Annu. Rev. Genet. 1998. 32:415–35

NATURAL SELECTION AT MAJOR HISTOCOMPATIBILITY COMPLEX LOCI OF VERTEBRATES

Austin L. Hughes and Meredith Yeager
Department of Biology and Institute of Molecular Evolutionary Genetics,
The Pennsylvania State University, University Park, Pennsylvania 16802;
e-mail: austin@hugaus3.bio.psu.edu

KEY WORDS: adaptive evolution, introns, major histocompatibility complex, overdominant selection

ABSTRACT

The loci of the vertebrate major histocompatibility complex encode cell-surface glycoproteins that present peptides to T cells. Certain of these loci are highly polymorphic, and the mechanisms responsible for this polymorphism have been intensely debated. Four independent lines of evidence support the hypothesis that MHC polymorphisms are selectively maintained: (*a*) The distribution of allelic frequencies does not fit the neutral expectation. (*b*) The rate of nonsynonymous nucleotide substitution significantly exceeds the rate of synonymous substitution in the codons encoding the peptide-binding region of the molecule. (*c*) Polymorphisms have been maintained for long periods of time ("trans-species polymorphism"). (*d*) Introns have been homogenized relative to exons over evolutionary time, suggesting that balancing selection acts to maintain diversity in the latter, in contrast to the former.

CONTENTS

0066-4197/98/1215-0415$08.00

INTRODUCTION

The major histocompatibility complex (MHC) of vertebrates is a multigene family whose products are cell-surface glycoproteins that play a key role in the immune system by presenting peptides to T cells (32). The MHC family includes two major subfamilies, called class I and class II. In most of the vertebrate species in which these genes have been mapped, the class I and class II families are linked together in a single gene complex. This complex is located on chromosome 6 in humans and is called the HLA complex (for "human leukocyte antigen"). In mammals, class I and class II genes are located in different regions of this complex, which are separated by a third region, sometimes called class III, that contains unrelated genes. Certain of the class I and class II genes have extraordinarily high levels of polymorphism, among the highest known in any organism (32). Furthermore, MHC polymorphisms are characterized by a large number of alleles of intermediate frequencies—a pattern of polymorphism inconsistent with selective neutrality but rather suggesting the action of some form of balancing selection (21).

To allow readers unfamilar with the MHC or with immunology to appreciate our discussion of the molecular evolution of MHC genes, we begin this review with an introduction to basic MHC biology. Then we discuss evidence that MHC polymorphism is maintained by balancing selection relating to the peptide-binding function of the MHC molecules and thus, ultimately, to disease resistance. Finally, we consider some unique aspects of MHC evolution that are ultimately consequences of this balancing selection.

STRUCTURE AND FUNCTION OF MHC MOLECULES

Class I Molecules

The polymorphic class I MHC molecules (called the class Ia molecules or class I classical molecules) are glycoproteins expressed on the surface of all nucleated somatic cells; they function to present peptides to cytotoxic T lymphocytes (CTL). The class I molecule is a heterodimer consisting of the following two chains: (*a*) an α chain or heavy chain, made up of three extracellular domains (designated α_1, α_2, and α_3), a transmembrane region, and a cytoplasmic domain; and (*b*) a molecule called β_2-microglobulin (β_2m), which consists of a single domain (Figure 1). β_2m is noncovalently linked to the α_3 domain. The α chains

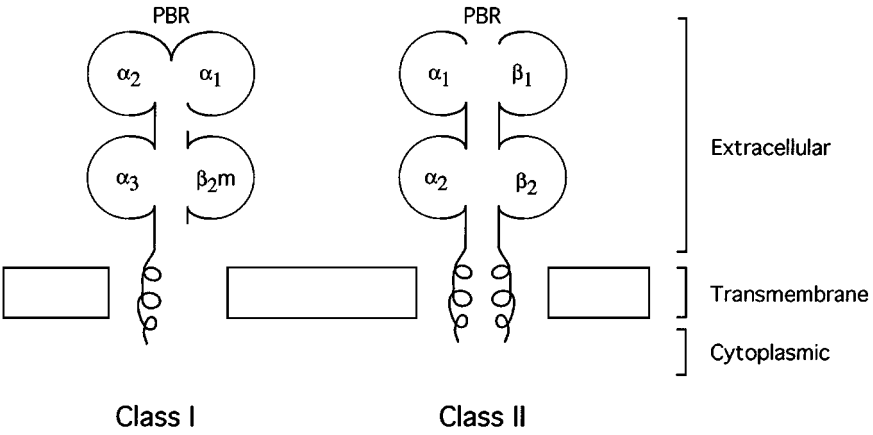


Figure 1 Schematic illustrations of MHC class I and class II molecular structure. PBR, peptide binding region.

are encoded within the MHC complex by the polymorphic class Ia loci, of which there are three in humans (*HLA-A*, *HLA-B*, and *HLA-C*). In mammals and probably in most other vertebrates, β_2m is encoded outside the MHC complex (on chromosome 15 in humans). β_2m shows evidence of a distant evolutionary relationship to class I α chains and to class II MHC molecules, but the locus encoding it is not polymorphic.

In all cells, there is a constant turnover of cellular proteins that are broken down into small peptides by a multimeric proteolytic complex in the cytoplasm known as the proteasome (46, 50). Proteasomes are present in all organisms, but their components are considerably more highly diversified in eukaryotes than they are in archaebacteria (22, 50). In mammals (and probably in most other vertebrates), there are two proteasome components encoded within the MHC class II region, called LMP2 and LMP7. (LMP is an abbreviation for low molecular mass polypeptide.) LMP2 and LMP7 are not expressed under all circumstances. Rather, two constitutively expressed components, called X and Y, are ordinarily expressed in their place (2). The cytokine γ-interferon enhances expression of both class I MHC molecules and LMP2 and LMP7. A proteasome containing LMP2 and LMP7 (called an LMP+ proteasome) has an altered specificity with regard to where it cleaves polypeptides; the LMP+ proteasome specifically produces peptides of a sort likely to be bound by class I MHC molecules (14, 17, 39).

These peptides are transported across the membrane of the endoplasmic reticulum (ER) by a dimeric transporter called TAP. The two subunits of TAP are themselves encoded in the MHC class II region. In the ER, a complex is

formed involving the class I MHC molecule, the peptide, and β_2m, which is then transported to the cell surface. When the first crystal structure of a class I MHC molecule was described, its most striking feature was a groove at the top of the molecule formed by two α helices bordering a β-pleated sheet (6, 7). Residues from both the α_1 and α_2 domains contribute to this groove (Figure 1). It seemed obvious that this groove was where the peptide is bound, a hypothesis later confirmed by crystallographic images of class I molecule complexed with peptides (18, 53). The class I peptide-binding region (PBR) consists of five pockets (pockets A–F) in which side-chains of the peptide residues fit (52).

In an uninfected cell, the peptides bound by a class I molecule are derived from the cell's own proteins (often called self peptides.). CTL exercise a continual surveillance in the body by means of their cell-surface receptors (T cell receptors or TCR). In the development of CTL in the thymus, TCR are selected so that the only CTL permitted to circulate are those that do not attack the complex of self class I MHC and self peptide (5). However, during infection by a virus or other intracellular parasite, some of the proteins broken down by the proteasome are of parasitic origin. Thus at least some of the class I molecules expressed on the surface of an infected cell will bind nonself or foreign peptides. When CTL encounter the complex of self class I MHC and foreign peptide, a cytotoxic reaction is initiated that kills the infected cell. CTL can only recognize foreign peptides in the context of self class I MHC; this phenomenon is known as class I MHC restriction of CTL. The CTL and class I MHC together thus provide a drastic solution to the problem of an intracellular parasite: killing all cells that harbor the infection (5). Mice that do not express β_2m and, thus, do not express class I MHC on their cell surfaces, suffer severe effects when exposed to intracellular pathogens. For example, these mice showed delayed viral clearance and increase mortality when infected with influenza virus (3) and 100% mortality when infected with the intracellular bacterium *Mycobacterium tuberculosis* (16). These results show that the class I MHC plays an essential role in immune defense against intracellular pathogens.

Class II Molecules

Class II MHC molecules have a much more restricted expression pattern than do class I molecules, in that they are expressed primarily on antigen-presenting cells of the immune system. The class II molecule presents peptides to helper T cells. In response to a foreign peptide, the helper T cells release cytokines that trigger an appropriate immune response (including the production of antibodies). The class II molecule is similar to the class I molecule in having four extracellular domains, but it achieves this structure in a rather different way (Figure 1). The class II molecule is a heterodimer consisting of an α chain

and a β chain, each of which in mammals is encoded in the class II region of the MHC complex. In placental mammals, the class II region is divided into subregions (designated DR, DP, and DQ in humans), each of which contains a functional α chain gene and one or more functional β chain genes. The α chain includes two extracellular domains (α_1 and α_2), a transmembrane region, and a cytoplasmic tail.

Like the class I molecule, the class II molecule binds peptides in a groove at the top of the molecule. As with class I, the class II peptide-binding groove consists of two α helices bordering a β-pleated sheet. The difference is that in class II, one of the α helices and about half of the β-pleated sheet are contributed by the α chain, whereas the other α helix and the other half of the β-pleated sheet come from the β chain (Figure 1). Unlike the peptides presented by class I, which are mainly 9 amino acids in length, the peptides presented by class II molecules can vary substantially in length, between about 11 and 17 residues (49). The reason for this difference is that in the case of class I the ends of the peptide are tucked down into the peptide-binding groove, limiting the peptide's length. In class II, the peptide's ends are free, which makes its length less constrained.

The complex between the class II molecule and its peptide ligand is created by a mechanism quite distinct from that of class I. Before transport to the cell surface, the class II dimer forms a complex with a polypeptide known as the invariant chain (Ii). This complex then travels to an acidic endosome-like compartment (47). There Ii is degraded, and the class II molecule binds the peptide which it transports to the cell surface. A molecule known as DM serves as a chaperone facilitating the loading of peptides by class II molecules (34). Interestingly, DM is clearly evolutionarily related to the class II molecule; it consists of an α chain and a β chain, each of which shows clear evidence of homology to the corresponding chains of the class II heterodimer (29).

EXPLAINING MHC POLYMORPHISM

The Overdominance Hypothesis

Zinkernagel & Doherty (59) first demonstrated class I MHC restriction of antigen (i.e. peptide) recognition by CTL. Soon afterwards, Doherty & Zinkernagel (13) proposed the first hypothesis for MHC polymorphism that took into account the actual biological function of these molecules. Doherty & Zinkernagel presented evidence that different class I MHC gene products differ with respect to the antigens that they can present. In other words, to express the concept in terms of our current knowledge of MHC function, different allelic products bind different arrays of peptides Thus, they argued, in a population exposed to an array of pathogens, it will be advantageous for an individual to be heterozygous

Table 1 Examples of anchor residues (boldface) and auxiliary anchor residues of peptides bound by the human class I loci MHC HLA-A and HLA-B

	Residue position								
Allele	1	2	3	4	5	6	7	8	9 (10)[a]
A1		T	D				L		Y
		S	E						
A*0201		L							V
		M							L
A68.1		V							R
		T							K
B*		H				I			
39011									
		R				V			L
						L			
B*4403		E							Y
									F
B53		P							

[a]The last residue position of the peptide is usually number 9 but may be number 10. Data are from Reference 49.

at MHC loci because a heterozygote will be able to present a broader array of antigens and thus resist a broader array of pathogens. Such a mechanism of heterozygote advantage (also known as overdominant selection) could account for the extraordinary polymorphism found at MHC loci.

Doherty & Zinkernagel's early evidence pointing to a difference between different MHC gene products with respect to the peptides they bind has been confirmed in recent years by sequencing peptides bound by MHC molecules. Comparisons of many such peptides have shown that the peptides bound by a specific MHC allelic product invariably contain one or more characteristic residues—called anchor residues, because they anchor the peptide into the binding groove (Table 1). In the case of the class I MHC, the anchor residues are usually the second residue of the peptide, which fits into the B pocket, and /or the ninth residue of the peptide, which fits into the F pocket (49). Because positions other than the anchor residues seem to be relatively free to vary, each allelic product can potentially bind thousands of different peptides. Nonetheless, because different allelic products have different anchor motifs, a heterozygote will presumably have much broader immune surveillance than a homozygote has. For example, HLA-B*39011 prefers the positively charged residues R or H in the second residue of the peptide, whereas HLA-B*4403 prefers the negatively charged residue E (Table 1). An individual heterozygous for these two alleles will be able to bind both types of peptides.

Doherty & Zinkernagel's hypothesis that overdominant selection maintains MHC polymorphism did not initially meet with wide acceptance. Some population geneticists had the mistaken impression that overdominant selection cannot maintain a polymorphism as extensive as that seen at MHC loci. This impression was based on theoretical models that did not take into account the role of mutation in incorporating new alleles. More realistic models that incorporated the role of mutation showed that overdominant selection is indeed capable of maintaining a high level of polymorphism (37).

An additional problem was the difficulty in testing the hypothesis of overdominant selection at MHC loci by means of a conventional population study. In an outbred species such as human or mouse, most individuals are heterozygous at most MHC loci. Thus, it would be necessary to survey many thousands of individuals to amass a large enough sample of homozygotes to compare their fitness with that of heterozygotes. Furthermore, if the selective advantage possessed by heterozygotes were small—say one or a few percent—then the sample size would have to be still larger to have the statistical power to test for a difference between homozygotes and heterozygotes.

Patterns of Nucleotide Substitution

Hughes & Nei (24) took a different approach to testing Doherty & Zinkernagel's hypothesis. By the late 1980s, a number of nucleotide sequences for MHC class I genes had become available. In most genes the rate of synonymous nucleotide substitution per site (d_S) exceeds that of nonsynonymous substitution per site (d_N). Theoretical study (36) had predicted that overdominant selection should enhance the rate of codon substitution; thus, if Doherty & Zinkernagel's hypothesis is true, d_N should be enhanced in the case of MHC genes. Furthermore, the first crystal structure of a class I MHC molecule had recently been published (6, 7), revealing the peptide-binding groove. On the hypothesis that MHC polymorphism is maintained by overdominant selection relating to peptide binding, Hughes & Nei (24) predicted that an enhanced nonsynonymous rate should be seen mainly in the codons encoding the PBR of the molecule.

The results dramatically confirmed this prediction. Figure 2 shows the results of recent analyses using many more sequences than were available to Hughes & Nei (24), but the results are essentially the same as they reported. In the 57 codons encoding the PBR, d_N significantly exceeds d_S (Figure 2). By contrast, in the non-PBR portions of the α_1 and α_2 domains and in the α_3 domain, d_S exceeds d_N, as is true of most genes (Figure 2). Note that d_S values do not differ greatly from one gene region to another. Since d_S is expected to reflect the mutation rate (the fraction of neutral mutations at synonymous sites being close to 100%), the uniform value of d_S indicates that the enhanced value of d_N in the PBR codons cannot be explained by a higher mutation rate in

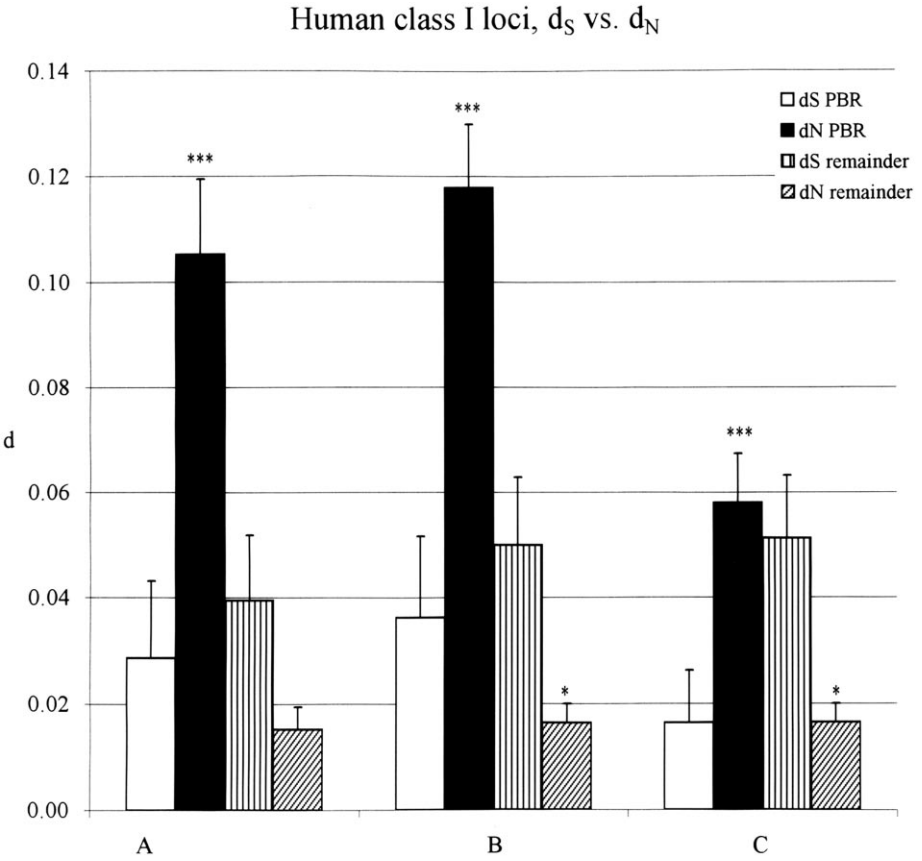


Figure 2 Mean numbers of synonymous (d_S) and nonsynonymous (d_N) nucleotide substitutions per site (41), with their standard errors (42), for pairwise comparisons among alleles at the *HLA-A* (*A*; 48 alleles). *HLA-B* (*B*; 99 alleles), and *HLA-C* (*C*; 35 alleles) loci. d_S and d_N were estimated separately for the peptide-binding region (PBR) codons and for the remainder of the α_1 and α_2 domains. Tests of they hypothesis that $d_S = d_N$: $^*P < 0.05$; $^{***}P < 0.001$.

those codons. Rather, the results strongly support the hypothesis that positive Darwinian selection has acted to enhance the rate of nonsynonymous substitution in the PBR codons and thus to enhance amino acid diversity in the PBR.

In the case of the class II MHC, a hypothetical structure was proposed by analogy with the known class I structure (8). Using this hypothetical structure, Hughes & Nei (25) found $d_N > d_S$ in the putative PBR. When a class II crystal structure was obtained (9), this finding was confirmed (Figure 3; 29).

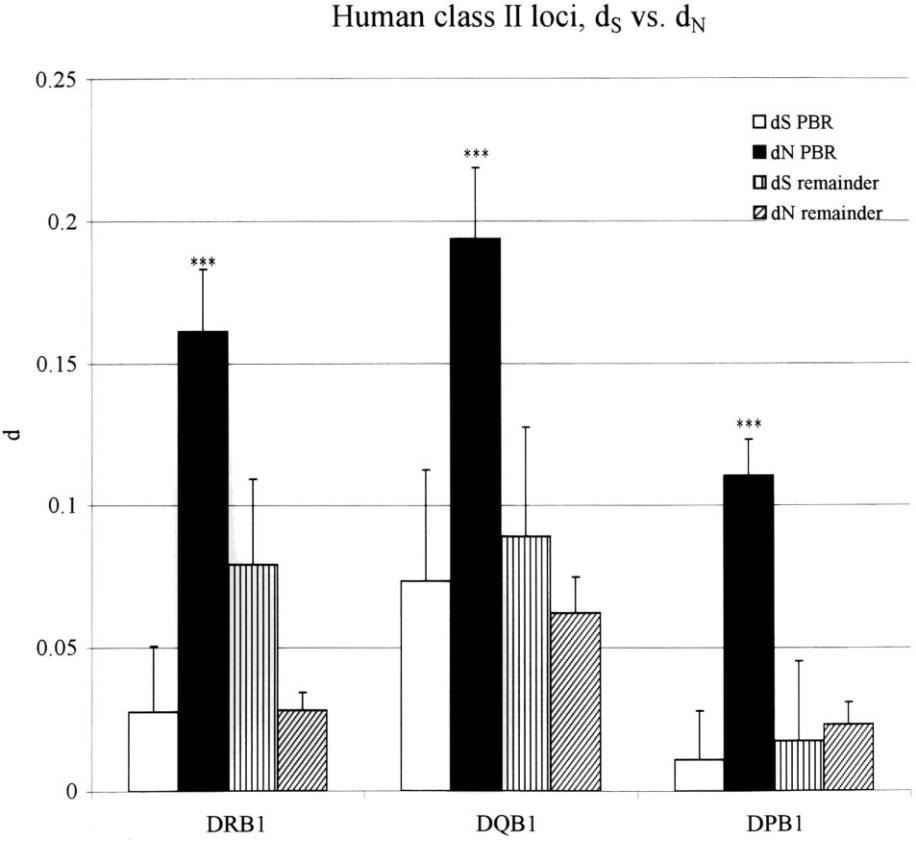


Figure 3 Mean numbers of synonymous (d_S) and nonsynonymous (d_N) nucleotide substitutions per site (41), with their standard errors (42), for pairwise comparisons among alleles at the *HLA-DRB1* (124 alleles), *HLA-DQB1* (14 alleles), and *HLA-DPB1* (49 alleles) loci. d_S and d_N were estimated separately for the peptide-binding region (PBR) codons and for the remainder of the β_1 domain. Tests of the hypothesis that $d_S = d_N$: $^{***}P < 0.001$.

Alternative Hypotheses

Comparison of rates of synonymous and nonsynonymous nucleotide substitution provided strong evidence that MHC polymorphism is maintained by some form of balancing selection; but it is still uncertain whether this selection is overdominant, as hypothesized by Doherty & Zinkernagel, or rather represents some other form of balancing selection. One form of balancing selection that has received a great deal of attention in the literature of theoretical population genetics is frequency-dependent selection. Of the several different models of

frequency-dependent selection, some are theoretically capable of maintaining a high level of polymorphism such as seen at MHC loci (55). In the case of the MHC, however, there is a clear rationale for overdominant selection based on the function of the molecules: namely, that heterozygotes have an advantage derived from a broader immune surveillance because a heterozygote can bind a broader spectrum of peptides than can a homozygote.

One early hypothesis to explain MHC polymorphism was independent of the molecules' function. This was the hypothesis that MHC loci have an unusually high mutation rate (1). DNA sequence data have made it possible to test this hypothesis rigorously. Because d_S is expected to reflect the mutation rate, d_S values for MHC genes can be compared with those of other genes to assess the comparative magnitude of the mutation rate at MHC loci. Such comparisons have shown that the mutation rates at MHC loci are below average for mammalian genes. For example, when d_S values between human and mouse were computed for 80 immunoglobulin superfamily C-type domains, the mean value was 0.654 ± 0.026 (23). The mean for class II β_2 domains, which are homologous to immunoglobulin C-type domains, was 0.526 ± 0.067 (23).

A more recent version of essentially the same hypothesis held that MHC polymorphism was enhanced by interlocus recombination (gene conversion) (36, 45). Theoretically, it is possible that if members of a gene family have diverged from each other at the sequence level and interlocus recombination subsequently occurs, polymorphism at each locus will be enhanced. However, gene conversion is expected to be an essentially random process; thus it cannot explain the very specific pattern of $d_N > d_S$ in the PBR codons that characterizes MHC loci (24, 25).

Because the function of MHC molecules was unknown for a long time, the earliest hypotheses to explain MHC polymorphism often attempted to explain both MHC function and polymorphism. Several popular hypotheses relied on analogies between the MHC and other biological systems. Most influential were hypotheses that saw an analogy between the MHC and the self-incompatibility systems of plants. The self-incompatibility loci are also extraordinarily polymorphic, and it was tempting to see the MHC as a vertebrate analogue. The result of this analogy was a proliferation of hypotheses relating the MHC to reproduction (e.g. 56). Even though the real function of the MHC is now known, some of these hypotheses have assumed a life of their own in the literature and continue to attract adherents.

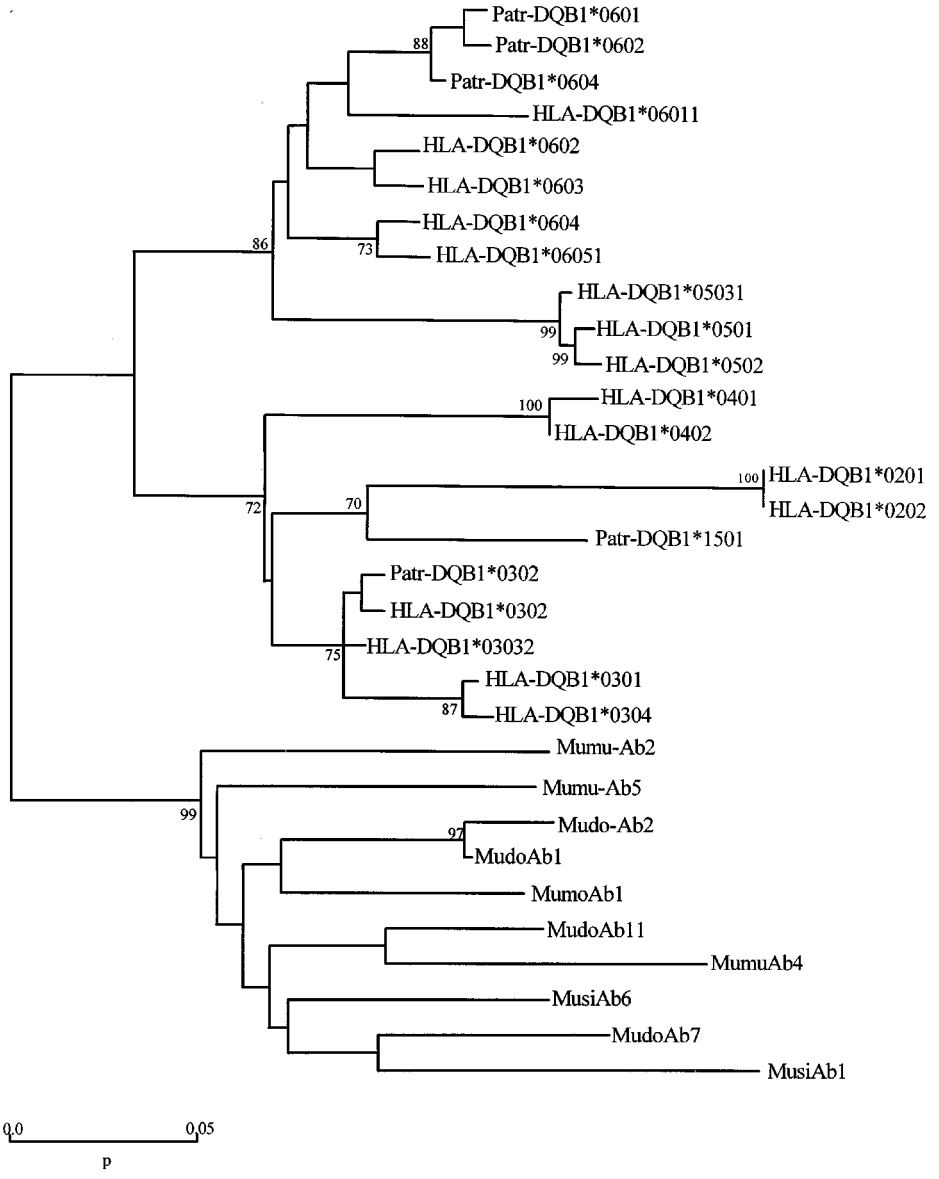
One hypothesis was that MHC polymorphism is maintained by maternal-fetal interactions (12). This hypothesis depends on the assumption that the production of maternal antibodies to fetal class I MHC molecules has a beneficial effect on fetal growth and survival. Although some early studies seemed to show such an effect, it has not been supported by subsequent work (11, 30, 32, 57).

Furthermore, it is hard to imagine how maternal-fetal interactions would lead to natural selection favoring diversity specifically in the PBR. On this hypothesis, one would predict that selection would enhance the rate of nonsynonymous substitution in epitopes for maternal antibodies. In the case of class I MHC molecules, these epitopes are scattered throughout the α_1 and α_2 domains, rather that being concentrated in the PBR (7, 40). Also, this hypothesis cannot account for the fact that MHC polymorphism is high in fish, amphibians, and birds, all of which lack maternal-fetal interactions. Finally, it cannot account for class II polymorphism since class II molecules are only expressed on antigen-presenting cells of the immune system and thus are unlikely to be involved in maternal-fetal interactions.

Another hypothesis is that MHC polymorphism is maintained by disassortative mating on the basis of MHC genotype, which is supposedly recognized via olfaction. Experiments cited as evidence for this phenomenon in mice (48, 58) lack relevant controls and are open to other interpretations (reviewed in 23). Data from the S-leut Hutterite religious isolate have been presented as evidence for MHC-based disassortative mate choice in humans (44). This is an endogamous population descended from a small number of founders, in which members avoid marriage to first cousins. In such a population, avoidance of closely consanguineous marriage alone will lead to a lower frequency of sharing of MHC haplotypes by spouses than would be expected under random mating. To test the hypothesis of MHC-based mate choice, it is necessary to compare the frequency of MHC-haplotype sharing between actual spouses with that between potential spouses (i.e. individuals of the appropriate sex, age, and degree of kinship to be chosen as spouses). So far, such a comparison has not been made (44). Furthermore, no evidence of MHC-associated mate choice was obtained in South Amerindians (20), whose population structure is probably more typical of human populations throughout history than are the highly inbred Hutterites. As with maternal-fetal incompatibility, it is hard to imagine how MHC-based mate choice would lead to natural selection focused specifically on the PBR.

TRANS-SPECIES POLYMORPHISM

A characteristic of MHC polymorphism that provides additional strong support for the hypothesis of balancing selection is the phenomenon called trans-species polymorphism. Often MHC polymorphisms are quite ancient, predating speciation events. For example, certain alleles at both class I and class II MHC alleles from human and chimpanzee belong to allelic lineages that have persisted since before these two species diverged 5–7 MYA (19, 35, 38). Figure 4 shows an example of trans-species polymorphism in the class II *DQB1* locus in



human and chimpanzee. In the phylogenetic tree of *DQB1* alleles, the human alleles *HLA-DQB1*0302* and *HLA-DQB1*03032* cluster with the chimpanzee allele *Patr-DQB1*0302* (Figure 4). On the other hand *HLA-DQB1*0501*, *HLA-DQB1*06011* and related human alleles cluster with the chimpanzee allele Patr-*DQB1*O601* and related chimpanzee alleles (Figure 4). These clusters of alleles apparently represent allelic lineages that were present in the common ancestor of chimpanzees and humans and have persisted in each population since their divergence.

Neutral polymorphisms are not expected to persist very long in populations. Coalescence theory predicts that, for pairs of neutral alleles selected at random from a locus in a randomly mating population, their mean coalescence time (that is, the time of their last common ancestor) will be $2N_e$ generations, where N_e is the long-term effective population size (55). Assuming a long-term effective population size of 10^4 for humans, the mean coalescence time for neutral alleles would be only 600,000 years. Thus, neutral polymorphism is, with respect to evolutionary time, a relatively transient phenomenon.

Under balancing selection, Takahata & Nei (55) showed that polymorphisms can persist much longer than in the neutral case. Using computer simulation, these authors studied overdominant selection and several models of frequency-dependent selection. They found that under overdominant selection and one type of frequency-dependent selection (which they called minority advantage), it was possible to maintain polymorphisms for very long times even with relatively modest selection coefficients. Thus either of these types of balancing selection can account for the long coalescence times of alleles at MHC loci.

In the model of minority advantage, it was assumed that a genotype had a selective advantage whenever it became rare in the population (54). Mathematically, this model turns out to be essentially the same as that of overdominant selection; however, from a biological point of view, it might well be questioned whether it is truly applicable to the MHC. Consider a parasite species that is easily eliminated by its host because most members of the host species bear an MHC allele (A1) whose product can bind and present a peptide from a given protein of the parasite. Then, suppose that a mutation occurs in the parasite so that this MHC allele can no longer bind the peptide, and the parasite is now able to infect most members of the host species with impunity. Suppose, however,

←

Figure 4 Phylogenetic tree of β_1 domain of human (HLA-) and chimpanzee (Patr-) DQB1 sequences, constructed by the neighbor-joining method (51) based on the proportion of amino acid difference (p). The tree is rooted with sequences from the orthologous *A* locus from *Mus* species: *M. domesticus* (Mudo-), *M. spicilegus* (Musi-), and *M. musculus* (Mumu-). The numbers on internal branches represent the percentage of 1000 bootstrap samples (15) supporting that branch; only values $>50\%$ are shown.

that there is a rare MHC allele in the host species (A2) whose product can bind another peptide from this parasite and protect against infection. Clearly, the A2 allele will have a selective advantage and will increase in frequency.

Now consider what will happen to the A1 allele. The minority advantage model assumes that, once the A2 allele becomes common, the A1 allele will again have a selective advantage. But how would this happen realistically in the case of the MHC? It might be that the parasite will mutate again so that the A2 allelic product can no longer efficiently bind a peptide from the parasite. But it seems rather unlikely that this new escape mutant will somehow restore binding by the A1 allelic product. Yet this is precisely what the minority advantage model requires. Therefore, this model may not be applicable to the case of the MHC.

On the other hand, after the parasite mutates so that the A2 allelic product no longer binds a peptide from its proteins, it might happen that still another new mutant MHC allele appears (A3), the product of which can efficiently bind a peptide from the parasite. If this happens, we can expect that the A3 allele will increase in frequency. This model is called the pathogen adaptation model by Takahata & Nei (55). However, this model cannot explain what is happening at MHC loci because rather than leading to a long-lasting polymorphism, this process will lead to a turnover of alleles over time (24, 55).

CLASS I INTRONS

Nucleotide Diversity in Exons and Introns

Recently, intron sequences have become available for a number of alleles at the human class I loci HLA-A, -B, and -C. The evolutionary dynamics of class I introns seems to differ strikingly from that of exons in ways that may seem surprising to some immunologists. Yet the properties of class I introns are in fact exactly what one would predict in the case of balancing selection. Thus, the analysis of class I introns has provided an additional, independent line of evidence that polymorphism at these loci is maintained by balancing selection.

Figure 5 shows plots of mean proportion of nucleotide different (p) for all pairwise comparisons among alleles at the HLA-A, -B, and -C loci. The regions of the gene compared are exons 2–3, which encode the α_1 and α_2 domains, including the PBR codons; exon 4, which encodes the conserved α_3 domain; and the first three introns of the gene (introns 1–3). One striking aspect of these

Figure 5 Proportion of nucleotide difference in a sliding window of 30 base pairs in all pairwise comparisons among alleles at the *HLA-A, -B,* and *-C* loci, from the beginning of intron 1 to the end of exon 4. *Horizontal bars* indicate the position of exons 2, 3, and 4. Reprinted from Reference 10.

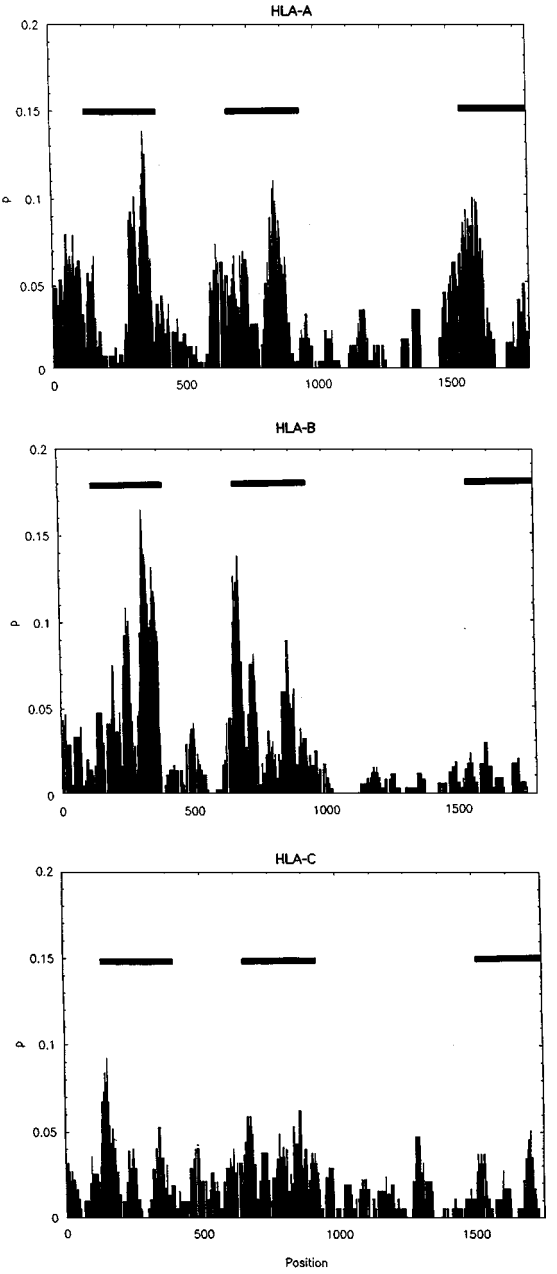


Table 2 Number of nucleotide substitutions per 100 sites (d) in introns 1, 2, and 3 and per 100 synonymous sites (d_S) in exons 2–3 for comparisons among HLA class I alleles

Comparison		d Intron 1	Intron 2	Intron 3	d_S Exons 2–3
Means for all pairwise comparisons (intralocus)[a]	*HLA-A* locus	4.2 ± 1.2	2.3 ± 0.6	2.2 ± 0.4	3.5 ± 0.8
	HLA-B locus	2.5 ± 0.9	1.6 ± 0.4***	0.7 ± 0.2***	4.9 ± 0.9
	HLA-C locus	1.6 ± 0.6	1.8 ± 0.5	1.4 ± 0.3**	3.6 ± 0.9
	All intralocus	2.7 ± 0.9	1.8 ± 0.5***	1.1 ± 0.3***	4.4 ± 0.9
Selected individual comparisons	*A*1101* vs *A*3002*	0.8 ± 0.8**	2.8 ± 1.0*	0.7 ± 0.4**	9.3 ± 2.8
	*A*2501* vs *A*2601*	0.0 ± 0.0	0.0 ± 0.0	2.6 ± 0.7***	0.0 ± 0.0
	B*0702 vs B*4201	0.0 ± 0.0	1.3 ± 0.7	0.7 ± 0.4	2.7 ± 1.4
	B*0702 vs B*5401	5.0 ± 0.2	0.0 ± 0.0	1.5 ± 0.5	6.8 ± 2.3
	Cw*0602 vs Cw*1203	1.6 ± 1.2	1.7 ± 0.9	0.4 ± 0.3*	4.6 ± 1.9

[a]Numbers of alleles compared for each locus are as follows: HLA-A, 15; HLA-B, 23; HLA-C, 12. Tests of the hypothesis that d in an intron equals d_S in exons 2–3: *P < 0.05; **P < 0.01; ***P < 0.001.

plots is that the mean p is generally lower in the introns than in the exons; this is particularly true of intron 3, which is much longer than either intron 1 or intron 2 (Figure 5). In most genes, this pattern would be reversed: p would be lower in exons than in introns because purifying selection eliminates most nonsynonymous mutations in exons (28).

A detailed examination of patterns of nucleotide substitution explains the usual results of the sliding window analysis. At each locus, the mean number of nucleotide substitutions per site (d) in introns 1–3 was compared with mean d_S in exons 2–3 (Table 2). Mean d_S in the exons was generally higher than mean d in the introns. This was particularly true for intron 3 and was most striking in the case of the *B* locus. At the *B* locus, mean d in intron 3 was less than 1%, whereas mean d_S in exons 2–3 was nearly 5% and seven times higher than mean d in intron 3 (Table 2). This result is very unusual because in the case of most genes d in introns and d_S in exons are about equal (28).

Hitch-Hiking under Balancing Selection

The most reasonable explanation for the fact that d in introns of human class I genes is often lower than d_S in exons is that introns are homogenized by

interallelic recombination and subsequent genetic drift (10). The exons of these genes—particularly those encoding the PBR—are quite ancient, having been maintained for millions of years by balancing selection. However, this selection does not apply to introns. Though intron sequences may hitch-hike along with exon sequences to some extent, if recombination and drift lead to loss of ancient polymorphism in an intron, this will be selectively neutral. Thus, introns of MHC genes are expected to be evolutionarily younger on average than are the exons encoding the PBR. Both d_S in exons and d in introns are expected to reflect the mutation rate, since most mutations at synonymous sites and at sites in introns are selectively neutral (28). When d_S in the exons is much higher than d in adjacent introns, the most straightforward interpretation is that the exons are older than the introns.

Population geneticists have extensively studied the problem of polymorphism at a locus linked to one under balancing selection (33, 43, 54). These studies predict that the degree of hitch-hiking—and thus the extent of polymorphism—at such a locus will be a function of the extent of recombination between that locus and the one under selection. Extending these predictions to the case of polymorphic class I MHC loci, we expect that introns more closely linked to exons 2–3 (encoding the PBR) will show higher levels of polymorphism than those less closely linked to exons 2–3. Introns 2 and 3 are relatively short (130 and 268 aligned nucleotide sites, respectively); and intron 1 is located just 5′ to exon 2, while intron 2 is located between exons 2 and 3. By contrast, intron 3, located 3′ to exon 3, contains 653 aligned sites; so, on average, nucleotide sites in intron 3 will be less closely linked to nonsynonymous PBR sites than will sites in introns 1 and 2. Thus, we might predict that intron 3 will be more likely to be homogenized by recombination and subsequent drift than will introns 1–2; the latter two introns are predicted to hitch-hike more closely with the PBR exons and thus to have higher levels of nucleotide diversity.

These predictions are supported by the data (Table 2). Intron 3 sequences show the lowest mean d for all three loci (Table 2). In addition, comparisons between individual sequences reveal some apparent recent cases of recombination. For example, the alleles *A*2501* and *A*2601* are identical in exons 2–3 and in introns 1 and 2, yet differ markedly in intron 3 (Table 2), which suggests that that this intron has been recently donated to one of these alleles by a more distantly related allele. On the other hand, *B*0702* and *B*5401*, though identical in intron 2, are highly divergent in the remainder of their sequence (Table 2). In this case, recombination seems to have caused the homogenization of intron 2 between these two alleles. These examples show that recombination in itself does not cause homogenization of introns among all alleles at a locus. Rather, over evolutionary time, genetic drift will lead to homogenization of introns, given that recombination occurs.

Bergstrom and colleagues (4) recently sequenced portions of the second intron from the human class II *DRB1* locus. They found that the intron sequences were much more similar to each other than were sequences of exon 2 (which includes the PBR); indeed, intron sequences were even more similar than were synonymous sites in exons. These results are very similar to the class I results, which suggests that introns are younger than exons because introns have been homogenized relative to exons by recombination and subsequent genetic drift. However, Bergstrom et al (4) favor a different interpretation of their data. They argue that in fact the introns reveal the true age of the MHC alleles. Thus, according to these authors, allelic lineages are not really as old as predicted under the hypothesis of trans-species polymorphism. Rather, they argue, *DRB1* alleles are really very recent.

Bergstrom et al (4) do not address the fact that their conclusions contradict those of previous studies that suggested that MHC alleles are very ancient. Moreover, if they are right, the class II exons have experienced very rapid recent evolution *at synonymous sites*. Positive selection will increase the rate of nucleotide substitution at nonsynonymous sites, but there is no known mechanism that will increase the rate of substitution at synonymous sites. Rather, we expect the rate of substitution at synonymous sites to be very similar to that at sites in introns; and this prediction is supported by comparison of mammalian introns and exons in the case of non-MHC genes (28). In spite of their interpretation, the Bergstrom et al (4) data clearly suggest that *DRB1* exon 2 sequences are much older than *DRB1* intron 2 sequences—as seen in class I and as predicted by population genetics theory (10).

It is important to distinguish the effects of hitch-hiking of one locus with another, linked locus in two different circumstances: (*a*) when the linked locus is under directional selection (which leads to fixation of a selectively favored allele); and (*b*) when the linked locus is under balancing selection. In the former case, both the locus under selection and a closely linked locus will show reduced polymorphism compared to neutral loci because of the recent fixation of the favorable allele. This phenomenon is often referred to as a selective sweep, because polymorphism linked to the favored mutation is swept out of the population as the new mutant goes to fixation. By contrast, a locus closely linked to a locus under balancing selection will show higher polymorphism than a neutral locus (43, 54). The polymorphism seen at such a linked locus will be a function of how tightly it is linked to the selected locus. In the case of *DRB1*, the selection is acting on nonsynonymous sites in the PBR codons in exon 2. Because sites in intron 2 of *DRB1* are less closely linked to PBR nonsynonymous sites than are synonymous sites in exon 2, the latter are expected to show a higher level of polymorphism because of their hitch-hiking with the PBR nonsynonymous sites, exactly as the *DRB1* data show (4).

ACKNOWLEDGMENTS

This research was supported by grants R01-GM34940 and K04-GM00614 from the National Institutes of Health.

Literature Cited

1. Baily DW, Kohn HI. 1965. Inherited histocompatibility changes in progeny of irradiated and unirradiated mice. *Genet. Res.* 6:330–40
2. Belich MP, Glynne RY, Senger G, Sheer D, Trowsdale J. 1994. Proteasome components with reciprocal expression to that of the MHC-encoded LMP proteins. *Curr. Biol.* 4:769–76
3. Bender BS, Crogham T, Zhang L, Small PA Jr. 1992. Transgenic mice lacking class I major histocompatibility complex restricted T cells have delayed viral clearance and increased mortality after influenza virus challenge. *J. Exp. Med.* 175: 1143–45
4. Bergstrom TF, Josefsson A, Erlich HA, Gyllensten UB. 1997. Analysis of intron sequences at the class II HLA-DRB1 locus: implications for the age of allelic diversity. *Hereditas* 127:1–5
5. Berke G. 1994. The binding and lysis of target cells by cytotoxic lymphocytes: molecular and cellular aspects. *Annu. Rev. Immunol.* 12:735–73
6. Bjorkman PJ, Saper MA, Samraoui B, Bennet WAS, Strominger JL, Wiley DC. 1987. Structure of the human class I histocompatibility antigen, HLA-A2. *Nature* 329:506–12
7. Bjorkman PJ, Saper MA, Samraoui B, Bennet WAS, Strominger JL, Wiley DC. 1987. The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens. *Nature* 329:512–18
8. Brown JH, Jardetzky T, Saper MA, Samraoui B, Bjorkman PJ, Wiley DC. 1988. A hypothetical model of the foreign antigen binding site of class II histocompatibility molecules. *Nature* 332:845–50
9. Brown JH, Jardetzky T, Saper MA, Samraoui B, Bjorkman PJ, Wiley DC. 1993. Three-dimensional structure of the human class II histo-compatibility antigen HLA-DR1. *Nature* 364:33–39
10. Cereb N, Hughes AL, Yang SY. 1997. Locus-specific conservation of the *HLA* class I introns by intra-locus homogenization. *Immunogenetics* 47:30–36
11. Clarke AG. 1971. The effects of maternal pre-immunization on pregnancy in the mouse. *J. Reprod. Fertil.* 24:369–75
12. Clarke B, Kirby DR. 1966. Maintenance of histocompatibility complex polymorphisms. *Nature* 211:999–1000
13. Doherty PC, Zinkernagel RM. 1975. Enhanced immunologic surveillance in mice heterozygous at the *H-2* gene complex. *Nature* 256:50–52
14. Driscoll J, Finley D. 1992. A controlled breakdown: antigen processing and the turnover of viral proteins. *Cell* 68:823–25
15. Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39:783–91
16. Flynn JL, Goldstein MM, Triebold KJ, Koller B, Bloom BR. 1992. Major histocompatibility complex class I-restricted T cells are required for resistance to *Mycobacterium tuberculosis* infection. *Proc. Natl. Acad. Sci. USA* 89:12013–17
17. Germain RN. 1994. MHC-dependent antigen processing and peptide presentation: providing ligands for T-lymphocyte activation. *Cell* 76:287–99
18. Guo H-C, Jardetzky TS, Garrett TPJ, Lane WS, Strominger JL, Wiley DC. 1992. Different length peptides bind to HLA-Aw68 similarly at their ends but bulge out in the middle. *Nature* 360:364–66
19. Gyllensten UB, Erlich HA. 1989. Ancient roots for polymorphism at the HLA-DQA locus in primates. *Proc. Natl. Acad. Sci. USA* 86:9986–90
20. Hedrick PW, Black FL. 1997. HLA and mate selection: no evidence in South Amerindians. *Am. J. Hum. Genet.* 61:494–96
21. Hedrick PW, Thomson G. 1983. Evidence for balancing selection at HLA. *Genetics* 104:449–56
22. Hughes AL. 1997. Evolution of the proteasome components. *Immunogenetics* 46: 82–92

23. Hughes AL, Hughes MK. 1995. Natural selection on the peptide-binding regions of major histocompatibility complex molecules. *Immunogenetics* 42:233–43
24. Hughes AL, Nei M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335:167–70
25. Hughes AL, Nei M. 1989. Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. *Proc. Natl. Acad. Sci. USA* 86:958–62
26. Hughes AL, Nei M. 1989. Evolution of the major histocompatibility complex: independent origin of nonclassical class I genes in different groups of mammals. *Mol. Biol. Evol.* 6:559–79
27. Hughes AL, Nei M. 1990. Evolutionary relationships of class II MHC genes in mammals. *Mol. Biol. Evol.* 7:491–514
28. Hughes AL, Yeager M. 1997 Comparative evolutionary rates of introns and exons in murine rodents. *J. Mol. Evol.* 45:125–30
29. Hughes AL, Hughes MK, Howell CY, Nei M. 1994. Natural selection at the class II major histocompatibility complex loci of mammals. *Phil. Trans. R. Soc. London Ser. B* 345:359–67
30. James DA. 1965. Effects of antigenic dissimilarity between mother and foetus on placental size in mice. *Nature* 205:613–14
31. James DA. 1967 Some effects of immunological factors on gestation in mice. *J. Reprod. Fertil.* 14:265–75
32. Klein J. 1986. *Natural History of the Major Histocompatibility Complex.* New York: Wiley
33. Kreitman M, Hudson RR. 1991. Inferring the evolutionary histories of the *Adh* and *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. *Genetics* 127:565–82
34. Kropshoffer H, Hammerling GJ, Vogt AB. 1997. How HLA-DM edits the MHC class II peptide repertoire: survival of the fittest? *Immunol. Today* 18:77–82
35. Lawlor DA, Ward FF, Ennis PD, Jackson AP, Parham P. 1988. HLA-A,-B polymorphisms predate the divergence of humans and chimpanzees. *Nature* 335:268–71
36. Lopez de Castro JA, Strominger JL, Strong DM, Orr HT. 1982. Structure of crossreactive human histocompatibility antigens HLA-A28 and HLA-A2: possible implications for the generation of HLA polymorphism. *Proc. Natl. Acad. Sci. USA* 79:3813–17
37. Maruyama T, Nei M. 1981. Genetic variability maintained by mutation and overdominant selection in finite populations. *Genetics* 98:441–59
38. Mayer WE, Jonker D, Klein D, Ivanyi P, van Seventer G, Klein J. 1988. Nucleotide sequence of chimpanzee MHC class I alleles: evidence for trans-species mode of evolution. *EMBO J.* 7:2765–74
39. Monaco JJ. 1992. A molecular model of MHC class-I-restricted antigen processing. *Immunol. Today* 13:173–78
40. Nathenson SG, Geliebter J, Pfaffenbach GM, Zaff RA. 1986. Murine major histocompatibility complex class-I mutants: molecular analysis and structure-function implications. *Annu. Rev. Immunol.* 4:471–502
41. Nei M, Gojobori T. 1986. Simple methods for estimating the number of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3:418–26
42. Nei M, Jin L. 1989. Variances of the average numbers of nucleotide substitutions within and between populations. *Mol. Biol. Evol.* 6:290–300
43. Nei M, Li W-H. 1980. Non-random association between electromorphs and inversion chromosomes in finite populations. *Genet. Res.* 35:65–83
44. Ober C, Weitkamp LR, Cox N, Dytch H, Kostyu D, Elias S. 1997. HLA and mate choice in humans. *Am. J. Hum. Genet.* 61: 497–504
45. Ohta T. 1982. Allelic and nonallelic homology of a supergene family *Proc. Natl. Acad. Sci. USA* 79:3251–54
46. Orlowski M. 1990. The multicatalytic proteinase complex: a major extralysosomal proteolytic system. *Biochemistry* 29: 10289–97
47. Peters PJ, Neefjes JJ, OOrschot V, Ploegh HL, Geuze HJ. 1991. Segregation of MHC class II molecules from MHC class I molecules in the Golgi complex for transport to lysosomal compartments. *Nature* 349:669–76
48. Potts WK, Manning CJ, Wakeland EK. 1991. Mating patterns in seminatural populations of mice influenced by MHC genotype. *Nature* 352:619–21
49. Rammensee H-G, Friede T, Stevanovic S. 1995. MHC ligands and peptide motifs: first listing. *Immunogenetics* 41:178–228
50. Rivett AJ. 1993. Proteasomes: multicatalytic proteinase complexes. *Biochem. J.* 291:1–10
51. Saitou N, Nei M. 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4:406–25
52. Saper MA, Bjorkman PJ, Wiley DC. 1991.

Refined structure of the human histocompatibility antigen HLA-A2 at 2.6 Å resolution. *J. Mol. Biol.* 219:277–319
53. Silver ML, Guo H-C, Strominger JL, Wiley DC. 1992. Atomic structure of a human MHC molecule presenting an influenza virus peptide. *Nature* 360:367–69
54. Strobeck C. 1983. Expected linkage disequilibrium for a neutral locus linked to a chromosomal arrangement. *Genetics* 103: 545–55
55. Takahata N, Nei M. 1990. Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics* 124:967–78
56. Thomas L. 1974. Biological signals for self-identification. In *Progress in Immunology II*, ed. L Brent, J Holborrow, pp. 239–47. Amsterdam: North-Holland
57. Wegmann TG. 1984. Foetal protection against abortion: is it immuno-suppression or immunostimulation? *Ann. Immunol.* 135B:307–12
58. Yamazaki K, Boyse EA, Mike V, Thaler HT, Mathieson BJ, et al. 1976. Control of mating preferences in mice by genes in the major histocompatibility complex. *J. Exp. Med.* 114:1324–35
59. Zinkernagel RM, Doherty PC. 1974. Immunological surveillance against altered self components by sensitized T lymphocytes in lymphocytic choriomeningitis. *Nature* 251:547–48

Annual Review of Genetics
Volume 32, 1998

CONTENTS