	HLA-K/70	C	8	HLA class I pseudogene
1583143-1583780	P5-5	C	1	P5 family pseudogene
1584066-1585125	HCGIV-6	+	1	anonymous, perfect match with HCGIV mRNA (X81005)
1591602-1592504	P5-6	+	1	P5 family pseudogene
1598316-1599481	3.8-1.3	+	1	3.8-1 pseudogene
1607145-1610758	HCGII-7	C	1	anonymous, 98.3% identity with HCGII mRNA (X81001)
1611600-1613141	HLA-16	C	5	HLA class I pseudogene
1616134-1616720	P5-7	+	1	P5 family pseudogene
1618618-1621934	HLA-H/54	C	8	HLA class I pseudogene
1621807-1622448	P5-9	C	1	P5 family pseudogene
1622675-1623571	HCGIV-7	+	1	HCGIV family pseudogene
1630053-1632705	P5-10	+	1	P5 family pseudogene
1636967-1638130	3.8-1.4	+	1	3.8-1 pseudogene
1650270-1652070	MICF	+	1	MIC family pseudogene
1672581-1673860	HCGVIII-2	+	1	HCGVIII pseudogene
1676419-1679795	HLA-G	C	8	HLA class I gene
1679562-1680257	P5-11	C	1	P5 family pseudogene
1680421-1681406	HCGIV-8	+	1	HCGIV family pseudogene
1688940-1691289	P5-12	+	1	P5 family pseudogene
1698793-1702407	HCGII-8	C	1	HCGII family pseudogene
1703504-1704393	RPL7B	+	1	ribosomal protein L7 pseudogene
1704736-1707110	HLA-90	C	6	HLA class I pseudogene
1707410-1709122	HCGIV-9	+	1	HCGIV family pseudogene

(*Table continues on the opposite page.*)

Table 1. (*Continued*)

Location	Name	Orientation	Exons	Homology or prominent features
1713090-1714204	P5-13	+	1	P5 family pseudogene
1714200-1715083	HLA-75	C	3	HLA class I pseudogene
1715846-1716732	HCGIV-10	+	1	HCGIV family pseudogene
1734444-1736782	P5-14	+	1	P5 family pseudogene
1741034-1742203	3.8-1.5	+	1	3.8-1 pseudogene
1754402-1761219	HCGIX-5	C	1	HCGIX family pseudogene
1758278-1765927	MICE	+	6	MIC family pseudogene
1780732-1783774	HLA-F	C	8	HLA class I gene
1785042-1786031	HCGIV-11	+	1	HCGIV family pseudogene
1789663-1791985	P5-15	+	1	P5 family pseudogene

Location is indicated by nucleotide positions numbered from the 5′ end of the *MICB* gene. The color code has been established as follows: orange indicates known expressed genes, green possibly expressed sequences, and red new expressed loci, i.e. for which cDNA clones were previously reported but where physical location was unknown or ambiguous. Regarding pseudogene, black depicts known pseudogenes and blue new ones. Gene orientation from centromere to telomere is shown by the letter C, and + depicts the opposite.

35-kb upstream regions of *MICA* and *MICB* genes display 80% and 85% nucleotide identity, respectively (5, 6), (*b*) around 39 kb upstream of *MICD* and 35 kb 5′ of *MICE* share significant nucleotide identity not only between themselves but also with (>50%) corresponding regions of *MICA* and *MICB* loci (Fig. 1*A* exemplifies the homology between *MICA* and *MICD* upstream segments), and (*c*) the 5′ segments of *MICD* and *MICF* are also homologous to each other (Fig. 1*B*). In sum, the upstream segments of all members of the *MIC* gene family (exception of *MICC*) display sequence homology to each other over distances longer than 15 kb. Interestingly all of these *MIC*-linked homologous segments share a unique mix of genes, all members of several multigene families. These are *HCGIX*, *3.8–1*, *P5*, *HCGIV*, *HLA class I*, and *HCGII* in this order and within the same gene orientation in most cases. These facts strongly imply that successive segmental duplication of this basic unit gave rise to the present human MHC class I system. In this respect, the analysis of the 300-kb telomeric end of the HLA class I region, which links *HLA-J/59* to *HLA-F* and includes *MICD, MICE*, and *MICF*, genes is staggering. Dot matrix analysis revealed this region to be filled with more than 30 pairs of homologous segments varying from 8 to 20 kb (Fig. 1*C*). This striking observation prompted us to look closely at each of these segments. Having done this, we realized that all 11 *HLA* class I genes (*HLA-J/59, -80, -A, -21, -K/70, -16, -H/54, -G, -90, -75*, and *-F*) are not only oriented in the same telomere to centromere orientation, but also display a high degree of homology within 8–20 kb of their upstream sequences. For example, the 20-kb upstream segment of the *HLA-K/70* gene (*HLA-70–P5–5–HCGIV-6–P5–6–3.8–1.3*) shows significant (75.3%) homology to the 20-kb upstream segment of the *HLA-H/54* gene (*HLA-54–P5–9–HCGIV-7–P5–10–3.8–1.4*).

Based on these results, Fig. 2 depicts a model that explains how through seven rounds of successive segmental duplications of the above-mentioned elementary unit, the HLA class I region was shaped. Of the 18 HLA class I genes and six MIC genes localized, 15 HLA (except *HLA-E, - 30*, and *-L/92*) and five MIC (except *MICC*) loci are associated with these conserved shared segments. These generally consist of the *HCGIX, 3.8–1, P5, HCGIV*, and *HCGII* family members in this order and gene orientation. Furthermore, it must be emphasized that the conserved segments around the *MIC* genes always contain the HLA class I genes (*MICB*: *HLA-X; MICA*: *HLA-17* and *HLA-B*; *MICD*: *HLA-J/59*; *MICE*: *HLA-75* and *HLA-90*; and *MICF*: *HLA-H/54* and *HLA-16*). Phylogenetic analysis of the HLA class I and MIC genes elegantly details the probable kinetics of these duplications. As shown in Fig. 4*A* (provided as supplemental data on the PNAS web site, www.pnas.org), the *HLA-F* and *MICE* genes first branched out from major clusters. Thus, *HLA-F* and *MICE*, which are in close proximity to each other at the telomeric end of the HLA class I region, possibly represent ancestral HLA class I and class I chain-related genes. According to this model, duplication of *HLA-F* gave birth to the *MICE* and *HLA-G* genes (from stage I-1 to I-2 in Fig. 2). The basic unit of *MIC–HCGIX–3.8–1–P5–HCGIV–HLA-class I–HCGII* therefore was created. Thereafter, two independent segmental duplications of this elementary unit simultaneously generated the *HLA-A--MICF* (from stage II-1 to II-2 in Fig. 2) as well as the *MICA–HLA-B* segments (from stage I-2 to II-3 in Fig. 2). The latter gave birth to both *MICB* and *HLA-C* genes after a single duplication (Fig. 2) (4, 5). The next partial segmental duplication gave rise to the *HLA-80–HCGIV-4* segment from the *HLA-A–HCGIV-5* segment (from stage III-1 to III-2 in Fig. 2). In a similar way, as illustrated in Fig. 2, four subsequent segmental duplications (stages IV to VII) including partial ones, led to the present-day gene organization of the HLA class I region. The order of the generation of each gene predicted by this model is supported by dendrograms of the *HLA-class I, MIC, HCGIX, P5, 3.8–1*, and *HCGIV* family members (Fig. 4 *A–F*). This model also is supported by the

Table 2. EST-matched sequences in the HLA class I region

Locus	Location	Orient.	Accession	Tissue derivation	Prominent features
EST1	97743-115640	+	N78217	fetal liver spleen	5' similar to contains Alu repetitive element
EST2	312427-313743	+	AA044209	uterus	5' similar to SW:POL_GALV P21414 POL POLYPROTEIN
EST3	341500-353833	+	AA100400	neuronal precursor	anonymous
EST4	381441-366622	C	AA045448	uterus	anonymous
EST5	377547-374854	C	D29144	keratinocyte	anonymous
EST6	573080-569839	C	H09683	infant brain	5' similar to SP:SYV_RAT Q04462 Valyl-tRNA Synthetase
EST7	673336-674718	+	AA491398	germinal center B cells	anonymous
EST8	753747-768671	+	AA199684	neurons	anonymous
EST9	805385-808319	+	AA126176	neuroepithelium	anonymous
EST10	816973-819974	+	H20209	brain	anonymous
EST11	852220-851709	C	H50176	brain	anonymous
EST12	878537-869997	C	AA115720	colon	anonymous
EST13	1151224-1149530	C	R96980	fetal liver spleen	anonymous
EST14	1154545-1154425	C	AA250908	germinal center B cells	similar to TR:G238611 XNF7=Zinc Finger Nuclear Phosphoprotein
EST15	1158519-1157314	C	N77592	multiple sclerosis lesions	anonymous
EST16	1170029-1204043	+	AA287977	germinal center B cells	anonymous
EST17	1204047-1204823	+	AA281507	germinal center B cells	anonymous
EST18	1338164-1339517	C	H80869	fetal liver spleen	anonymous
EST19	1355812-1356339	+	Y13935	unknown	anonymous
EST20	1357654-1358280	C	AF075031	placenta	anonymous
EST21	1437482-1437948	C	AI168183	pooled organ	similar to TR:O00475 O00475 BUTYROPHILIN
EST22	1658785-1660014	+	W65500	fetal heart	anonymous

GENETICS

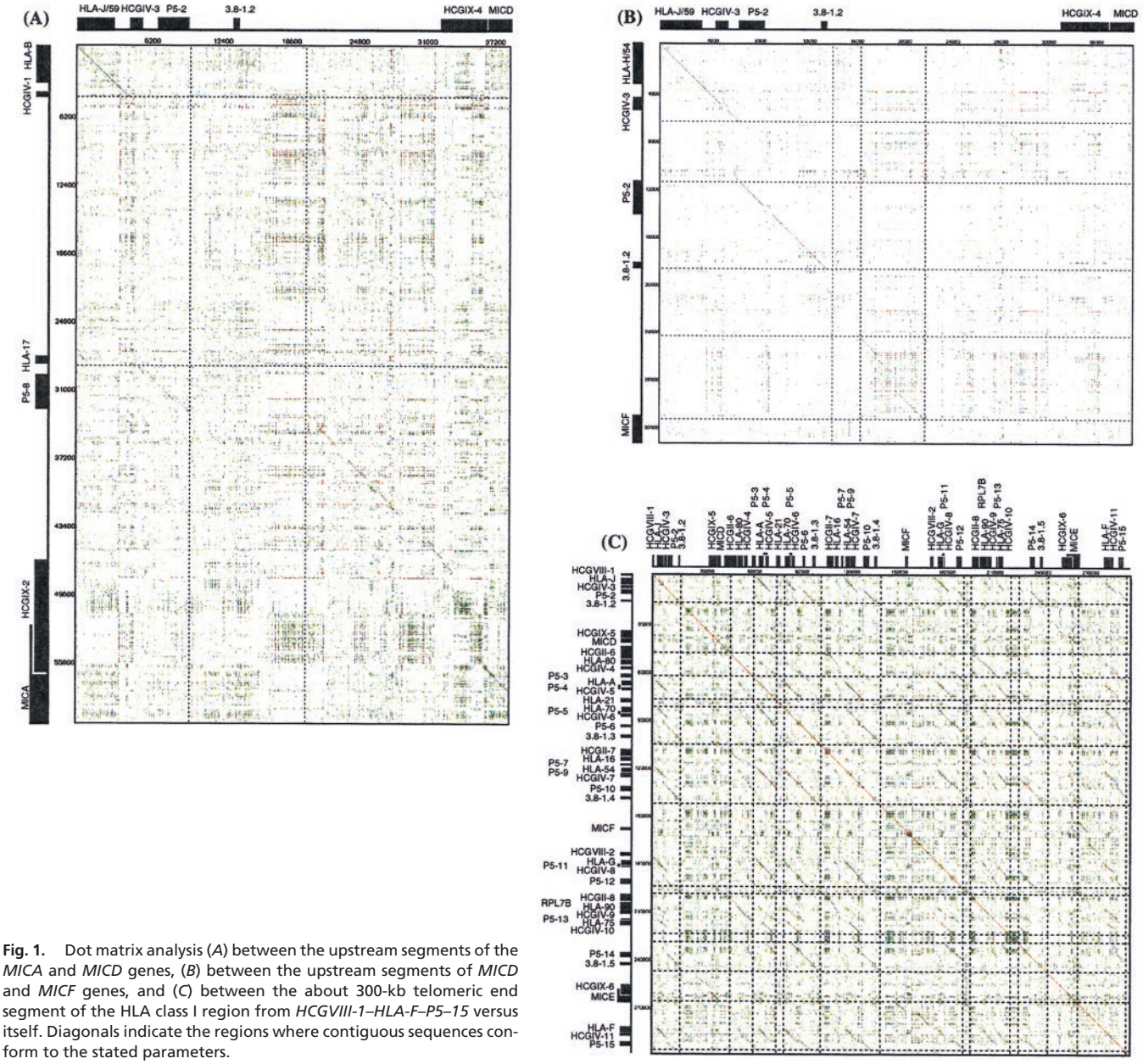

Fig. 1. Dot matrix analysis (*A*) between the upstream segments of the *MICA* and *MICD* genes, (*B*) between the upstream segments of *MICD* and *MICF* genes, and (*C*) between the about 300-kb telomeric end segment of the HLA class I region from *HCGVIII-1–HLA-F–P5–15* versus itself. Diagonals indicate the regions where contiguous sequences conform to the stated parameters.

chronological order of the generation of the *Alu*, *LINE*, and *LTR* (long terminal repeat) subfamily members linked to respective repeat units (data not shown), which serve as a molecular clock to define the generation time of genomic segments of interest (for example, *Alu J* is older than *Alu S*).

Discussion

The present work has established the nucleotide sequence of the human MHC class I region. Investigation of the obtained sequence unveiled the presence and exact location of a large number of genes, among which 37 were novel. This, combined with the identification of numerous highly polymorphic microsatellite repeats, will dramatically ease positional cloning approaches aimed at defining the molecular basis of HLA class I-associated diseases.

The molecular path taken across vertebrate evolution leading to the present-day MHC has been strongly debated for a number of years. Cloning and sequence analysis of MHC genes from various species, despite having brought a wealth of information, has failed to alleviate the confusion. The sequence data presented here resolves, at molecular level, this long controversy. It seems that *HLA-F* was the ancestor MHC class I gene that upon duplication gave rise to *MICE* and *HLA-G*. This contention is corroborated by substantial physical evidence: the phylogenetic tree analysis, the short genomic distance between *HLA-G*, *MICE*, and *HLA-F*, and the fact that the present-day *MICE* is the "least problematic" *MIC* pseudogene (in contrast to *MICC*, *MICD*, and *MICF*, which lack one or several exons/introns, *MICE* shows an intact genomic organization despite several nucleotidic defects). Moreover, no *Alu*, *LINE*, or *LTR* elements are linked to the *MICE–HLA-F* unit, whereas all the other MIC-HLA class I units share the oldest *Alu*, *LINE*, and *LTR* subfamily members. Therefore, these repetitive elements were inserted into these repeated basic units after duplication of *MICE–HLA-F* unit (after stage III-1 in Fig. 2). In fact, the ages of the *Alu*, *LINE*, and *LTR* subfamily members identified within the MIC–HLA class I basic units created after stage III-1 follow the order of the generation of these basic units predicted from our evolutionary model in Fig. 3. This primordial building block was already flanked by a number of structural units that upon a series of duplication and diversification gave birth to present-day MHC. This fact is evident in comparison of the upstream sequences

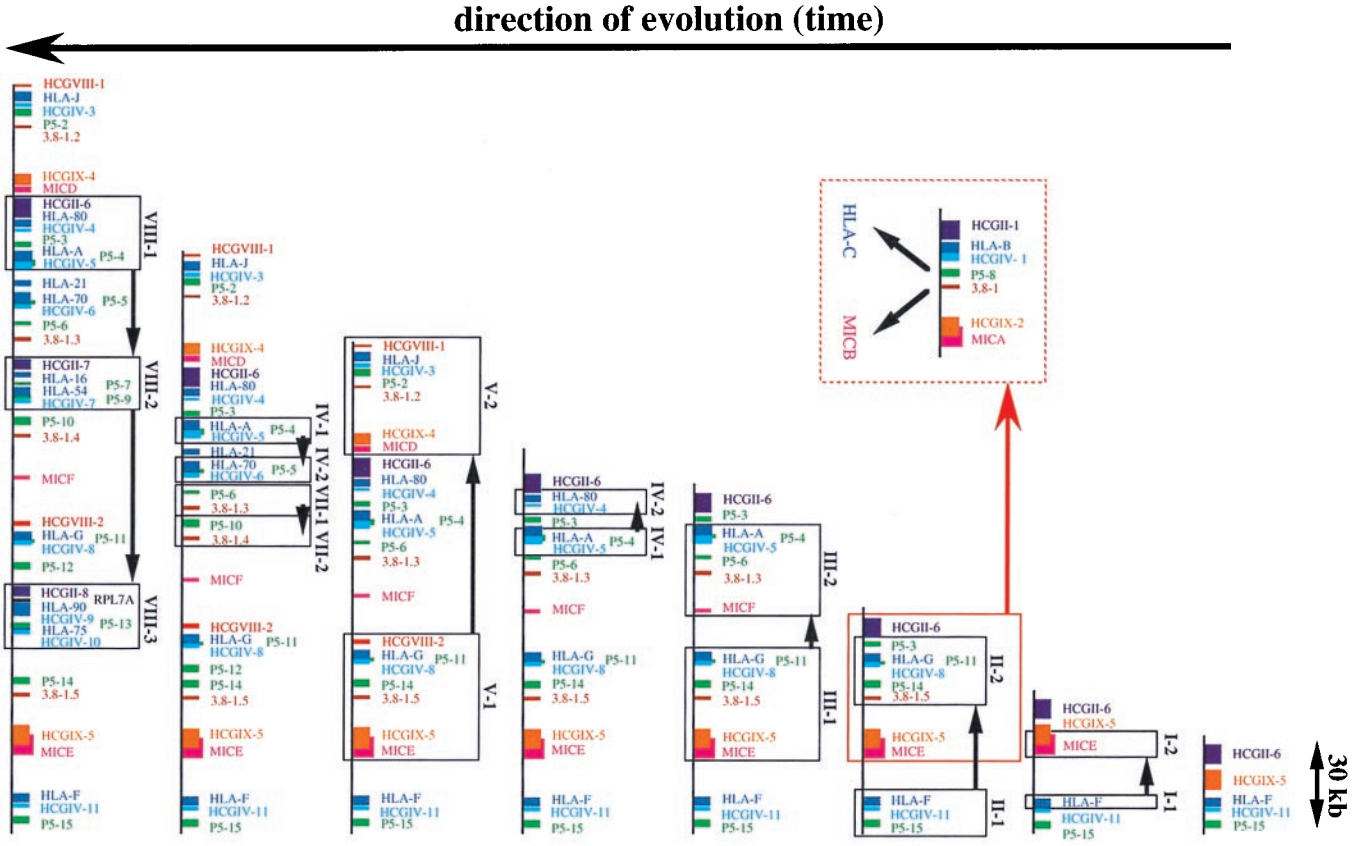

Fig. 2. A model that explains how the HLA class I region was shaped by seven rounds of successive segmental duplications of a basic unit, *MIC–HCGIX–3.8–1–P5–HCGIV–HLA class I*. Arrows indicate the evolutionary path of the class I region paved by segmental duplications.

of *MIC* and *HLA* genes. This scenario also explains the large number of pseudogenes present within the class I region. Indeed, among the 127 genes or gene candidates, more than half (67 genes) are pseudogenes. This accumulation of pseudogenes is unprecedented in the neighboring HLA class II and class III regions, paralleled by their respective "quiescent" evolutionary journey and is possibly reminiscent of the process of "birth and death process" (20) operational at the outset of MHC genesis. The large number of retrotransposon elements encountered throughout the class I region and some of the very structural genes part of the basic replicated unit (i.e., 3.8–1, P5, and HCGIV), which harbor LTR-like sequences, may have facilitated the entire process (21).

Among the 36 new expressed genes or potentially coding sequences around the *S* and *HLA-E* gene segment, six [*TFIIH, DDR, PRG1 (IEX-1), DBP2, ABC50*, and *TC4*] are involved in the process of DNA repair or cell proliferation. In this respect, the fact that numerous cancer cells exhibit decreased expression of HLA class I antigens as a result of deletion or loss of heterozygosity (LOH) of the HLA class I region is quite intriguing (22). This decline in the expression of HLA class I antigens is believed to provide an escape mechanism from the host immune system. Microsatellite alleles identified here provide the means to conclusively test this possibility, by narrowing candidate regions through the definition of LOH boundaries and subsequent investigation of tumor-associated mutations.

In sum, this highly accurate genomic sequence derived from a carefully tilled contig provides through a molecular blueprint of MHC structure, not only a platform for positional cloning experiments, but also a detailed case of evolution by "birth and death" relevant to our understanding of vertebrate genomics.

Grants from the Japan Science and Technology Corporation, an arm of the Science and Technology Agency; the Ministry of Education, Science, Sports and Culture, Japan, and the Tokai University School of Medicine supported this work. S.B. and H.I. are grateful for a French-Japanese collaborative grant awarded jointly by Institut National de la Santé et de la Recherche Médicale and Japan Society for the Promotion of Science. S.B. acknowledges the Fondation pour la Recherche Médicale (ARS2000) and the Association pour la Recherche sur le Cancer for additional support.

1. Campbell, R. D. & Trowsdale, J. (1997) *Immunol. Today* **18**, Suppl.
2. Bjorkman, P. J. & Parham, P. (1990) *Annu. Rev. Biochem.* **59,** 253–288.
3. Svejgaard, A., Buus, S. & Fugger, L., eds. (1996) *HLA and Disease: The Molecular Basis* (Alfred Benzon Symposium 40) (Munksgaard International Publishers, Copenhagen).
4. Bahram, S., Bresnahan, M., Geraghty, D. E. & Spies, T. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 6259–6263.
5. Mizuki, N., Ando, H., Kimura, M., Ohno, S., Miyata, S., Yamazaki, M., Tashiro, H., Watanabe, K., Ono, A., Taguchi, S., *et al.* (1997) *Genomics* **42,** 55–66.
6. Shiina, T., Tamiya, G., Oka, A., Yamagata, T., Yamagata, N., Kikkawa, E., Goto, K., Mizuki, N., Watanabe, K., Fukuzumi, Y., *et al.* (1998) *Genomics* **47,** 372–382.
7. Geraghty, D. E., Koller, B. H., Pei, J. & Hansen, J. A. (1992) *J. Immunol.* **149,** 1947–1956.
8. Osoegawa, K., Susukida, R., Okano, S., Kudoh, J., Minoshima, S., Shimizu, N., de Jong, P., Groet, J., Ives, J., Lehrach, H., *et al.* (1996) *Genomics* **32,** 375–387.
9. Shiina, T., Kikkawa, E., Yamagata, T., Saito, W., Tamiya, G., Oka, A., Watanabe, K., Yamazaki, M., Tashiro, H., Okumura, K., *et al.* (1998) *Immunogenetics* **48,** 402–407.
10. Albertsen, H. M., Abderrahim, H., Cann, H. M., Dausset, J., Paslier, D. L. & Cohen, D. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 4256–4260.
11. Imai, T. & Olson, M. V. (1990) *Genomics* **8,** 297–303.
12. Deininger, P. L. (1983) *Anal. Biochem.* **129,** 216–223.
13. Abderrahim, H., Sambucy, J. L., Iris, F., Ougen, P., Billault, A., Chumakov, I. M., Dausset, J., Cohen, D. & Le Paslier, D. (1994) *Genomics* **23,** 520–527.
14. Bernardi, G. (1995) *Annu. Rev. Genet.* **29,** 443–476.
15. Tenzen, T., Yamagata, T., Fukagawa, T., Sugaya, K., Ando, A., Inoko, H., Gojobori, T., Fujiyama, A., Okumura, K. & Ikemura, T., *et al.* (1997) *Mol. Cell. Biol.* **17,** 4043–4050.
16. Kazazian, H. H. & Moran, J. V. (1998) *Nat. Genet.* **19,** 19–24.
17. Schmid, T. (1996) *Prog. Nucleic Acid Res. Mol. Biol.* **53,** 283–319.
18. Beckman, J. S. & Weber, J. L. (1992) *Genomics* **12,** 627–631.
19. Tamiya, G., Shiina, T., Oka, A., Tomizawa, M., Ota, M., Katsuyama, Y., Yoshitome, M., Makino, S., Kimura, M., Inoko, H., *et al.* (1999) *Tissue Antigens* **54,** 221–228.
20. Nei, M. (1969) *Nature (London)* **221,** 40–42.
21. Van Arsdell, S. W., Denison, R. A., Bernstein, L. B., Weiner, A. M., Manser, T. & Gesteland, R. F. (1981) *Cell* **26,** 11–17.
22. Garrido, F., Ruiz-Cabello, F., Cabrera, T., Perez-Villar, J. J., Lopez-Botet, M., Duggan-Keen, M. & Stern, P. L. (1997) *Immunol. Today* **18,** 89–95.