Published in final edited form as:

Tissue Antigens. 2011 August ; 78(2): 102–114. doi:10.1111/j.1399-0039.2011.01710.x.

Characterization and polymorphic analysis of 4.5 kb genomic full-length HLA-C in the Chinese Han population

Y. Xu¹, Z. Deng¹, C. O'hUigin², D. Wang¹, S. Gao¹, J. Zeng¹, B. Yang¹, S. Jin¹, H. Zou¹

¹Immunogenetics Laboratory, Shenzhen Blood Center, Shenzhen, Guangdong, China

²Genetics Core, Cancer and Inflammation Program, National Cancer Institute, SAIC-Frederick, Frederick, MD, USA

Abstract

This study used long-range polymerase chain reaction to sequence 4.5 or 4.3 kb of genomic DNA covering human leukocyte antigen C (HLA-C) and its flanks in 45 Chinese Han subjects to better characterize variation in the gene in a single population. Sequences of 35 HLA-C alleles were obtained from the population, including major alleles of 13 lineages of HLA-C. Four novel alleles, C*03:04:01:02, C*04:01:01:03, C*08:22, and C*17:01:01:02, were identified, and complete full-length sequences of 18 HLA-C alleles were obtained for the first time. All sequences herein reported also represent extensions through the promoter region and the 3'-untranslated region. Fourteen 5'-nucleotide sequences and 14 3'-nucleotide sequences were detected outside the coding region. In total, 316 single-nucleotide polymorphisms unequally distributed in HLA-C subregions were observed. In addition to exons 2 and 3, nucleotide variability was found to be particularly high in exon 5, which encodes the transmembrane region. The differentiation of the C*07 and C*17 lineages in this region accounts for the high variability. The congruence of phylogeny across most regions of the gene suggests that gene Conversion or recombination has not markedly influenced divergence between lineages in the evolution of HLA-C.

Keywords

evolution; genomic full-length; high-fidelity polymerase chain reaction amplification; human leukocyte antigen C; single-nucleotide polymorphisms

Introduction

Human leukocyte antigen (HLA) class Ia genes (HLA-A, HLA-B, and HLA-C) encode cell-surface glycoproteins that present endogenous peptides in the recognition by cytotoxic T cells (1). HLA-B and HLA-C are derived from a gene duplication that occurred in the ancestral ape lineage approximately 22 million years ago (2), and important clues to the

Correspondence: Zhihui Deng, PhD, Immunogenetics Laboratory, Shenzhen Blood Center, Shenzhen, Guangdong 518035, China, Tel: +86 755 83242567, Fax: +86 755 83221000, zhihui_deng@yahoo.com.cn and Dr Colm O'hUigin, Genetics Core, Cancer and Inflammation Program, National Cancer Institute, SAIC-Frederick, Frederick, MD 21702, USA, Tel: +1 301 846 1717, Fax: +1 301 846 1909, ohuiginc@mail.nih.gov.

Conflict of interest

The authors have declared no conflict of interests.

functional divergence of HLA-C have been inferred by studies of its polymorphism (3, 4). HLA-C molecules have been definitively defined as transplant antigens with strong relevance in unrelated stem-cell transplantation (5–8), and their interaction with natural killer cells indicates that they play an important role in cellular immune responses (9–11).

The primary site of polymorphism and allelic differentiation is the antigen-binding region (ABR) of exons 2 and 3, which have been the prime targets of sequencing and functional studies. Consequently, most available HLA-C sequences cover only exons 2 and 3, with a limited number extending to the genomic full-length sequence. However, strong effects in tissue compatibility and disease associations map to HLA-C outside of the ABR (12). Polymorphisms within exon 1 of the HLA-A, HLA-B, and HLA-C alleles, which encode the leader sequence, can influence the presentation of leader sequence-derived peptides by HLA-E (13). The $\alpha 3$ domain encoded by exon 4 of class I alleles also contributes to the $\beta 2m$ association (14) and interactions with the CD8 molecule (15). Mutations within the transmembrane-encoding exon (exon 5 in HLA class I genes) potentially influence associations with other cell-surface molecules (16), mRNA levels (17), and alloantigenic recognition (18). Although the role of intracellular signals (exons 6–8 in class I genes) has been previously investigated, it has not been fully defined (19). A large survey of variability across entire genes may indicate whether polymorphisms of coding regions outside the ABR contribute to functional differentiation.

Among the 829 HLA-C alleles included in the July 2010 release [2.28.0] of the IMGT/HLA Database (<http://www.ebi.ac.uk/imgt/hla>) (20), the polymorphisms of 19 groups including 66 alleles lie outside exons 2 and 3. For example, *15:05:01 and *15:05:02 show polymorphisms in exon 1; *01:02:01 and *01:02:02 show polymorphisms in exon 4; *07:01:01, *07:01:02, *07:06, and *07:18 show polymorphisms in exons 5 and 6; and *04:01:01:01 and *04:09N show polymorphisms in exon 7. Moreover, some allele groups are differentiated by polymorphisms outside the coding regions (e.g. C*04:01:01:01 and 0*04:01:01:02). Previous characterizations of noncoding regions in HLA class I genes initially focused on introns flanking or encompassed by the ABR (21–23). Subsequent research has incorporated sequencing of entire genes (24, 25), but the conducted surveys were not comprehensive for specific populations and the sequenced segments were generally limited to coding exons and their immediate flanks. In addition, large elements of the noncoding regions of common HLA-C alleles remain unsequenced, presenting the risk of polymerase chain reaction (PCR) dropout to sequence-based typing (SBT) genotyping protocols (26, 27). The accurate determination of genomic full-length sequences from the 5'-upstream region to the 3'-downstream region for HLA-C alleles in a single population can offer insights into possible functional variation and noncoding differentiation of populations as well as facilitate rapid and accurate genotyping.

A noncoding sequence is particularly relevant in elucidating the mechanism of generation and maintenance of diversity in HLA-C sequences. Gene conversion and recombination are the widely accepted explanations for the observed shared motif and high variability patterns of the ABR (28, 29). Cereb et al. (22) proposed an intronic gene conversion process that homogenizes introns 1–3 relative to the variable ABR exons using partial gene sequences. The lack of intronic variability has also been subsequently documented in downstream

introns (30). Recombination and gene conversions should blur phylogenetic signals because the phylogeny of novel sequences will reflect that of both their donor and recipient gene components. Kriener et al. (31) examined the congruity of phylogenies in different gene regions to evaluate the roles of recombination and conversion in class II DRB genes in various primates.

Contiguous intronic, exonic, and intergenic sequences provide opportunities for analysis not available via sequence coding alone. The objectives of this extensive study on HLA-C sequences were to obtain a fuller insight into HLA-C variation in a specific population than afforded by the coding segments, to evaluate the variability of coding segments outside the ABR, and to determine the indications for prevalent conversion or recombination events in the generation of HLA-C diversity.

Materials and methods

Samples

Forty-five unrelated bone marrow registry donors from Shenzhen, China, were enrolled into this study after they provided their informed consent. All subjects were Chinese Han. Of the 45 peripheral blood samples, 28 were randomly selected and 17 were selected from 1280 donors with known HLA-C genotypes that had been typed by performing SBT to clone more alleles from different HLA-C lineages. The typing was performed using a commercial kit (AlleleSEQR HLA-C Plus, Atria Genetics, San Francisco, CA).

DNA extraction

Genomic DNA was extracted from peripheral blood using a modified salting-out method, as recommended by the International Histocompatibility Work Group (<http://www.ihwg.org/> protocols).

PCR amplification, cloning, and sequencing

The full-length HLA-C target sequence, with a fragment size of 4.5 kb (from positions -962 to 3576 when numbered from the first base of the start codon), encompasses the 5'-promoter region up to the 3'-untranslated region (UTR). The sequences were amplified by long-range PCR using *Pfu* high-fidelity polymerase (*PfuUltra* II Fusion HS DNA polymerase, Stratagene, La Jolla, CA) and our in-house forward (5'-CGCAACTTTGAGGTGATGACT-3') and reverse (5'-TTGTCTCAGAAAGCACAGGGA-3') primers.

The PCR product was purified using a gel extraction kit (Axygen, Union City, CA), and the 3'-A overhangs were added in the reactions with a volume of 20 µl [17 µl of PCR products, 0.5 µl of dATP (25 mM), 2.0 µl of 10× Taq buffer, and 0.5 µl of rTaq polymerase (5 U/µl; Takara, Dalian, China)]. After incubation at 37°C for 30 min, the products conjugated with dATP at the 3' ends of the target sequence were ligated into a pGEM-T Easy Vector (Promega, Madison, WI). The ligated products were then transformed into JM109 competent cells and cultivated for 24 h in 2× YT medium with ampicillin. For every sample, 12 positive clones were harvested to ensure that each allele had at least three clones. Three plasmids of

each allele were sequenced in both directions using an ABI PRISM BigDye Terminator Cycle Sequencing Ready Reaction Kit (PE Applied Biosystems, Foster City, CA) and our in-house sequencing primers (Table S1; Supporting Information) on an ABI 3730 DNA Sequencer (Applied Biosystems, Foster City, CA).

Up to 22 4.5-kb genomic HLA-C alleles were identified from the initial 28 randomly selected samples. On the basis of these sequences, another pair of inner degenerate primers was designed to ensure the detection of all alleles from the research subjects. The other 17 purposely selected DNA samples were amplified using the RCwF 5'-TGGACTCACACARGGAACTC-3' and RCwR 5'-GGGACAAGGACAATGGAGCAG-3' primers. The target sequence with a fragment size of 4.3 kb was then subjected to molecular cloning and haplotype sequencing using the method described above.

Population data analysis

The sequence fractions sequenced using all forward and reverse primers from more than six positive clones were automatically assembled using the SeqManII module of the dnastar software package (laseroeni version 5, DNA Star, Madison, WI). Three clones with identical sequences were obtained after assembly for each allele. The checked sequences of all samples were aligned with the CLUSTAL-x program (32) and then manually rechecked using bioedit version 6.0.7 (<http://www.mbio.ncsu.edu/BioEdit/bioedit.html>) to ensure the accuracy of the sequences.

All the diploid sequences of the research subjects were aligned and imported into DNASP 4.0 (33) to calculate the degree of polymorphisms in all coding and noncoding regions based on the number of haplotypes (H), the average number of pairwise nucleotide differences per site (π) (34), the average number of pairwise synonymous substitutions per site ($\pi - s$), and the average number of pairwise nonsynonymous substitutions per site ($\pi - a$) (35). Moreover, the sliding window (window length, 100 bp; step size, 25 bp) method was used to calculate the T_t value across the entire region. Likewise, the ratio of nonsynonymous-to-synonymous substitution (γ) in the coding regions was estimated using a sliding window of 50 codons. Phylogenetic reconstruction of the sequences was performed via the neighbor-joining method based on Kimura's two-parameter distances using the MEGA 4.0 package (36).

HLA-C allele nomenclature

The genomic sequences of 35 HLA-C alleles were submitted to GenBank (<http://www.ncbi.nlm.nih.gov/Genbank/>) and IMGT/HLA Database (<http://www.ebi.ac.uk/imgt/hla/>), with the accession numbers and submission numbers, respectively, listed in Table 1. The alignment of the 35 sequences with C*01:02:01 is shown in Figure S1.

The allele names have been officially assigned by the World Health Organization (WHO) Nomenclature Committee for Factors of the HLA System. This follows the agreed policy that, subject to the conditions stated in the most recent Nomenclature Report (37), names will be assigned to new sequences as they are identified. Such new names will be published in the next WHO Nomenclature Report.

Results

Additions and extensions to the database of HLA-C allele sequences

In July 2010, 54 genomic full-length alleles including sequences from position –283 (in the 5'-promoter region) to position 3066 (in the 3'-UTR) of HLA-C were released in the IMGT/HLA Database; 18 of those were first extended or identified in the present study. Moreover, the 35 alleles herein reported were all extended beyond the IMGT/HLA representations in both the 5'-promoter region and the 3'-UTR (from positions –961 to 3576 and/or from positions –859 to 3477). Among these alleles, four novel alleles, namely, C*03:04:01:02, C*04:01:01:03, C*08:22, and C*17:01:01:02, were first identified, with the names assigned by the WHO Nomenclature Committee. The unique substitutions of these novel alleles that differentiate them from their closest relatives are listed in Table 1. C*08:22 differs from C*08:01:01 by A>G at position 2557 in exon 6. The nucleotide substitutions differentiating C*04:01:01:03, C*17:01:01:02, and C*03:04:01:02 are located in intron 3, intron 6, and 3'-UTR, respectively.

The intron sequences of 17 alleles, namely, C*01:03, C*01:08, C*03:03:01, C*03:04:01:02, C*04:01:01:03, C*04:03, C*07:06, C*07:66, C*07:67, C*08:03:01, C*08:20, C*08:21, C*08:22, C*14:02:01, C*15:05:02, C*16:04:01, and C*17:01:01:02, are reported here for the first time. As Table 2 shows, the introns of C*01:03 and C*01:08 are identical with those of C*01:02:01. Similarly, both C*03:03:01 and C*03:04:01:02 have introns identical with those of C*03:04:01:01. The introns of C*07:66, C*07:67, C*16:04:01, as well as C*08:03:01, C*08:20, C*08:21, and C*08:22 are identical with those of C*07:01:01:01, C*07:01:01:03, C*16:01:01, and C*08:01:01, respectively. However, the introns of six other alleles have unique substitutions compared with their closest relatives. The introns of C*04:01:01:03 and C*04:03 differ from C*04:01:01:01 at positions 1111 (A>G) and 1307 (C>G) of intron 3, respectively. C*07:06 has two unique substitutions in intron 3 (1280, T>C) and intron 4 (1971, T>C) compared with C*07:01:01. The introns of C*14:02:01 differ from those of C*14:03 at positions 2488 (G>A) and 2522 (G>A) in intron 5. C*15:05:02 differs from C*15:02:01 at position 1924 (C>T) in intron 4, whereas C*17:01:01:02 differs from C*17:01:01:01 at position 1478 (C>T) in intron 3 and at position 2582 (G>C) in intron 6.

The 5'-promoter region and 3'-UTR sequences of all alleles had been extended or added as the IMGT/HLA Database only released sequences from positions –283 to 3066. Fourteen promoter haplotypes (859 bp) were detected from 13 lineages (Table 3). First, most alleles in the same lineages share their unique promoter, such as those in the C*01 (C*01:02:01, C*01:03, and C*01:08), C*03 (C*03:02:02, C*03:03:01, and C*03:04:01:02), C*04 (C*04:01:01:01 and C*04:01:01:03), C*08 (C*08:01:01, C*08:02:01, C*08:20, C*08:21, C*08:22, and C*08:03:01), and C*15 (C*15:02:01 and C*15:05:02) lineages. Second, some alleles in different lineages share the same promoter. For example, alleles in the C*06 (C*06:02:01:01 and C*06:02:01:02), C*12 (C*12:02:02 and C*12:03:01:01), and C*16 (C*16:04:01) lineages are of the same promoter. Similarly, C*05:01:01:02 shares the same promoter with alleles in the C*08 lineage. Third, C*02:02:02, C*04:03, C*14:02:01, and C*17:01:01:02 each has a unique promoter. Lastly, four promoters were found in the C*07

lineage. C*07:01:01 and C*07:06 share the same promoter, and the promoter of C*07:01:02 differs from the rest at position –257. C*07:02:01:01, C*07:66, C*07:02:01:03, and C*07:67 share the same promoter, which is different from that of C*07:01:01 at positions –500 (T>G) and –810 (A>T). Compared with other alleles in the C*07 lineage, the 5'-promoter region of C*07:04:01 has three unique substitutions at positions –196 (A>G), –215 (C>T), and –761 (A>G). Moreover, some alleles with identical exon and intron sequences in different individuals show variant sites in the 5'-upstream region. For example, 0*01:02:01 (FJ827032) differs from 0*01:02:01 (FJ515900) at position –923 (C>T).

Polymorphism analyses in HLA-C subregions

All 90 full-length HLA-C allele sequences of 4.5 kb (56 from 28 individuals) or 4.3 kb (34 from 17 individuals) from the 45 Chinese Han research subjects were obtained using an in-house cloning and haplotype sequencing technique. Among these individuals, three were homozygous, with the rest being heterozygous. The obtained diploid data enabled analysis of the polymorphism structure in the Chinese population. Among 1000 individuals who were typed, 35 alleles representing most of the major lineages were identified. This is similar to the findings of similarly sized studies in comparable populations (<http://allelefrequencies.net>).

All overlapping sequences with the fragment size of 4.3 kb were aligned and divided into 16 subregions (Table 4). In total, 315 single-nucleotide polymorphisms (SNPs) and 17 insertion/deletion polymorphisms (indels) were found to be unequally distributed in different subregions (Table 4). As expected, the nucleotide diversity within exon 2 ($\pi = 24.72/\text{kb}$) and that within exon 3 ($\pi = 32.04/\text{kb}$) were higher than those of surrounding introns, and much of the diversity was attributable to nonsynonymous nucleotide substitutions ($\pi - a = 32.11/\text{kb}$ in exon 3), which are characteristic of balancing selection on these regions (38). Consistent with the conjecture of Cereb et al. (22) that the introns flanking exons 2 and 3 are homogenized, introns 1–3 exhibited much lower diversity ($\pi < 20/\text{kb}$) than exons 2 and 3 in this study. However, these introns are not less diverse than other introns on noncoding regions. The average noncoding diversity in sequenced segments is 14.3 substitutions per kilobase, whereas that of introns alone is 16.9 substitutions per kilobase. Introns 1–3 have above-average intronic diversity, with only intron 4 showing higher levels of diversity. These data show that the increased diversity in exons 2 and 3, rather than diminished flanking intronic diversity, is responsible for the differences in diversity around the region encoding the antigen-binding domains.

The nucleotide diversities in intron 4, exon 5, and the 3'-UTR are particularly striking. As shown in Table 2, the nucleotide diversity in intron 4 is notable among introns ($\pi = 33.70/\text{kb}$) and that in exon 5 is the highest among exons ($\pi = 35.06/\text{kb}$, $\pi - s = 44.92/\text{kb}$, and $\pi - a = 30.31/\text{kb}$). Available sequences for HLA-C and HLA-B across the 4.5-kb region were collated to characterize this variation further. Sliding window analyses of diversity (π) across the alignment for both HLA-C (39 sequences) and HLA-B (15 sequences) were then conducted. The relative contributions of nonsynonymous and synonymous substitutions to diversity were measured by subjecting their ratio (γ) in coding regions across the genes to sliding window analysis. Figure 1 shows the distribution of variation across the sequenced

segments for both HLA families. Among the 15 HLA-B alleles, diversity was particularly focused upstream of the first exon, in the second and third exons, and in intron 7. The highest diversity (in excess of 10% per kilobase) was found, as expected, in the exons encoding the ABR. On the other hand, for HLA-C, peaks in the ABR were lower (~4% per kilobase) and higher levels of diversity were found in intron 4, exon 5, and the middle of the 3'-UTR.

A closer examination of sequences showed C*07 and C*17 lineages divergence was responsible for the high diversity levels: the peaks in intron 4 and exon 5 were as a result of multiple substitutions distinguishing the C*07 and C*17 lineages from all other lineages. An 18 bp in-frame insertion occurred in exon 5 of the C*17 lineage, but such insertion events are not measured by the standard estimates of diversity used here. However, it appears that much of the diversity observed in exon 5 was synonymous. The ratio of nonsynonymous-to-synonymous substitution (ν) for exon 5 was <1 (Figure 2), in contrast with exons 2 and 3, where ν was generally >1 .

Gene conversion or recombination analysis

Where gene conversion or recombination is extensive, the phylogeny of exchanged segments may be inconsistent with that of the recipient gene allowing their identification. Additional ape sequences extracted from DNA databases were aligned in this study, giving a total of 43 sequences across a 4.5-kb region, to examine the roles of conversion and recombination in generating diversity in the evolutionary history of HLA-C. The phylogenetic trees of individual exon or intron sequence segments for congruence with trees of neighboring segments and those of the whole region were examined following the method described by Kriener et al. (31). First, nine HLA-C clades based on the tree of the complete alignment were defined (Figure 3). The nine clades corresponded to differently named lineages or related pairs of lineages and were recovered by maximum likelihood and parsimony methods of phylogenetic reconstruction (data not shown), as well as by neighbor joining. (The clades were *01, *02/15, *03, *04, *05/08, *06/12, *07, *14, and *16. The single HLA-C*17 lineage sequence was too distinct to be reliably pooled with other sequences and is, by itself, uninformative for this analysis.) Neighbor-joining trees for different segments of the alignment were then systematically generated and examined.

A clade can be recovered in phylogenetic trees in either of two topological conditions: In the first case, all sequences in the clade are recovered monophyletically (i.e. without interruption by intervening sequences of other clades). Alternatively, this clade (hereafter referred to as the *primary clade*) may be recovered interspersed with sequences from other (secondary) clades. In the latter case, a distinction can be made between instances when the secondary clade sequences are identical with some or all sequences in the primary clade and cases where the secondary clade sequences are different from all primary clade sequences. This distinction is relevant for short introns or exons, where identical ancestral states are shared by many sequences, such that the effects of recombination or homogenization cannot be gauged. Our group considers a tree in which any of the primary clades is partitioned into separate elements by intervening sequences (from secondary clades), where the intervening sequences are themselves not identical with any of the partitioned elements, to be

incongruent. Table 5 shows the recovery conditions for the nine primary clades in trees made with different segments of the HLA-C alignment. A simple number corresponding to the number of sequences in the clade is given where sequences are recovered monophyletically. Brackets surround that number when other sequences from secondary clades are identical with those in the primary clade. Finally, where the clade is incongruent (split into nonadjacent parts), the number of sequences in each constituent part is indicated and separated by a plus (+) sign.

Despite the small sizes of several segments, the trees obtained from each intron or exon are generally congruent with the tree obtained from the whole region. The least consistent lineage was *02/15, which is incongruent in 2 of 12 gene segments tested. Lineages *01, *07, and *14 were monophyletic across their entire length. Even the most diversified exons (Figure S2A) and their flanking introns (Figure S2B) produced generally monophyletic trees. As frequent recombination or homogenization should lead to an alteration in trees reflecting the phylogeny of exchanged segments, the results indicate that the homogenized segments either occur infrequently or are much smaller than the individual domains examined in this study.

Discussion

Research on HLA-C polymorphisms across the entire gene and its flanks in the Chinese population has shown issues of both clinical and evolutionary relevance. As increased numbers of HLA alleles are characterized, typing ambiguities proliferate for both the heterozygous combinations and the allelic differences outside the analyzed region. Characterization of the genomic full-length sequences of more alleles can help define their unique substitutions and decrease the number of ambiguities; however, the genomic full-length sequences of HLA-C in the IMGT/HLA Database were relatively limited (36 in 829 HLA-C alleles) prior to inclusion of the data presented here.

This study indicates that substitutions are common in the exons outside the ABR, such as exons 5 and 6. The novel allele C*08:22, which differs by a nonsynonymous substitution from C*08:01:01 in exon 6, was found in 2 of the 45 research subjects and attained a frequency of 0.92% in unrelated Chinese Han individuals (unpublished results), thus meriting inclusion in clinical high-resolution HLA-C typing. Recent studies have shown that non-ABR polymorphisms can be relevant in tissue typing. The surface density of class I molecules has been suggested to be mediated by physical association with other cell-surface molecules via the transmembrane domain (16), and peptides derived from the transmembrane domain (exon 5) in HLA-A2 can potentially be alloantigenically recognized by T cells (18). Thus, the polymorphisms in exon 5 may be relevant to the expression regulation and immune responses to alloantigens.

The polymorphism of introns can be crucial in designing appropriate primers for SBT, and unidentified nucleotide substitutions in noncoding sequences may result in allele dropout by SBT (26, 27), especially in substitutions in introns 1–4, where PCR or sequencing primers usually locate. Six common and well-documented alleles, C*04:01:01:03, C*04:03, C*07:06, C* 14:02:01, C*15:05:02, and C*17:01:01:02, with unique substitutions in

intronic sequences were elucidated here for the first time. The intronic SNP information may help in the design of appropriate primers for developing more accurate HLA-C SBT kits.

Differential expression of HLA-C is largely determined by the polymorphisms in the promoter region. This study detected 18 additional SNPs and three indels in the promoter region by extending the sequences of alleles included in the IMGT/HLA Database. C*17 showed two indels and eight unique substitutions, of which two locate in the *cis*-element *Enhancer A* (39). Furthermore, the promoter of C*07:04:01 was found to differ from other C*07 allele promoters by three unique substitutions, one of which is in *Enhancer A* (Table 3). Polymorphisms of the 3'-UTR in class I genes may be important in posttranscriptional control of gene expression (40, 41). This study characterized 14 3'-UTR haplotypes that may serve such purpose.

In addition to their clinical implications, this study used the sequences to determine issues of relevance to the generation of diversity at HLA-C. The ABR was found to have higher diversity than its flanking introns. However, because these introns have higher diversities than most of the noncoding segments of the gene, no evidence for their homogenization, as previously suggested by Cereb et al. (22), was seen. Furthermore, this study's examination of individual exons and introns showed that the phylogeny of HLA-C lineages is generally congruent from domain to domain, with little evidence showing that homogenization or recombination has shaped HLA-C diversity.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Sample preparation, sequencing analysis, and all experimental work were conducted at the Shenzhen Blood Center, Shenzhen, Guangdong, China, which was supported by the Research Fund of Guangdong Science and Technology Department (Project No. 2008B030301277) and the Guangdong Natural Science Fund (Project No. 9451803501004124). Additional phylogenetic analysis was conducted by Dr CO, who was supported by the National Cancer Institute, National Institutes of Health, Frederick, MD, USA, with federal funds under HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government.

References

1. Townsend A, Bodmer H. Antigen recognition by class I restricted T lymphocytes. *Annu Rev Immunol* 1989; 7: 601–24. [PubMed: 2469442]
2. Fukami-Kobayashi K, Shiina T, Anzai T et al. Genomic evolution of MHC class I region in primates. *Proc Natl Acad Sci USA* 2005; 102: 9230–4. [PubMed: 15967992]
3. Colonna M, Brooks EG, Falco M, Ferrara GB, Strominger JL. Generation of allospecific natural killer cells by stimulation across a polymorphism of F1LA-C *Science* 1993; 260: 1121–4. [PubMed: 8493555]
4. Katz G, Markel G, Mizrahi S, Arnon TI, Mandelboim O. Recognition of F1LA-Cw4 but not F1LA-Cw6 by the NK cell receptor killer cell Ig-like receptor two-domain short tail number 4. *J Immunol* 2001; 166: 7260–7. [PubMed: 11390475]
5. Petersdorf EW, Longton GM, Anasetti C et al. Association of HLA-C disparity with graft failure after marrow transplantation from unrelated donors. *Blood* 1997; 89: 1818–23. [PubMed: 9057668]

6. Sasazuki T, Juji T, Morishima Y et al. Effect of matching of class I HLA alleles on clinical outcome after transplantation of hematopoietic stem cells from unrelated donors. *N Engl J Med* 1998; 339: 1177–85. [PubMed: 9780337]
7. Flomenberg N, Baxter-Lowe LA, Confer D et al. Impact of HLA class I and class II high-resolution matching on outcomes of unrelated donor bone marrow transplantation: HLA-C mismatching is associated with a strong adverse effect on transplantation outcome. *Blood* 2004; 104: 1923–30. [PubMed: 15191952]
8. Kawase T, Morishima Y, Matsuo K et al. Japan Marrow Donor Program. High-risk HLA allele mismatch combinations responsible for severe acute graft-versus-host disease and implication for its molecular mechanism. *Blood* 2007; 110: 2235–41. [PubMed: 17554059]
9. Winter CC, Gumperz JE, Parham P, Long EO, Wagtmann N. Direct binding and functional transfer of NK cell inhibitory receptors reveal novel patterns of HLA-C allotype recognition. *J Immunol* 1998; 161: 571–7. [PubMed: 9670929]
10. Brooks AG, Boyington JC, Sun PD. Natural killer cell recognition of HLA class I molecules. *Rev Immunogenet* 2000; 2: 433–48. [PubMed: 11256749]
11. Boyington JC, Sun PD. A structural perspective on MHC class I recognition by killer cell immunoglobulin-like receptors. *Mol Immunol* 2002; 38: 1007–21. [PubMed: 11955593]
12. Thomas R, Apps R, Qi Y et al. HLA-C cell surface expression and control of HIV/AIDS correlate with a variant upstream of HLA-C. *Nat Genet* 2009; 41: 1290–4. [PubMed: 19935663]
13. Braud VM, Allan DS, Wilson D, McMichael AJ. TAP- and tapasin-dependent HLA-E surface expression correlates with the binding of an MHC class I leader peptide. *Curr Biol* 1998; 8: 1–10. [PubMed: 9427624]
14. Hebert AM, Strohmaier J, Whitman MC et al. Kinetics and thermodynamics of beta 2-macroglobulin binding to the alpha 3 domain of major histocompatibility complex class I heavy chain *Biochemistry*. *Biochemistry* 2001; 40: 5233–42. [PubMed: 11318646]
15. Salter R, Benjamin RJ, Wesley PK et al. A binding site for the T-cell co-receptor CDS on the alpha 3 domain of HLA-A2. *Nature* 1990; 345: 41–6. [PubMed: 2109837]
16. Bremond A, Meynet O, Mahiddine K et al. Regulation of HLA class I surface expression requires CD99 and p230/golggin-245 interaction. *Blood* 2009; 113: 347–57. [PubMed: 18849489]
17. Watanabe Y, Magor KE, Parham P. Exon 5 encoding the transmembrane region of HLA-A contains a transitional region for the induction of nonsense-mediated mRNA decay. *J Immunol* 2001; 167: 6901–11. [PubMed: 11739508]
18. Hanvesakul R, Maillere B, Briggs D, Baker R, Larche M, Ball S. Indirect recognition of T-cell epitopes derived from the alpha 3 and transmembrane domain of HLA-A2. *Am J Transplant* 2007; 7: 1148–57. [PubMed: 17355239]
19. Schaefer MR, Williams M, Kulpa DA, Blakely PK, Yaffee AQ, Collins KL. A novel trafficking signal within the HLA-C cytoplasmic tail allows regulated expression upon differentiation of macrophages. *J Immunol* 2008; 180: 7804–17. [PubMed: 18523244]
20. Robinson J, Malik A, Parham P, Bodmer JG, Marsh SGE. IMGT/HLA - a sequence database for the human major histocompatibility complex. *Tissue Antigens* 2000; 55: 280–7. [PubMed: 10777106]
21. Cereb JSf, Kong Y, Lee S, Maye P, Yang SY. Nucleotide sequences of MHC class I introns 1, 2 and 3 in humans and intron 2 in non human primates. *Tissue Antigens* 1996; 47: 498–511. [PubMed: 8813739]
22. Cereb N, Hughes AL, Yang SY. Locus-specific conservation of the HLA class I introns by intra-locus homogenization. *Immunogenetics* 1997; 47: 30–6.
23. Blasczyk R, Kotsch K, Wehling J. The nature of polymorphism of the HLA class I non-coding regions and their contribution to the diversification of HLA. *Hereditas* 1997; 127: 7–9. [PubMed: 9471425]
24. Cox ST, McWhinnie AJ, Robinson J et al. Cloning and sequencing full-length HLA-B and -C genes. *Tissue Antigens* 2003; 61: 20–48. [PubMed: 12622774]
25. Beck S, Trowsdale J. The human major histocompatibility complex: lessons from the DNA sequence. *Annu Rev Genomics Hum Genet* 2000; 1: 117–37. [PubMed: 11701627]

26. Deng Z, Wang D, Xu Y et al. HLA-C polymorphisms and PCR dropout in exons 2 and 3 of the Cw*0706 allele in sequence-based typing for unrelated Chinese marrow donors. *Hum Immunol* 2010; 71: 577–81. [PubMed: 20226825]
27. Heinold A, Schaller-Suefling E, Opelz G, Scherer S, Tran TH. Identification of two novel HLA-alleles, HLA-A*02010103 and HLA-B*4455, and characterization of the complete genomic sequence of HLA-A*290201. *Tissue Antigens* 2008; 72: 397–400. [PubMed: 18647362]
28. Parham P, Ohta T. Population biology of antigen presentation by MHC class I molecules. *Science* 1996; 272: 67–74. [PubMed: 8600539]
29. Martinsohn JT, Sousa AB, Guethlein LA, Howard JC. The gene conversion hypothesis of MHC evolution: a review. *Immunogenetics* 1999; 50: 168–200.
30. ElSner HA, Blasczyk R. The phylogenies of introns 4–7 demonstrate an inconsistent pattern between human leukocyte antigen-C group topologies. *Tissue Antigens* 2004; 63: 109–21. [PubMed: 14705982]
31. Kriener K, O’Hugin C, Tichy H, Klein J. Convergent evolution of major histocompatibility complex molecules in humans and New World monkeys. *Immunogenetics* 2000; 51: 169–78.
32. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The CLUSTAL_X windows interface: flexible strategies for multiple sequences alignment aided by quality analysis tools. *Nucleic Acids Res* 1997; 25: 4876–82. [PubMed: 9396791]
33. Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R. DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 2003; 19: 2496–7.
34. Nei M. *Molecular Evolutionary Genetics*. New York: Columbia University Press, 1987.
35. Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* 1986; 3: 418–26. [PubMed: 3444411]
36. Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* 2007; 24: 1596–9. [PubMed: 17488738]
37. Marsh SGE, Albert ED, Bodmer WF et al. Nomenclature for factors of the HLA system, 2010. *Tissue Antigens* 2010; 75: 291–455. [PubMed: 20356336]
38. Hughes AL, Nei M. Pattern of nucleotide substitution at MHC class I loci reveals overdominant selection. *Nature* 1988; 235: 167–70.
39. Van den Elsen PJ, Gobin SJ, van Eggermond MC, Peijnenburg A. Regulation of MHC class I and II gene transcription: differences and similarities. *Immunogenetics* 1998; 48: 208–21. [PubMed: 9683666]
40. McCutcheon JA, Gumperz J, Smith KD, Lutz CT, Parham P. Low HLA-C expression at cell surfaces correlates with increased turnover of heavy chain mRNA. *J Exp Med* 1995; 181: 2085–95. [PubMed: 7760000]
41. Rousseau P, Le Discorde M, Mouillot G, Marcou C, Carosella ED, Moreau P. The 14 bp deletion-insertion polymorphism in the 3’UT region of the HLA-G gene influences HLA-G mRNA stability. *Hum Immunol* 2003; 64: 1005–10. [PubMed: 14602228]

Figure 1. Sliding window analysis of nucleotide diversities in human leukocyte antigen (HLA)-C and HLA-B. Nucleotide diversities (π) measured for pairwise combinations of 40 HLA-C sequences (solid line) and 15 HLA-B sequences (broken line). The ordinate indicates pairwise nucleotide diversity (percent). The window length and step size are 100 and 25 bp, respectively, and the horizontal axis indicates the position in the alignment. Boxes above the plot indicate the approximate size and position of coding exons.

Figure 2. Ratio of nonsynonymous-to-synonymous substitution (γ) in coding regions of human leukocyte antigen (HLA)-C and HLA-B genes. The mean γ value of the ratio is measured for pairwise comparisons of the coding regions of the 40 HLA-C genes (solid line) and the 15 HLA-B genes (broken line). The ordinate indicates the mean estimates of γ , whereas the horizontal axis indicates the position in the alignment. Boxes above the plot indicate the approximate position of exons. A window size of 50 codons was used, with a step size of 1 codon.

Figure 3.

Neighbor-joining tree of major histocompatibility complex class (MHC-C) genes from humans and their close relatives. Nucleotide distances were estimated using Kimura's two-parameter method over 4.5 kb of aligned sequence for sites shared by all sequences. The percentage bootstrap recovery over 500 replicates is indicated by the number above the nodes. The scale indicates the number of substitutions over the specified length. Additional four human sequences extracted from DNA database are followed by accession codes. Sequences from gorillas, chimpanzees, and orangutans are indicated with species

abbreviations (Gogo, Patr, and Popy, respectively) followed by accession codes. The orangutan sequence was obtained from a draft genome assembly (WUSTL Pongo_abelii-2.02).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

GenBank accession numbers and IMGT/HLA submission numbers of the 35 HLA-C alleles studied

Allele	GenBank accession number	IMGT/HLA submission number	Length (bp)	Number of appearances	Comparison with IMGT/HLA Database
C*01:02:01	FJ515900	HWS10006562	4537	9	Extended sequences of 5'-promoter region and 3'-UTR
C*01:02:01	FJ827032	HWS10006744	4537	1	Extended sequences of 5'-promoter region and 3'-UTR
C*01:03	GQ472834	HWS10006680	4336	1	Added sequences of 5'-promoter region, 3'-UTR, and all introns
C*01:08	GQ472835	HWS10006681	4336	1	Added sequences of exon 1, exons 4–8, 5'-promoter region, 3'-UTR, and all introns
C*02:02:02	GQ472836	HWS10006682	4329	1	Extended sequences of 5'-promoter region and 3'-UTR
C*03:02:02	FJ827033	HWS10006563	4536	2	Extended sequences of 5'-promoter region and 3'-UTR
C*03:03:01	FJ515902	HWS10006152	4537	9	Added sequences of 5'-promoter region, 3'-UTR, and all introns
C*03:04:01:02	FJ515903	HWS10005879	4537	8	Novel with the unique substitution at NT3033 (C>G) in 3'-UTR
C*04:01:01:01	FJ515901	HWS10006561	4537	2	Extended sequences of 5'-promoter region and 3'-UTR
C*04:01:01:03	FJ515899	HWS10005875	4537	1	Novel with the substitution at NT1111 (A>G) in intron 3
C*04:03	GQ472837	HWS10006684	4336	4	Added sequences of 5'-promoter region, 3'-UTR, and all introns
C*05:01:01:02	GQ895733	HWS10006742	4335	1	Extended sequences of 5'-promoter region and 3'-UTR
C*06:02:01:01	FJ785734	HWS10006557	4536	2	Extended sequences of 5'-promoter region and 3'-UTR
C*06:02:01:02	FJ785735	HWS10006558	4536	1	Extended sequences of 5'-promoter region and 3'-UTR
C*07:01:01	GQ472838	HWS10006686	4350	1	Extended sequences of 5'-promoter region and 3'-UTR
C*07:01:02	GQ472839	HWS10006687	4348	1	Extended sequences of 5'-promoter region and 3'-UTR
C*07:02:01:01	FJ515904	HWS10006548	4551	6	Extended sequences of 5'-promoter region and 3'-UTR
C*07:02:01:03	FJ785731	HWS10006549	4551	2	Extended sequences of 5'-promoter region and 3'-UTR
C*07:04:01	FJ785728	HWS10006550	4551	1	Added sequences of 5'-promoter region
C*07:06	FJ785732	HWS10006238	4551	5	Added sequences of 5'-promoter region, 3'-UTR, and all introns
C*07:66	FJ785729	HWS10006551	4551	1	Added sequences of exons 5–8, 5'-promoter region, 3'-UTR, and all introns
C*07:67	FJ785730	HWS10006552	4551	1	Added sequences of exons 5–8, 5'-promoter region, 3'-UTR, and all introns
C*08:01:01	FJ785724	HWS10006554	4536	6	Extended sequences of 5'-promoter region and 3'-UTR
C*08:02:01	FJ785733	HWS10006555	4536	1	Extended sequences of 5'-promoter region and 3'-UTR

Allele	GenBank accession number	IMGT/HLA submission number	Length (bp)	Number of appearances	Comparison with IMGT/HLA Database
C*08:03:01	GQ472840	HWS10006688	4335	2	Added sequences of 5'-promoter region, 3'-UTR, and all introns
C*08:20	FJ785725	HWS10005896	4536	1	Added sequences of exons 5–8, 5'-promoter region, 3'-UTR, and all introns
C*08:21	FJ785726	HWS10006556	4536	1	Added sequences of exons 5–8, 5'-promoter region, 3'-UTR, and all introns
C*08:22	FJ785727	HWS10006188	4536	2	Novel with substitution at NT2557 (A>G) in exon 6
C*12:02:02	GQ472841	HWS10006689	4335	2	Extended sequences of 5'-promoter region and 3'-UTR
C*12:03:01:01	GQ472842	HWS10006690	4335	3	Extended sequences of 5'-promoter region and 3'-UTR
C*14:02:01	FJ785736	HWS10006239	4537	3	Added sequences of 5'-promoter region, 3'-UTR, and all introns
C*15:02:01	FJ785737	HWS10006559	4537	5	Extended sequences of 5'-promoter region and 3'-UTR
C*15:05:02	GQ895734	HWS10006743	4336	1	Added sequences of 5'-promoter region, 3'-UTR, and all introns
C*16:04:01	GQ472843	HWS10006691	4335	1	Added sequences of exons 6–8, 5'-promoter region, 3'-UTR, and all introns
C*17:01:01:02	GQ472844	HWS10006692	4359	1	Novel with substitution at NT1478 (C>T) in intron 3 and NT2582 (G>C) in intron 6

HLA, human leukocyte antigen; UTR, untranslated region.

Table 2

Closest relatives and differences of the 17 introns first reported

Allele	Closest relatives and differences
C*01:03	Identical with C*01:02:01
C*01:08	Identical with C*01:02:01
C*03:01:01	Identical with C*03:04:01:01
C*03:04:01:02	Identical with C*03:04:01:01
C*04:01:01:03	Differs from C*04:01:01:01 at position 1111 (A>G) in intron 3
C*04:03	Differs from C*04:01:01:01 at position 1307 (C>G) in intron 3
C*07:06	Differs from C*07:01:01 at positions 1280 (T>C) in intron 3 and 1971 (T>C) in intron 4
C*07:66	Identical with C*07:01:01:01
C*07:67	Identical with C*07:01:01:03
C*08:03:01	Identical with C*08:01:01
C*08:20	Identical with C*08:01:01
C*08:21	Identical with C*08:01:01
C*08:22	Identical with C*08:01:01
C*14:02:01	Differs from C*14:03 at positions 2488 (G>A) and 2522 (G>A) in intron 5
C*15:05:02	Differs from C*15:02:01 at position 1924 (C>T) in intron 4
C*16:04:01	Identical with C*16:01:01
C*17:01:01:02	Differs from C*17:01:01:01 at position 1478 (C>T) in intron 3 and position 2582 (G>C) in intron 6

Table 3

Differences in the 14 haplotypes of the 5'-promoter region^a

	Nucleotide variation						Shared Allele
	-5 21/	-	-	-	-	-	
Promoter	-8-8-8-7-7-6-5-5-4	-4-4-402/- 225-2-2-2-2-1-1	-3-3-3-2-	-1-1-1-1-1-8-6-	4-4-4-4-4-3-3-3- 2-2		
Haplotype	5939108382612678220059	-451/-450	4510 401 7135018357 1 46373115039689	-181/-180	6564636251 9 4 8 3 2 1 0 9 8 7 1 0-7		
H1	G T T G G G C C - G A - - - - AT - - - - GGTGGTAGACCAT - - - - TCTCTGCACCCAGTAGC						*01:02:01, *01:03 and *01:08
H2		...C.A.-.-.-.-.-.C.....-.-.-.-.A.....G..					*02:02:02
H3		TC..C.AA.-.-.-.-.-.A.C.....-.-.-.-.A.G.....					*03:02:02, *03:03:01 and *03:04:01
H4		 A . - - - - - . - - - - . A . C G . - - - - - . - - - - .					*04:01:01:01 and *04:01:01:03
H5		A.-.-.-.-.-.C....G.-.-.-.-.-					*04:03
H6		.C..C.A.-.-.-.-.-.C.....-.-.-.-.T.....G. .					*05:01:01:02, *08:01:01, *08:02, *08:20, *08:21, *08:22, and *08:03:01
H7		...C.A.-.-.-.-.-.C.....-.-.-.-.-.G. .					*06:02:01:01, *06:02:01:02, *12:02:02, *12:03:01:01, and *16:04:01
H8		T C . . C . A . - G T A A G G G . G A G T T . . C . . T . G . - - - - - C G					*07:01:01 and *07:06
H9		T C . . C . A . - G T A A G G G . G A G T T . A C . . T . G . - - - - - C G					*07:01:02
H10		T C A . C . A . - T G T A A G G G . G A G T T . . C . . T . G . - - - - - C G					*07:02:01:01, *07:66, *07:02:01:03, and *07:67
H11		T C . . C A A . - G T A A G G G . G A G T T . . C - - - - - C G					*07:04:01
H12		...C.A.-.-.-.-.-.C.....-.-.-.-.-					*14:02:01
H13		...AC.A.G.-.-.-.-.-.A.C.....-.-.-.-.A.....G..					*15:02:01 and *15:05:02
H14		TC..C.A.-.-.-.-.-.AGAG.....CTACTAGCACTCCC.....AG-----...					*17:01:01:02

^aDifferences in actual promoter elements (*EnhancerA*) are highlighted in gray.

Table 4Polymorphism levels of subregions in the HLA-C gene^a

Region	Length (bp)	H	S	Indels	π ($\times 1000$)	$\pi - s$ ($\times 1000$)	$\pi - a$ ($\times 1000$)
5'-Promoter	859	14	33	6	8.92	—	—
Exon 1	73	5	6	0	21.17	0	22.54
Intron 1	130	10	11	0	19.28	—	—
Exon 2	270	18	22	0	24.72	30.05	22.44
Intron 2	251	11	19	3	19.86	—	—
Exon 3	276	22	35	0	32.04	32.87	32.11
Intron 3	587	17	39	0	13.48	—	—
Exon 4	276	12	22	0	19.70	37.11	13.76
Intron 4	124	7	15	1	33.70	—	—
Exon 5	120	10	22	1	35.06	44.92	30.31
Intron 5	441	11	28	1	14.03	—	—
Exons 6–7–8	86	4	4	0	9.27	0	12.81
Intron 6	107	6	6	0	10.91	—	—
Intron 7	164	8	10	1	14.96	—	—
3'-UTR	416	13	32	3	20.75	—	—
3'-Downstream	168	6	12	1	13.87	—	—

H indicates the number of unique nucleotide sequences; HLA-C, human leukocyte antigen C; *S*, number of segregating sites; indels, number of insertion/deletion polymorphisms; UTR, untranslated region; π , average number of pairwise nucleotide differences per site; $\pi - s$, number of synonymous nucleotide substitutions per site; and $\pi - a$, number of nonsynonymous nucleotide substitutions per site.

^aThe regions with the highest polymorphism levels are highlighted in gray.

Table 5

Phylogenetic congruence of HLA-C lineages by domain

Clade segment	*01	*02/15	*03	*04	*05/08	*06/12	*07	*14	*16	Segment size (bp)
5'-UTR	4	3	5	3	8	4	8	2	2	982
Exon 1	4	3	5	3	8	4	8	2	2	73
Intron 1	4	3	5	3	8	3	8	2	2	131
Exon 2	4	3	5	2 + 1	8	2 + 2	8	2	2	270
Intron 2	4	(3)	5	3	(8)	4	8	2	2	259
Exon 3	4	2 + 1	5	3	8	4	8	2	1 + 1	276
Intron 3	4	2 + 1	5	(3)	8	4	8	2	2	588
Exon 4	4	3	5	3	(8)	4	8	2	2	276
Intron 4	4	3	5	3	8	4	8	2	2	124
Exons 5–8	4	3	5	3	8	4	8	2	2	233
Intron 5–7	4	3	5	3	8	4	8	2	2	733
3'-UTR	4	(3)	3 + 1+1	(3)	8	(4)	8	2	2	689
All	4	3	5	3	8	4	8	2	2	4634

HLA-C, human leukocyte antigen C; UTR, untranslated region.