THE NEUTRAL THEORY AND NATURAL SELECTION IN THE HLA REGION

Yoko Satta, Yi-Ju Li and Naoyuki Takahata

Department of Biosystems Science, The Graduate University for Advanced Studies, Hayama, Kanagawa 240-0193, Japan

Received 4/3/98 Accepted 4/7/98

TABLE OF CONTENTS

1. Abstract
2. Introduction
3. Genes and nucleotide differences in HLA
4. Intergenic recombination
5. DRB1 allelic lineages and disease association
6. Perspective
7. Acknowledgments
8. References

1. ABSTRACT

Based on available DNA sequence data in the HLA region of 4 Mb, we review the degree of polymorphism at 39 loci of which most are involved in the immune system. The extent of nucleotide differences per silent site differs greatly from locus to locus. It is exceptionally high at classical MHC loci, intermediate at six MHC-related pseudogenes as well as at some loci in class I and II regions, and low in the class III region. Different exons of individual MHC loci show also different degrees of silent polymorphism; high in the exons encoding for the peptide binding region (PBR) and low in the exons encoding for trans-membranes and cytoplasmic tails. The degree of polymorphism within MHC allelic lineages is not much smaller than that between allelic lineages, contrary to the expectation where intra-allelic sequence exchanges are restricted. The observation that many allelic lineages at the HLA-DRB1 locus are combinations of distinct motifs in the beta pleated sheet and alpha helix of PBR indicates that sequence exchanges occur even within exon 2. Semi-quantitative analysis is presented about the rate of sequence exchanges between selected and linked neutral regions, although more sequence information is necessary to make definite conclusions. The extraordinary MHC polymorphism is viewed from the dual function of MHC molecules that controls the acquired immune system.

2. INTRODUCTION

The pattern and degree of polymorphism across the genome are a useful indicator for identifying genes or genomic regions which are subjected to different types of natural selection. In some circumstances, polymorphism may also be used to infer the function of unknown genes. Without the action of natural selection, polymorphism (often measured by the pairwise nucleotide differences at the DNA level) must have evolved in much the same way as what the neutral theory of molecular evolution depicts (1, 2). The larger the effective population size \(N_e\) and the higher the neutral mutation rate, the more polymorphic. In genomic regions experiencing either positive or negative selection (purifying selection), polymorphism is lowered (3, 4), while in those experiencing diversifying or balancing selection, it is enhanced (5-9). These contrasting effects of natural selection are not necessarily confined in the target region per se, but they should be manifest also in a neighboring genomic region in linkage. An important factor is obviously the rate (c) of recombination between two linked regions under study (5, 9).

A pair of neutral genes in an autosomal region segregate for \(2N_e\) generations on average (10), so that if recombination occurs rarely or at rate \(c < 1/N_e\) or less with the target region of natural selection, the linked neutral polymorphism will be affected (11). Since \(N_e\) is estimated as about \(10^8\) for the human population over the past one million years (7, 12), the indirect effect of natural selection extends to the neighboring region with \(c < 0.01\%\) or physical map distance \(\leq 10\) kb if \(1\) cM = \(1\) Mb (13). Thus, if reduced polymorphism is observed in a neutral region, purifying selection might have occurred somewhere in the surrounding left or right 10 kb region during a much shorter period of time than the last \(2 \times 10^9\) generations. On the other hand, if enhanced polymorphism is observed, balancing selection might have been operating throughout a much longer period of time than \(2 \times 10^9\) generations. In this case, the candidate region may be broader than of 20 kb because the efficiency of recombination is reduced in proportion to the number of alleles that are maintained by balancing selection (9).

In humans, the nucleotide differences are 0.08% over 18,844 silent (synonymous) sites for 48 pairs of carefully checked autosomal sequences (“standard”) (12, 14). For 12 additional pairs of unchecked sequences (14), the silent differences become high (0.31% over 6,071 sites), yet they are much smaller than 1%. Although no comparable estimates except for mitochondrial DNA are available in non-human primates (15), the silent differences (1.3% on average) in Drosophila melanogaster and its sibling species suggested that the degree of DNA polymorphism in humans is exceedingly low (14). Notwithstanding, the silent differences at some functional HLA class I and II loci are 50 to 100-fold greater than the standard (16). Convincing evidence accumulated to support the notion that this results from the long lasting operation of balancing selection for non-silent substitutions in the peptide binding region (PBR) of major histocompatibility complex (MHC) molecules (17 - 20).

This review first summarizes the polymorphism at 39 loci dispersed in the HLA region of 4 Mb; some are MHC proper or their pseudogenes and others encode for proteins involved in the immune system. Second, the silent
Polymorphism in HLA

Figure 1. The silent nucleotide differences (p_s) and chromosomal locations of 39 loci in the HLA region of 4 Mb (41 - 43).

The remaining 30 include immunity-related loci in addition to two class I and four class II pseudogenes in which all nucleotide sites are treated as silent. The overall p_s value at these 30 loci reduces to 127/15,132 = 0.84%. Yet, it is ten-fold larger than the standard (P < 0.01). Also, the p_s value of 1.2% over the four class II pseudogenes (DPB2, DQB2, DQA2 and DQB3) is much greater than the standard (P < 0.01). This enhancement may be due to recent cessation of balancing selection, inter-locus sequence exchanges (unequal crossover or gene conversion), or tight linkage to classical class II loci. The first possibility is unlikely because the presence of DPB2 or DQB2 orthologs in various primates (26) suggests that the two loci were inactivated long before their alleles were generated. The second possibility is inconsistent with the monophyletic relationships between these pseudogenes and functional paralogs (26); with inter-locus sequence exchanges, the relationships must be para- or polyphyletic. Thus, the relatively large p_s value for these pseudogenes is attributed to indirect effects of balancing selection operating on nearby polymorphic class II loci.

The LMP2 and LMP7 genes encode for subunits of a proteasome, while the TAP1 and TAP2 genes encode for ABC transporter proteins, all being involved in processing of proteins into peptides that are loaded onto class I molecules (27). Unlike the rat TAP ortholog (28, 29), these four loci in HLA are rather monomorphic: The p_s value of 0.19% over the 1,492 silent sites is not different from the standard. The DNA and DOB genes as well as DMA and DMB genes encode for alpha and beta polypeptide chains, respectively, each pair of chains forming heterodimers like classical class II chains (30 - 32). The function of the DM heterodimer is thought to facilitate the exchange of class II associated invariant chains for peptides that are generated in lysosomes from self and non-self proteins (33), while the DQ heterodimer inhibits such a catalytic action of DM (34, 35).

Among 13 loci in the class III region, CYP21B (steroid 21-hydroxylase involved in biosynthesis of cortisol and aldosterone) and HSP70-2 (the major heat-inducible chaperone of the HSP70 group) are shown to have undergone frequent sequence exchanges with nearby paralogs (36, 37). As a result, the nucleotide differences are rather large so that in what follows, these two loci are excluded from consideration. In the remaining 11 loci, the average p_s value of 0.22% at 3,329 sites is not different from the standard. However, the non-silent differences (p_n) of 0.24% are greater than 0.024% at the standard loci (P < 0.01). It turns out that MICA (MHC class I chain-related A) is unusual in that among 16 alleles at the locus, there are 22 non-silent segregating sites out of 618 and the alleles are different from each other by seven such sites on average. The MICA and MICB genes encode for non-classical (or class Ib) MHC molecules (38 - 41) and may be recognized mainly in intestinal epithelium by T lymphocytes with gamma-delta T cell receptors (Tcr) (42). The significantly large p_n value of MICA may be associated with this putative function. The class III region also encodes for several other immunity-related genes (27) such as in the complement cascade (C4β, Bf) and the regulation of T lymphocyte development and function (TNF).
Polymorphism in HLA

Table 1. The per-site nucleotide differences (p) and the number of segregating sites (S) over L sites at each of 39 loci in the HLA coding region.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Silent</th>
<th>Non-silent</th>
<th>References or accession numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(#genes)</td>
<td>K_s/L_s = p_s (%)</td>
<td>S_s</td>
</tr>
<tr>
<td>DPB2 (2)</td>
<td>37/1977 = 1.9</td>
<td>37</td>
<td>–</td>
</tr>
<tr>
<td>DPB1 (67)</td>
<td>1/63 = 1.6</td>
<td>6.5</td>
<td>9.1/191 = 4.8</td>
</tr>
<tr>
<td>DPA1 (8)</td>
<td>3.5/59 = 5.9</td>
<td>8.7</td>
<td>5.2/185 = 2.8</td>
</tr>
<tr>
<td>DNA (3)</td>
<td>2/204 = 1.0</td>
<td>3</td>
<td>0/547 = 0</td>
</tr>
<tr>
<td>DMA (4)</td>
<td>0/71 = 0</td>
<td>0</td>
<td>3/208 = 1.5</td>
</tr>
<tr>
<td>DMB (7)</td>
<td>0.5/60 = 0.83</td>
<td>2</td>
<td>2/177 = 1.1</td>
</tr>
<tr>
<td>LMP2 (3)</td>
<td>0.5/173 = 0.24</td>
<td>1</td>
<td>0.4/463 = 0.09</td>
</tr>
<tr>
<td>TAP1 (5)</td>
<td>0.3/631 = 0.05</td>
<td>1</td>
<td>1.8/610 = 0.11</td>
</tr>
<tr>
<td>LMP7 (5)</td>
<td>0/172 = 0</td>
<td>0</td>
<td>0/509 = 0</td>
</tr>
<tr>
<td>TAP2 (3)</td>
<td>2/516 = 0.39</td>
<td>3</td>
<td>2/1350 = 0.15</td>
</tr>
<tr>
<td>DOB (2)</td>
<td>0/214 = 0</td>
<td>0</td>
<td>1.2/605 = 0.17</td>
</tr>
<tr>
<td>DQB2 (8)</td>
<td>1.4/201 = 0.69</td>
<td>3</td>
<td>–</td>
</tr>
<tr>
<td>DQA2 (2)</td>
<td>2/2504 = 0.84</td>
<td>21</td>
<td>–</td>
</tr>
<tr>
<td>DQB3 (2)</td>
<td>15/1407 = 1.1</td>
<td>15</td>
<td>–</td>
</tr>
<tr>
<td>DQB1 (27)</td>
<td>6.2/73 = 8.6</td>
<td>17.5</td>
<td>17/212 = 8.1</td>
</tr>
<tr>
<td>DQA1 (15)</td>
<td>9.0/150 = 6.0</td>
<td>26.5</td>
<td>19/426 = 4.5</td>
</tr>
<tr>
<td>DRB1 (135)</td>
<td>3.6/60 = 6.0</td>
<td>24</td>
<td>12/183 = 6.5</td>
</tr>
<tr>
<td>DRA (2)</td>
<td>1/174 = 0.57</td>
<td>1</td>
<td>1.0/504 = 0.20</td>
</tr>
<tr>
<td>PBX-2 (3)</td>
<td>0/338 = 0</td>
<td>0</td>
<td>0.7/953 = 0.07</td>
</tr>
<tr>
<td>RAGE (2)</td>
<td>0.5/338 = 0.15</td>
<td>0.5</td>
<td>3.5/874 = 0.40</td>
</tr>
<tr>
<td>G13 (2)</td>
<td>2/578 = 0.35</td>
<td>2</td>
<td>3/1520 = 0.20</td>
</tr>
<tr>
<td>CYP2B1 (2)</td>
<td>5/405 = 1.23</td>
<td>5</td>
<td>3/1080 = 0.28</td>
</tr>
<tr>
<td>C4B (2)</td>
<td>1/310 = 0.32</td>
<td>1</td>
<td>1/833 = 0.12</td>
</tr>
<tr>
<td>G11 (2)</td>
<td>0/214 = 0</td>
<td>0</td>
<td>0.5/600 = 0</td>
</tr>
<tr>
<td>RD (4)</td>
<td>1.5/274 = 0.54</td>
<td>3</td>
<td>2.6/778 = 0.33</td>
</tr>
<tr>
<td>Bf (2)</td>
<td>0/467 = 0</td>
<td>0</td>
<td>1/1350 = 0.07</td>
</tr>
<tr>
<td>HSP70-2 (2)</td>
<td>7/508 = 1.4</td>
<td>7</td>
<td>2/1410 = 0.14</td>
</tr>
<tr>
<td>TNF (3)</td>
<td>0/191 = 0</td>
<td>0</td>
<td>0.7/508 = 0.13</td>
</tr>
<tr>
<td>BAT1 (3)</td>
<td>0/211 = 0</td>
<td>0</td>
<td>0.7/639 = 0.11</td>
</tr>
<tr>
<td>MICB (7)</td>
<td>0.3/204 = 0.14</td>
<td>1</td>
<td>1.7/618 = 0.28</td>
</tr>
<tr>
<td>MICA (16)</td>
<td>2/204 = 0.96</td>
<td>5</td>
<td>7.2/618 = 1.2</td>
</tr>
<tr>
<td>B (91)</td>
<td>9.7/289 = 3.4</td>
<td>45.4</td>
<td>25/796 = 3.2</td>
</tr>
<tr>
<td>C (35)</td>
<td>8.3/281 = 3.0</td>
<td>46.8</td>
<td>18/779 = 2.3</td>
</tr>
<tr>
<td>SC1 (2)</td>
<td>0/301 = 0</td>
<td>0</td>
<td>8/767 = 1.0</td>
</tr>
<tr>
<td>E (5)</td>
<td>0.5/127 = 0.43</td>
<td>3</td>
<td>1.8/346 = 0.51</td>
</tr>
<tr>
<td>J (3)</td>
<td>7.5/1046 = 0.72</td>
<td>11</td>
<td>–</td>
</tr>
<tr>
<td>A (50)</td>
<td>10/261 = 3.8</td>
<td>38</td>
<td>25/758 = 3.3</td>
</tr>
<tr>
<td>H (6)</td>
<td>18.7/1091 = 1.71</td>
<td>45</td>
<td>–</td>
</tr>
<tr>
<td>G (6)</td>
<td>1.1/195 = 0.56</td>
<td>5</td>
<td>1.3/522 = 0.25</td>
</tr>
</tbody>
</table>

The K and L are the average number of nucleotide differences and the average number of sites in all pairwise comparisons, respectively, and the subscripts stand for silent and non-silent nucleotide substitutions.
Polymorphism in HLA

The average \( p_s \) value over five loci (SC1, E, J, H and G) in the class I region is 0.98% which is greater than the standard (P \( \times 0.01 \)). The SC1 gene is thought to control the cell cycle, while the E and G genes encode for class Ib molecules. The relatively high \( p_s \) value in the class I region is largely attributable to processed pseudogene HLA-H (not to be confused with the renamed gene (HFE) that is responsible for hereditary hemochromatosis (43), more than 3 Mb telomeric from the HLA). Compared with the case of HLA-J which is also a pseudogene, both \( p_s = 1.7\% \) and \( S = 43 \) over 1,067 silent sites at the HLA-H locus appear to be too large. However, there is evidence that a telomeric region of the HLA-A locus is somewhat suppressed in recombination (H. Inoko, personal communication). Although the reason is poorly understood, the reduced recombination may well account for the elevated polymorphism at the HLA-H locus as well as the slightly increased polymorphism at the HLA-G locus.

4. INTERGENIC RECOMBINATION

In order to study the relationships between silent differences (\( p_s \)) and recombination rates (\( c \)), each \( p_s \) value is plotted against the physical map distance (44 - 46) which is measured from the nearest highly polymorphic HLA locus (figure 2). As expected, the longer the distance, the smaller the \( p_s \) value. The following theoretical model (9, 47, 48) is used to examine the relationships more quantitatively. The model assumes that the neutral region and the PBR encoding exon are recombined with rate \( c \) per generation. The model also assumes that any PBR non-silent nucleotide substitution always generates a new allelic lineage which, together with all other pre-existing alleles, is subjected to random genetic drift and balancing selection. The mutation rate per PBR is assumed to be \( 2.7 \times 10^{-6} \) for class I and \( 0.8 \times 10^{-6} \) for class II locus owing to the difference in the number of non-silent sites (25). The per-site mutation rate is \( 2 \times 10^{-9} \) per generation and the generation time is 20 years. If a population has not been demographically stable over time, \( N_e \) is dependent of a time period during which polymorphism has been generated (9). This time period for neutral polymorphism is relatively short, while that for HLA polymorphism is relatively long. The estimate of \( N_e = 10^7 \) is made based on the short-lived neutral polymorphism, whereas that of \( N_e = 10^6 \) is made based on the enhanced polymorphism due to long-lived HLA allelic lineages (7). The reduction in the effective population size might have begun when *Homo erectus* first migrated from Africa about one million years ago (7, 9, 48), although there are alternatives. Figure 2 also depicts the expected level of linked neutral polymorphism. The indirect effect of balancing selection is remarkable in linked neutral regions with \( c < 0.01\% \), or only within 10 kb if 1 cM = 1Mb is postulated (13). Thus, in order to account for large \( p_s \) values by linkage, tightly linked regions must be considered. The relatively large \( p_s \) value at DPB2, DQB3, DNA, DQA2 and DQB2 locus as well as that at MICA, J, H and G or by DPB2, DQB3, DNA, DQA2 and DQB2.

Figure 2. The observed silent nucleotide differences at individual loci are plotted against the physical map distance from the nearest highly polymorphic HLA locus. The solid and dotted curves are computed by the theoretical formulas (9, 45) under the assumption of 1 cM = 1 Mb and 0.1 cM = 1 Mb, respectively. It is assumed that selection intensity is 2% and the number of breeding individuals is \( 10^3 \) and \( 10^4 \) before and after 50,000 generations ago. The values of the per-site neutral mutation rate and the non-silent substitutions rate per PBR per generation are given in text. The discrepancy from the 1 cM = 1 Mb curve is caused by substitutions rate per PBR per generation are given in text. The per-site neutral mutation rate is \( 2 \times 10^{-8} \) per generation and the non-silent substitution rate is \( 2 \times 10^{-6} \) for class II locus owing to the postulate of 1 cM = 1Mb is postulated (13).

5. DRB1 ALLELIC LINEAGES AND DISEASE ASSOCIATION

Long lasting allelic lineages at classical MHC loci permit us to glean insight into molecular mechanisms of the polymorphism, in particular roles of intra-exonic sequence exchanges (recombination or gene conversion). The silent differences within (\( p_b \)) and between (\( p_a \)) allelic lineages are particularly useful for this purpose. Without sequence exchanges among allelic lineages, the silent differences within lineages should be much smaller than those between lineages (figure 3). However, in the absence of information on associations between alleles at a locus and the nearby MHC locus, figure 3 cannot be used. An exception is MHC loci themselves at which the linkage relationships between non-PBR coding exons or introns and the PBR coding exon are certain in some data sets (50, 51).

Both \( p_a \) and \( p_b \) values depend critically on the definition of allelic lineages. In addition to serological methods, sequence motifs as well as phylogenetic analyses can provide reasonable classifications of MHC allelic lineages (9). As an example, we take a close look at...
Theoretically, the best classification of allelic lineages may be either beta or alpha motifs, p\(_s\) is maximized and p\(_w\) is close to p\(_s\) by definition. Theoretically, the best classification of allelic lineages may be obtained when p\(_s\) is maximized and p\(_w\) is minimized. In the beta motif-based classification of allelic lineages, p\(_w\) is smaller than 0.1% in the beta region, which is in accord with the expectation (figure 3), whereas p\(_s\) is more than 3.8% in the alpha region. This contrasting pattern of p\(_w\) in the two regions becomes less conspicuous in the alpha motif-based classification of allelic lineages (3.3% in the beta region and 1.2% in the alpha region). This asymmetry may indicate that the whole or partial alpha motifs have been recombined with distinct beta motifs and that this shuffling generated an enormously large number of DRB1 alleles (figure 5). An alternative is to invoke convergent evolution in alpha motifs (52, 53). In this respect, it is important to note that five (two non-silent associated and three solitary) silent substitutions in the alpha region are shared by distinct allelic lineages defined by beta motifs (data not shown), resulting in the elevated p\(_w\) value in the alpha region. Since it is difficult to invoke convergence for these shared silent substitutions, it is reasonable to conclude that new combinations between beta and alpha motifs have been generated by intra-exonic recombination or gene conversion (22, 23). Nonetheless, it should be kept in mind that if sequence exchanges always occur at the same boundary, the number of alleles maintained in a population becomes rather limited (54). Moreover, were some sequence exchanges not allowed within the alpha region, it does become difficult to account for the relatively large p\(_w\) value even in the alpha-based classification (23). The shared silent substitutions among different alpha motifs support that the tract of sequence exchanges involving the alpha region is at variance. This conclusion also agrees with the finding that although p\(_w\) is larger in the alpha than in the beta region, the phylogenetic relationships in the former are more star-like and more compact than in the latter (figure 4).

Based on the serological classification (22), the p\(_b\) (p\(_w\)) value averaged over about 20 complete DRB1 sequences becomes 8.4% (1.7%) in exon 2, 8.6% (0.07%) in intron 1, 4.8% (0.06%) in intron 2, 6.0% (1.8%) in exon 3 and 1.3% (0.4%) throughout exon 4-6. The extremely small p\(_w\) value in intron 1-2 (22) is due partly to the exclusion of "recombinant" DRB1*0806 as well as partly to the distinction between DRB1*1602 and *15011/*15021 although these alleles are regarded as identical in the present beta motif-based classification. More importantly, it may be noted that the p\(_w\) value tends to be smaller in non-PBR coding regions and this tendency is observed at other classical MHC loci as well (9). Thus, even though intralocus recombination is rare, it might have acted so as to decrease the level of silent polymorphism in non-PBR coding exons. For such reduction to be observed in non-PBR coding exons which are located within a few kilobases away from the PBR coding exon, the rate (c) of intra-exonic sequence exchanges may be of the order of 0.001%, as expected from 1 cM = 1Mb.

Interestingly, the beta and alpha motifs defined above play distinct roles in shaping the physico-chemical environment of peptide binding pockets of MHC molecules (55 - 57) and that some alpha motifs are associated with autoimmune diseases. Rheumatoid arthritis is associated with DR4 subtypes (58). These MHC molecules possess a set of amino acids (Q70/R71A74 by the one letter amino acid code) which participate in forming the P4 pocket. According to the present classification of DRB1 allelic lineages, this alpha motif occurs in some DR1 and DR14 subtypes as well, so that their susceptibility to the disease is also worth being investigated. Insulin-dependent diabetes mellitus (IDDM) is associated with DR3/DR4 (59), while a reduced incidence is found in associations with DR2 (60). In this case too, the DRB1 amino acid residue implicated in susceptibility or resistance is somehow related to that specificity of the P4 pocket (61). It is also reported that tuberculosis leprosy is associated with
Figure 4. The relationships among the major beta or alpha motifs of DRB1 peptide chains. The relationships are reconstructed from non-silent nucleotide differences in each of the beta and alpha region. Each circle represents one non-silent nucleotide substitution. Codon positions used for the beta motifs are 11, 13, 30 and 37 and those for the alpha motifs are 57, 67, 70, 71 and 74.

DR2 in Asia (62 - 64) and DR3 in Venezuela and Surinam (65, 66). Recently, an Arginine (R) or a negatively charged amino acid (D or E) in the P4 pocket is hypothesized to be critical to susceptibility (67). There are only two alpha motifs which do not have any of R, D and E in the P4 pocket: D57L67Q70K71A74 in some DRB1*04 and *14 as well as D57I67Q70A71A74 in most of DRB1*15 and *1309. It is observed that charged residues at position 71 control the charge permitted in the P4 pocket, and that peptides with charged residues in the P4 pocket are capable of binding only if an opposing charge, or no charge, is present at position 71 (68). The same is applied to position 57 in relation to the P9 pocket. This observation lends itself to the biological implication for the present motif-based classification of allelic lineages.

6. PERSPECTIVE

Unfortunately, the loci used are insufficient to make an accurate estimate of recombination rate in HLA. The present analysis shows that it is crucial to examine regions that are tightly linked to MHC loci. If 1 cM = 1 Mb holds true, c is 0.1% between an MHC locus and a region 100 kb apart from it, and recombination with this c value is too frequent for the present data set to demonstrate slightly enhanced polymorphism by indirect effects of balancing selection. Sampling as well as sequencing errors in the estimate of silent polymorphism must also be minimized. An alternative approach may be to use population data of haplotypes.

Recent survey of Siberian populations for class II haplotypes showed that more haplotypes are found in Siberian than in any other population and that most of them have been generated by recombination since the colonization of the subcontinent 40,000 years ago (24). The finding of new combinations between DQA1 and DQB1 molecules that are rarely found in other areas indicated frequent occurrences of recombination during the past 2,000 generations. This is consistent with the expected value of c = 0.02% under 1 cM = 1 Mb, since the two loci are located 20 kb apart (44). Thus, this kind of population data will certainly complement DNA sequence information.

Immunoglobulin (Ig), T cell receptor (Tcr) and MHC genes are encoded for by multigene families (27). A relatively small number of loci and extremely high polymorphism in MHC contrast with a large number of rather monomorphic segmented genes and their somatic rearrangements in Ig and Tcr. It is argued that unlike other multigene families in which concerted evolution is commonly invoked, these immunity-related multigene families have undergone contraction, expansion and subsequent diversification of loci (69 - 71). However, expansion and subsequent diversification of MHC loci must be restrictive owing to the dual roles of MHC molecules; elimination of auto-reactive T cell clones and presentation of non-self peptides to mature T cell repertoire (72 - 74). Because of this dual function in thymus and peripheral, MHC could not have evolved as a multigene family consisting of a large number of functional loci. If individuals express so many different MHC molecules encoded for by a number of diversified duplicated loci, either virtually all T cell clones will be capable of reacting self peptides and be eliminated or the cell surface density of particular MHC molecules will become too low to interact with Tcr. In either case, the immune system will be unable to work properly. On the other hand, appropriate amounts of MHC diversity are necessary to deal with various pathogens that individuals encounter (74). With a relatively small number of functional MHC loci, natural selection has found its way to operate and favored polymorphism. The effect is by no means dramatic every generation (75) and it is only after millions of generations of its operation that the MHC polymorphism has become truly remarkable. It appears that the polymorphism is a sort of molecular compromise in the acquired immune system which is controlled by the dual function of MHC molecules. Couldn’t any better alternative have been invented?
Polymorphism in HLA

Figure 5. Combinations between some beta and alpha motifs that occur in DRB1 peptide chains.

7. ACKNOWLEDGMENTS

This is contribution no. 4 from Department of Biosystems Science and supported by grants from the Monbusho and the Graduate University for Advanced Studies. We thank Drs J. Klein, H. Inoko and Y. Obata for their helpful comments.

8. REFERENCES

Polymorphism in HLA


30. Trowsdale J & A. Kelly: The human HLA class II alpha chain gene DZ alpha is distinct from genes in the DP, DQ and DR subregions. EMBO J 4, 2231-2237 (1985)


Polymorphism in HLA


Key words: Polymorphism, Nucleotide Differences, Balancing Selection, Allelic Lineage, Sequence Motifs, Recombination, Gene Conversion, Population Genetics, Major Histocompatibility Complex

Send correspondence to: Dr. Naoyuki Takahata, Department of Biosystems Science, The Graduate University for Advanced Studies, Hayama, Kanagawa 240-0193, Japan. TEL: +81 468 58 1504, FAX: +81 468 58 1542, E-mail: takahata@soken.ac.jp