Genomic Sequence Analysis of the MHC Class I G/F Segment in Common Marmoset (Callithrix jacchus)

Azumi Kono, Markus Brameier, Christian Roos, Shingo Suzuki, Atsuko Shigenari, Yoshie Kametani, Kazutaka Kitaura, Takaji Matsutani, Ryuji Suzuki, Hidetoshi Inoko, Lutz Walter and Takashi Shiina

*J Immunol* 2014; 192:3239-3246; Prepublished online 5 March 2014;
doi: 10.4049/jimmunol.1302745
http://www.jimmunol.org/content/192/7/3239

**Supplementary Material**
http://www.jimmunol.org/content/suppl/2014/03/05/jimmunol.1302745.DCSupplemental

**References**
This article cites 28 articles, 6 of which you can access for free at:
http://www.jimmunol.org/content/192/7/3239.full#ref-list-1

**Subscription**
Information about subscribing to *The Journal of Immunology* is online at:
http://jimmunol.org/subscription

**Permissions**
Submit copyright permission requests at:
http://www.aai.org/About/Publications/JI/copyright.html

**Email Alerts**
Receive free email-alerts when new articles cite this article. Sign up at:
http://jimmunol.org/alerts
Genomic Sequence Analysis of the MHC Class I G/F Segment in Common Marmoset (Callithrix jacchus)

Azumi Kono,* Markus Brameyer, † Christian Roos, ‡,‡ Shingo Suzuki,* Atsuko Shigenari,* Yoshie Kametani, § Kazutaka Kitaura, ‡ Takaji Matsutani, ‡ Ryuji Suzuki, ‡ Hidetoshi Inoko,* Lutz Walter, †,‡,§ and Takashi Shiina*

The common marmoset (Callithrix jacchus) is a New World monkey that is used frequently as a model for various human diseases. However, detailed knowledge about the MHC is still lacking. In this study, we sequenced and annotated a total of 854 kb of the common marmoset MHC region that corresponds to the HLA-A/G/F segment (Caja-G/F) between the Caja-G1 and RNF39 genes. The sequenced region contains 19 MHC class I genes, of which 14 are of the MHC-G (Caja-G) type and 5 are of the MHC-F (Caja-F) type. Six putatively functional Caja-G and Caja-F genes (Caja-G1, Caja-G3, Caja-G7, Caja-G12, Caja-G13, and Caja-F4), 13 pseudogenes related either to Caja-G or Caja-F, three non-MHC genes (ZNRD1, PPP1R11, and RNF39), two miscRNA genes (ZNRD1-AS1 and HCG8), and one non-MHC pseudogene (ETF1P1) were identified. Phylogenetic analysis suggests segmental duplications of units consisting of basically five (four Caja-G and one Caja-F) MHC class I genes, with subsequent expansion/deletion of genes. A similar genomic organization of the Caja-G/F segment has not been observed in catarrhine primates, indicating that this genomic segment was formed in New World monkeys after the split of New World and Old World monkeys. The Journal of Immunology, 2014, 192: 3239–3246.

Common marmosets (Callithrix jacchus) are small New World monkeys that have gained much attention and importance as experimental animals because of their smaller body size, easier handling, shorter generation time, and lower costs compared with the more commonly studied Old World monkeys, such as the rhesus (Macaca mulatta) and cynomolgus (Macaca fascicularis) macaques. Marmosets are used in different biomedical research fields, such as Parkinson’s disease (1), drug toxicology (2, 3), transplantation (4), transgenic techniques and biomedical research fields, such as Parkinson’s disease (1), drug toxicology (2, 3), transplantation (4), transgenic techniques and stem cell research (5, 6), immunity and infectious diseases (7–9), and autoimmune diseases (10). Many of these diseases are essentially influenced by polymorphisms of the MHC region (11).

Hence, detailed knowledge of the experimental animals’ MHC genomic region is a prerequisite to understand the role of MHC genes in these diseases and to refine the animal model.

Detailed knowledge of the marmoset MHC (Caja) is only starting to emerge; because of the duplicated nature of this region, only fragmented sequences are available from the C. jacchus whole-genome shotgun-sequencing draft assembly. The overall structure of the MHC class I region is largely similar to the HLA, common chimpanzee (Patt), and rhesus macaque (Mamu) corresponding regions (12). The Caja class I region is divided into three segments: HLA-B/C corresponding Caja-B segment, HLA-E corresponding Caja-E segment, and HLA-A/G/F corresponding Caja-G/F segment. From the MHC-B/C segment, which has been under permanent selective pressure in the evolution of primates, a genomic sequence of 1179 kb was determined that includes nine duplicated Caja-B genes (12). In contrast, although four to seven Caja-G-like alleles were detected in each common marmoset from a recent population study (13), the genomic structure of the Caja-G/F segment between GABBR1 and ZNRD1 is still not solved, and 54 alleles of Caja-G and Caja-E are currently listed in the Immuno Polymorphism Database (IPD)-MHC database (14).

In this study, we used seven bacterial artificial chromosome (BAC) clones to determine and annotate a total 854-kb genomic sequence between the Caja-G1 and RNF39 genes, including the Caja-G/F segment, and we characterized Caja-G and Caja-F (Caja-G/F) genes and the evolutionary process of the Caja-G/F segment by phylogenetic analysis.

Materials and Methods

Construction of a contiguous map of overlapping BAC clones of the Caja-G/F segment

BAC library CHORI-259 (constructed from kidney cells obtained from a male common marmoset) was obtained from the BACPAC Resource Center at the Children’s Hospital Oakland Research Institute (Oakland, CA: http://bapac.chori.org/home.htm). Hybridization screening was performed following the recommended protocols. As probes for BAC library screening, we used PCR products of 800–900 bp from the common

---

*Division of Basic Medical Science and Molecular Medicine, Department of Molecular Life Science, Tokai University School of Medicine, Isehara, Kanagawa 259-1143, Japan; †Primate Genetics Laboratory, German Primate Center, Leibniz Institute for Primate Research, 37077 Göttingen, Germany; ‡Gene Bank of Primates, German Primate Center, Leibniz Institute for Primate Research, 37077 Göttingen, Germany; §Division of Basic Medical Science and Molecular Medicine, Department of Immunology, Tokai University School of Medicine, Isehara, Kanagawa 259-1143, Japan; and ‡Department of Rheumatology and Clinical Immunology, Clinical Immunology Research Center for Allergy and Rheumatology, Sagamihara, Kanagawa 228-0815, Japan

Received for publication October 11, 2013. Accepted for publication January 17, 2014.

This work was supported by Scientific Research on Science and Technology of Japan and a Grant-in-Aid for Scientific Research (B) (24300155) from the Japan Society for the Promotion of Science, the program “Pakt für Forschung und Innovation” Grant “Biodiversity” of the Leibniz Society, and institutional support from the German Primate Center.

Address correspondence and reprint requests to Dr. Takashi Shiina, Department of Molecular Life Science, Division of Basic Medical Science and Molecular Medicine, Tokai University School of Medicine, 143 Shimokasuya, Isehara, Kanagawa 259-1143, Japan. E-mail address: shiina@is.isc.cc.tokai.ac.jp

The online version of this article contains supplemental material.

Abbreviations used in this article: BAC, bacterial artificial chromosome; BLAST, basic local alignment search tool; BLAT, BLAST-like alignment tool; contig, contiguous; IPD, Immuno Polymorphism Database; LINE, long interspersed nuclear element; LTR, long terminal repeat; MID, multiple identifier; NJ, neighbor joining; ver., version.

Copyright © 2014 by The American Association of Immunologists, Inc. 0022-1767/14/$16.00

www.jimmunol.org/cgi/doi/10.4049/jimmunol.1302745

The Journal of Immunology
marmoset MHC-I genes (exons 2–4), as well as one unique microRNA, ZNRD1-ASI (Supplemental Table I). Positive BAC clones were ordered by comparison of EcoRI and HindIII restriction fragments and PCR-based mapping, with 11 locus-specific primer sets obtained by BAC-end sequencing (Supplemental Table I). All PCR reactions were performed using Ex Taq DNA polymerase (TaKaRa, Shiga, Japan) and the GeneAmp PCR System 9700 (Life Technologies, Carlsbad, CA).

Nucleotide sequencing

BAC clones 205N13, 62H24, 195P8, 88K6, 246G18, 41G14, and 211G14 were subjected to nucleotide sequence determination by combination of massively parallel pyrosequencing (15) and/or bidirectional shotgun sequencing (16). Pyrosequencing was performed following the recommended protocol for the Roche 454 GS Junior Bench Top System (Roche, Basel, Switzerland). Briefly, titanium rapid libraries of fragmented DNA linked to AMPure beads (Beckman Coulter Genomics, Danvers, MA) were prepared for the Roche GS Junior Bench Top System by nebulization, fragment end repair, and multiple identifier (MID)-labeled adaptor ligation and were subjected to emulsion PCR and emulsion breaking, which were performed according to the manufacturer’s protocol (Roche) (17). After the emulsion-breaking step, the beads carrying the ssDNA templates were enriched and counted, and 0.5 million beads were deposited into a PicoTiterPlate to obtain sequence reads (17). After the sequencing run, image processing, signal correction, and base-calling were performed using GS Run Processor version 2.5 (Roche), with full processing for shotgun or paired-end filter analysis. Quality-filter sequence reads that passed the assembler software (single SFF file) were binned on the basis of the MID labels into four separate sequence SFF files using SFF file software (Roche). These files were further quality trimmed to remove poor sequence at the end of the reads with quality values < 20. The trimmed and MID-labeled sequence reads were assembled at 95% matched parameters using GS De Novo Assembler ver. 2.5 (Roche).

GS Junior sequencing also was performed with cDNA from various organs derived from a single common marmoset. Primers and sequencing were essentially as described by O’Leary et al. (18). Obtained sequences were compared by basic local alignment search tool (BLAST) search with Caja-G and Caja-E CDNA sequences available from public databases (IPD), as well as with Caja-B1, Caja-B3, Caja-B4, Caja-B6, and Caja-B7 sequences described by Shinya et al. (12). Sequence reads from individual transcripts were counted, and the frequencies of reads matching a database sequence were calculated. Because Caja-B sequences are very similar, we considered only those sequencing reads that were identical to reported Caja-B sequences.

Shotgun sequencing also was performed by the cycle-sequencing method using AmpliTaq DNA Polymerase FS fluorescein labeled BigDye terminators in the GeneAmp PCR System 9700 (Life Technologies). A 3130xl Genetic Analyzer was used for automated Sanger sequencing (Life Technologies). Individual sequences were minimally edited to remove vector sequences and were assembled into Sequencher ver. 5.0.1 (Gene Codes, Ann Arbor, MI), along with the assembled consensus sequences obtained by the next-generation sequencing method. Remaining gaps or ambiguous nucleotides were determined by the direct sequencing of PCR products obtained with appropriate PCR primers or by nucleotide sequence determination of shotgun clones.

Sequence analysis and annotation

General sequence analysis was performed using GENETYX-MAC software ver. 12.2.7 (Genetyx, Tokyo, Japan). Nucleotide similarities between sequences were calculated by Sequeacher; those with nucleotide sequences in GenBank/European Molecular Biology Laboratory/DNA Data Base in Japan were searched by the BLAST program (http://www.ncbi.nlm.nih.gov/BLAST/), whereas those with C. jacchus whole-genome shotgun-sequencing draft assembly (WUGSC 3.2 (GCA_00004665.1)) generated by the Washington University Genome Sequencing Center (St. Louis, MO) and the Baylor College of Medicine (Houston, TX) were searched with BLAST-like alignement tool (BLAT) (http://genome.ucsc.edu/cgi-bin/hgBlat?command=start). Prediction of coding sequences was performed using GENSCAN (http://genes.mit.edu/GENSCAN.html). Identification and classification of repeat sequences were performed using RepeatMasker2 (http://repeatmasker.genome.washington.edu/).

Phylogenetic analysis

Multiple-sequence alignment was created using the ClustalW Sequence Alignment program of Molecular Evolution Genetics Analysis software 5 (http://www.megasoftware.net/) (19). Phylogenetic trees of the MHC-I genes were constructed by the neighbor-joining (NJ) method (Molecular Evolution Genetics Analysis software 5) (20), using genomic sequences of exons 3–8 (alignment length: 1539 bp, excluding gap sites). NJ trees were constructed by the Maximum Composite Likelihood model and assessed using 10,000 bootstrap replicates. We used the following MHC-I sequences (DNA accession numbers) for phylogenetic analyses: HLA-A (NG_029217); HLA-B (NG_023187); HLA-C (NG_029422); HLA-E and HLA-F (NT_113891); HLA-G (NG_029038); Patr-A, Patr-B, Patr-C, Patr-E, Patr-F, and Patr-G (BA000041); Mamu-A1 (AB128833); Mamu-B1 (AB128860); Mamu-E (AB128840); Mamu-F (AB128841); Mamu-GA3 (AB128837); Mamu-G5 (AB128049); and Caja-B1, Caja-B3, Caja-B4, Caja-B6, and Caja-B7 (AB600201, AB600202).

Results

Construction of a BAC contiguous map covering the Caja-G/F segment

A total of 22 BAC clones, covering the genomic segment between the GABABR1 and RNF39 genes, was isolated by hybridization screening and assembled into three clusters (tentatively named contigs 1–3) based on comparison of restriction fragments and PCR-based mapping (Fig. 1A). In this process, 205N13 overlapping clone 171K8 was thought to be suitable for genomic sequencing, but we excluded this clone from the following sequencing process because the draft sequence was already released (accession number: GL286223). Therefore BAC clones 205N13, 62H24, 195P8, 88K6, 246G18, 41G14, and 211G14 were subjected to complete sequencing using a combination of massively parallel pyrosequencing and bidirectional shotgun-sequencing methods (Fig. 1A).

Genomic sequence information of the Caja-G/F segment

Nucleotide sequencing of the seven selected BAC clones was performed independently with the Roche GS Junior System. Table I lists all relevant parameters obtained from Roche GS Junior sequencing. After shotgun sequencing, gap filling, and validation of the assembled sequences, we determined the complete genomic sequences ranging between 183,844 bp (88K6) and 212,178 bp (205N13) (Table I).

Overlaps between all BAC clones were ascertained at the sequence level. Four overlaps between the BAC clones are present, and two overlaps (62H24/195P8 and 195P8/88K6) are not identical: the 67-kb overlap between 62H24 and 195P8 has 99.0% nucleotide identity, and the 124-kb overlap between BAC clones 195P8 and 88K6 shows 99.6% nucleotide identity. Therefore, these data support allelic relationships of all junctions. Finally, we constructed the 470,958-bp contiguous sequence using 274,255 bp of 205N13 and 62H24, 12,859 bp of 195P8, and 183,844 bp of 88K6. Contig 2 contains 191,375 bp derived from BAC clones 246G18 and 41G14 and contig 3 contains 192,325-bp sequence derived from BAC clone 211G14. In total, a 845,658-bp genomic sequence (contig 1: 470,958 bp, contig 2: 191,375 bp, and contig 3: 192,325 bp; accession numbers AB809558–AB809560) in the DNA Data Base in Japan, European Molecular Biology Laboratory, and GenBank nucleotide sequence databases [http://www.ncbi.nlm.nih.gov/genbank/] was determined.

Comparison of the Caja-G/F segment with C. jaccus draft assembly WUGSC 3.2

To compare the Caja-G/F segment with the C. jaccus draft assembly, we performed BLAT searches of our three clusters with WUGSC 3.2 sequence data using a sliding window of 25 kb. A high degree of sequence identity, ranging between 97.6 and 100%, was found for 150 kb of contig 1, 50 kb of contig 2, and 92 kb of contig 3, with draft sequences of Chr.4_GL285015, Chr.4_GL285016, and Chr.4_GL285017 scaffolds (Table II). The other sequences from the clusters show high similarities with some short parts of
draft sequences that are derived from chr. 4 or other or unknown chromosomal locations.

**Genomic structure of the Caja-G/F segment**

The GC content of the Caja-G/F segment is almost the same as in other primate MHC-A/G/F segments, ranging from 43.9% in Caja to 45.2% in HLA (Table III). An analysis of the segment using the RepeatMasker 2 program unveiled the following frequencies of interspersed repeats: 5.1% short interspersed nuclear elements (Alus + mammalian interspersed repeats [MIRs]), 31.7% long interspersed nuclear elements (LINEs) (LINE1 + LINE2 + L3/CR1), 13.0% long terminal repeat (LTR) elements, and 2.6% DNA elements. These repeats collectively occupied 53.7% of the Caja-G/F segment. The densities of the LTR and DNA elements of the Caja-G/F segment are much lower than for other simian MHC-A/G/F segments, but the LINE density of the segment is much higher than for other simian MHC-A/G/F segments (Table III).

The 854-kb genomic sequence stretching from Caja-G1 to RNF39 was subjected to gene-identification analysis using BLAST and GENSCAN, which revealed the presence of 25 genes within this segment (Fig. 1B, Table IV). Among them are 19 Caja-G/F genes—6 putatively functional Caja-G/F genes (Caja-G1, Caja-G3, Caja-G7, Caja-G12, Caja-G13, and Caja-F4), 8 Caja-G/F pseudogenes along with frequently observed nonsense mutations and/or indels on their coding regions (Caja-G2, Caja-G6, Caja-G8, Caja-G11, Caja-G14, Caja-F1, Caja-F2, and Caja-F3), and 5 truncated-type Caja-G pseudogenes with deletion of exons 1 and 2 (Caja-G4, Caja-G5, Caja-G9, and Caja-G10) or exons 6 and 7 (Caja-F3); 3 non-MHC genes (ZNRD1, PPP1R11, and RNF39); 2 miscRNA genes (ZNRD1-AS1 and HCG8); and 1 non-MHC pseudogene (ETF1P1) (Fig. 1B, Table IV). Nucleotide similarities on coding sequences of the three non-MHC class I genes with their human orthologs were 95%, on average, and ranged from 94% (RNF39) to 96% (PPP1R11) (Table IV). Finally, sequences showing similarity to common marsupial MHC genes (MHC1 and MHC2 genes (12)) were not observed in this segment.

**Structural characteristics of the Caja-G/F genes**

Six Caja-G/F genes (Caja-G1, Caja-G3, Caja-G7, Caja-G12, Caja-G13, and Caja-F4) have open reading frames coding for 360 (Caja-G4) to 365 (Caja-G1) amino acids, similar to the expressed
MHC-I genes of other primates. The transcriptional orientation of the genes is the same as for MHC-A, MHC-G, and MHC-F genes in HLA, Patr, and Manu. Nucleotide and amino acid similarities of the coding sequences among the six Caja-G/F genes range from 74.3 to 96.9% and 74.2 to 93.7%, respectively, with MHC05027, respectively; are identical to Caja-G13 (IPD Accession Number: MHC04993), Caja-G*07:01:01 (MHC04991), and Caja-G*20:01 (MHC04982), with one nucleotide differences, respectively. The formal allele names of Caja-G3 and Caja-G7 were assigned to Caja-G*17:03 and Caja-G*08:06, respectively. In contrast, Caja class I sequences showing high similarities (>85%) to Caja-F4 were not observed in known Caja class I sequences.

All of the deduced Caja-G/F amino acid sequences contain the highly conserved and structurally essential amino acid residues, such as N-glycosylation sites (amino acid positions 86–88), four cysteine residues involved in disulfide bonding (positions 101, 164, 203, and 259) and CD8 binding, as well as most of the α-chain and β2 microglobulin contact sites, compared with the three-dimensional structure of HLA-A2 molecule (21, 22) (Supplemental Fig. 1). In addition, α helices, β sheets, and transmembrane structures of all six Caja-G/F sequences are similar to other MHC-A class I proteins (data not shown). Therefore, the six Caja-G/F amino acid

differences, respectively. The formal allele names of Caja-G3 and Caja-G7 were assigned to Caja-G*17:03 and Caja-G*08:06, respectively. In contrast, Caja class I sequences showing high similarities (>85%) to Caja-F4 were not observed in known Caja class I sequences.

All of the deduced Caja-G/F amino acid sequences contain the highly conserved and structurally essential amino acid residues, such as N-glycosylation sites (amino acid positions 86–88), four cysteine residues involved in disulfide bonding (positions 101, 164, 203, and 259) and CD8 binding, as well as most of the α-chain and β2 microglobulin contact sites, compared with the three-dimensional structure of HLA-A2 molecule (21, 22) (Supplemental Fig. 1). In addition, α helices, β sheets, and transmembrane structures of all six Caja-G/F sequences are similar to other MHC-A class I proteins (data not shown). Therefore, the six Caja-G/F amino acid

development.
sequences are expected to have a typical MHC class I protein structure. Moreover, 36 residues of the 57 Ag-binding sites are diverged among the Caja-G/F sequences (Fig. 2), suggesting different Ag-binding characteristics of the six Caja-G/F proteins. In the Caja-F4 sequence, 22 amino acid residues (19 amino acid changes, 3 amino acid deletions) were uniquely different from the other sequences, and changes in the amino acid polarity were observed in five sites (positions 66, 69, 116, 150, and 151) (Fig. 2), clearly distinguishing Caja-F sequences from Caja-G sequences.

Phylogenetic relationship of the Caja-G/F genes

To study the phylogenetic relationship of the Caja-G/F genes, we constructed a phylogenetic tree using the NJ method based on 42 nucleotide sequences covering exons 3–8. The phylogenetic tree shows five major lineages: MHC-A/AG, MHC-B/C, MHC-E, MHC-F, and MHC-G (Fig. 3). As expected, the HLA, Patr, and Mamu genes show close relationships, such as HLA-A, Patr-G, and Mamu-G5 in MHC-G lineage; HLA-A, Patr-A, Mamu-A1, and Mamu-AG3 in MHC-A lineage; HLA-F, Patr-F, and Mamu-F in MHC-F lineage; HLA-E, Patr-E, and Mamu-E in MHC-E lineage; and HLA-B, HLA-C, Patr-B, Patr-C, and Mamu-B1 in MHC-B/C lineage.

Of the 19 Caja MHC class I sequences, 14 show close relationships with MHC-G genes (Caja-G groups 1–4), and 5 show close relationships with MHC-F genes (Caja-F group). The phylogenetic relationship supports that Caja-G groups 1–4 evolved after divergence from HLA, Patr, and Mamu MHC-G genes and diverged in the following order: group 4, group 3, and group 1/2. Caja-G groups 1 and 2 include putatively functional Caja-G genes (Caja-G1, Caja-G3, Caja-G7, Caja-G12, and Caja-G13), whereas Caja-G groups 3 and 4 contain only Caja-G pseudogenes. The Caja-F group includes only one putatively functional Caja-F gene (Caja-F4) and four Caja-F pseudogenes (Figs. 3, 4).

Discussion

In this study, we determined 854 kb of genomic sequence of the common marmoset MHC-G/F segment that includes 14 Caja-G and 5 Caja-F genes by using seven BAC clones. As a result of naturally occurring bone marrow chimerism in marmosets, individuals may contain up to four alleles of genetic loci. Therefore,

### Table III. Features of interspersed repeat sequences in the MHC-A/G/F segment

<table>
<thead>
<tr>
<th>Species</th>
<th>Contig</th>
<th>Reference Sequence</th>
<th>Nucleotide Length (bp)</th>
<th>GC Content (%)</th>
<th>SINEs (%)</th>
<th>LINEs (%)</th>
<th>LTRs (%)</th>
<th>DNA Elements (%)</th>
<th>Total (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caja</td>
<td>Contig 1</td>
<td>AB809558</td>
<td>470,958</td>
<td>44.0</td>
<td>4.7</td>
<td>33.5</td>
<td>12.9</td>
<td>2.4</td>
<td>55.3</td>
</tr>
<tr>
<td>Caja</td>
<td>Contig 2</td>
<td>AB809559</td>
<td>191,375</td>
<td>44.2</td>
<td>5.1</td>
<td>28.9</td>
<td>13.4</td>
<td>2.4</td>
<td>51.4</td>
</tr>
<tr>
<td>Caja</td>
<td>Contig 3</td>
<td>AB809560</td>
<td>192,325</td>
<td>43.4</td>
<td>6.0</td>
<td>29.8</td>
<td>12.5</td>
<td>3.0</td>
<td>52.4</td>
</tr>
<tr>
<td>In total</td>
<td></td>
<td></td>
<td>854,658</td>
<td>43.9</td>
<td>5.1</td>
<td>31.7</td>
<td>13.0</td>
<td>2.6</td>
<td>53.7</td>
</tr>
<tr>
<td>HLA</td>
<td></td>
<td>BA000025</td>
<td>361,129</td>
<td>45.2</td>
<td>7.4</td>
<td>20.8</td>
<td>21.2</td>
<td>4.8</td>
<td>54.1</td>
</tr>
<tr>
<td>Patr</td>
<td></td>
<td>BA0000041</td>
<td>323,419</td>
<td>45.0</td>
<td>7.4</td>
<td>15.4</td>
<td>23.3</td>
<td>4.3</td>
<td>50.4</td>
</tr>
<tr>
<td>Mamu</td>
<td></td>
<td>AB128049</td>
<td>937,292</td>
<td>44.5</td>
<td>6.5</td>
<td>18.8</td>
<td>21.6</td>
<td>4.5</td>
<td>51.4</td>
</tr>
</tbody>
</table>

GC, percentage of guanine and cytosine; SINE, short interspersed nuclear element.

### Table IV. Gene content of the Caja-G/F segment

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Locus Type</th>
<th>Direction</th>
<th>No. of Exons</th>
<th>Prominent Features</th>
<th>Reference Sequence</th>
<th>Identity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caja-G1</td>
<td>Gene (+)</td>
<td>8</td>
<td>MHC, class I, G1</td>
<td>NM_002116</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>Caja-G2</td>
<td>Pseudo (+)</td>
<td>8</td>
<td>MHC, class I, G2</td>
<td>NM_002116</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>Caja-F1</td>
<td>Pseudo (+)</td>
<td>8</td>
<td>MHC, class I, F1</td>
<td>NM_001098479</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>Caja-G3</td>
<td>Gene (+)</td>
<td>8</td>
<td>MHC, class I, G3</td>
<td>NM_002116</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>Caja-G4</td>
<td>Pseudo (+)</td>
<td>6</td>
<td>MHC, class I, G4</td>
<td>NM_002116</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>Caja-G5</td>
<td>Pseudo (+)</td>
<td>6</td>
<td>MHC, class I, G5</td>
<td>NM_002116</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>Caja-G6</td>
<td>Pseudo (+)</td>
<td>8</td>
<td>MHC, class I, G6</td>
<td>NM_005514</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>Caja-F2</td>
<td>Pseudo (+)</td>
<td>7</td>
<td>MHC, class I, F2</td>
<td>NM_001098479</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>Caja-G7</td>
<td>Gene (+)</td>
<td>8</td>
<td>MHC, class I, G7</td>
<td>NM_002116</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>Caja-G8</td>
<td>Pseudo (+)</td>
<td>8</td>
<td>MHC, class I, G8</td>
<td>NM_002116</td>
<td>81</td>
<td></td>
</tr>
<tr>
<td>Caja-G9</td>
<td>Pseudo (-)</td>
<td>6</td>
<td>MHC, class I, G9</td>
<td>NM_002116</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>Caja-G10</td>
<td>Pseudo (-)</td>
<td>6</td>
<td>MHC, class I, G10</td>
<td>NR_001434</td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>Caja-G11</td>
<td>Pseudo (+)</td>
<td>8</td>
<td>MHC, class I, G11</td>
<td>NM_002116</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>Caja-F3</td>
<td>Pseudo (+)</td>
<td>5</td>
<td>MHC, class I, F3</td>
<td>NM_002116</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>Caja-F4</td>
<td>Gene (+)</td>
<td>7</td>
<td>MHC, class I, F4</td>
<td>NM_001098479</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>Caja-G12</td>
<td>Gene (+)</td>
<td>8</td>
<td>MHC, class I, G16</td>
<td>NM_002116</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>Caja-F5</td>
<td>Pseudo (+)</td>
<td>7</td>
<td>MHC, class I, F5</td>
<td>NM_001098479</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>Caja-G13</td>
<td>Gene (+)</td>
<td>8</td>
<td>MHC, class I, G13</td>
<td>NM_002116</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>Caja-G14</td>
<td>Pseudo (+)</td>
<td>8</td>
<td>MHC, class I, G14</td>
<td>NM_002116</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td>ZNRD1-AS1</td>
<td>Misc RNA (-)</td>
<td>6</td>
<td>ZNRD1 antisense RNA 1</td>
<td>NR_026751</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>HCG8</td>
<td>Misc RNA (-)</td>
<td>1</td>
<td>HLA complex group 8 pseudogene 1 ortholog</td>
<td>NR_103542</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>ETF1P1</td>
<td>Pseudo (+)</td>
<td>1</td>
<td>Eukaryotic translation factor 1 pseudogene</td>
<td>NM_004730</td>
<td>92</td>
<td></td>
</tr>
<tr>
<td>ZNRD1</td>
<td>Gene (+)</td>
<td>4</td>
<td>Zinc ribbon domain containing 1</td>
<td>NM_170873</td>
<td>94</td>
<td></td>
</tr>
<tr>
<td>PPP1R1I1</td>
<td>Gene (+)</td>
<td>3</td>
<td>Protein phosphatase 1, regulatory subunit 11</td>
<td>NM_021959</td>
<td>96</td>
<td></td>
</tr>
<tr>
<td>RNF39</td>
<td>Gene (-)</td>
<td>4</td>
<td>Ring finger protein 39</td>
<td>NM_025236</td>
<td>94</td>
<td></td>
</tr>
</tbody>
</table>

Caja class I loci are shown in bold type.
overlapping BAC clone sequences of 854 kb might be a composite of up to four haplotypes.

Comparison of our sequence with that from the current draft assembly of the common marmoset genome WUGSC 3.2 suggests allelic variation ranging from 0 to 2.4% in the overlapping 150-kb segment of contig 1 (the MOG side) and ranging from 0.2 to 2.4% in the overlapping 92-kb segment of contig 3 (the RNF39 side) (Table II). Overlaps among the BAC clones in contigs 1 and 3 are supported by comparison of restriction fragments and PCR-based mapping (data not shown). Therefore, we determined the nucleotide sequences using the representative BAC clones to obtain consensus sequences of the segment, because these well-conserved segments, locating near non-MHC genes, are thought to have similar genomic structures among the Caja-G haplotypes. Furthermore, gene order and gene content of the Caja-G/F segment are consistent with those of the common marmoset draft assembly and the HLA region. The remaining 562-kb segment (321 kb in contig 1, 141 kb in contig 2, and 100 kb in contig 3), including the duplicated Caja-G/F genes, is represented only by several mini contigs in the WUGSC 3.2 sequence (Table II), which is most likely due to the current quality (draft assembly and not finished sequence) of the common marmoset draft genome sequence. Genomic diversities in the Caja-G loci were suggested to be generated considerably after duplication based on the population study of the Caja-G genes (13). This finding supports that genomic structures

FIGURE 2. Comparison of 57 Ag binding sites derived from the deduced amino acid sequences of Caja-G/F genes. Dots and dashes indicate identical amino acids with Caja-G1 sequence and deleted sites, respectively. Gray and black backgrounds indicate conserved sites and unique amino acids observed in Caja-F4, respectively. For additional comparison, the respective amino acid residues of HLA-A*02:01 are included.

FIGURE 3. Nucleotide sequence–based phylogenetic tree of MHC class I genes constructed by the NJ method. Bold type indicates putatively functional Caja-G and Caja-F genes. Numbers at branches indicate bootstrap values.

FIGURE 4. Schematic gene structures showing segmental duplications involved in Caja-G and Caja-F genes. White, black, and gray boxes indicate putatively functional genes, pseudogenes, and missing genes.
of the Caja-G/F segment are different among the Caja-G/F haplotypes. In fact, nucleotide similarity between 10 kb of contig 1 (BAC clone 88K6 side) and 10 kb of contig 2 (BAC clone 246G18 side) was extremely low (<50%), although Caja-G9 in contig 1 is closely related to Caja-G10 in contig 2 with a high nucleotide similarity (98.9%) between them and a close phylogenetic relationship (Fig. 3). Hence, we tentatively described the location of contig 2 as being between contig 1 and contig 3 (Fig. 1B). Such difficulties in draft genomes are a known problem and are particularly evident in highly duplicated regions, such as the one analyzed in this study. Therefore, our determined genomic sequences provide additional and informative data for the complete determination of the common marmoset MHC region and provide researchers with an excellent basis to search for polymorphisms and to activate many fields of biomedical research involving the MHC of the common marmoset.

Phylogenetic analysis using 42 primate MHC class I genes and using common marmoset MHC-G/F allele sequences supported the classification of the common marmoset MHC class I genes of the Caja-G/F region into the MHC-G and MHC-F lineages. Interestingly, these genes were substantially expanded: four groups contain a total of 14 Caja-G genes, and one group contains 5 Caja-F genes (Fig. 3). The phylogenetic relationships, together with the positional relationships among the Caja-G and Caja-F genes (Fig. 1), also supported the presence of at least four duplicated units (Fig. 4). Unit 3 includes only the Caja-G9 pseudogene from Caja-G group 1, and we speculate that Caja-G genes belonging to Caja-G groups 2–4, as well as the Caja-F group, were deleted. Similarly, the Caja-G group 4 gene in unit 4 and genes from Caja-G groups 1 and 3 in unit 5 are missing. These missing MHC class I genes are either deleted or map to regions that are not covered by our BAC contigs (Fig. 1B). Therefore, we cannot exclude the possibility that Caja-G9 may belong to unit 4 as well, because Caja-G9 and Caja-G10 show a close relationship with Caja-G group 1 (Figs. 3, 4) and might have been duplicated after establishment of the different Caja-G groups.

Comparative genomic analysis of sequences from the duplication units of the HLA region revealed that the MIC, HCG9, HCG2, HCP5, and HCG26 sequences are contained within discrete duplication units, or duplons (23). The hypothetical ancestral duplication unit appears to have formed the duplicated class I genomic structures seen within the HLA-A/G/F and HLA-B/C segment as the result of a series of imperfect serial duplications during primate evolution (23, 24). This segmental duplication is also observed in the Patr-A/G/F and Mamu-A/G/F segments (25). However, genomic traits for MIC, HCG2, HCG9, HCP5, and HCG26 were not found in the Caja-G/F and Caja-B segments (12), suggesting that the Caja-G/F segment (or New World monkey’s MHC-G/F segment) was formed by species-specific segmental duplication of an MHC-A/G–MHC-F ancestral unit after separation from the other simian primates by birth-and-death evolution of MHC class I genes to adapt to environmental pathogens (26).

Among the 14 Caja-G and 5 Caja-F genes, 6 putatively functional genes (Caja-G1, Caja-G3, Caja-G7, Caja-G12, Caja-G13, and Caja-F4) were identified in this study. All of the deduced Caja class 1 amino acid sequences contain the highly conserved and structurally essential amino acid residues (Supplemental Fig. 1), and partial Caja-G1, Caja-G3, Caja-G7, Caja-G12, Caja-G13, and Caja-F4 cDNA sequences matched perfectly with all six Caja-G genes from our RT-PCR analysis of exons 2–4 (Caja-G: 472 bp, and Caja-F: 321 bp) using 10 RNA samples derived from Callithrix jacchus spleen tissues (Caja-G1: AB855795, Caja-G3: AB855796, Caja-G7: AB855797, Caja-G12: AB855798, Caja-G13: AB855799, and Caja-F4: AB855800) (T. Shiina, Y. Kametani, R. Suzuki, L. Walter, manuscript in preparation).

Interestingly, we found, using the deep-sequencing approach, that Caja-B3 and Caja-B4 are transcribed in various tissues at low levels, confirming our previous hypothesis that these MHC class I genes might indeed be actively transcribed. Thus, the common marmoset contains and transcribes all six MHC class I gene lineages, namely MHC-B/C, MHC-E, MHC-F, and MHC-G. The only exception is MHC-A, which is not present in the common marmoset or (most probably) in other New World monkey genomes. The loss of MHC-A and its function as a classical MHC class I gene of MHC-G either happened in the lineage leading to platyrhine primates (New World monkeys) or, alternatively, the specialized function of MHC-G and the classical function of MHC-A genes evolved in the lineage leading to catarrhine primates (Old World monkeys, apes, and humans). Unfortunately, strepsirrhine primates, such as lemurs, are not informative because the MHC class I genes of these species do not show a particularly close relationship with any of the simian MHC class I lineages (27).

HLA-G is involved in the modulation of immunological responses through inhibition of activity and mediating apoptosis of cytotoxic CD8+ T cells and NK cells by interacting with their inhibitory receptors and inhibiting the alloreactivity by CD4+ T cells (28). Although we tried to compare the transcriptional regulatory factor sequences of the 14 Caja-G genes, we could not identify a single Caja-G gene with a function similar to HLA-G, based on the genomic information provided in this study. Future expression analyses using trophoblast cells will show whether any of the Caja MHC class I genes reveal a typical MHC-G transcription pattern.

Acknowledgments

We thank Nico Westphal for excellent technical support.

Disclosures

The authors have no financial conflicts of interest.

References


