[Frontiers in Bioscience, 3, d459-467, April 27, 1998]

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 10^4 for the human population over the past one million years (7, 12), the indirect effect of natural selection extends to the neighboring region with $c \leq 0.01\%$ or physical map distance ≤ 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 2×10^4 generations. On the other hand, if enhanced polymorphism is observed, balancing selection might have been operating throughout a much longer period of time than 2×10^4 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

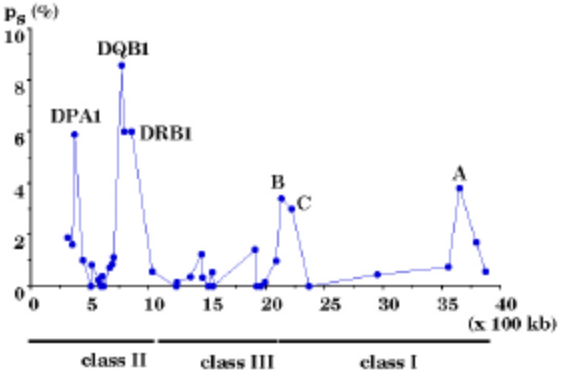

Figure 1. The silent nucleotide differences (p_S) and chromosomal locations of 39 loci in the HLA region of 4 Mb (41 - 43).

differences at a particular locus are plotted against the physical map distance from the nearest highly polymorphic MHC locus and this relationship is used to examine semi-quantitatively whether 1 cM = 1 Mb holds true in the HLA region. Third, the silent differences within and between class II DRB1 allelic lineages are presented to discuss the possibility of intra-exonic recombination or gene conversion (21 - 23). Relevant multi-locus haplotype data from Siberian populations (24) are briefly mentioned. Finally, we provide short comments on MHC-mediated thymic selection in T cell repertoire and some perspectives for the HLA study.

3. GENES AND NUCLEOTIDE DIFFERENCES IN HLA

The degree of polymorphism at 39 loci in HLA is summarized in table 1 in terms of the average number of nucleotide differences and the number of segregating sites at both silent (S_s) and non-silent sites (S_n). It is clear that the degree differs greatly from locus to locus (figure 1). Without direct and indirect effects of natural selection, the silent polymorphism should be relatively uniform over the loci, while the non-silent polymorphism may vary from locus to locus owing to different degrees of functional constraints (2). Nevertheless, the non-silent as well as silent polymorphism at classical class I or class Ia (A, B and C) and class II (DPB1, DPA1, DQB1, DQA1, DRB1 and DRA) loci is enhanced considerably from the standard value of 0.08%. The overall silent differences (p_s) at these class Ia and II loci are 3.37% and 4.20% per site, respectively. If the neutral mutation rate is 10^{-9} per site per year (25), it must have taken about 20 million years or some millions of generations for these silent differences to have accumulated. The different GC content in the HLA region may well affect the neutral mutation rate, but this does not seem sufficient to account for the observed heterogeneity in the extent of polymorphism. No doubt, the enhanced silent polymorphism has resulted largely from the action of balancing selection on the PBR and its indirect effects on tightly linked silent sites. The overall non-silent polymorphism is also as great as the silent polymorphism. The absence of any appreciable difference between silent and non-silent polymorphism of MHC genes is similar to what is expected for pseudogenes (2), but for very different reasons.

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 DO heterodimer inhibits such a catalytic action of DM (34, 35). The average p_s value at these loci is 0.46%. Because of the small number of sites compared (549 bp), the difference of p_s = 0.46% from the standard is only marginally significant ($0.05 < P < 0.1$).

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 (C4B, Bf) and the regulation of T lymphocyte development and function (TNF).

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.

	Locus	Silent		Non-silent		References
	(#genes)	$K_s / L_s = p_s$ (%)	S_s	$K_n / L_n = p_n$ (%)	S_n	or accession numbers
1	DPB2 (2)	37/1977 = 1.9	37	–	–	(73), (74)
2	DPB1 (67)	1/63 = 1.6	6.5	9.1/191 = 4.8	24.5	(47)
3	DPA1 (8)	3.5/59 = 5.9	8.7	5.2/185 = 2.8	12.3	(47)
4	DNA (3)	2/204 = 1.0	3	0/547 = 0	0	M31525, M26039, X02882
5	DMA (4)	0/71 = 0	0	3/208 = 1.5	6	X62744, X76775, U04878, U04877
6	DMB (7)	0.5/60 = 0.83	2	2/177 = 1.1	5	Y14395, U00700, U31743, AF00482, U16762, U32663, X76776
7	LMP2 (3)	0.5/173 = 0.24	1	0.4/463 = 0.09	1	X62741, S75169, U01025
8	TAP1 (5)	0.3/631 = 0.05	1	1.8/1610 = 0.11	4	X57522, L21205, L21206, L21207, L21208
9	LMP7 (5)	0/172 = 0	0	0/509 = 0	0	X62598, L11045, U17496, U17497, X66401
10	TAP2 (3)	2/516 = 0.39	3	2/1350 = 0.15	3	U07844, Z22935, Z22936
11	DOB (2)	0/214 = 0	0	1.2/605 = 0.17	1	M26040, L29472
12	DQB2 (8)	1.4/201 = 0.69	3	–	–	M83889, M83890, M83891, M24921, M24920, M24922, M24923, M95729
13	DQA2 (2)	21/2504 = 0.84	21	–	–	Z84490, M29615
14	DQB3 (2)	15/1407 = 1.1	15	–	–	Z84490, M26577
15	DQB1 (27)	6.2/73 = 8.6	17.5	17/212 = 8.1	42	(47)
16	DQA1 (15)	9.0/150 = 6.0	26.5	19/426 = 4.5	47.5	(47)
17	DRB1 (135)	3.6/60 = 6.0	24	12/183 = 6.5	55.5	(47)
18	DRA (2)	1/174 = 0.57	1	1.0/504 = 0.20	1	(47)
19	PBX-2 (3)	0/338 = 0	0	0.7/953 = 0.07	1	X59842, X80700, D28769
20	RAGE (2)	0.5/338 = 0.15	0.5	3.5/874 = 0.40	3.5	M91211, D28769
21	G13 (2)	2/578 = 0.35	2	3/1520 = 0.20	3	X98054, U89337
22	CYP21B (2)	5/405 = 1.23	5	3/1080 = 0.28	3	M31022, AF019413
23	C4B (2)	1/310 = 0.32	1	1/833 = 0.12	1	K02404, AF019413
24	G11 (2)	0/214 = 0	0	0/560 = 0	0	X77836, (75)
25	RD (4)	1.5/274 = 0.54	3	2.6/778 = 0.33	5	L03411, M32275, X16105, AF019413
26	Bf (2)	0/467 = 0	0	1/1350 = 0.07	1	X72875, AF0149413
27	HSP70-2 (2)	7/508 = 1.4	7	2/1410 = 0.14	2	M59830, M11717
28	TNF (3)	0/191 = 0	0	0.7/508 = 0.13	1	X01394, X02910, M10988
29	BAT1 (3)	0/211 = 0	0	0.7/639 = 0.11	1	Z37166, AF029062, A02961
30	MICB (7)	0.3/204 = 0.14	1	1.7/618 = 0.28	6	(36), (37)
31	MICA (16)	2/204 = 0.96	5	7.2/618 = 1.2	22	(38), (39)
32	B (91)	9.7/289 = 3.4	45.4	25/796 = 3.2	91.5	(48)
33	C (35)	8.3/281 = 3.0	46.8	18/779 = 2.3	89.8	(48)
34	SC1 (2)	0/301 = 0	0	8/767 = 1.0	8	S53374, U25826
35	E (5)	0.5/127 = 0.43	3	1.8/346 = 0.51	6	(76)
36	J (3)	7.5/1046 = 0.72	11	–	–	(76)
37	A (50)	10/261 = 3.8	38	25/758 = 3.3	90.5	(48)
38	H (6)	18.7/1091 = 1.71	45	–	–	(76)
39	G (6)	1.1/195 = 0.56	5	1.3/522 = 0.25	4	(76)

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.

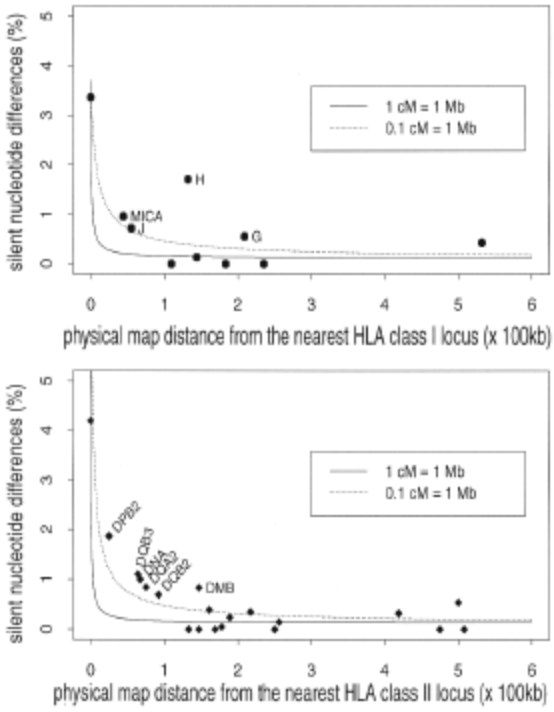

Figure 2. The observed silent nucleotide differences at individual loci are plotted against the physical map distance from the nearest highly polymorphic HLA locus. The solid and dotted curves are computed by the theoretical formulas (9, 45) under the assumption of 1 cM = 1 Mb and 0.1 cM = 1 Mb, respectively. It is assumed that selection intensity is 2% and the number of breeding individuals is 10^5 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 MICA, J, H and G or by DPB2, DQB3, DNA, DQA2 and DQB2.

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 ($P < 0.01$). 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_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 2.7×10^{-6} for class I and 0.8×10^{-6} for class II locus owing to the difference in the number of non-silent sites (25). The per-site mutation rate is 2×10^{-8} 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 $N_e = 10^4$ is made based on the short-lived neutral polymorphism, whereas that of $N_e = 10^5$ 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 locus requires that the c value is smaller than 0.01% (figure 2). Since these loci are located about 25 kb to more than 200 kb apart from the nearest polymorphic MHC locus (44), recombination may be rarer than expected from 1 cM = 1 Mb. One can claim that proper alignment necessary for recombination between the homologous chromosomes is hindered by highly diversified loci (49). However, the small p_s value at the remaining loci is by and large consistent with the postulate of 1 cM = 1Mb. In addition, a reduction of silent polymorphism in non-PBR coding exons at classical MHC loci indicated that recombination is not fully suppressed even within a locus (9).

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_w) and between (p_b) 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_w 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

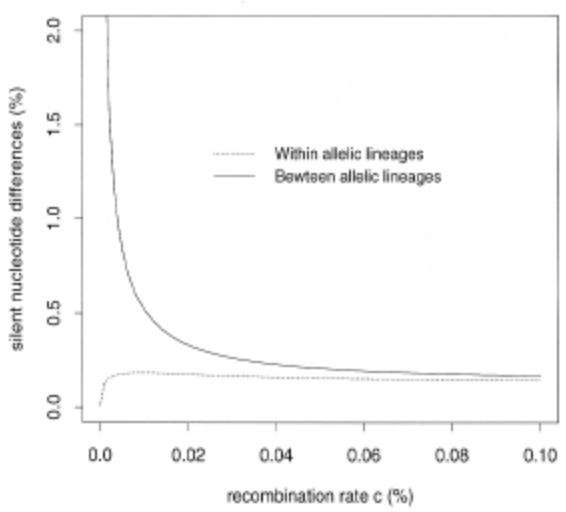

Figure 3. The silent nucleotide differences within (p_w) and between (p_b) allelic lineages based on the theoretical formulas (9, 45). The same set of parameter values are used as in figure 2.

the DRB1 locus at which the reported number of alleles is about 180 and the average p_s value at exon 2 is 8.0%, amounting to the sequence divergence of 40 million years. Based on the observation that the 5' exon 2 region encoding for the beta pleated sheet and the 3' exon 2 region encoding for the alpha helix are phylogenetically incongruent, it is proposed that the chi-like sequence at codon 51-55 somehow mediates sequence exchanges (21). Recently, DRB1 intron 1 and 2 sequences revealed that the phylogenetic relationships are consistent with those of the beta region, and therefore inconsistent with those of the alpha region (22). In terms of linkage disequilibrium, the sequences of these introns are more strongly associated with those in the beta region than in the alpha region. It is therefore suggested that the alpha region (or conversely, though less likely, the beta+intron region) has been shuffled among DRB1 alleles by either double crossover or gene conversion (21 - 23).

Among 103 DRB1 exon 2 sequences available for codon 9 through codon 86 (50), there are 28 and 45 different non-silent motifs in the beta and alpha region, respectively. Of these, 17 beta or 38 alpha motifs can be characterized simply by a set of codon 11, 13, 30 and 37 or by a set of codon 57, 67, 70, 71 and 74. If we exclude minor motifs which differ by single unique substitutions only, there remain ten distinct beta and 16 alpha motifs (see figure 4 for their relationships drawn by the phylogenetic analysis). Importantly, some of these distinct motifs in each region are shared by chimpanzees, gorillas, or Old World Monkeys (51, 53), suggesting that they are of ancient origins. This antiquity of sequence motifs is consistent with the observed large p_s values; 9.3% in the alpha region and 4.4% in the beta region.

If DRB1 sequences in each region are classified by either beta or alpha motifs, p_s will be divided into p_w and p_b although p_b is close to p_s by definition. Theoretically, the best classification of allelic lineages may be obtained when p_b 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_w 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_s 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, iIt may be noted that the p_b 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 intra-locus 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 regions 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 (Q70R71A74 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 the specificity of the P4 pocket (61). It is also reported that tuberculoid leprosy is associated with

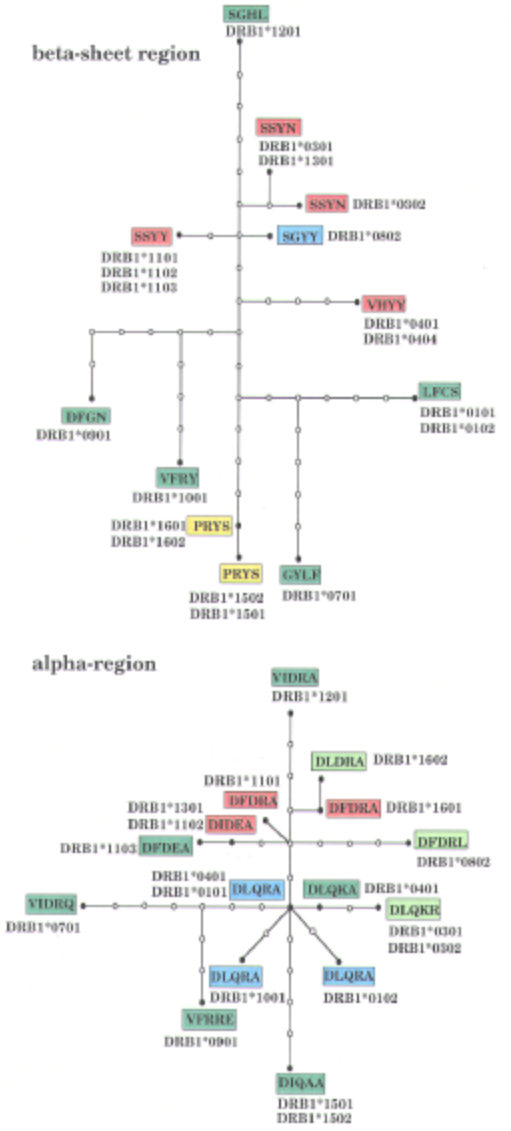

Figure 4. The relationships among the major beta or alpha motifs of DRB1 peptide chains. The relationships are reconstructed from non-silent nucleotide differences in each of the beta and alpha region. Each circle represents one non-silent nucleotide substitution. Codon positions used for the beta motifs 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 change, or no charge, is present at position 71 (68). The same is applied to position 57 in relation to the P9 pocket. This observation based on MHC-associated disease susceptibility and the structural characterization of MHC-peptide complexes (68) 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. When two loci are recombined with c = 0.05%, it takes about 2,000 generations or 40,000 years for the linkage to be broken. 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?

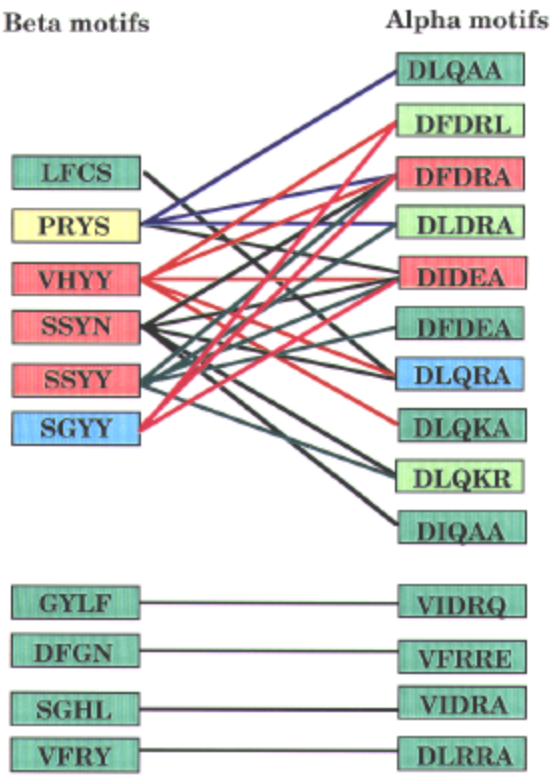

Figure 5. Combinations between some beta and alpha motifs that occur in DRB1 peptide chains.

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

1. Kimura M: Evolutionary rate at the molecular level. *Nature* 217, 624-626 (1968)

2. Kimura M: The neutral theory of molecular evolution. Cambridge Univ. Press, Cambridge (1983)

3. Aquadro C. F, D. J. Begun & E. C. Kindahl: Selection, recombination and DNA polymorphism. In: Non-neutral evolution. Ed: Golding B, Chapmann & Hall, NY (1995)

4. Magunus N, B. Charlesworth & D. Charlesworth: The effects of recombination on background selection. *Genet Res* 67, 159-174 (1996)

5. Hudson R. R. & N. L. Kaplan: The coalescent process in models with selection and recombination. *Genetics* 120, 831-840 (1988)

6. Takahata N: A simple genealogical structure of strongly balanced allelic lines and trans-specific evolution of polymorphism. *Proc Natl Acad Sci, USA* 87, 2419-23 (1990)

7. Takahata N: Allelic genealogy and human evolution. *Mol Biol Evol* 10, 2-22 (1993)

8. Vekemans X. & M. Slatkin: Gene and allelic genealogies at a gametophytic self-incompatibility locus. *Genetics* 137, 1157-1165 (1994)

9. Takahata N. & Y. Satta: Footprints of intragenic recombination at HLA loci. *Immunogenetics* (1998 in press)

10. Kingman J. F. C: On the genealogy of large populations. *J Appl Prob* 19A, 27-43 (1982)

11. Wright S: Evolution in mendelian populations. *Genetics* 16, 97-159 (1931)

12. Takahata N. & Y. Satta: Evolution of the primate lineage leading to modern humans: phylogenetic and demographic inferences from DNA sequences. *Proc Natl Acad Sci, USA* 94, 4811-4815 (1997)

13. Vogel F. & A. G. Motulsky: Human genetics: problems and approaches (3rd edn), Springer, Heidelberg (1997)

14. Li W.-H. & L. A. Saddler: Low nucleotide diversity in man. *Genetics* 129, 513-523 (1991)

15. Ruvolo M, D. Pan, S. Zehr, T. Goldberg, T. R. Disotell & M. von Dornum: Gene tree and hominoid phylogeny. *Proc Natl Acd Sci, USA* 91, 8900-8904 (1994)

16. Satta Y: Balancing selection at HLA loci. In: Mechanisms of molecular evolution. Eds: Takahata N, Clark A. G, Springer, Tokyo (1992)

17. Hughes A. L. & M. Nei: Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335, 167-170 (1988)

18. Hughes A. L. & M. Nei: Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. *Proc Natl Acad Sci, USA* 86, 958-962 (1989)

19. Klein J, Y. Satta,, C. O'hUigin & N. Takahata: The molecular decent of the major histocompatibility complex. *Annu Rev Immunol* 11, 269-295 (1993)

20. Takahata N, Y. Satta & J. Klein: Polymorphism and balancing selection at major histocompatibility complex loci. *Genetics* 130, 925-938 (1992)

21. Gyllensten U. B, M. Sundvall & H. A. Erlich: Allelic diversity is generated by intraexon sequence exchange at the DRB1 locus of primates. *Proc Natl Acad Sci, USA* 88, 3686-3690 (1991)

22. Bergstrom T. F, A. Josefsson, H. A. Erlich & U. Gyllensten: Recent origin of HLA-DRB1 alleles and implications for human evolution. *Nat Genet* 18, 237-242 (1998)

23. Takahata N. & Y. Satta: Improbable truth in human MHC diversity? *Nat Genet* 18, 204-206 (1998)

24. Grahovac B, R. I. Sukernik, C. O'hUigin, Z. Zaleska-Rutczynska, N. Blagitko, O. Raldugina, T. Kosutic, Y. Satta, F. Figueroa, N. Takahata & J. Klein: Polymorphism of the HLA class II loci in Siberian populations. *Hum Genet* 102, 27-43 (1998)

25. Satta Y, C. O'hUigin, N. Takahata & J. Klein: The synonymous substitution rate at the primate MHC loci. *Proc Natl Acad Sci, USA* 90, 7480-7484 (1993)
26. Otting N, M. Kenter, P. van Weeren, M. Jonker & R. E. Bontrop: MHC-DQB repertoire variation in hominoid and Old World primate species. *J Immunol* 149, 461-470 (1992)

27. Klein J. & V. Horejsi: Immunology, Blackwell, MA (1997)

28. Joly E, E. V. Deverson, W. J. Coadwell, E. Guenther, J. C. Howard & G. W. Butcher: The distribution of Tap2 alleles among laboratory rat RT1 haplotypes. *Immunogenetics* 40, 45-53 (1994)

29. Joly E, A-F. Le Rolle, A. L. Gonzalez, B. Mehling, J. Stevens, W. J. Coadwell, T. Hunig, J. C. Howard & G. W. Butcher: Co-evolution of rat TAP transporters and MHC class I RT1-A molecules. *Curr Biol* 8, 169-172 (1998)

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)

31. Servenius B, L. Rask & P. A. Peterson: Class II genes of the human major histocompatibility complex. The DO beta gene is a divergent member of the class II beta gene family. *J Biol Chem* 25, 8759-8766 (1987)

32. Kelly A. P, J. J. Monaco, S. G. Cho & J. Trowsdale: A new human HLA class II-related locus, DM. *Nature* 353, 571-573 (1991)

33. Sloan V. S, P. Cameron, G. Porter, M. Gammon, M. Amaya, E. Mellins & D. M. Zaller: Mediation by HLA-DM of dissociation of peptides from HLA-DR. *Nature* 375, 802-806 (1995)

34. van Ham S. M, E. P. M. Tjin, B. F. Lillemeier, U. Gruneberg, K. E. van Meijgaarden, L. Pastoors, D. Verwoerd, A. Tulp, B. Canas, D. Rahman, T. H. Ottenhoff, D. J. Pappin, J. Trowsdale and J. Neefjes: HLA-DO is a negative modulator of HLA-DM-mediated MHC class II peptide loading. *Curr Biol* 7, 950-957 (1997)

35. Jensen P. E: Antigen processing: HLA-DO-a hitchhiking inhibitor of HLA-DM. *Curr Biol* 8, R128-R131 (1998)

36. Tusie-Luna M. T. & P. C. White: Gene conversions and unequal crossovers between CYP21 (steroid 21-hydroxylase gene) and CYP21P involve different mechanisms. *Proc Natl Acad Sci, USA* 92, 10796-10800 (1995)

37. Milner C. M. & R. D. Campbell: Structure and expression of the three MHC-linked HSP70 genes. *Immunogenetics* 32, 242-251 (1990)

38. Bahram S, T. Shiina, A. Oka, G. Tamiya & H. Inoko: Genomic structure of the human MHC class I MICB gene. *Immunogenetics* 45: 161-162 (1996)

39. Pellet P, M. Renaud, N. Fodil, L. Laloux, H. Inoko, G. Hauptman, P. Debre, S. Bahram & I. Theodorou: Allelic repertoire of the human MICB gene. *Immunogenetics* 46, 434-436 (1997)

40. Bahram S, M. Bresnahan, D. E. Geraghty & T. Spies: A second lineage of mammalian major histocompatibility complex class I genes. *Proc Natl Acad Sci, USA* 91, 6259-6263 (1994)

41. Fodil N, L. Laloux, V. Wanner, P. Pellet, G. Hauptman, N. Mizuki, H. Inoko, T. Spies, I. Theodorou & S. Bahram: Allelic repertoire of the human MHC class I MICA gene. *Immunogenetics* 44, 351-357 (1996)
42. Groh V, A. Steinle, S. Bauer & T. Spies: Recognition of stress-induced MHC molecules by intestinal epithelial gamma delta T cells. *Science* 279, 1737-1740 (1998)

43. Feder J. N, A. Gnirke, W. Thomas, Z. Tsuchihashi, D. A. Ruddy, A. Basava, F. Dormishian, R. Jr. Domingo, M. C. Ellis, A. Fullan, L. M. Hinton, N. L. Jones, B. E. Kimmel, G. S. Kronmal, P. Lauer, V. K. Lee, D. B. Loeb, F. A. Mapa, E. McClelland, N. C. Meyer, G. A. Mintier, N. Moeller, T. Moore, E. Morikang, C. E. Prass, L. Quintana, S. M. Starnes, R. C. Schatzman, K. J. Brunke, D. T. Drayna, N. J. Risch, B. R. Bacon & R. K. Wolff : A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nat Genet* 13, 399-408 (1996)

44. Campbell R. D. & J. Trowsdale: A map of the human major histocompatibility complex. *Immunol Today* 18, suppl. (1997)

45. Shiina T, G. Tamiya, A. Oka, T. Yamagata, N. Yamagata, E. Kikkawa, K. Goto, N. Mizuki, K. Watanabe, Y. Fukuzumi, S. Taguchi, C. Sugawara, A. Ono, L. Chen, M. Yamazaki, H. Tashiro, A. Ando, T. Ikemura, M. Kimura & H. Inoko: Nucleotide sequencing analysis of the 146-kilobase segment around the IkBL and MICA genes at the centromeric end of the HLA class I region. *Genomics* 47, 372-382 (1998)

46. Mizuki N, H. Ando, M. Kimura, S. Ohno, S. Miyata, M. Yamazaki, H. Tashiro, K. Watanabe, A. Ono, S. Taguchi, C. Sugawara, Y. Fukuzumi, K. Okumura, K. Goto, M. Ishihara, S. Nakamura, J. Yonemoto, Y. Y. Kikuti, T. Shiina, L. Chen, A. Ando, T. Ikemura & H. Inoko: Nucleotide sequence analysis of the HLA class I region spanning the 237-kb segment around the HLA-B and -C genes. *Genomics* 15, 55-66 (1997)

47. Takahata N: Repeated failures that led to the eventual success in human evolution. *Mol Biol Evol* 11, 803-805 (1994)

48. Takahata N: A genetic perspective on the origin and history of humans. *Ann Rev Ecol Syst* 26, 343-372 (1995)

49. Klein J, C. O'hUigin, M. Kasahara, V. Vincek, D. Klein & F. Figueroa: Frozen haplotypes in Mhc evolution. In: Molecular evolution of the major histocompatibility complex. Eds: Klein J, Klein D, Springer, Berlin (1992)

50. Marsh S. G. E. & J. G. Bodmer: HLA class II region nucleotide sequences 1995. *Tissue Antigens* 46, 258-280 (1995)

51. Arnett K. L. & P. Parham: HLA class I nucleotide sequences 1995. *Tissue Antigens* 46, 217-257(1995)

52. Klein J. & C. O'hUigin: Class II B Mhc motifs in an evolutionary perspectives. *Immunol Rev* 143, 89-112 (1995)

53. O'hUigin C: Quantifying the degree of convergence in primate Mhc-DRB genes. *Immunol Rev* 143, 123-140 (1995)

54. Satta Y: Effects of intra-locus recombination of HLA polymorphism. *Hereditas* 127, 105-112 (1997)

55. Bjorkman P. J, M. A. Saper, B. Samraoui, W. S. Bennett, J. L. Strominger & D. C. Wiley: Structure of the human class I histocompatibility antigen, HLA-A2. *Nature* 329, 506-512 (1987)

56. Brown J. H, T. S. Jardetzky, J. C. Gorga, L. J. Stern, R. G. Urban, J. L. Strominger & D. C. Wiley: Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. *Nature* 364, 33-39 (1993)

57. Stern L. J, J. H. Brown, T. S. Jardetzky, J. C. Gorga, R. G. Urban, J. L. Strominger & D. C. Wiley: Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. *Nature* 368, 215-221 (1994)

58. Gregersen P. K, M. Shen, Q. L. Song, P. Merryman, S. Degar, T. Seki, J. Maccari, D. Goldberg, H. Murphy & J. Schwenzer: Molecular diversity of HLA-DR4 haplotypes. *Proc Natl Acad Sci, USA* 83, 2642-6 (1986)

59. Tiwari J. L. & P. I. Terasaki: Endocrinology. In: HLA and disease associations. Eds: Tiwari JL, Terasaki PI, Springer, NY (1985)

60. Noble J. A, A. M. Valdes, M. Cook, W. Klitz, G. Thomson & H. A. Erlich: The role of HLA class II genes in insulin-dependent diabetes mellitus: molecular analysis of 180 Caucasian, multiplex families. *Am J Hum Genet* 59, 1134-1148 (1996)

61. Zamani M. & J. J. Cassiman: Reevaluation of the importance of polymorphic HLA class II alleles and amino acids in the susceptibility of individuals of different populations to type I diabetes. *Am J Med Genet* 76, 183-194 (1998)

62. van Eden W, R. R. de Vries, N. K. Mehra, M. C. Vaidya, J. D'Amaro and J. J. van Rood: HLA segregation of tuberculoid leprosy: confirmation of the DR2 marker. *J Infect Dis* 141, 693-701 (1980)

63. Mehra N. K: Role of HLA linked factors in governing susceptibility to leprosy and tuberculosis. *Trop Med Parasitol* 41, 352-354 (1990)

64. Todd J. R, B. C. West & J. C. McDonald: Human leukocyte antigen and leprosy: study in northern Louisiana and review. *Rev Infect Dis* 12, 63-74 (1990)

65. van Eden W, R. R. de Vries, J. D'Amaro, I. D. Schreuder, L. Leiker & J. J. van Rood: HLA-DR-associated genetic control of the type of leprosy in a population from Surinam. *Hum Immunol* 4, 343-350 (1982)

66. van Eden W, N. M. Gonzalez, R. R. de Vries, J. Convit & J. J. van Rood: HLA-linked control of predisposition to lepromatous leprosy. *J Infect Dis* 151, 9-14 (1985)

67. Zerva L, B. Cizman, N. K. Mehra, S. K. Alahari, R. Murali, C. M. Zmijewski, M. Kamoun & D. S. Monos: Arginine at positions 13 or 70-71 in pocket 4 of HLA-DRB1 alleles is associated with susceptibility to tuberculoid leprosy. *J Exp Med* 183, 829-836 (1996)

68. Wucherpfennig K. W. & J. L. Strominger: Selective binding of self peptides to disease-associated major histocompatibility complex (MHC) molecules: A mechanism for MHC-linked susceptibility to human autoimmune diseases. *J Exp Med* 181, 1597-1601 (1997)

69. Ota T. & M. Nei: Divergent evolution and evolution by the birth-and-death process in the immunoglobulin VH gene family. *Mol Biol Evol* 11, 469-482 (1994)

70. Nei M, X. Gu & T. Sitnikova: Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proc Natl Acad Sci, USA* 94, 7799-7806 (1997)

71. Sitnikova T. & C. Su: Coevolution of immunoglobulin heavy and light chain variable region gene families. *Genetics* (1998 in press)

72. Lawlor D. A, J. Zemmour, P. D. Ennis & P. Parham: Evolution of class-I MHC genes and proteins: from natural selection to thymic selection. *Annu Rev Immunol* 8, 23-63 (1990)

73. Takahata N: Polymorphism at Mhc loci and isolation by the immune system in vertebrates. In: Non-neutral evolution. Ed: Golding B, Chapmann & Hall, NY (1995)

74. Klein J: Of HLA, trps, and selection: an essay on coevolution of MHC and parasites. *Hum Immunol* 30, 247-258 (1991)

75. Satta Y, C. O'hUigin, N. Takahata & J. Klein: Intensity of natural; selection at the major histocompatibility complex loci. *Proc Natl Acad Sci, USA* 91, 7184-7188 (1994)

76. Kappes D. J, D. Arnot, K. Okada & J. L. Strominger: Structure and polymorphism of the HLA class II SB light chain genes. *EMBO J* 3, 2985-2993 (1984)

77. Gustafsson K, E. Widmark, A. K. Jonsson, B. Servenius, D. H. Sachs, D. Larhammar, L. Rask & P. A. Peterson: Class II genes of the human major histocompatibility complex: evolution of the DP region as deduced from nucleotide sequences of the four genes. *J Biol Chem* 262, 8778-8786 (1987)

78. Sargent C. A, M. J. Anderson, S. L. Hsieh, E. Kendall, N. Gomez-Escobar & R. D. Campbell: Characterization of the novel gene G11 lying adjacent to the complement C4A gene in the human major histocompatibility complex. *Hum Mol Genet* 3, 481-488 (1994)

79. Parham P: ASHI information. In: http://www.swmed.edu (1997)

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