Polymorphism and Balancing Selection at Major Histocompatibility Complex Loci

Naoyuki Takahata,* Yoko Satta* and Jan Klein†

**Department of Population Genetics, National Institute of Genetics, Mishima 411, Japan, and †Max-Planck-Institut für Biologie, Abteilung Immungenetik, Corrensstr. 42, W-7400 Tübingen, Germany, and Department of Microbiology and Immunology, University of Miami School of Medicine, Miami, Florida 33101*

Manuscript received September 5, 1991
Accepted for publication December 13, 1991

ABSTRACT

Amino acid replacements in the peptide-binding region (PBR) of the functional major histocompatibility complex (*Mhc*) genes appear to be driven by balancing selection. Of the various types of balancing selection, we have examined a model equivalent to overdominance that confers heterozygote advantage. As discussed by A. Robertson, overdominance selection tends to maintain alleles that have more or less the same degree of heterozygote advantage. Because of this symmetry, the model makes various testable predictions about the genealogical relationships among different alleles and provides ways of analyzing DNA sequences of *Mhc* alleles. In this paper, we analyze DNA sequences of 85 alleles at the *HLA-A, -B, -C, -DRB1* and *-DQB1* loci with respect to the number of alleles and extent of nucleotide differences at the PBR, as well as at the synonymous (presumably neutral) sites. Theory suggests that the number of alleles that differ at the sites targeted by selection (presumably the nonsynonymous sites in the PBR) should be equal to the mean number of nucleotide substitutions among pairs of alleles. We also demonstrate that the nucleotide substitution rate at the targeted sites relative to that of neutral sites may be much larger than 1. The predictions of the presented model are in surprisingly good agreement with the actual data and thus provide means for inferring certain population parameters. For overdominance selection in a finite population at equilibrium, the product of selection intensity (s) against homozygotes and the effective population size (N) is estimated to be 350–3000, being largest at the B locus and smallest at the C locus. We argue that N is of the order of 10^5 and s is several percent at most, if the mutation rate per site per generation is 10^{-8}.

THE functional major histocompatibility complex (*Mhc*) genes, or more specifically the parts coding for the peptide (antigen)-binding region (PBR) of the molecule, are believed to be subjected to balancing selection (DOHERTY and ZINKERNAGEL 1975; HUGHES and NEI 1988, 1989). The fundamental observations that led to this conclusion are (1) that there exist a large number of alleles at functional *Mhc* loci (KLEIN 1986); (2) allele frequencies are distributed more or less evenly (HEDRICK and THOMSON 1983); (3) alleles often differ at a number of nucleotide sites (KLEIN 1986); (4) certain alleles of one species are generally more similar to certain alleles of another species than they are to other alleles of the first species (the so-called *trans*-species mode of polymorphism in KLEIN 1980; FIGUEROA, GÜNTHER and KLEIN 1988; LAWLOR *et al.* 1988; MAYER *et al.* 1988; MCCONNELL *et al.* 1988); (5) the rate of nonsynonymous substitutions at the PBR is higher than that of the synonymous substitutions (HUGHES and NEI 1988, 1989); and (6) many alleles are in strong linkage disequilibrium (KLITZ and THOMSON 1987). There are some other observations which are suggestive of balancing selection, but are controversial. These include deficiency of *Mhc* homozygotes (DEGOS *et al.* 1974; BLACK and SALZANO 1981; RITTE *et al.* 1991) and association of certain *Mhc* alleles with diseases (KLEIN 1986, 1990; TIWARI and TERASAKI 1985; THOMSON 1988; HILL *et al.* 1991).

By balancing selection population geneticists mean either overdominance (heterozygote superiority in fitness), frequency-dependent selection, or diversifying selection that favors different genotypes in different environments (DOBZHANSKY 1970). Although the three forms of balancing selection are biologically distinguishable in some cases (NEI and HUGHES 1991), there are types of frequency-dependent and diversifying selection that are theoretically equivalent to overdominance (TAKAHATA and NEI 1990; DENNISTON and CROW 1990). These forms may be collectively referred to as the overdominance-type selection. Heterozygote superiority of *Mhc* genes has thus far not been proven experimentally (but see RITTE *et al.* 1991).

WRIGHT (1939) developed a theory of self-sterility alleles in plants and derived formulas for the equilibrium allele frequencies, the expected number of segregating alleles, and heterozygosity. His theory and results can be applied directly to the case of overdom-

inance selection, although homozygote lethality is not necessarily assumed in the latter. It was remarked that overdominance selection is an inefficient mechanism for maintaining a large number of alleles at a locus (LEWONTIN 1985). LEWONTIN, GINZBURG and TULJAPURKAR (1978) studied stable, feasible equilibria under heterosis or overdominance selection and concluded that the proportion of fitness arrays leading to such equilibria of more than 6 or 7 alleles is vanishingly small. However, this conclusion changes when new mutations are taken into account. Indeed, the authors themselves have argued that "As new mutations occur, they will be lost to the population if their fitnesses in homozygous and heterozygous condition do not lie in the appropriate region, while the new alleles will be maintained in the population if they have the appropriate fitnesses. Thus, although few new mutations may have the appropriate fitnesses, those that do will be accumulated, and it is these that we see in nature." This point was actually made by ROBERTSON (1962) and it was supported by computer simulations (MARUYAMA and NEI 1981; TAKAHATA and NEI 1990; SPENCER and MARKS 1988 for the deterministic case). Hence, even when new mutations have asymmetric fitnesses, overdominance selection tends to maintain only a particular set of alleles so that those alleles that we observe will have symmetric fitness arrays. Moreover, the number of such alleles needs not be small if one assumes a continuous mutation pressure (KIMURA and CROW 1964; Table 14.3 in WRIGHT 1969), and this assumption is essential in studying the long-term evolution of balanced alleles which could persist in a population for tens of million years (KLEIN 1980; FIGUEROA, GÜNTHER and KLEIN 1988; LAWLOR *et al.* 1988; MAYER *et al.* 1988). Therefore the symmetric overdominance-type model can potentially be compatible with *Mhc* data comprised by DNA sequences of many alleles (*e.g.*, MARSH and BODMER 1991; ZEMMOUR and PARHAM 1991).

To interpret the currently available DNA sequence data, it is essential to understand the allelic relationships (allelic genealogy). One way of studying the allelic genealogy was suggested by TAKAHATA and NEI (1990) and TAKAHATA (1990) who based their arguments on KIMURA and CROW's (1964) model of infinitely many-allele mutations. This model is more appropriate than for example, the classical two-allele mutation scheme used by HUDSON and KAPLAN (1988) with the purpose of applying the coalescent process to neutral sites partially linked to a selected site. In this paper, we examine DNA sequences of human *Mhc* alleles using the theory of allelic genealogy under symmetric overdominance-type selection and the model of infinitely many alleles. The allelic genealogy has a mathematically simple structure, in particular about the relationship between the number of segregating alleles and the number of nucleotide differences between alleles. The simplicity results from random extinction of existing alleles and the assumption that mutations always produce new allelic lines of descent.

ALLELIC GENEALOGY

If we are not concerned about the time scale of allelic genealogy, the following consideration is sufficient to construct the allelic relationships at a given locus. Under symmetric overdominance-type selection, the allele frequencies tend to be evenly distributed and different alleles are equivalent in their fate. When a new descendant allele (*DA*) is produced from one of the parental alleles (*PA*) and is incorporated into a population, the *DA* becomes *PA* and one of the previous *PA*s becomes extinct (allelic turnover). The extinct line can be the parent itself that produced *DA*. In a population at equilibrium, the incorporation of *DA* and the extinction of *PA* alleles are in balance and the number of different alleles (n) remains more or less constant. Because every *PA* has an equal probability of extinction, $1/n$, the genealogical relationships among alleles (allelic genealogy) are simple (TAKAHATA 1990) (Figure 1), and are similar to those of randomly sampled neutral genes (KINGMAN 1982; TAVARÉ 1984; WATTERSON 1984). If the parental allele of *DA* becomes extinct, there is no way of learning anything about the bifurcation of *PA* and *DA*. Only when both *PA* and *DA* or their direct lines of descent survive and are present in a sample, can we make inferences about their divergence or coalescence. The *DA* is always one mutational step away from the *PA* under the infinitely many-allele model. In what follows, allelic genealogy analysis will focus on the divergence of alleles at target sites of selection only. Such allele divergences are due to amino acid replacements in the PBR of Mhc molecules (nonsynonymous substitutions at the corresponding parts of the *Mhc* genes).

Suppose that we have sampled i different alleles from n existing alleles ($i \leq n$) and have determined their sequences, whereby by *different* we mean alleles that differ at the nonsynonymous sites in the PBR. We restrict our sequence analyses to pairwise comparisons that do not require precise knowledge of the ancestral relationships among the different alleles. We compute the number (K_N) of nonsynonymous changes in the PBR between any pair of alleles. For i sampled alleles, there are $i(i-1)/2$ values of K_N from which we define the pairwise mean number (K_N^P) and the largest number (K_N^L). If $i = 2$, $K_N^P = K_N^L$ by definition. Here and subsequently, the superscript (P or L) stands for the pairwise mean number and the largest number of changes unless otherwise specified, while the subscript (N or S) stands for nonsynonymous and synonymous changes. The theoretical calculation assumes

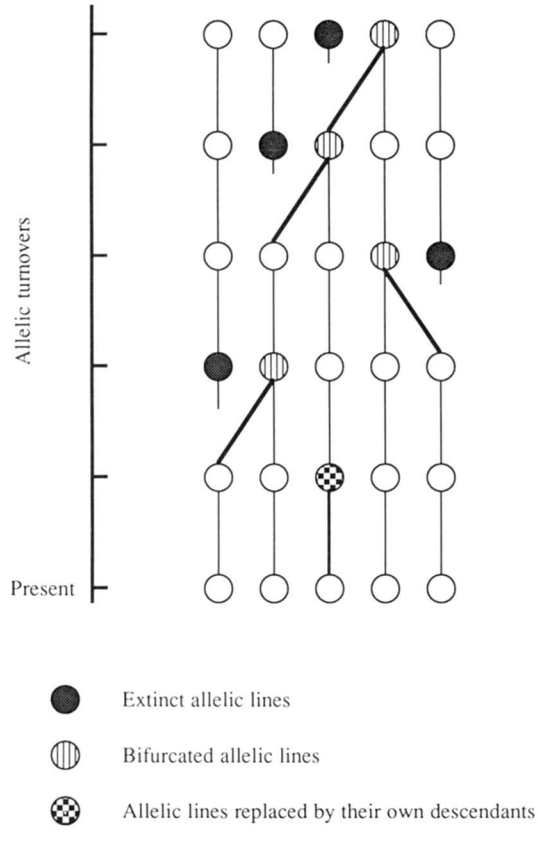

FIGURE 1.—Allelic genealogy. In this figure, there are five alleles which are equivalent in terms of reproduction. In each allelic turnover, one of five parental alleles (*PA*) produces a descendant allele (*DA*) and one *PA* goes to extinction. The allelic turnover rate is not specified here, but the waiting time for each allelic turnover is exponentially distributed with the mean of α/n generations where n is the number of alleles in a population (TAKAHATA 1990).

effectively the infinite site model without recombination (WATTERSON 1975) and makes no correction for multiple hits. The correction is made to the data if at all.

The expected values of K_N^P and K_N^L are independent of the topology of allelic genealogy and are given by the formulas

$$E\{K_N^P\} = n \tag{1}$$

and

$$\begin{aligned} E\{K_N^L\} &= 2n\left(1 - \frac{1}{i}\right) - \sum_{j=3}^{i} \frac{2}{j} \\ &\approx 2n\left(1 - \frac{1}{i}\right) \text{ for large } n \end{aligned} \tag{2}$$

in which $E\{X\}$ stands for the expectation of random variable X taken by its distribution (see APPENDIX). That is, the expected number of nonsynonymous changes in the PBR is the same as the number of different alleles in a population and the largest difference between a pair of alleles in the sample roughly equals $2(1 - 1/i)$ times the number of alleles n or $E\{K_N^P\}$.

We can also derive the expected relationship between K_N^P (K_N^L) and the corresponding synonymous changes K_S^P (K_S^L) among *Mhc* genes. (Furthermore, the same relationship can be derived for the nonsynonymous changes outside the PBR if they are selectively neutral.) Assume that the linkage between the synonymous sites and PBR within a gene is complete and that the total rate of synonymous changes is v per gene per generation. To relate K_S^P (K_S^L) to K_N^P (K_N^L), we must define our model more precisely and know how rapidly nonsynonymous changes occur relative to synonymous changes. We base our considerations on the model of symmetric overdominance-type selection studied by WRIGHT (1939), KIMURA and CROW (1964), MARUYAMA and NEI (1981), TAKAHATA (1990) and others. Applying this selection model to a finite population, KIMURA and CROW (1964) derived an approximate formula for the number of alleles (n). For our purpose, the formula may be written as

$$M = \sqrt{\frac{S}{16\pi}} \exp\left\{-\frac{S}{n^2}\right\} \tag{3}$$

in which $S = 2Ns$ and $M = Nu$ where N is the number of breeding individuals in the population, s the selective disadvantage of homozygotes in the overdominance selection model, and u the nonsynonymous mutation rate per PBR per generation. TAKAHATA (1990) demonstrated that the divergence time of a pair of alleles is exponentially distributed with mean

$$\alpha = \frac{n^3}{2\sqrt{2}uS} \tag{4}$$

in units of generations. During this divergence time, a pair of alleles accumulates n nonsynonymous changes on average (see APPENDIX and Equation 1), whereas the synonymous sites accumulate $2\alpha v$ changes. For a given divergence time of alleles, the distribution of the number of both changes is approximately Poisson [see APPENDIX, also WATTERSON (1975) for the neutral case]. Similarly to the divergence time between a pair of alleles, there are $2\alpha(1 - 1/i)$ generations between the two most distantly related alleles in the i sampled alleles. Thus we have

$$E\{K_S^P\} = 2\alpha v \text{ and } E\{K_S^L\} = 4\alpha v\left(1 - \frac{1}{i}\right). \tag{5}$$

For convenience, we define

$$\theta = 2\alpha v. \tag{6}$$

If u and v are proportional to the number of nonsynonymous (L_N) and synonymous (L_S) sites and the per-site mutation rate (μ) is the same for L_N and L_S, θ can be rewritten as $L_S n^3/(\sqrt{2}L_N S)$ by Equation 4. From

Equations 1, 2 and 5, the ratio of nonsynonymous to synonymous changes is

$$\frac{E\{K_N^Y\}}{E\{K_S^Y\}} = \frac{n}{\theta} = \frac{\sqrt{2}L_N S}{L_S n^2} \tag{7}$$

where superscript Y stands for either P or L. If the synonymous and nonsynonymous changes per site are defined by

$$k_S^Y = \frac{E\{K_S^Y\}}{L_S}$$

and

$$k_N^Y = \frac{E\{K_N^Y\}}{L_N},$$

respectively, the ratio (γ) of k_N^Y to k_S^Y becomes

$$\gamma = \frac{\sqrt{2}S}{n^2}.$$

We may rewrite the above equation as

$$S = \frac{n^2\gamma}{\sqrt{2}}. \tag{8}$$

If n is the observed number of alleles or estimated from Equation 1 or 2 and if k_N^Y and k_S^Y are computed from sequence data, Equation 8 gives an estimator of $S = 2Ns$. An estimate of $M = Nu$ can then be obtained from Equation 3.

From intraspecies variation of the *Mhc*, HUGHES and NEI (1988) estimated the ratio of the nonsynonymous rate in the PBR to the synonymous rate to be about 3. The ratio is equivalent to γ. Here we consider the possibility that calculating γ on the basis of *all* pairwise comparisons may lead to an underestimate because of a difficulty in inferring extensive nonsynonymous changes in the PBR that have accumulated *trans*-specifically. An alternative method we propose here is based on the fact that young alleles do not differ much in terms of K_S. For pairs of such alleles, the K_N can be expected to be relatively small and multiple hits to be rare so that errors in the multiple hit correction can be minimized.

Suppose that we have identified allelic pairs with small values of $K_S = m$ and that we also know their K_N. For these pairs, we obtain $K_N = j$ and $K_S = m$ ($j = 1, 2, 3, \ldots$ and $m = 0, 1, 2, \ldots$). Our task is to find the conditional probability of K_N when $K_S = m$, and we denote this probability as $P\{K_N^m = j\}$. Under symmetric overdominance-type selection the conditional probability becomes a negative binomial distribution

$$P\{K_N^m = j\} = \frac{(j+m)!}{j!m!}\left\{\frac{1+\theta}{1+n+\theta}\right\}^{m+1}\left\{\frac{n}{1+n+\theta}\right\}^{j} \tag{9}$$

(see APPENDIX). The conditional mean of K_N^m is then given by

$$E\{K_N^m\} = \frac{n(1+m)}{1+\theta}, \tag{10}$$

from which we have

$$n = \frac{1+\theta}{1+m}E\{K_N^m\}, \tag{11}$$

and hence another estimator of n. For instance, if we happen to find pairs of alleles that are identical at the synonymous sites, we set $m = 0$ in Equation 11, and we have $n = E\{K_N^0\}(1+\theta)$. Alternatively and more desirable statistically, we may use pairs of alleles whose synonymous differences are in a certain range of K_S. For pairs of alleles with $0 \le K_S \le k$, we have

$$n = \frac{E\{K_N \mid 0 \le K_S \le k\}}{1 - \dfrac{(k+1)\theta^{k+1}}{(1+\theta)^{k+2} - (1+\theta)\theta^{k+1}}} \tag{12}$$

(see APPENDIX).

As K_S becomes large, the actual number of nonsynonymous changes may not increase in such a way that Equation 10 predicts. The leveling-off signifies that, since alleles with large values of K_S diverged long ago, nonsynonymous differences in their PBR may be saturated. For this reason, we use Equation 12 with small values of K_S to estimate n.

APPLICATION TO *HLA* LOCI

All pairwise comparisons: We have applied the above theoretical results to the sequence data obtained for the human *HLA-A*, *-B*, *-C*, *-DRB1* and *-DQB1* loci. The numbers of alleles examined at these loci are 19, 26, 6, 19 and 15, respectively, giving a total number of 85. To estimate the actual number of synonymous and nonsynonymous substitutions, we used the JUKES and CANTOR (1969) method. In general the synonymous differences were small so that no substantial correction was made. The correction for nonsynonymous differences in the PBR was a different matter and requires caution. For instance, in *HLA-B*, K_N^m does not increase linearly but levels off at about 20 as $K_S = m$ becomes large (Figure 2A). Although this number is small compared to the total number of nonsynonymous sites in the PBR ($L_N \approx 135$), it may be near the saturation level because some sites in the PBR are well conserved (see also HUGHES and NEI 1988). If this were the case, some sites would have undergone many

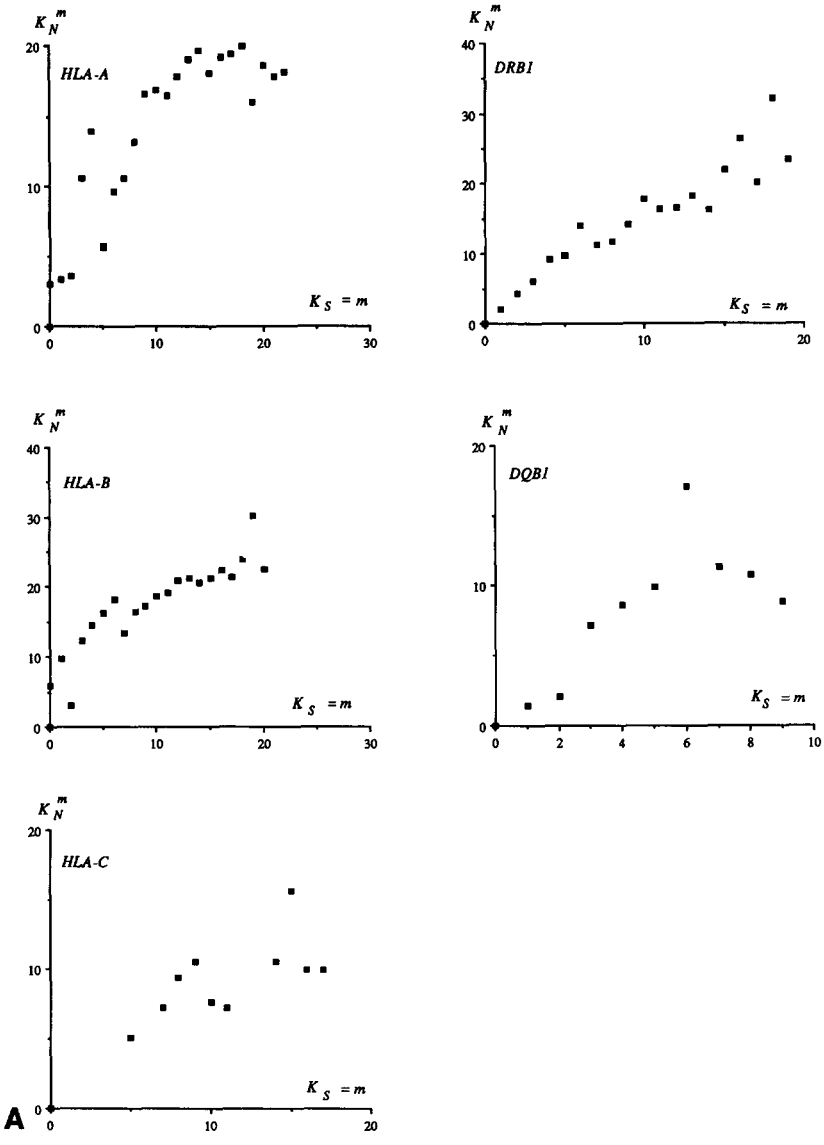

FIGURE 2.—A, The pairwise mean number of nonsynonymous substitutions K_N in the PBR conditioned on the number of synonymous changes $K_S = m$.

and others no substitutions. In order to examine this possibility, we carried out a maximum parsimony analysis of the phylogeny of the *HLA* alleles by DNAPARS (in PHYLIP, version 3.3, provided by J. FELSENSTEIN). It turns out that the number of nonsynonymous substitutions varies greatly from site to site, ranging from 0 to 7, and that these substitutions at various sites are not distributed according to the Poisson distribution [see UZZELL and CORBIN (1971) for a similar finding). Since the Poisson (including JUKES and CANTOR's) method failed to provide the necessary correction for multiple hit substitutions, discrepancies among estimates of n based on Equations 1, 11 or 12 could be expected. To avoid these discrepancies, we could estimate K_N by a non-Poisson correction method (JIN and NEI 1990; TAKAHATA 1991b). Alternatively, we could use a set of relatively young alleles for which both Poisson and non-Poisson methods are expected to make similar estimates for the actual number of substitutions per PBR (TAKAHATA 1991b).

Disregarding the possibility that both K_N^L and K_N^P may be underestimated (see the next subsection), we first examined whether the ratio of these two values is close to $2(1 - 1/i)$ for i sampled alleles. If we could sample all existing alleles, the sample size would be n and the ratio would become equal to twice the expected heterozygosity (H) generated by nonsynonymous changes in the PBR. This is because under strong symmetric overdominance-type selection n is equivalent to the effective number of alleles (KIMURA and CROW 1964), so that $1/n$ becomes the expected homozygosity (F). For class I loci, the ratio is 1.8 to 1.9 so that $H \approx$ 90–95%. These H values are very close to those estimated from allele frequencies (see Table 8.10-12 in KLEIN 1986). For class II loci, on the other hand, the ratio is somewhat larger than 2, suggesting some abnormalities in nonsynonymous substitutions in the putative PBR or simply reflecting large sampling errors (the ratio of K_N^L and K_N^P is shown to have a skewed distribution; SATTA 1992). Nonetheless, for both class I and II loci, the estimated ratio of K_S^L to K_S^P at the synonymous sites is very close to the

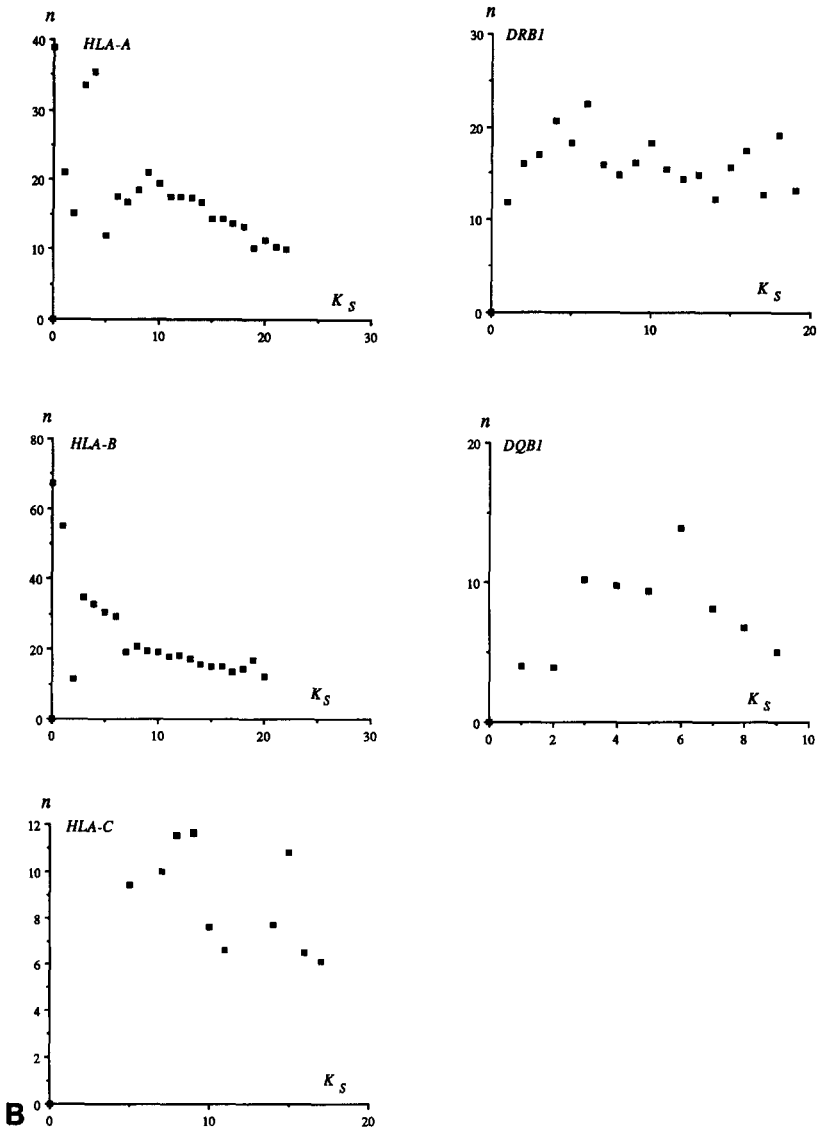

FIGURE 2.—B, Estimates of n based on Equation 11. The value of θ for each locus used is given in Table 2. Since the estimates decrease as $K_S = m$ increases, we used pairs of young alleles whose K_S values are no greater than k. The k value ranges from 6 to 10, depending on the locus (Table 3).

expected value (Table 1). In other words, the actual allelic genealogy is similar to that expected from the model used to derive Equations 1 and 2. This may in turn suggest that nonsynonymous changes in the class II PBR are somewhat evolutionarily different from those in the class I PBR (NEI and HUGHES 1991) due presumably to the differences in their definition and structure (KLEIN 1986; BJORKMAN *et al.* 1987a,b; BROWN *et al.* 1988).

The estimated value of n based on K_S^P is slightly larger than the number of different alleles at the C locus. At the *DRB1* locus, the estimated n from K_S^P values is about 16 although there are actually more than 25 different alleles (KLEIN, GUTKNECHT and FISCHER 1990). The estimate of n is 16 at the A and 18 at the B loci, while the number of electrophoretically or serologically detected alleles is 19 and 37, respectively (Tables 8.10-11 in KLEIN 1986). These discordances may be caused by the following factors: First, since the values of K_S and K_N for a given allelic pair are geometrically distributed (APPENDIX), their variances can be quite large. Second, K_N may be underestimated for the reason mentioned above. Third, in contrast to the class I loci, at the class II loci K_N is subjected to larger sampling errors because of the relatively small number of nonsynonymous sites ($L_N = 39$). Moreover, the class II PBR is putative and there is a possibility that other sites may be subjected to balancing selection or that nonselected sites have been included in the putative PBR. A failure to include all selected sites would lead to an underestimate of the number of alleles. This has apparently happened to the *DQB1* alleles for which only 10 of the 16 putative PBR codons were present in the partial sequences shown.

Conditional pairwise comparisons: In order to test whether K_N is really underestimated, we examined Equations 11 and 12. Figure 2B shows that, except at the *DRB1* locus, the n estimated from Equation 11 tends to decrease with large values of K_S. Such a decrease can be expected for rapidly evolving non-

TABLE 1

The pairwise mean and largest changes at the *HLA* loci

Sample	*A*	*B*	*C*	*DRB1*	*DQB1*
Nonsynonymous changes in PBR					
Sample size i	19	26	6	19	15
$2(1 - 1/i)$	1.89	1.92	1.67	1.89	1.87
$\hat{K}_N^P$	15.8 ± 2.5	18.3 ± 2.7	8.6 ± 2.0	15.6 ± 4.1	9.0 ± 3.5 (14.4)
$\hat{K}_N^L$	30.1 ± 6.2	34.2 ± 6.7	15.6 ± 4.2	39.4 ± 12.2	22.9 ± 8.5 (36.6)
Ratio[a]	1.91	1.87	1.81	2.53	2.54
Synonymous changes in the whole coding regions					
$\hat{K}_S^P$	11.7 ± 2.1	10.3 ± 1.6	10.0 ± 2.1	10.3 ± 2.0	4.7 ± 1.4
$\hat{K}_S^L$	22.2 ± 4.9	19.8 ± 4.6	16.7 ± 4.2	18.7 ± 4.6	9.3 ± 3.5
Ratio[a]	1.90	1.92	1.67	1.82	1.98

Each sample consists of alleles that differ at least by one nonsynonymous change in the PBR (a required condition to apply the theory of allelic genealogy). The sequences of *DQB1* cover only 10 out of the 16 exon 2 PBR codons, so that the K_N value is small (an extrapolation to the whole PBR is in parentheses). Symbol ^ indicates an estimate.

[a] The ratio is defined as $\hat{K}^L$ divided by $\hat{K}^P$. The (maximum) sampling errors (±) were computed from TAKAHATA and TAJIMA (1991).

synonymous sites in the PBR. For instance, for the value of K_S at a class I locus to be 10, $10/(2v) = 5/(L_S\mu)$ generations would be required on average since the two alleles diverged from each other. Recently, SATTA *et al.* (1991) estimated μ as 1.5×10^{-8} per site per generation assuming the generation time of primates to be 15 years. If we use this estimate of μ, we can compute the above allele divergence time as 1.3×10^6 generations or 20 million years (MY). This time period would be long enough for a number of nonsynonymous substitutions to occur in the PBR of these alleles, so that it might be difficult to estimate K_N by the usual correction methods (NEI 1987). We therefore used pairs of alleles whose K_S values were relatively small. We chose the values of k in Equation 12 as 6 to 10 because for higher values the conditional pairwise mean of K_N started to decrease at all class I loci. We did not choose lower values because they would result in a small number of pairs compared and thus large sampling errors.

The n thus estimated at each locus (Table 2) becomes about twice as large as that in Table 1. Also, it is close to or larger than the number of known alleles: a large difference occurs at the *C* locus at which the frequency of unidentified (*blank*) alleles is highest among class I loci (KLEIN 1986). KLEIN, GUTKNECHT and FISCHER (1990) listed more alleles that are distinguishable by T cell typing. However, alleles identified by serology or T cell typing do not always differ from each other at the nonsynonymous sites in the PBR. For instance, *DRB1*0701* and *0702* listed in KLEIN, GUTKNECHT and FISCHER (1990) are different by the typing but they do not have any change at the nonsynonymous sites in the PBR. Considering many other uncertainties about the number of alleles detected by different methods, the agreement between the number of alleles and that of the nonsynonymous substitutions is impressive.

TABLE 2

Estimates of n based on allele pairs with $K_S \leq k$ and the ratio (γ) of the nonsynonymous substitution rate to the synonymous rate

Parameter	*A* ($k = 10$)	*B* ($k = 9$)	*C* ($k = 9$)	*DRB1* ($k = 9$)	*DQB1* ($k = 6$)
n*	19	37	9	27	17
$\hat{n}$	27.3	36.2	17.0	23.5	15.1
$\hat{\theta}$	11.7	10.3	10.0	10.3	4.7
L_N	134.3	134.8	135.6	38.9	23.6
L_S	257.0	260.7	261.7	128.7	41.0
$\hat{\gamma}$	4.5	6.8	3.3	7.5	5.6
$\hat{h}$	0.044	0.038	0.037	0.074	0.103

The $\hat{\theta}$ is the same as $\hat{K}_S^P$ in Table 1. n*, the observed number of alleles taken after KLEIN (1986) for class I (Unidentified alleles are counted as one), KLEIN, GUTKNECHT and FISCHER (1990) for *DRB1* and MARSH and BODMER (1991) for *DQB1*. Note that the values of $\hat{\gamma} = \hat{k}_N/\hat{k}_S = \hat{n}L_s/\hat{\theta}L_N$ are larger than those of HUGHES and NEI (1988). The $\hat{h}$ corresponds to Equation 13. The sampling error for this n cannot be obtained by usual methods. Roughly speaking, however, it is of the order of square root of $\hat{n}$ under the Poisson approximation and this assumption is valid for a small number of nucleotide differences (TAKAHATA 1991b).

The ratio (γ) of k_N to k_S was more than 6 for *B* and *DRB1*, about 5 for *A* and *DQB1*, and about 3 for *C* (Table 2). These values are much larger than those obtained by HUGHES and NEI (1988, 1989) who used all allelic pairs in smaller samples. For the overdominance-type selection to be compatible with the observed large number of alleles, the value of S must be fairly large (WRIGHT 1939; KIMURA and CROW 1964; YOKOYAMA and NEI 1979; MARUYAMA and NEI 1981; TAKAHATA and NEI 1990; TAKAHATA 1990). In this case, the nonsynonymous substitution rate is expected

TABLE 3

Estimates of population parameters $S = 2Ns$ and $M = Nu$ in the symmetric overdominance-type model

Parameter	*A*	*B*	*C*	*DRB1*	*DQB1*
$\hat{S}$	2371	6301	674	2929	2320
$\hat{M}$	0.29	0.09	0.36	0.04	0.13
$\hat{N}$	1.5×10^5	4.5×10^4	1.8×10^5	6.9×10^4	2.2×10^5

The u is the per-generation mutation rate of nonsynonymous changes in the PBR; in the case of class I, $u = \mu L_N = 135\mu$. If our previous estimate of μ is used (SATTA *et al.* 1991) and one generation amounts to 15 years, then $u \approx 10^{-9} \times 15 \times 135 = 2.0 \times 10^{-6}$. For $\hat{M}$ to be about 0.4, N must be as large as $1 \sim 2 \times 10^5$. If $N = 10^5$, s ranges from 0.4% to 3.2%. The theory then predicts that the longest allelic divergence in the sample could be as much as $2\alpha = 1\text{–}3 \times 10^6$ generations which amounts to 20–60 million years.

to be much higher than the synonymous substitution rate (TAKAHATA 1990, 1991b). Therefore the large value of γ is consistent with the overdominance-type model.

The values of γ were used in Equation 8 to estimate the value of S for each locus. The selection intensity is different at different loci, being strongest at the *B* locus ($S \approx 6000$), intermediate at the *A*, *DQB1* and *DRB1* loci ($S \approx 2000\text{–}3000$), and weakest at the *C* locus ($S \approx 700$). These figures correlate well with the extents of polymorphism at the loci. The M estimated from Equation 3 ranges from 0.04 to 0.36 (Table 3).

If we knew the nonsynonymous mutation rate (u) at the PBR, we could estimate N and s from the estimated values of M and S (Table 3). If $\mu = 1.5 \times 10^{-8}$ (*vide supra*) and therefore $u = \mu L_N$ is 2×10^{-6} at the class I PBR and 6×10^{-7} at the class II PBR, the mean value of N becomes approximately 10^5 (*e.g.*, KLEIN, GUTKNECHT and FISCHER 1990; TAKAHATA 1990) and s becomes 0.4% to 3.2%, depending on the locus. The estimate of N is about 10 times larger than that based on other protein polymorphisms (NEI 1987). TAKAHATA (1991a) discussed this discrepancy on the basis that *Mhc* polymorphism has lasted for tens of million years while other protein polymorphism (which is largely neutral) is a reflection of the relatively recent history of human populations.

We can also examine the length of time alleles persist and hence whether most alleles are *trans*-specific. The mean divergence time between two alleles (α) is about 10^6 generations. If we again take 15 years as the generation time of primates, one million generations amount to 15 MY. The mean time until i alleles coalesce to j ancestors can be computed by

$$2\alpha\left(\frac{1}{j} - \frac{1}{i}\right) \text{ for } 1 \leq j \leq i$$

(TAKAHATA 1990) which amounts to $30(1/j - 1/i)$ MY. Hence, several alleles could have predated the human-chimpanzee splitting, which is consistent with the concept of *trans*-species polymorphism of *HLA* alleles (KLEIN 1980; FIGUEROA, GÜNTHER and KLEIN 1988; LAWLOR *et al.* 1988; MAYER *et al.* 1988). Of course, these estimates depend on the assumption that the population has been at equilibrium for a sufficiently long time. Whether or not this assumption is valid will be considered elsewhere (SATTA 1992).

FURTHER ANALYSES AND COMPUTER SIMULATION

In the previous section we were concerned with the expected pairwise mean distances (the number of nucleotide substitutions) at the selected sites in the PBR and synonymous (presumably neutral) sites. In this section, we study the distribution of the distances by computer simulation. Such a pairwise distribution does not have the usual probabilistic meaning unless it is applied to pairs of alleles sampled from unlinked, independent loci; the pairwise distances computed for a single locus are not independent and are necessarily correlated in their ancestry. Nevertheless, we may use this distribution to examine the internal consistency of the proposed model.

Following the simulation method of TAKAHATA (1990), we generated 10^4 independent genealogies for a sample of alleles. To determine the rate of allelic turnover, we used an estimated pairwise mean distance at the nonsynonymous sites in the PBR of each *HLA* gene. In each replicate, we computed the largest difference (d_{max}) between the observed and computer-generated distributions of pairwise distances using a statistical method similar to the Kolmogorov-Smirnov test (LEDERMANN 1984). These 10^4 d_{max} values were divided into 25 classes, each 0.04 wide, and a histogram was drawn. The distribution of d_{max} is broad, with the highest peak in the interval (0.44, 0.48). In this case, because the critical value of d_{max} at a 95% confidence level is approximately 1.36/5 = 0.272 (Figure 3; see also LEDERMANN 1984, pp. 194), it is concluded that the observed distributions for all loci are compatible with the model of symmetric overdominance-type selection.

As mentioned earlier, *Mhc* alleles differ greatly not only at the nonsynonymous sites in the PBR but also at the synonymous sites. In fact, some pairs of alleles differ at as many as 22 synonymous sites. This is a large number compared to that observed at other loci. The fact that there is only one synonymous change between the human and gorilla β-globin genes (SAVATIER *et al.* 1987) underscores this point. To demonstrate the correlation between synonymous and nonsynonymous changes, we produced a scatter diagram in which individual points represent the synonymous and nonsynonymous changes for pairs of alleles at a given locus. Again these variables are not statistically independent and there is a positive correlation

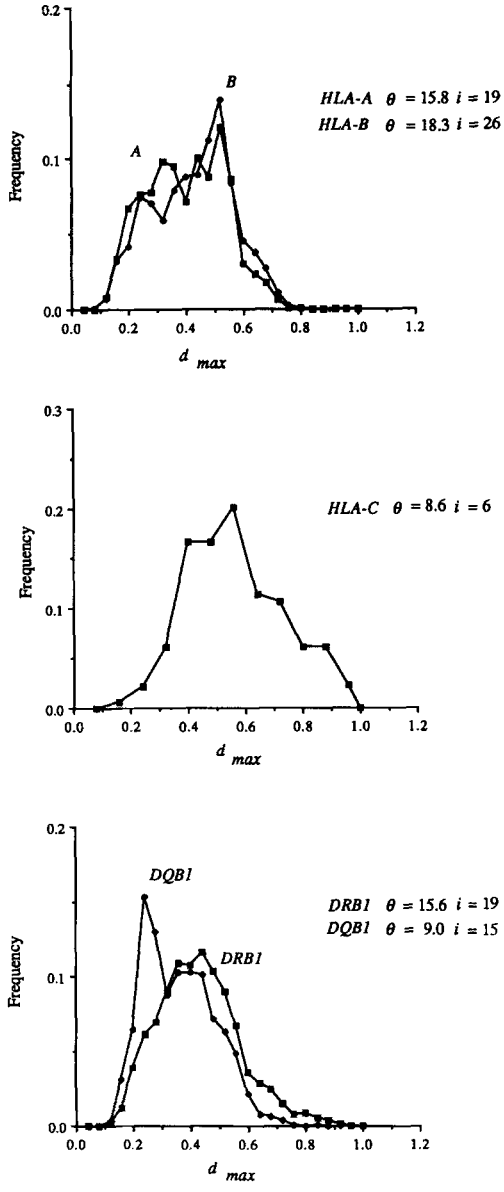

FIGURE 3.—Frequency spectrum of d_{max} obtained from 10^4 repeated simulations. The average number of pairwise nucleotide differences (θ) is set as the estimated number; 15.8 for *A*, 18.3 for *B*, 8.6 for *C*, 15.6 for *DRB1*, and 9.0 for *DQB1* alleles. The value of i indicates the number of alleles sampled from each locus (Table 1).

between them. To show that such a correlation is in part due to linkage between synonymous and nonsynonymous sites, we generated an allelic genealogy, superimposed on it two types of mutation, synonymous and nonsynonymous, and computed the correlation coefficient R. Repeating this process 10^4 times, we obtained the distribution of R which we could then compare to the observed values. Figure 4 shows that in the case of complete linkage the value of R ranges mostly from 0.5 to 1.0. By contrast, in the case of free recombination the value of R is concentrated in the range of (0, 0.3). Therefore a high value of R (0.744 in the case *DRB1*, for example) cannot be fortuitous,

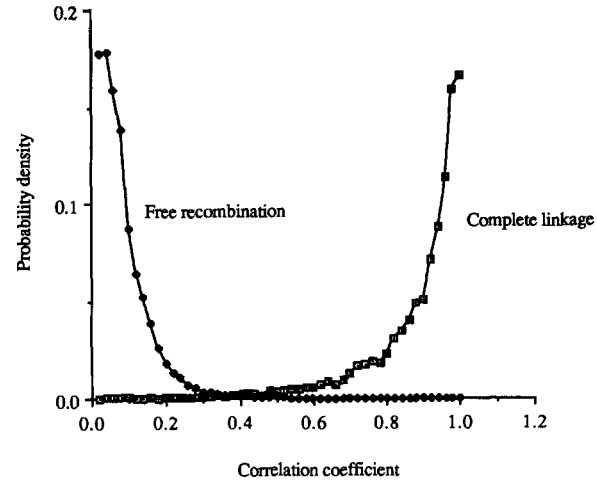

FIGURE 4.—Histograms obtained by computer simulation and showing the correlation between the pairwise differences at the nonsynonymous sites in the PBR and those at the synonymous sites. These two types of sites were assumed either to be in complete linkage or to recombine freely. The mean pairwise differences were chosen as 12.1 or 10.5, depending on synonymous or nonsynonymous changes. Each histogram was obtained by 10^4 repeats.

and balancing selection at nonsynonymous sites in the PBR also has an effect on the synonymous sites.

If linkage has such substantial effects on the accumulation of nucleotide changes at the synonymous sites, we can expect that heterozygosity (h) per nucleotide or amino acid site in regions near the PBR will be higher than that in the more remote regions capable of recombining with the PBR [see HUDSON, KREITMAN and AGUADÉ (1987) and HUDSON and KAPLAN (1988) for a related problem). However, since the mutation rate per synonymous site (μ) is as low as 10^{-8} per generation (HAYASHIDA and MIYATA 1983; LI, LUO and WU 1985; SATTA *et al.* 1991), h may not be very large. Quantitatively, in the absence of recombination the expected h is given by

$$E\{h\} = \frac{2\alpha\mu}{1 + 2\alpha\mu} \qquad (13)$$

(TAKAHATA 1991a), and the $E\{h\}$ ranges from 3.7% to 10%, depending on the locus (Table 2). Table 3 in HEDRICK *et al.* (1991) shows that the h values for *HLA-A* and *-B* loci averaged for the non-PBR sites are lower than those for the PBR sites, but that the observed h value in the non-PBR is 10–25 times higher than the 0.2–0.4% h value of human insulin, β-globin, growth hormone, and mitochondrial DNA (NEI 1987). Hence the h at *HLA-A* and *-B* loci is substantially increased in comparison to non-*HLA* loci and is in good agreement with Equation 13. The variance of h is, however, so large that highly polymorphic as well as monomorphic sites can exist near the PBR solely by chance (KIMURA 1983; NEI 1987).

DISCUSSION AND CONCLUSION

We have shown that the nonsynonymous substitution rate in the PBR may be at least twice as fast as that estimated by HUGHES and NEI (1988, 1989) and therefore much faster than the synonymous substitution rate (Table 3). The previous underestimates of the nonsynonymous substitution rate in the PBR are likely due to the difficulty in correcting extensive multiple hits at this site. The present results reinforce HUGHES and NEI's conclusion: the fact that the nonsynonymous substitution rate is higher than the synonymous rate provides strong evidence for balancing selection on *Mhc* polymorphism.

SERJEANTSON (1989) argues that the *Mhc* polymorphism can be accounted for by the neutral theory [see KIMURA (1968) and (1983) for review] and disassortative mating. As noted in CROW and KIMURA (1970), however, it is difficult to imagine strongly disassortative mating without selection. This is exemplified by self-sterility alleles (*e.g.*, CLARK and KAO 1991). Selection against homo-allelic pollination is the cause of disassortative mating despite no preference in prefertilization. The recent demonstration of female disassortative mating preference in mice (POTTS, MANNING and WAKELAND 1991) may be accounted for as a consequence of avoiding deleterious effects of inbreeding. However, assortative or disassortative mating causes less increase in homozygosity than inbreeding. Furthermore any mating preference based on a similar or dissimilar phenotype affects only segregating loci related directly to that trait (CROW and KIMURA 1970). Therefore the *Mhc* itself is likely to be the genomic locus whose heterozygosity is of principal evolutionary concern (HOWARD 1991). In this view, selection is the primary cause of *Mhc* polymorphism rather than secondarily developed mating preference. In any case, disassortative mating responsible for a particular set of *trans*-specific alleles must also be *trans*-specific; it would have to last for as long as 10–20 MY because some pairs of alleles are this old.

HILL *et al.* (1991) have provided evidence that *HLA-Bw53* allele and *DRB1*1302-DQB1*0501* haplotype are independently associated with protection from severe malaria. Like mating preference, however, if coevolution between *Mhc* alleles and pathogens is the main cause of *Mhc* polymorphism, it must have been transmitted through many speciation events (KLEIN 1991). Unfortunately the association of *HLA-Bw53* allele and *DRB1*1302-DQB1*0501* haplotype with malaria was estimated to be no more than 10,000 years old so that it does not account for the *trans*-species mode of *Mhc* polymorphism.

We have ignored the nonsynonymous changes outside the PBR against which purifying selection is thought to operate (HUGHES and NEI 1988, 1989). If the degree of selective constraint against alleles has not changed throughout the course of evolution, the rate of nonsynonymous substitutions can be expected to correlate with the synonymous rate. In fact the number of synonymous substitutions and that of nonsynonymous substitutions outside the PBR are nearly the same in many pairs of alleles, implying that the nonsynonymous sites in the non-PBR are conserved as in other loci (KIMURA 1983) and that the degree of selective constraint has been 1/3 for a long time. This is because there are about 2.7 times more nonsynonymous than synonymous sites per locus. For example, when *A3* at the *HLA-A* locus is compared to *A24*, there are about 10 synonymous and 9 nonsynonymous substitutions (in addition to 21 nonsynonymous substitutions in the PBR). Thus, the ratio of nonsynonymous to synonymous changes per site is about 1/3. However, there are some exceptional pairs in which the nonsynonymous substitutions are significantly larger or lower than the synonymous substitutions: in comparison of *A3* (*A2*) and *A32* (*A28*), there are about 10 (7) synonymous substitutions, but there are 20 (1) nonsynonymous substitutions outside the PBR. From a view of the neutral theory (KIMURA 1983), these deviations may be caused by changes in the degree of selective constraint. Further analysis of PBR-linked nonsynonymous changes will be made elsewhere.

Under the overdominance-type hypothesis, there may be deficiency of homozygotes (deviation from Hardy-Weinberg proportion) and there may exist substantial segregation load, L_g (CROW and KIMURA 1970). This load can be expressed by $L_g = sF$ and becomes

$$L_g = \frac{s}{n} = \sqrt{\frac{s}{4N} \ln \frac{s}{8\pi N u^2}}$$

in which equation the right hand side can be derived from KIMURA and CROW (1964). If we use the observed number of alleles for n, then $L_g \approx 0.05s$ for both *DRB1* and *DQB1*. Even when selection is as strong as $s = 0.1$, the L_g is no greater than one percent at either of these loci and the total genetic load for k such loci with $L_g = 0.01$ becomes $1 - (1 - L_g)^k = 1 - 0.99^k$. Therefore, once such unparalleled polymorphic loci as *Mhc* have evolved, constant segregation of homozygotes does not produce any substantial genetic load. It was noted long ago by CROW (1958) that the population can reduce L_g by increasing the number of alleles that are maintained under overdominance selection.

We have assumed that individuals in a population mate at random. However, since any natural population is to some extent structured geographically, we must consider the effects of population subdivision. If we assume neutral variation and if $4Nm > 1$ where N stands for the number of breeding individuals in each subpopulation and m is the gene migration rate per

generation, the whole population can be regarded as randomly mating and there is little local genetic differentiation (WRIGHT 1931). For balanced alleles, the ancestry is elongated by the factor $f_s = \alpha/(2N)$ relative to the neutral one so that there is a high probability that allelic lineages migrated over various subpopulations since they diverged from a common ancestor. In effect, operation of overdominance-type selection is equivalent to increasing N by f_s (TAKAHATA 1990) so that the condition for random mating would be $4Nf_sm > 1$. Since f_s can be much larger than 1, the extent of local genetic differentiation at balanced loci becomes relatively low even if the value of $4Nm$ does not greatly exceed 1. Compilation of gene frequency data (ROYCHOUDHURY and NEI 1988) shows a relatively small extent of geographic differentiation at *HLA* loci among human populations, although this does not necessarily deny the possibility of subdivided population structure in the early history of hominids.

All of these considerations lead to the conclusion that the overdominance-type model, which is also appropriate to the case of disassortative mating (KARLIN and FELDMAN 1968; POTTS, MANNING and WAKELAND 1991), is consistent with the main features of *Mhc* polymorphism and that currently there is no reason to reject it. As noted by DOHERTY and ZINKERNAGEL (1975) and KLEIN (1986), the biological cause of selection at the *Mhc* loci must be related to the fact that the T cells have dual specificity and that they simultaneously recognize viral (as well as other) antigens (nonself) and Mhc molecules of the stimulating and target cells (self).

We thank two anonymous reviewers and the corresponding editor for valuable suggestions. This work was supported by a grant from the Ministry of Education, Science and Culture, Japan, and grant R01 AI23667 from the National Institutes of Health, Bethesda, Maryland.

LITERATURE CITED

BLACK, F. E., and F. M. SALZANO, 1981 Evidence for heterosis in the HLA system. Am. J. Hum. Genet. **33:** 894–899.

BJORKMAN, P. J., M. A. SAPER, B. SAMRAOUI, W. S. BENNET, J. L. STROMINGER and D. C. WILEY, 1987a Structure of the human class I histocompatibility antigen, HLA-A2. Nature **329:** 506–512.

BJORKMAN, P. J., M. A. SAPER, B. SAMRAOUI, W. S. BENNET, J. L. STROMINGER and D. C. WILEY, 1987b The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens. Nature **329:** 512–518.

BROWN, J. H., T. JARDETZKY, M. A. SAPER, B. SAMRAOUI, P. J. BJORKMAN and D. C. WILEY, 1988 A hypothetical model of the foreign antigen binding site of class II histocompatibility molecules. Nature **332:** 845–850.

CLARK, A. G., and T.-H. KAO, 1991 Excess nonsynonymous substitution at shared polymorphic sites among self-incompatibility alleles of Solanaceae. Proc. Natl. Acad. Sci. USA **88:** 9823–9827.

CROW, J. F., 1958 Some possibilities for measuring selection intensities in man. Hum. Biol. **30:** 1–13.

CROW, J. F., and M. KIMURA, 1970 *An Introduction to Population Genetics Theory.* Harper & Row, New York.

DEGOS, L., J. COLOMBANI, A. CHAVENTRE, B. BENGTSON and A. JACQUARD, 1974 Selective pressure on HLA-A polymorphism. Nature **249:** 62–63.

DENNISTON, C., and J. F. CROW, 1990 Alternative fitness models with the same allele frequency dynamics. Genetics **125:** 201–205.

DOBZHANSKY, TH., 1970 *Genetics of the Evolutionary Process.* Columbia University Press, New York.

DOHERTY, P. C., and R. M. ZINKERNAGEL, 1975 Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. Nature **256:** 50–52.

FIGUEROA, F., E. GÜNTHER and J. KLEIN, 1988 MHC polymorphism pre-dating speciation. Nature **335:** 265–267.

HAYASHIDA, H., and T. MIYATA, 1983 Unusual evolutionary conservation and frequent DNA segment exchange in class I genes of the major histocompatibility complex. Proc. Natl. Acad. Sci. USA **80:** 2671–2675.

HEDRICK, P. W., and G. THOMSON, 1983 Evidence for balancing selection at HLA. Genetics **104:** 449–456.

HEDRICK, P. W., W. KLITZ, W. P. ROBINSON, M. K. KUHNER and G. THOMSON, 1991 Population genetics of HLA, pp. 248–271 in *Evolution at the Molecular Level,* edited by R. K. SELANDER, A. G. CLARK and T. S. WHITTAM. Sinauer Associates, Sunderland, Mass.

HILL, S. A., C. E. M. ALLSOPP, D. KWIATKOWSKI, N. M. ANSTEY, P. TWUMASI, P. A. ROWE, S. BENNETT, D. BREWSTER, A. J. MCMICHAEL and B. M. GREEWOOD, 1991 Common West African HLA antigens are associated with protection from severe malaria. Nature **352:** 595–600.

HOWARD, J. C., 1991 Disease and evolution. Nature **352:** 565–567.

HUDSON, R. R., and N. L. KAPLAN, 1988 The coalescent process in models with selection and recombination. Genetics **120:** 831–840.

HUDSON, R. R., M. KREITMAN and M. AGUADÉ, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

HUGHES, A. L., and M. NEI, 1988 Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335:** 167–170.

HUGHES, A. L., and M. NEI, 1989 Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. Proc. Natl. Acad. Sci. USA **86:** 958–962.

JIN, L., and M. NEI, 1990 Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol. Biol. Evol. **7:** 82–102.

JUKES, T. H., and C. R. CANTOR, 1969 Evolution of protein molecules, pp. 21–32 in *Mammalian Protein Metabolism,* edited by H. N. MUNRO. Academic Press, New York.

KARLIN, S., and M. W. FELDMAN, 1968 Further analysis of negative assortative mating. Genetics **59:** 117–136.

KIMURA, M., 1968 Evolutionary rate at the molecular level. Nature **217:** 624–626.

KIMURA, M., 1983 *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge.

KIMURA, M., and J. F. CROW, 1964 The number of alleles that can be maintained in a finite population. Genetics **49:** 725–738.

KINGMAN, J. F. C., 1982 On the genealogy of large populations. J. Appl. Probab. **19A:** 27–43.

KLEIN, J., 1980 Generation of diversity at MHC loci: implications for T cell receptor repertories, pp. 239–253 in *Immunology 80,* edited by M. FOUGEREAU and J. DAUSSET. Academic, London.

KLEIN, J., 1986 *Natural History of the Major Histocompatibility Complex.* John Wiley & Sons, New York.

KLEIN, J., 1990 *Immunology*. Blackwell Scientific Publications, Oxford.

KLEIN, J., 1991 Of HLA, tryps, and selection: an essay on coevolution of MHC and parasites. Hum. Immunol. **30:** 247–258.

KLEIN, J., J. GUTKNECHT and N. FISCHER, 1990 The major histocompatibility complex and human evolution. Trends Genet. **6:** 7–11.

KLITZ, W., and G. THOMSON, 1987 Disequilibrium pattern analysis. II. Application of Danish HLA *A* and *B* locus data. Genetics **116:** 633–643.

LAWLOR, D. A., J. ZEMMOUR, P. P. ENNIS and P. PARHAM, 1988 Evolution of class I MHC genes and proteins: From natural selection to thymic selection. Nature **335:** 268–271.

LEDERMANN, W. (Editor), 1984 *Handbook of Applicable Mathematics*, Vol. VI (Part A). John Wiley & Sons, New York.

LEWONTIN, R. C., 1985 Population genetics. Annu. Rev. Genet. **19:** 81–102.

LEWONTIN, R. C., L. R. GINZBURG and S. D. TULJAPURKAR, 1978 Heterosis as an explanation for large amounts of genic polymorphism. Genetics **88:** 149–170.

LI, W.-H., C.-C. LUO and C.-I. WU, 1985 Evolution of DNA sequences, pp. 1–94, in *Molecular Evolutionary Genetics*, edited by R. J. MACINTYRE. Plenum, New York.

MARSH, S. G. E., and J. G. BODMER, 1991 HLA Class II nucleotide sequences, 1991. Immunogenetics **33:** 321–334.

MARUYAMA, T., and M. NEI, 1981 Genetic variability maintained by mutation and overdominant selection in finite populations. Genetics **98:** 441–459.

MAYER, W. E., M. JONKER, D. KLEIN, P. IVANYI, G. VAN SEVENTER and J. KLEIN, 1988 Nucleotide sequences of chimpanzee MHC class I alleles: evidence for *trans*-species mode of evolution. EMBO J. **7:** 2765–2774.

MCCONNELL, T. J., W. S. TALBOT, R. A. MCINDOE and E. K. WAKELAND, 1988 The origin of MHC class II gene polymorphism within the genus *Mus*. Nature **332:** 651–654.

NEI, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.

NEI, M., and A. L. HUGHES, 1991 Polymorphism and evolution of the major histocompatibility complex loci in mammals, pp. 222–247 in *Evolution at the Molecular Level*, edited by R. K. SELANDER, A. G. CLARK and T. S. WHITTAM. Sinauer Associates, Sunderland, Mass.

POTTS, W. K., C. JO MANNING and E. K. WAKELAND, 1991 Mating patterns in seminatural populations of mice influenced by MHC genotype. Nature **352:** 619–621.

RITTE, U., E. NEUFELD, C. O'HUIGIN, U. MORTO, F. FIGUEROA and J. KLEIN, 1991 Possible selection for H-2 heterozygotes in natural populations of the house mouse, pp. 435–440 in *Molecular Evolution of the Major Histocompatibility Complex*, edited by J. KLEIN and D. KLEIN. Springer-Verlag, Heidelberg.

ROBERTSON, A., 1962 Selection for heterozygotes in small populations. Genetics **47:** 1291–1300.

ROYCHOUDHURY, A. K., and M. NEI, 1988 *Human Polymorphic Genes: World Distribution*. Oxford University Press, Oxford.

SAVATIER, P., G. TRABUCHET, Y. CHEBLOUNE, C. FAURE, G. VERDIER and V. M. NIGO, 1987 Nucleotide sequences of the β-globin genes in gorilla and macaque: the origin of nucleotide polymorphisms in human. J. Mol. Evol. **24:** 309–318.

SATTA, Y., 1992 Balancing selection at *HLA* loci, in *The Proceedings of the 17th Taniguch Symposium*, edited by N. TAKAHATA. Japan Scientific Societies Press, Tokyo (in press).

SATTA, Y., N. TAKAHATA, C. SCHÖNBACH, J. GUTKNECHT and J. KLEIN, 1991 Calibrating evolutionary rates at the major histocompatibility complex loci, pp. 51–62 in *Molecular Evolution of the Major Histocompatibility Complex*, edited by J. KLEIN and D. KLEIN. Springer-Verlag, Heidelberg.

SERJEANTSON, S. W., 1989 The reasons for MHC polymorphism in man. Transplant. Proc. **21:** 598–601.

SPENCER, H. G., and R. W. MRAKS, 1988 The maintenance of single-locus polymorphism. I. Numerical studies of a viability selection model. Genetics **120:** 605–613.

TAJIMA, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437–460.

TAKAHATA, N., 1990 A simple genealogical structure of strongly balanced allelic lines and *trans*-species evolution of polymorphism. Proc. Natl. Acad. Sci. USA **87:** 2419–2423.

TAKAHATA, N., 1991a *Trans*-species polymorphism of HLA molecules, founder principle, and human evolution, pp. 29–49 in *Molecular Evolution of the Major Histocompatibility Complex*, edited by J. KLEIN and D. KLEIN. Springer-Verlag, Heidelberg.

TAKAHATA, N., 1991b Overdispersed molecular clock at the major histocompatibility complex loci. Phil. Trans. R. Soc. Lond. B **243:** 13–18.

TAKAHATA, N., and M. NEI, 1985 Gene genealogy and variance of interpopulational nucleotide differences. Genetics **110:** 325–344.

TAKAHATA, N., and M. NEI, 1990 Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. Genetics **124:** 967–978.

TAKAHATA, N., and F. TAJIMA, 1991 Sampling errors in phylogeny. Mol. Evol. Biol. **8:** 494–502.

TAVARÉ, S., 1984 Lines-of-descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. **26:** 119–164.

THOMSON, G., 1988 HLA disease associations: models for insulin dependent diabetes mellitus and the study of complex human genetic disorders. Annu. Rev. Genet. **22:** 31–50.

TIWARI, J. L., and P. TERASAKI, 1985 *HLA and Disease Associations*. Springer-Verlag, New York.

UZZELL, T., and K. W. CORBIN, 1971 Fitting discrete probability distributions to evolutionary events. Science **172:** 1089–1096.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

WATTERSON, G. A., 1984 Lines-of-descent and the coalescent. Theor. Popul. Biol. **26:** 77–92.

WRIGHT, S., 1931 Evolution in Mendelian populations. Genetics **16:** 97–159.

WRIGHT, S., 1939 The distribution of self-sterility alleles in populations. Genetics **24:** 538–552.

WRIGHT, S., 1969 *Evolution and the Genetics of Populations, Vol. 2. The Theory of Gene Frequencies*. University of Chicago Press, Chicago.

YOKOYAMA, S., and M. NEI, 1979 Population dynamics of sex-determining alleles in honey bees and self-incompatibility alleles in plants. Genetics **91:** 609–626.

ZEMMOUR, J., and P. PARHAM, 1991 HLA Class I nucleotide sequences, 1991. Immunogenetics **33:** 310–320.

Communicating editor: A. G. CLARK

APPENDIX

To derive Equations 1 and 2, we assume equivalence of alleles in an allele turnover, by which we mean that it is equally likely that all existing alleles can produce a descendant allele or become extinct. In each allele turnover, one of two things can happen. In some cases, the parental allele *PA* becomes extinct after producing descendant allele *DA* which is one mutational step away from the *PA* (event *A*). This event occurs with the probability of $1/n$ where n is

the total number of segregating alleles in a population. In other cases, *PA* and *DA* both survive while other *PA* goes to extinction (event *B*). This event occurs with the probability of $1 - 1/n$. We sample a set of i alleles at random from n alleles segregating in a population and denote the sample as S_i. In order to consider the number of nonsynonymous changes (K_N) for a random pair of alleles in the sample, S_i is divided into two mutually exclusive subsets, S_2 and S_{i-2}. The S_2 subset contains a particular pair of alleles and S_{i-2} contains the remaining alleles of S_i. The alleles that are not sampled are denoted collectively by S_{n-i}.

We can compute the conditional probabilities that under event *A* the descendant allele *DA* is included in S_2, S_{i-2}, or S_{n-i}

$$P\{DA \in S_2 \mid A\} = \frac{2}{n},\ P\{DA \in S_{i-2} \mid A\} = \frac{i-2}{n},\ \text{and}\ P\{DA \in S_{n-i} \mid A\} = \frac{n-i}{n}.$$

For event *B*, the two alleles, *PA* and *DA*, may be included in the same subset or in different subsets. The conditional probability that S_2 includes both *PA* and *DA* is

$$P\{PA,\ DA \in S_2 \mid B\} = \frac{2}{n(n-1)},$$

and that S_2 and S_{i-2} each include only one of the two alleles is

$$P\{PA \in S_2,\ DA \in S_{i-2} \mid B\} = P\{DA \in S_2,\ PA \in S_{i-2} \mid B\} = \frac{2(i-2)}{n(n-1)}.$$

Likewise, we have the conditional probabilities that other subsets include *PA* and *DA*

$$P\{PA, DA \in S_{i-2} \mid B\} = \frac{(i-2)(i-3)}{n(n-1)},$$

$$P\{PA, DA \in S_{n-i} \mid B\} = \frac{(n-i)(n-i-1)}{n(n-1)},$$

$$P\{PA \in S_2, DA \in S_{n-i} \mid B\} = P\{DA \in S_2, PA \in S_{n-i} \mid B\} = \frac{2(n-i)}{n(n-1)}$$

and

$$P\{PA \in S_{i-2},\ DA \in S_{n-i} \mid B\} = P\{DA \in S_{i-2},\ PA \in S_{n-i} \mid B\} = \frac{(i-2)(n-i)}{n(n-1)}.$$

The divergence or coalescence of *PA* and *DA* is observed only when both *PA* and *DA* or their direct descendants are included in a given subset of alleles.

To compute the mean of K_N^P, we first note that the unconditional probability that *PA* and *DA* are included in S_i and therefore that coalescence occurs in event *B* is given by

$$P\{PA,\ DA \in S_i\} = \frac{i(i-1)}{n^2}.$$

The probability that the coalescence occurs in the $(k+1)$-th event *B* for the first time is geometrically distributed

$$P_k = \frac{i(i-1)}{n^2}\left\{1 - \frac{i(i-1)}{n^2}\right\}^k.$$

The number of nucleotide changes in S_i increases by one only if S_i contains *DA*. If there is no coalescence in S_i in one allele turnover, the probability of $DA \in S_i$ (a nucleotide change observed in S_i) is given by

$$P\{DA \in S_i\} = \frac{i(n-i+1)}{n^2}$$

and the conditional probability that S_i includes *DA* becomes

$$P\{DA \in S_i\} = \frac{i(n-i+1)}{n^2 - i(i-1)}.$$

Hence, during k allele turnovers, the number of nucleotide changes in S_i is binomially distributed and has the probability generating function (*pgf*)

$$\left\{\frac{i(n-i+1)z}{n^2 - i(i-1)} + \frac{n(n-i)}{n^2 - i(i-1)}\right\}^k,$$

in which z is a dummy variable. Taking the expectation of the above *pgf* with respect to P_k, we have

$$\frac{i-1}{n-(n-i+1)z}.$$

This is the *pgf* of K_N in the sample of size i immediately before two genes in S_i diverged from a common ancestral allele. The expected value of K_N is $n/(i-1) - 1$. However, this value does not include the final allelic divergence under event *B* at which the number of K_N necessarily increases by one. To obtain the expectation of the mean pairwise, $E\{K_N^P\}$, we set $i = 2$ and add one to $n/(i-1) - 1$. Thus we have

$$E\{K_N^P\} = n.$$

Note that whereas $E\{K_N^P\}$ is the same as the mean for a random pair of alleles, their variances differ from each other. To obtain the variance of K_N^P, we must consider the ancestral relationships of four alleles [see TAJIMA 1983; TAKAHATA and NEI (1985) in the case of neutrality].

To study K_N^L, we consider two genes in S_2 that coalesce last in the sample S_i. Suppose that there are k allele turnovers before any coalescence in S_i occurs. In this situation, the conditional probability that K_N^L

in S_2 does or does not increase in each allelic turnover is given by

$$q = \frac{2(n - i + 1)}{n^2 - i(i - 1)}$$

or

$$r = \frac{n(n - 2) - (i - 1)(i - 2)}{n^2 - i(i - 1)}.$$

Thus, the *pgf* of K_N^L conditioned on no coalescence during k allele turnovers becomes

$$(qz + r)^k.$$

Again taking the expectation of the above *pgf* with respect to P_k, the unconditional *pgf* becomes

$$\frac{i(i - 1)}{2n - 2(n - i + 1)z + (i - 1)(i - 2)}$$

and the mean becomes $2(n - i + 1)/\{i(i - 1)\}$. After the first coalescence in S_i, the sample size is reduced to $i - 1$ and the process continues until the size becomes 3. The mean of K_N^L in the whole process becomes

$$E\{K_N^L\} = \sum_{j=3}^{i} \frac{2(n - j + 1)}{j(j - 1)} + n$$
$$= 2n\left(1 - \frac{1}{i}\right) - \sum_{j=3}^{i} \frac{2}{j}.$$

Finally, we derive the formulas relating the number of nonsynonymous changes (K_N) to that of synonymous changes (K_S). We consider two randomly sampled alleles ($i = 2$). The divergence time of such alleles was shown to be exponentially distributed with mean α (TAKAHATA 1990). The mutation rate per generation is u and v for the selected sites and synonymous sites, respectively. For a given allelic divergence time t, K_N and K_S follow Poisson distributions with mean nt/α and $2vt$ (TAKAHATA 1990). Taking the expectation of random variable t with respect to the exponential distribution, we have the *pgf* of K_N and K_S to be geometric

$$Q(z_N, z_S) = \frac{1}{1 + n(1 - z_N) + \theta(1 - z_S)}$$

in which $\theta = 2\alpha v$ as in the text, and z_N and z_S are dummy variables. The probability of $K_N = j$ and $K_S = m$ is given by the coefficient of z_N^j and z_S^m in $Q(z_N, z_S)$. Since from

$$P\{K_N = j, K_S = m\} = \frac{(j + m)!}{(1 + \theta + n)j!m!} x^j y^m,$$

$$x = \frac{n}{1 + \theta + n} \text{ and } y = \frac{\theta}{1 + \theta + n},$$

the probability of $\{K_N = j, 0 \le K_S \le k\}$ is

$$P\{K_N = j, 0 \le K_S \le k\} = \sum_{m=0}^{k} \frac{(j + m)!}{(1 + \theta + n)j!m!} x^j y^m$$

and

$$P\{0 \le K_S \le k\} = 1 - \left\{\frac{\theta}{1 + \theta}\right\}^{k+1}.$$

For given values of K_S, the conditional probability of $K_N = j$ can be computed by dividing $P\{K_N = j, 0 \le K_S \le k\}$ by $P\{0 \le K_S \le k\}$, and the conditional mean of K_N is given by

$$E\{K_N \mid 0 \le K_S \le k\}$$
$$= n\left\{1 - \frac{(k + 1)\theta^{k+1}}{(1 + \theta)^{k+2} - (1 + \theta)\theta^{k+1}}\right\}.$$