Journal of Human Genetics (2009) 54, 224–229
© 2009 The Japan Society of Human Genetics All rights reserved 1434-5161/09 $32.00
www.nature.com/jhg

ORIGINAL ARTICLE

Complex divergence at a microsatellite marker *C1_2_5* in the lineage of *HLA-Cw/-B* haplotype

Daisuke Shichi[1], Masao Ota[2], Yoshihiko Katsuyama[3], Hidetoshi Inoko[4], Taeko K Naruse[1] and Akinori Kimura[1,5]

The human leukocyte antigen (HLA) complex locus has shaped a framework for evolutionary processes because of the dense clustering and strong linkage disequilibrium (LD) of polymorphic genes. Although the landscape of LD among conventional single-nucleotide polymorphisms (SNPs) has been described, the data on the lineage of major histocompatibility complex (MHC) haplotype are limited to pairwise comparisons of several haplotypes in Caucasoid populations. Multi-allelic markers, including microsatellite markers, may provide us with a larger power to analyze the MHC haplotype lineage because the mutation rate of microsatellite exceeds that of SNPs by several orders of magnitude. In this study, we investigated the complex structure of repeat motifs in a microsatellite to figure out the structural lineage of *HLA-Cw/-B* segments in Japanese. It was found that the genetic differences of *HLA-Cw/-B* haplotype lineage were reflected by repeat motif patterns at *C1_2_5* locus, suggesting that unique mutational dynamics of microsatellites may be a useful marker to chase the haplotype lineage.

Journal of Human Genetics (2009) **54,** 224–229; doi:10.1038/jhg.2009.15; published online 27 February 2009

Keywords: C1_2_5; HLA; microsatellite; repeat motif

INTRODUCTION

The human major histocompatibility complex (MHC), the human leukocyte antigen locus (HLA) on chromosome 6p21.3, spans about 4 Mb and contains many polymorphic genes relevant to the adaptive immune system.[1] Among them, genes for classical HLA molecules play pivotal roles in the immunological recognition of self versus non-self through presentation of antigenic peptides from either intracellular or extracellular origin.[2] Most of the extensive polymorphisms in the *HLA* genes were found at the peptide-binding groove of HLA molecules, thereby defining the bound peptides.[3] The HLA alleles at a given locus differ from each other by 1–30 amino acids at the protein level[4] and have been designated by the four-digit number or more according to the patterns of single-nucleotide polymorphisms (SNPs) and insertion–deletion polymorphisms within the coding sequence. The difference in allele distribution among different ethnic groups may be shaped by selective and demographic history.[5] It is well known that there is a strong linkage disequilibrium (LD) among alleles of genes in *HLA* locus, and combination of these alleles in LD form specific haplotypes.[6] Owing to the functional significance of classical *HLA* genes, the MHC haplotypes have been defined by using classical *HLA* alleles as highly polymorphic markers, and the *HLA* haplotypes served as a model system for high-resolution mapping of disease susceptibility genes,[7] evolution[8] and population structure.[9]

Detailed information on allelic diversity, recombination hotspot and profiles of LD within the MHC region are available,[6] but data on the lineage of the MHC haplotype and its evolution are not complete. The MHC Haplotype Sequencing Project was designed to elucidate the complete MHC genetic maps of several common Caucasian MHC haplotypes,[10] but little information is available for MHC haplotypes from different ethnic groups other than that from the Caucasians. Using selected genomic variation, including SNPs, individual MHC haplotypes can be characterized. This strategy has been used extensively to resolve the structure of the *HLA* allelic composition of SNPs and to determine new *HLA* alleles.[11] However, conventional SNP-based tagging could not adequately provide a resolution to capture the characteristics of variations of the MHC region at the worldwide population level. In other words, genetic markers other than SNPs, including copy number variations (CNVs) and microsatellites, might provide additional information in tracing the differentiation of the MHC haplotype.

Microsatellites, in general, undergo rapid change because of the insertion or deletion of one or multiple repeat units, primarily through replication slippage.[12] Moreover, the mutation rate of microsatellites (10^{-5}–10^{-3} per generation) exceeds that of SNPs and CNVs by several orders of magnitude. The difference in the mutational dynamics suggested that the microsatellites may be useful in tracing recent divergence in the structure of MHC haplotypes. More than

[1]Department of Molecular Pathogenesis, Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan; [2]Department of Legal Medicine, Shinshu University School of Medicine, Matsumoto, Japan; [3]Department of Pharmacy, Shinshu University Hospital, Matsumoto, Japan; [4]Department of Molecular Life Science, Tokai University School of Medicine, Isehara, Japan and [5]Laboratory of Genome Diversity, School of Biomedical Science, Tokyo Medical and Dental University, Tokyo, Japan
Correspondence: Professor A Kimura, Department of Molecular Pathogenesis, Medical Research Institute, Tokyo Medical and Dental University, 1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8510, Japan.
E-mail: akitis@mri.tmd.ac.jp

Received 17 November 2008; revised 3 February 2009; accepted 4 February 2009; published online 27 February 2009

1000 polymorphic microsatellite markers have been described within the *HLA* region.[13–16] The microsatellite markers showed considerable polymorphism and strong LD with particular alleles of classical *HLA* loci composing of well-defined extended *HLA* haplotypes.[17] The *HLA* haplotypes can be separated into several blocks, including a haplotype block containing the *HLA-Cw* and *-B* genes, just centromeric to the MHC class I region, which is known to be one of the highest polymorphic loci in the human genome.[18] In this study, we analyzed the microsatellite diversity surrounding the *HLA-Cw/-B* loci to investigate the haplotype lineages in a Japanese population.

MATERIALS AND METHODS

Study population and genotyping methods

The study population consisted of 261 Japanese individuals selected at random. All subjects gave informed consent and the study was approved by the Research Ethics Committee of Medical Research Institute, Tokyo Medical and Dental University and Tokai University School of Medicine. Complete genotyping was achieved for classical *HLA* genes and nine microsatellite markers from all individuals were enrolled in this study. Deviation from the Hardy–Weinberg equilibrium was tested for each *HLA* locus and each microsatellite marker. None of the selected markers showed significant ($P<0.05$) deviation from the Hardy–Weinberg equilibrium. High-resolution *HLA* genotyping (at four-digit allele resolution) was carried out with a sequence-based typing method at the class I genes (*HLA-A*, *-B* and *-Cw*) as recommended by the 13th International Histocompatibility Workshop protocols (http://www.ihwg.org/) and/or manufacturer's instructions (Forensic Analytical, Hayward, CA, USA). When an ambiguity in the genotype assignment was observed in the sequence trace data, genotype was predicted from the allele frequency and LD information in the Japanese.[19] DNA regions spanning the microsatellite polymorphisms were amplified by PCR using primer pairs under the conditions listed in Table 1, and the sequenced reference B-cell line samples, COX and PGX, were used as standard for sizing assignment of microsatellite.[20] In addition, to show the motif variation at *C1_2_5* locus, we sequenced the PCR products obtained from each subject on both strands. The number of repeat units was determined by the direct sequencing along with the fragment length analysis. A part (about 38%) of the subjects was also investigated for the *C1_2_5* allele by cloning the PCR products using the TA cloning kit (Invitrogen, Carlsbad, CA, USA). Data from the cloning of *C1_2_5* were completely consistent with the genotyping data obtained from the direct sequencing method.

Phylogenetic analysis

Sequence data on the *HLA-Cw* alleles (exon 4) were obtained from the IMGT/HLA sequence database (http://www.ebi.ac.uk/imgt/hla/index.html). Sequence alignments of the alleles were created by using GENETYX version 8.1.2 (GENETYX CORPORATION, Tokyo, Japan). Phylogenetic analyses were performed using the unweighted pair group method using arithmetic average (UPGMA) by the MEGA software Version 4.0 (http://www.megasoftware.net/).

Statistical analysis

Deviation from the Hardy–Weinberg equilibrium was tested for all marker loci by using the PyPop v.0.6.0 software package (http://www.pypop.org/).[21] The expectation–maximization algorithm implemented in the 'haplo_stat' package for R statistics software (http://www.r-project.org/)[22] was used to construct haplotypes and estimate their frequencies. The strength of pairwise LD between the alleles of classical *HLA* genes and/or microsatellite markers was quantified by two LD coefficients, D' and r^2, through the add-on R software package 'genetics'.[23] We also evaluated the associations between the *HLA-B* and *HLA-Cw* alleles by sensitivity and specificity; sensitivity was defined as the probability of observing the *HLA-B* allele when a particular *HLA-Cw* allele was observed, whereas specificity was the probability of not observing the *HLA-B* allele in the absence of the particular *HLA-Cw* allele. The long-range association was investigated by the extended haplotype homozygosity (EHH) statistic that was calculated according to the formula developed by Sabeti *et al*.[24] Overall LDs between two loci were estimated by using two statistics, Hedrick's multi-allelic D'[25] and Cramer's *V*.[26] When there are only two alleles per locus, Cramer's *V* is equivalent to the correlation coefficient between the two loci. Statistical significance of the LD between pairs of loci was tested using a permutation test with 1000 permutations for each locus pair.

RESULTS

Association between *HLA-B* and *-Cw* gene loci

Significant associations between the alleles of *HLA-Cw* and *HLA-B* genes were found among 261 Japanese individuals as expected from the physical proximity of *HLA-Cw* and *-B* (85 kb). Of the 75 different

Table 1 Primer sets for microsatellite genotyping around the *HLA-B/-Cw* loci

Marker name	*Position*[a]	*Repeat unit*	*PCR product size [bp]*	*Primer sequence (5′–3′)*	*Dye*	*PCR condition*[b]
C2_4_4	31697425–31697663	(GAAA)	181–281	GGCTTGACTTGAAACTCAGAGACC	Hex	i
				TTATCTACTTATAGTCTATCACGG	—	
C1_3_1	31884120–31884408	(TTG)	279–345	CAGTGACAAGCACCTGGCAC	Tet	i
				GCCAGATGTGGTGGCATGC	—	
C1_2_5	31367081–31367280	(CA)	178–220	CAGTAGTAAGCCAGAAGCTATTAC	6-Fam	i
				AAGTCAAGCATATCTGCCATTTGG	—	
C1_4_1	31439129–31439353	(AAAC)	171–271	CGAGAGAACAACTGGCAGGACTG	6-Fam	i
				GACAGTCCTCATTAGCGCTGAGG	—	
MIB	31457335–31457670	(CA)	326–356	CTACCATGACCCCCTTCCCC	Hex	i
				CCACAGTCTCTATCAGTCCA	—	
STR_MICA	31488069–31488251	(GCT)	179–194	CCTTTTTTTCAGGGAAAGTGC	6-Fam	i
				CCTTACCATCTCCAGAAACTGC	—	
C1_2_A	31579685–31579926	(CA)	234–264	AATAGCCATGAGAAGCTATGTGGGGGAG	6-Fam	ii
				CTACCTCCTTGCCAAACTTGCTGTTTGTG	—	
TNFa	31643387–31643503	(AC)	61–161	CCTCTCTCCCCTGCAACACACA	6-Fam	i
				GCCTCTAGATTTCATCCAGCCACA	—	
TNFd	31664102–31664231	(TC)	131–137	AGATCCTTCCCTGTGAGTTCTGCT	Hex	i
				CATAGTGGGACTCTGTCTCCAAAG	—	

[a] The chromosome 6 genomic sequences was used as a reference.
[b] The PCR was carried out in an ABI9700 thermal cycler under the following conditions: (i) 12 min at 95 °C followed by 35 cycles of 95 °C for 30 s, 55 °C for 45 s, 72 °C for 1 min and final extension for 10 min at 72 °C; (ii) 2 min at 94 °C followed by 30 cycles of 94 °C for 1 min, 55 °C for 1 min, 72 °C for 2 min, with an additional 5 min final extension at 72 °C.

Table 2 Association performance of *HLA-Cw/-B* haplotypes in a Japanese population

HLA-B/-Cw haplotype	*Haplotype frequency*	*D′*	*Hill's r^2*	*Sensitivity[a] (%)*	*Specificity[b] (%)*
*Cw*1202-B*5201*	0.144	0.98	0.93	97.4	98.9
*Cw*0102-B*5401*	0.092	0.94	0.32	94.1	84.9
*Cw*0303-B*1501*	0.036	0.40	0.12	45.2	92.3
*Cw*0304-B*4002*	0.054	0.77	0.34	77.8	93.4
*Cw*0102-B*4601*	0.057	0.91	0.19	90.9	81.8
*Cw*0702-B*0702*	0.061	0.96	0.43	97.0	93.0
*Cw*0303-B*3501*	0.042	0.66	0.24	68.8	93.1
*Cw*0304-B*4001*	0.034	0.53	0.14	58.1	91.4
*Cw*1403-B*4403*	0.057	1.00	0.93	96.8	99.8
*Cw*1402-B*5101*	0.048	0.06	0.76	80.6	99.8

[a]Sensitivity: the probability of observing the particular *HLA-Cw* allele given the presence of the particular *HLA-B* allele.
[b]Specificity: the probability of not observing the particular *HLA-Cw* allele given the absence of the particular *HLA-B* allele.

HLA-B and *-Cw* allele combinations observed, 10 were relatively common with haplotype frequency above 3%, by which 63% of the Japanese panels could be explained (Table 2). Two combinations, *Cw*1202-B*5201* and *Cw*1403-B*4403*, showed high correlations (over 0.90) for sensitivity, specificity, *D′* and r^2. In contrast, *HLA-Cw/-B* haplotypes containing *Cw*0102*, *Cw*0303* and *Cw*0304* showed less association, although these haplotypes could account for a considerable part in the Japanese population, because these *HLA-Cw* alleles composed of several haplotypes with different *HLA-B* alleles.

Long-range haplotype around the *HLA-B* and *-Cw* genes

To analyze a long-range structure of the region, EHH analysis was performed, which enabled us to estimate the length of LD from the alleles of a landmark locus. As illustrated in Figure 1, each EHH profile within the 300 kb from the landmark tended to decline the LD with increasing distance from the landmark as expected. However, the pattern of EHH varied substantially, depending on the allele at the landmark locus and on the two-locus haplotype. The haplotypes landmarked by the alleles of *HLA-Cw* and *-B* genes extended longer to telomeric side (MHC class I region) and centromeric side (MHC class II region), respectively (Figures 1a–d). Nevertheless, *HLA-Cw/-B* haplotypic combinations (for example, *Cw*1202* and *B*5201*, *Cw*1403* and *B*4403*) formed by almost one-to-one correspondence showed the long-range LD. In clear contrast, others (for example, *Cw*0102*, *Cw *0303* and *Cw *0304*) with highly diverged combinations rapidly diminished the EHH score even within approximately 100 kb around the landmark locus (Figure 1b). As a rapid EHH decay was found at the *C1_2_5* locus around 22.1 kb centromeric to the *HLA-Cw* locus, we examined the EHH pattern from the landmark of two-locus haplotype extended from *C1_2_5* to either *HLA-B* or *-Cw* locus. The degree of EHH decay from the haplotypes of *HLA-B* or *-Cw* coupled with *C1_2_5* showed a similar tendency to that obtained from the landmark of *HLA-B* or *-Cw* alone. Interestingly, it was found that the EHH pattern was different between *Cw*0702* and *Cw*0303* even though these two *HLA-Cw* alleles were linked to the identical allele, *C1_2_5*200* (Figure 1d). As expected, the EHH scores of *HLA-Cw/-B* haplotypes tended to maintain a long-range LD extending centromeric and teromeric to the landmark locus (Figure 1e). These observations suggested that the diversity of *C1_2_5* locus at the nucleotide level was well correlated with the lineage of the *HLA-Cw/-B* haplotype.

Structural analysis of *C1_2_5* marker

To further delineate the haplotypic structure of the *HLA-Cw/-B* region, we focused on the motif structure of a microsatellite marker, *C1_2_5*, which was located between *HLA-B* and *HLA-Cw*. Sequencing analysis of *C1_2_5* revealed four motifs consisting of nucleotide substitutions in addition to gain or loss of CA repeat units (Figure 2a). These substitutions *per se* were observed within the repeat tract and hence did not change the size of PCR fragments, whereas the differences of the motif structure provided us with the additional information on diversity, as exemplified by *C1_2_5*200* and *C1_2_5*218*. Using these data, genetic associations between *C1_2_5* alleles and individual *HLA-Cw/-B* alleles were investigated to characterize the diversity of HLA haplotypes. The $(CA)_n$CTCA and $(CA)_4$AA$(CA)_5$AA$(CA)_n$CTCA motifs were in tight LD with *Cw*0801* and *Cw*0102*, respectively, and the majority of *C1_2_5* alleles showed strong LD, with particularly *HLA-Cw* alleles, but there were several exceptions. For example, *Cw*0304* was in LD with three different *C1_2_5* variations, $(CA)_4$AA$(CA)_{19}$CTCA, $(CA)_4$AA$(CA)_{21}$CTCA and $(CA)_4$AA$(CA)_{23}$TACACTCA. The former two variations forming the identical *HLA-Cw/-B* haplotype, *Cw*0304-B*4002*, should be derived from the same repeat motif. In contrast, the third variation with different motif was linked to a different HLA-B allele, *B*4001*, forming the *Cw*0304-B*4001* haplotype.

Phylogenetic relationship between alleles of *C1_2_5* and *HLA-Cw*

The mutation rate of SNP was estimated to be 10^{-8} per generation, whereas that of microsatellite was between 10^{-5} and 10^{-3} per generation.[27] Relationships among *C1_2_5* alleles with four motifs were phylogenetically analyzed (Figure 2a). Of four major motifs, the simplest structure was $(CA)_n$CTCA, observed in short alleles of both *C1_2_5*188* and **192*. All other *C1_2_5* alleles had a (CA) to (AA) change at the fifth CA unit, resulting in a motif sequences $(CA)_4\underline{A}A$ $(CA)_n$, interrupting the CA repeat array. In addition, *C1_2_5* alleles containing the interrupting sequence, $(CA)_4$AA$(CA)_n$, can be subdivided into two different motifs as follows; $(CA)_4$AA$(CA)_n$ $\underline{T}$ACACTCA resulted from a (CA) to (TA) change in 3′-side of the CA repeat and $(CA)_4$AA$(CA)_5\underline{A}$A$(CA)_n$CTCA resulted from CA-to-AA change at the 11th unit. As all microsatellite motifs shared the simplest motif, it was assumed that $(CA)_n$CTCA was the core structure of the *C1_2_5* microsatellite. On the other hand, to investigate the relationships of lineage between the *C1_2_5* motif structures and the neighboring SNPs, we constructed a phylogenetic tree using the exon 4 sequences of *HLA-Cw* alleles, which encoded the α3 domain, to exclude the effects of selective pressure acting on the peptide-binding domain (Figure 2b). It was found that the relationship among *C1_2_5* alleles (microsatellite lineage) was not always concordant with the relationship of *HLA-Cw* alleles (SNPs lineage).

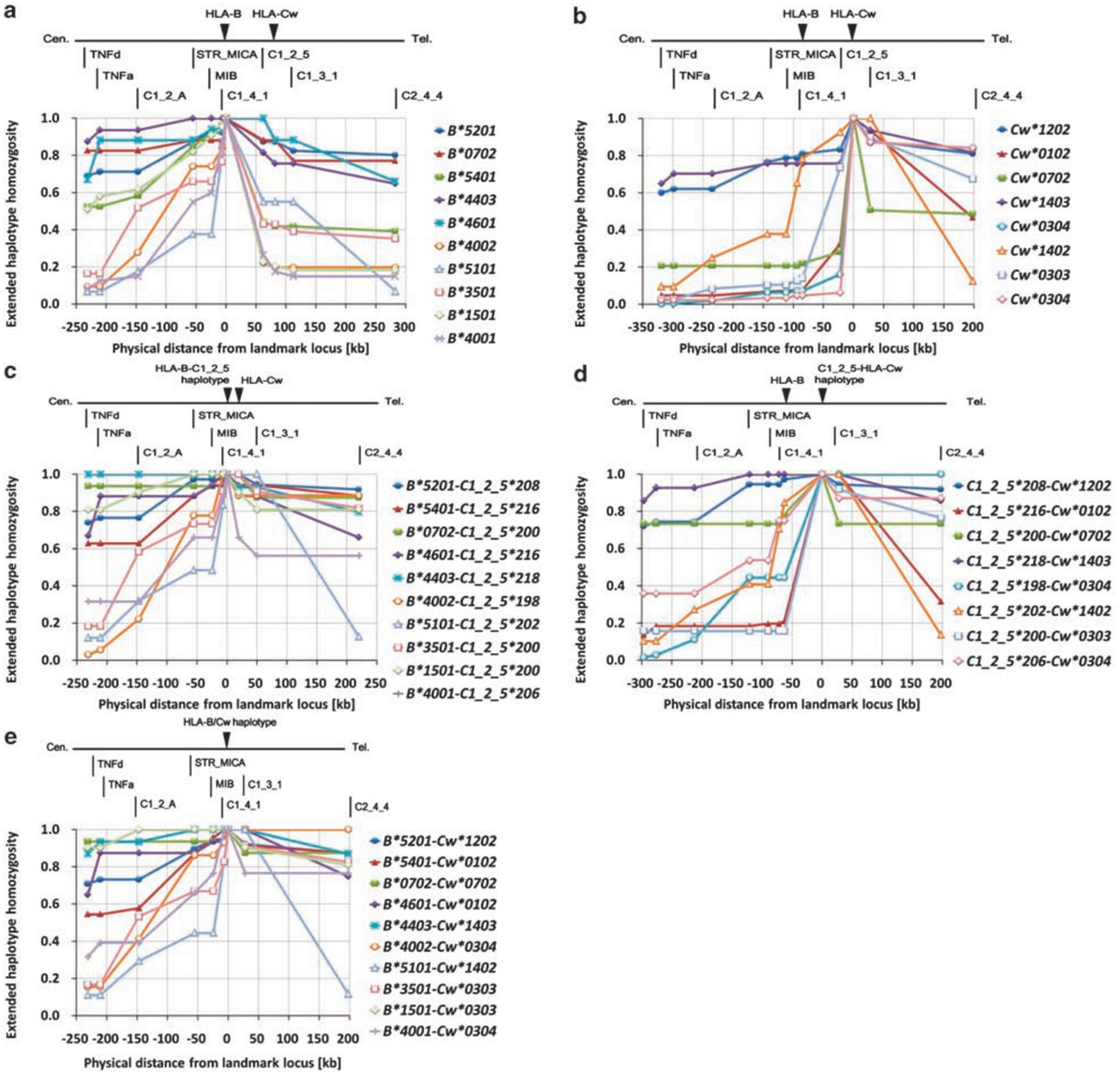


Figure 1 Long-range haplotype test using classical *HLA* genes and microsatellite markers. Each plot represents the extended haplotype homozygosity (EHH) values spanning about 200–300 kb from alleles at two landmark loci, (**a**) *HLA-B* and (**b**) *HLA-Cw*, and three two-locus haplotypes, (**c**) *HLA-B-C_1_2_5*, (**d**) *C1_2_5-HLA-Cw* and (**e**) *HLA-B-HLA-Cw*, in both directions. Vertical lines and arrowheads over the map indicate the locations of microsatellite markers and *HLA* loci, respectively. The gene map was obtained from the Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/HGP/Chr6/MHC.shtml). The physical distances are given in kb, with negative and positive numbers used for locations proximal to and distal from the landmark, respectively.

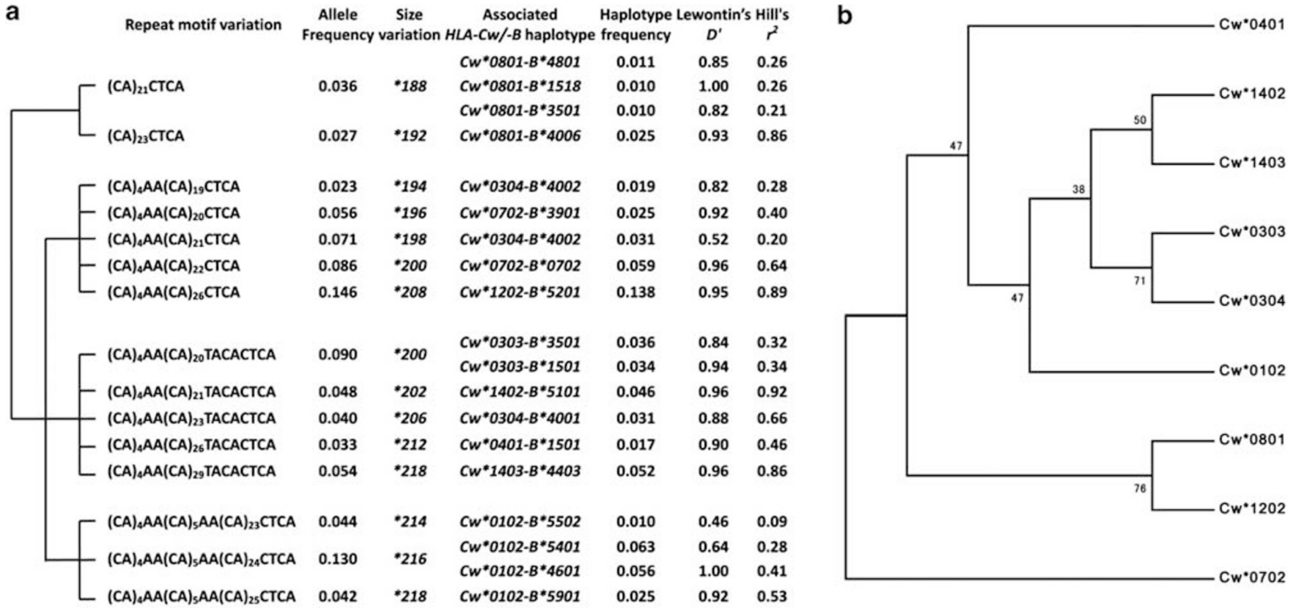
Multiallelic analysis of LD between *C1_2_5* and its franking *HLA* genes

As the EHH analysis was focused on the LD among specific pairs of alleles and haplotypes with relatively high frequency ($>3\%$), we also evaluated overall LDs between two loci among *HLA-B*, *-Cw* and *C1_2_5* to figure out the overall nature of the LD structure in this region (Table 3). It was found that the *C1_2_5* locus, at both the allele level and the motif level, showed stronger LD with *HLA-Cw/-B* haplotype than with either *HLA-B* or *-Cw* locus. These observations suggested that the divergence of *C1_2_5* locus reflected its tight association with the *HLA-Cw/-B* haplotype rather than the association with *HLA-Cw* alleles or *HLA-B* alleles.

DISCUSSION

In this study, we investigated whether a microsatellite marker adjacent to the most polymorphic *HLA-Cw/-B* loci could provide us with information to delineate the haplotype lineage. We found that the *C1_2_5* microsatellite was highly variable by three substitutions within the CA repeat array in addition to the number of CA repeats. The unique polymorphic patterns at *C1_2_5* locus were well correlated with the *HLA-Cw/-B* haplotypes. It was also shown that the simple analysis of fragment-size variation should overlook the nature of microsatellite variations.

The structure of repeat motif was attractive from the evolutional viewpoint because the microsatellite and *HLA-Cw* alleles appeared to



Repeat motif variation	Allele Frequency	Size variation	Associated *HLA-Cw/-B* haplotype	Haplotype frequency	Lewontin's D'	Hill's r^2
			*Cw*0801-B*4801*	0.011	0.85	0.26
$(CA)_{21}$CTCA	0.036	**188*	*Cw*0801-B*1518*	0.010	1.00	0.26
			*Cw*0801-B*3501*	0.010	0.82	0.21
$(CA)_{23}$CTCA	0.027	**192*	*Cw*0801-B*4006*	0.025	0.93	0.86
$(CA)_4AA(CA)_{19}$CTCA	0.023	**194*	*Cw*0304-B*4002*	0.019	0.82	0.28
$(CA)_4AA(CA)_{20}$CTCA	0.056	**196*	*Cw*0702-B*3901*	0.025	0.92	0.40
$(CA)_4AA(CA)_{21}$CTCA	0.071	**198*	*Cw*0304-B*4002*	0.031	0.52	0.20
$(CA)_4AA(CA)_{22}$CTCA	0.086	**200*	*Cw*0702-B*0702*	0.059	0.96	0.64
$(CA)_4AA(CA)_{26}$CTCA	0.146	**208*	*Cw*1202-B*5201*	0.138	0.95	0.89
$(CA)_4AA(CA)_{20}$TACACTCA	0.090	**200*	*Cw*0303-B*3501*	0.036	0.84	0.32
			*Cw*0303-B*1501*	0.034	0.94	0.34
$(CA)_4AA(CA)_{21}$TACACTCA	0.048	**202*	*Cw*1402-B*5101*	0.046	0.96	0.92
$(CA)_4AA(CA)_{23}$TACACTCA	0.040	**206*	*Cw*0304-B*4001*	0.031	0.88	0.66
$(CA)_4AA(CA)_{26}$TACACTCA	0.033	**212*	*Cw*0401-B*1501*	0.017	0.90	0.46
$(CA)_4AA(CA)_{29}$TACACTCA	0.054	**218*	*Cw*1403-B*4403*	0.052	0.96	0.86
$(CA)_4AA(CA)_5AA(CA)_{23}$CTCA	0.044	**214*	*Cw*0102-B*5502*	0.010	0.46	0.09
$(CA)_4AA(CA)_5AA(CA)_{24}$CTCA	0.130	**216*	*Cw*0102-B*5401*	0.063	0.64	0.28
			*Cw*0102-B*4601*	0.056	1.00	0.41
$(CA)_4AA(CA)_5AA(CA)_{25}$CTCA	0.042	**218*	*Cw*0102-B*5901*	0.025	0.92	0.53

Figure 2 Phylogenetic relationship between *C1_2_5* motif and *HLA-Cw/-B* haplotype. (**a**) Phylogenetic tree predicted from the repeat motif structures at the *C1_2_5* locus. *HLA-Cw/-B* haplotypes associated with *C1_2_5* alleles were indicated with two LD coefficients, D' and r^2. Representative *HLA-B* alleles were shown. (**b**) Phylogenetic analysis of exon 4 sequences of *HLA-Cw*. Phylogenetic tree was constructed by the UPGMA method. The numbers for interior branches refer to the bootstrap values in percentage with 1000 replications.

Table 3 Overall LD among *HLA-B*, *HLA-Cw* and *C1_2_5* loci

LD pair[a]	*Hedrick's multiallelic D′*	*Cramer's V*
HLA-B locus and C1_2_5 locus	0.85	0.70
HLA-B locus and C1_2_5 motif	0.88	0.60
C1_2_5 locus and HLA-Cw locus	0.87	0.64
C1_2_5 motif and HLA-Cw locus	0.91	0.73
HLA-Cw locus and HLA -B locus	0.88	0.82
HLA-Cw/-B haplotype and C1_2_5 locus	0.94	0.85
HLA-Cw/-B haplotype and C1_2_5 motif	0.95	0.74

Overall LDs for each pair were statistically significant ($P<0.05$).
[a] C1_2_5 locus and C1_2_5 motif indicate allele (fragment size) and repeat motif structure, respectively.

co-evolve. For example, the change of *C1_2_5* found in the *Cw*0304-B*4002* haplotypes was attributable to the differences in the number of repetitive units, which can be explained by a strand-slippage mechanism. On the other hand, the difference of repeat motifs in *C1_2_5*198* and *C1_2_5*206* associated with the identical allele, *Cw*0304*, was characterized by distinct *HLA-B* alleles, *B*4002* and *B*4001*, respectively. It was unlikely that these two *Cw*0304*-linked haplotypes were shaped by a simple recombination event between *HLA-Cw* and *-B* loci, as the motif structures of *C1_2_5* were different between them. Instead, *Cw*0304* might originally exist in two different haplotype lineages.

Comparison of EHH profile showed that the length of LD varied depending on the HLA haplotypes. One possible explanation for the variation includes the diversity of pairing between the alleles of *HLA-B* and *-Cw*. Indeed, the *HLA* allele with a short-range LD profile showed larger diversity due to the repeated recombination events over time, thereby providing the LD decay between the landmark allele and the linked markers. On the other hand, haplotypes with a long-range LD profile might be of recent origin. In general, human genetic geography showed high continuity, and it is well known that the MHC haplotypes in neighboring populations were introduced to Japan through multiple routes.[28] Therefore, the MHC haplotype structures in the Japanese population might be shaped by multiple immigrations.

Each repeat motif observed in the *C1_2_5* locus was in tight LD with a particular *HLA-Cw* allele and in part with an *HLA-B* allele, which consisted of *HLA-Cw/-B* haplotypes. The mutation rate at a microsatellite is known to depend on the intrinsic features, including repeat number, length and motif size.[29] For example, microsatellites with greater number of repeats showed higher mutation rates due to the increased probability of slippage.[30] In contrast, interruption of perfect repeat array had a great impact on the stability of microsatellite alleles.[31] Indeed, interrupted motif within repeat tracts that were correlated with *HLA-DR/-DQ* haplotypes was described for *DQCAR*.[32]

In conclusion, we revealed that unique mutational dynamics at *C1_2_5* locus could serve as a useful resource for tracing haplotype lineage in the Japanese population. Analysis of *C1_2_5* structures along with *HLA-Cw/-B* haplotypes in other ethnic groups will show the lineages of haplotypes. Statistical methodology for predicting the *HLA* allele and its haplotype carried on the chromosome have been established using informative SNPs inside and/or outside the *HLA* genes.[33,34] However, the use of bi-allelic SNPs as a marker requires more efforts to obtain the information than the use of multi-allelic microsatellite markers, because many *HLA* alleles show a mosaic structure shaped by multiple polymorphic backgrounds. Microsatellite markers will shed light on the haplotype lineage in a different perspective from the SNP-based tagging approach.

ACKNOWLEDGEMENTS

We are grateful to Atsuko Shigenari, Eri F Kikkawa, Hisako Kawata, Norihisa Sugita, Daisuke Sumiyama, Ryoko Ohkubo, Rui-Ping Dong, Shuji Hoshino, Masayuki Yoshida, Shin-ichiro Yasunaga and Yukiji Date for their contributions in genotyping. This work was supported in part by Grant-in-Aids from the Ministry of Education, Science, Sports, Culture and Technology (MEXT) of Japan, the program of Founding Research Centers for Emerging and Reemerging Infection Disease supported by MEXT, Japan, and research grants from the Ministry of Health, Labor and Welfare, Japan, and from the Japan Health Sciences Foundation.

1 Horton, R., Wilming, L., Rand, V., Lovering, R. C., Bruford, E. A., Khodiyar, V. K. *et al.* Gene map of the extended human MHC. *Nat. Rev. Genet.* **5,** 889–899 (2004).
2 Cooke, G. S. & Hill, A. V. Genetics of susceptibility to human infectious disease. *Nat. Rev. Genet.* **2,** 967–977 (2001).
3 Rudolph, M. G., Stanfield, R. L. & Wilson, I. A. How TCRs bind MHCs, peptides, and coreceptors. *Annu. Rev. Immunol.* **24,** 419–466 (2006).
4 Little, A. M. & Parham, P. Polymorphism and evolution of HLA class I and II genes and molecules. *Rev. Immunogenet.* **1,** 105–123 (1999).
5 Meyer, D., Single, R. M., Mack, S. J., Erlich, H. A. & Thomson, G. Signatures of demographic history and natural selection in the human major histocompatibility complex loci. *Genetics.* **173,** 2121–2142 (2006).
6 Miretti, M. M., Walsh, E. C., Ke, X., Delgado, M., Griffiths, M., Hunt, S. *et al.* A high-resolution linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **76,** 634–646 (2005).
7 Shichi, D., Kikkawa, E. F., Ota, M., Katsuyama, Y., Kimura, A., Matsumori, A. *et al.* The haplotype block, NFKBIL1-ATP6V1G2-BAT1-MICB-MICA, within the class III-class I boundary region of the human major histocompatibility complex may control susceptibility to hepatitis C virus-associated dilated cardiomyopathy. *Tissue Antigens* **66,** 200–208 (2005).
8 Shiina, T., Ota, M., Shimizu, S., Katsuyama, Y., Hashimoto, N., Takasu, M. *et al.* Rapid evolution of major histocompatibility complex class I genes in primates generates new disease alleles in humans via hitchhiking diversity. *Genetics* **173,** 1555–1570 (2006).
9 Ahmad, T., Neville, M., Marshall, S. E., Armuzzi, A., Mulcahy-Hawes, K., Crawshaw, J. *et al.* Haplotype-specific linkage disequilibrium patterns define the genetic topography of the human MHC. *Hum. Mol. Genet.* **12,** 647–656 (2003).
10 Horton, R., Gibson, R., Coggill, P., Miretti, M., Allcock, R. J., Almeida, J. *et al.* Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project. *Immunogenetics* **60,** 1–18 (2008).
11 Nagy, M., Entz, P., Otremba, P., Schoenemann, C., Murphy, N. & Dapprich, J. Haplotype-specific extraction: a universal method to resolve ambiguous genotypes and detect new alleles—demonstrated on HLA-B. *Tissue Antigens* **69,** 176–180 (2007).
12 Subirana, J. A. & Messeguer, X. Structural families of genomic microsatellites. *Gene* **408,** 124–132 (2008).
13 Matsuzaka, Y., Makino, S., Nakajima, K., Tomizawa, M., Oka, A., Bahram, S. *et al.* New polymorphic microsatellite markers in the human MHC class III region. *Tissue Antigens* **57,** 397–404 (2001).
14 Matsuzaka, Y., Makino, S., Nakajima, K., Tomizawa, M., Oka, A., Kimura, M. *et al.* New polymorphic microsatellite markers in the human MHC class II region. *Tissue Antigens* **56,** 492–500 (2000).
15 Foissac, A., Salhi, M. & Cambon-Thomsen, A. Microsatellites in the HLA region: 1999 update. *Tissue Antigens* **55,** 477–509 (2000).
16 Cullen, M., Malasky, M., Harding, A. & Carrington, M. High-density map of short tandem repeats across the human major histocompatibility complex. *Immunogenetics* **54,** 900–910 (2003).
17 Malkki, M., Single, R., Carrington, M., Thomson, G. & Petersdorf, E. MHC microsatellite diversity and linkage disequilibrium among common HLA-A, HLA-B, DRB1 haplotypes: implications for unrelated donor hematopoietic transplantation and disease association studies. *Tissue Antigens* **66,** 114–124 (2005).
18 Mungall, A. J., Palmer, S. A., Sims, S. K., Edwards, C. A., Ashurst, J. L., Wilming, L. *et al.* The DNA sequence and analysis of human chromosome 6. *Nature* **425,** 805–811 (2003).
19 Saito, S., Ota, S., Yamada, E., Inoko, H. & Ota, M. Allele frequencies and haplotypic associations defined by allelic DNA typing at HLA class I and class II loci in the Japanese population. *Tissue Antigens* **56,** 522–529 (2000).
20 Stewart, C. A., Horton, R., Allcock, R. J., Ashurst, J. L., Atrazhev, A. M., Coggill, P. *et al.* Complete MHC haplotype sequencing for common disease gene mapping. *Genome Res.* **14,** 1176–1187 (2004).
21 Lancaster, A. K., Single, R. M., Solberg, O. D., Nelson, M. P. & Thomson, G. PyPop update-software pipeline for large-scale multilocus population genomics. *Tissue Antigens* **69**(Suppl 1), 192–197 (2007).
22 Warnes, G. R. The genetics package. *R News* **3,** 9–13 (2003).
23 Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M. & Poland, G. A. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet.* **70,** 425–434 (2002).
24 Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419,** 832–837 (2002).
25 Hedrick, P. W. Gametic disequilibrium measures: proceed with caution. *Genetics* **117,** 331–341 (1987).
26 Cramer, H. *Mathematical Methods of Statistics* (Princeton University Press, Princeton, New Jersey, 1946).
27 Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
28 Tokunaga, K., Ishikawa, Y., Ogawa, A., Wang, H., Mitsunaga, S., Moriyama, S. *et al.* Sequence-based association analysis of HLA class I and II alleles in Japanese supports conservation of common haplotypes. *Immunogenetics* **46,** 199–205 (1997).
29 Ellegren, H. Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.* **5,** 435–445 (2004).
30 Lai, Y. & Sun, F. The relationship between microsatellite slippage mutation rate and the number of repeat units. *Mol. Biol. Evol.* **20,** 2123–2131 (2003).
31 Boyer, J. C., Hawk, J. D., Stefanovic, L. & Farber, R. A. Sequence-dependent effect of interruptions on microsatellite mutation rate in mismatch repair-deficient human cells. *Mutat. Res.* **640,** 89–96 (2008).
32 Macaubas, C., Jin, L., Hallmayer, J., Kimura, A. & Mignot, E. The complex mutation pattern of a microsatellite. *Genome Res.* **7,** 635–641 (1997).
33 Leslie, S., Donnelly, P. & McVean, G. A Statistical method for predicting classical HLA alleles from SNP data. *Am. J. Hum. Genet.* **82,** 48–56 (2008).
34 de Bakker, P. I., McVean, G., Sabeti, P. C., Miretti, M. M., Green, T., Marchini, J. *et al.* A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat. Genet.* **38,** 1166–1172 (2006).