Evolution and ecology of MHC molecules: from genomics to sexual selection
Scott V. Edwards and Philip W. Hedrick

In the past few years the DNA sequence database for molecules of the MHC (major histocompatibility complex) has expanded greatly, yielding a more complete picture of the long-term rates and patterns of evolution of the MHC in vertebrates. Sharing of MHC allelic lineages between long-diverged species (ancestral-species evolution) has been detected virtually wherever it is sought, but new analyses of linked neutral regions and the complexities of sequence convergence and microrecombination in the peptide binding region challenge traditional phylogenetic analyses. Methods for estimating the intensity of selection on MHC genes suggest that viability is important, but recent studies in natural populations of mammals give inconsistent results concerning mate choice. The complex and interacting roles of microrecombination, parasite-mediated selection and mating preferences for maintaining the extraordinary levels of MHC polymorphism observed are still difficult to evaluate.

Genomics and long-term evolution
One way to appreciate the extraordinary diversity associated with the MHC is through the examination of its structure and proximate function in immune recognition. Although the molecular details of antigen-binding and processing continue to be refined, the basic function of class I and II MHC molecules in presenting degraded foreign and self-peptides to T-cells has remained stable (Box 1, Fig. 1). What will probably have more immediate impact on zoologists and human geneticists interested in MHC evolution are new discoveries emanating from the Human Genome Project. Currently, several teams in the US, Europe and Japan are poised to sequence the human and mouse MHCs in their entirety (approximately four million base pairs of DNA in each species). In addition, efforts to sequence the cosmid clusters mapped for the chicken MHC (or ‘B-complex’) are currently in progress (J. Kaufman, C. Audrey, L. Hood, pers. commun.). The first reports of such projects are just appearing \( ^2,^3 \). These will initiate several aspects of research on MHC ecology and evolution, including data on comparative organization, discovery of new genes, mapping of genomic regions influencing disease resistance, and the dynamics of duplications and isochores.

For example, in sequencing part of the human class I region, Mizuki et al. \( ^5 \) discovered a shared 40-kb segment downstream of both the \( HLA \) and \( C \) genes, providing detailed information on the sequences of boundary duplications that could have given rise to these paralogous genes (i.e. genes on the same chromosome that are homologous by duplication). Guillaudeau et al. \( ^1 \) compared 40-kb of noncoding DNA between the \( HLA \) and \( C \) genes for two haplotypes and in some regions detected differences in excess of 7% of sites—a very high level of polymorphism for human noncoding DNA, perhaps because of linkage to genes under balancing selection. In addition, they discovered a novel class I homology unit possessing conserved features of several class I genes, which they speculated could be involved in generating \( HLA \) diversity via gene conversion. Although many of the large-scale MHC sequencing efforts are still in a descriptive phase, they are expected to provide the most substantial database to date for testing molecular evolutionary hypotheses. Similar projects have already been done so for several completed genome projects.

In most vertebrates investigated, MHC genes of all three classes (I–II: Box 1, Fig. 1) are linked in a cluster on the same chromosome, and in chickens the two sections of the MHC (the B-complex and the \( Dp \)-gene) that appeared to segregate independently in pedigrees turn out to be linked on the same microchromosome but separated by a region of high recombination. However, in \( Xenopus \) frogs, a family of
nonclassical class I genes (MHC genes with low levels of expression and polymorphism) exists outside and is unlimited to the main MHC genes. MHC mapping in embryonic mouse tissues shows that class I and II genes are indeed found on different linkage groups, raising the possibility that the clustering of MHC genes in a single chromosome region in mammals could be derived from a formerly dispersed condition. In contrast, Kasahara et al. proposed a model in which a major chromosomal region harboring many genes duplicated as a block to another chromosome to give rise to the vertebrate MHC. This model was based on the finding that duplicated copies of several genes involved with antigen processing and other MHC-related functions (the gene pool of the proteasome, HSP70, a heat shock protein, and ABC transporters, Box 1) map to a non-MHC bearing chromosome in mice and humans. Consideration of the types and taxonomic distribution of genes participating in this duplication suggested that it may have occurred at or before the emergence of cartilaginous fish.

Recent attempts to detect MHC or proteasome subunit sequences (low molecular mass protein, LMP; Box 1) in primitive chordates have added to the controversy over the timing of the origin of MHC molecules. Several laboratories examining expressed genes in cyclostomes (jawless fish) have attempted but failed to detect MHC genes. In addition, although genes such as LMP2 and LMP7, which are thought to confer class II specificity to peptides generated by the proteasome (Box 1), are known in sharks, only their constitutively expressed and proteolytically more generalized counterparts (e.g. LMPX and LMPY) are known from jawless fish. A phylogenetic and molecular clock analysis of these and other LMP genes suggests that the divergence of the LMP7 and LMPX genes occurred some 600 million years ago, before the divergence of jawed and jawless fish. This in turn implies that the origin of class I MHC molecules also preceded the fish divergence and predicts that the MHC could have been lost in jawless fish. LMP genes are among those with paralogous copies on non-MHC bearing chromosomes.

**Box 1. The ABCs of the MHC**

The MHC is the "center of the immune universe". MHC molecules are members of the large immunoglobulin superfamily. The extracellular components of MHC molecules are composed of two main parts, a "stalk" (the immunoglobulin domain) that anchors the molecule in the surface of cells, and a receptor, shaped like a groove or basket, called the peptide binding region (PBR). The function of class I and II MHC molecules in immune recognition is to "display" short protein fragments ("peptides") of viruses and other pathogens on the surface of immune system cells, such as macrophages, or other cell types. Complement genes encode proteins that provide many effector functions critical to antibody-mediated immunity and inflammation. Class I PBRs, composed of α1 and α2 domains, are encoded by two exons in a single gene, are expressed on all cell types and bind peptides synthesized within cells, whereas class II PBRs are encoded by two genes (α and β chains), are expressed primarily on lymphocytes and bind peptides that enter such cells via phagocytosis. Most PBRs of class I and II genes are polymorphic but some, particularly of nonclassical MHC genes, are not. These nonpolymorphic genes can be on their way to becoming nonfunctional or may be specialized to bind certain phylogenetically conserved motifs or structures of bacterial peptides, thereby ensuring a constant ability to detect a wide range of related pathogens. The PBRs of both class I and II molecules might have originated from domains in heat shock proteins such as HSP70, which also bind peptides. In a process known as antigen processing, these peptides are, in the case of class I molecules, transported by polymorphic TAP ("transporter associated with antigen processing") molecules to a collection of protein-degrading enzymes (including low molecular mass proteins, or LMP) tightly bound in a structure called the proteasome in the endoplasmic reticulum. In the proteasome they are then cut into appropriate sizes, usually 9–20 amino acids, depending on the peptide source and ultimate destination. During or after this process, they are loaded onto empty MHC molecules. The peptide–MHC pair are then shuttled to the surface of cells where they can be interrogated by populations of T-cells, which have on their surface T-cell receptors, also members of the immunoglobulin superfamily. An appropriate match between peptide, MHC molecule and T-cell receptor is necessary to produce an immune response consisting of rapid proliferation of highly specific lineages of T-cells programmed to kill cells infected with the foreign peptide, or proliferation of cells that in response consisting of rapid proliferation of highly specific lineages of T-cells programmed to kill cells infected with the foreign peptide, or proliferation of cells that in turn produce antibodies that can smother and prepare such cells for phagocytosis.
Therefore, the interpretation of gene maps, which favors an origin of the MHC after jawless fish branched off, and that from phylogeny and molecular dating of duplicated LMP genes are in conflict. Further investigation of the immune systems of jawless fish (which appear to lack a fully-fledged MHC region), T-cell receptors and immunoglobulins, as well as scrutiny of the constancy of evolutionary rates among proteasome subunit genes, will clarify the picture.

Box 2. Multigene family evolution and the MHC

Three models that could potentially explain the long-term pattern of gene duplications in the MHC and other multigene families have recently been formalized.15,16 Each of these three models has distinct predictions of the shape of phylogenetic trees linking different genes together. In each tree in the figure, α, β, γ, and δ refer to different genes, and numbers refer to species in which those genes are found. Open circles are functional genes and filled circles are pseudogenes. The divergent evolution model (a) is the simplest model, in which gene duplication is followed by maintenance of paralogous in descendant species by selection. In the case of the MHC, there would be an advantage to an increased number of MHC receptors with which to combat pathogens, but theoretically there is an optimum number of genes, above which too many immune cells potentially recognizing self would be deleted during ontogeny of the immune system, resulting in lowered fitness.

The concerted evolution model (b), known for other multigene families, such as ribosomal genes, is marked by replacement of all gene family members in a genome by a single variant gene copy in the ancestor gene conversion and divergence of entire arrays in different species, resulting in groups of genes clustered by sequences between major vertebrate groups, genes within each lineage cluster together and orthologous relationships (i.e. homologies due to direct ancestry) are not retained, suggesting that ancestral MHC genes have duplicated independently in different lineages in rounds of expansion and contraction.17 The rates and timescales over which such duplications and contractions take place need to be estimated with greater rigor in more diverse lineages. The fitness advantages of families of comparatively closely related MHC genes, as opposed to the long-diverged ones exemplified by several mammalian class II genes, remain obscure and appear to contradict the paradigm that MHC genes are diverse at both the level of individual genes and gene families.

Box 2. Multigene family evolution and the MHC

Surprises continue to emerge from investigations of MHC structure in model and nonmodel organisms. Figueroa et al.18 found that the MHC class II B genes of cichlid fish and other Perciformes, as well as those of representative actinopterygian (rayfinned) fish, possess an extra intron in the immunoglobulin-like domain. This intron not only serves as a useful phylogenetic marker uniting rayfinned fish, but also provides further evidence for rounds of duplication of MHC genes after the divergence of the major groups of fish. Investigation of new MHCs inevitably reveals variation in intron length among species, and microsatellites occur in many MHC introns and flanking regions. Meager and Potts19 found that mouse microsatellites within the MHC were more polymorphic than those flanking the MHC, suggesting increased microsatellite diversity owing to hitchhiking with selected MHC genes.

Generation of new variants

Those of MHC genes that encode the peptide-binding region (PB) are the locus of much of the population genetic research on MHC diversity (Box 1). Under the neutral theory, the rate of synonymous nucleotide substitutions is predicted to be larger than the rate of nonsynonymous substitutions. However, as originally shown by Hughes and Nei20, the rate of nonsynonymous substitutions is generally higher than that for synonymous ones for PBR sites of MHC genes, suggesting the action of balancing selection.

There is agreement that evolution of immune response genes is an important role for gene conversion. However, the relative importance of point mutation versus conversion is still controversial. Indirect evidence for microrecombination comes from alleles that differ from one another only by small segments, and the demonstration that the majority of interlocus sequences on both sides of the PBR have similar evolutionary histories, a pattern that makes single-crossover recombination unlikely.21 The average size of the putative exchanges among human class I alleles is estimated to be between 15 and 20 nucleotides, and a recent global survey found evidence for only four single-point mutations, whereas at least 80 sequences suggested gene conversion.22 Recent analysis of class II MHC genes in mamman12 show that this is not the case, and that birth and death is a more appropriate model for these loci. The ‘birth-and-death model’ (c) posits frequent retention of ancestrally duplicated genes, as in the divergent evolution model, but also incorporates the evolution of pseudogenes, and deletion of genes such that the total gene number is approximately maintained.
and several orders of magnitude higher than the estimated rate of silent substitution (Box 3). Hogstrand et al.25 estimated the rate of interlocus gene conversion between two loci in mice as 0.5 × 10⁻⁶, much lower than the rate of interallelic gene conversion but high enough to influence variation in the genes involved. Although the amount of microrecombination appears to vary over MHC loci, these observations suggest a stronger role for microrecombination in generating diversity and reveal the inadequacy of models assuming independence of nucleotide changes, an assumption fundamental to many calculations of molecular evolution. In addition, they raise the question of whether these elevated estimates of microrecombination rates are in part due to the ease of detectability of such events in the MHC, which are apparently because of the long timescale of neutral and balanced polymorphisms in this region26.

Trans-species evolution: testing assumptions

Trans-species evolution refers to the retention of alleles across multiple species divergences over a long time, and its hallmark is a discordance between gene and species trees that cannot be explained solely by the retention of neutral polymorphisms over relatively short timespans after speciation. Indeed, the population dynamics of MHC alleles closely match those for one other system only: the self-incompatibility alleles in plants27. It is now common for surveys of sequence variability in the PBRs of the MHC to reveal trans-species patterns when multiple species are investigated. As more data and analyses have accumulated, however, the causes of these allelic patterns are becoming less obvious. An increasing number of examples of convergence in the PBR of mammalian MHC genes at both the nucleotide and amino acid sequence levels raises doubts about some claims for extremely long persistence times, particularly from analyses based solely on PBR sites16,27. Phylogenies of PBRs and adjacent (and presumably neutral) introns and exons invariably conflict, with neutral sites implying a more attenuated picture of trans-species evolution14,28.

In principle, the persistence of MHC alleles over long timespans provides unique opportunities for probing organismal histories far back in time—much longer than would be possible with neutral genes, such as mtDNA. This property combined with novel analytic methods have been used to estimate ancestral effective population sizes of humans29,30, chinchilla (Cichlidae)30 and most recently, Darwin’s finches (Geospiza)31. Each of these studies implied less extreme founder effects in the origin of these species than previously assumed because current levels of MHC diversity in these species were too high to be explained by chaotic bottlenecks in the past, assuming mutation rates of the order of 10⁻⁶ per site per year in the PBR (Box 3) and various selection coefficients. As these authors acknowledge, however, the large number of variables and assumptions required to implement such analyses (mutation rates, selection coefficients and absence of recombination) and difficulties in determining orthologous relationships among sequences for nonmammalian vertebrates make application to data from diverse creatures and environments challenging. For example, the much higher rates of change by non-neutral mechanisms32 and significant levels of post-speciation diversification suggested for some MHC loci might suggest alternative scenarios. Bergström et al.25 sequenced the two introns flanking the polymorphic second exon of 28 class II DRB1 alleles in 13 major allelic lineages of humans.

Surprisingly, within each allelic lineage, the sequence divergence of alleles implied a coalescence time of about 250 000 years, a common ancestry that is recent for a human nuclear gene and which contrasts with the more ancient coalescence implied by variation in the PBR itself26 or by the intra-diversity across allelic lineages at that locus33. Their data suggest recent origins for the majority of alleles at DRB1, after the divergence of clumps and humans, rather than prespeciation origins as suggested by coding sequences. Combined with the ancient divergences of allelic lineages at DRB1, as well as analyses of neutral genes25, the data suggest a scenario in which population size in the lineage leading to modern humans has oscillated over time, although the details of these oscillations, particularly in the past million years, will no doubt be refined as the molecular diversity for noncoding regions of MHC alleles is further investigated and compared with that of coding regions.

Box 3. Measuring evolutionary rates at MHC loci

Difficulties in measuring long-term rates of substitution arise in MHC genes because of the phylogenetic discordances with species trees created by trans-species evolution (see text). Trans-species evolution renders traditional methods of inferring relative or absolute rates of molecular evolution useless when applied to MHC variation because we cannot use species trees as a proxy for the gene trees we wish to study. Satta et al.34 helped overcome these problems by noting that, even in the absence of trans-species evolution, the closest pair of alleles found in two different species are likely to have diverged close to the time of the species split. This allows traditional rate tests to be used. The authors estimated rates of synonymous (silent) substitution in primates MHC genes at about 1 × 10⁻⁹ substitutions per site per year, a surprisingly low rate given the large genetic distances between primate alleles. The long persistence time of alleles is one way to reconcile such large divergences at silent sites with a low rate of silent substitution.

Edwards et al.35 used Satta et al.’s method and data on times of species divergence to show that silent rates in class II MHC DRB1 were as much as four times higher than those measured in primates (Box 3). The open symbols in each graph indicate a set of divergence times that is ambiguous. Solid and dotted lines indicate the rates estimated depending on whether the traditional or novel divergence times, respectively, are accepted for these groups. Relative rate tests, which do not depend on absolute divergence times, also suggested increased silent rates of substitution in rodents, implying that mammalian MHC genes exhibit the same types of generation-time effect as do neutrally evolving genes. Despite this consistency, estimating rates of silent substitution in this way requires assumptions of no recombination or gene conversions, the latter of which is thought to elevate synonymous substitution rates. Reproduced, with permission, from Ref. 53.
Sources of balancing selection

The types of selection envisioned for MHC genes have traditionally included heterozygote or rare-allele advantage. Although some types of frequency-dependent selection can be ruled out, both mechanisms are compatible with patterns of MHC polymorphism and various selective scenarios involving disease resistance, mate choice and/or reproductive incompatibilities between individuals. (In humans, the number of individuals that share MHC alleles is that associated with the main autoimmune diseases, but most of these diseases are uncommon and they tend to lower fitness in life, making unlikely players in maintaining MHC variation.) Recent surveys still reveal a significant excess of couples with a history of spontaneous abortions sharing a single or more MHC loci. (The immunological explanation for these observations, which would only apply to mammalian MHC diversity, is that the presence of an immune response occurring when the mother and father differ at HLA loci is necessary for proper implantation and fetal growth, a type of selection that could maintain many alleles at a single locus.

Following on the experiments of Yamazaki, Poole and co-workers implicating dissatisfactory mating as a mechanism to elevate MHC diversity, further experiments in mice demonstrating an effect of early experience on imprinting and mate choice. However, norman mating without respect to MHC type was not detected in sheep, and the data on human populations are contradictory, with MHC-based mate preferences detectable in some populations but not others. Studies indicating that female university students preferred the odors of MHC-disimilar males in some situations (but see Refs 37 and 39) and that body odor preferences in men and women are influenced by the degree to which they share MHC alleles suggest an olfactory mechanism for mate-choice similar to that known for mice. Singer et al. recently showed via gas chromatography that levels of volatile acids in mouse urine, in particular phenylacetic acid, differ between MHC types and that such quantitative differences can be discriminated by mice. Although it has still to be determined whether the evolution of chemosen-sory mate choice in mammals is a cause or a consequence of ancient MHC polymorphisms, the presence of urine of compounds such as phenylacetic acid, which are naturally con-jugated with amino acids in mouse urine and which therefore could be bound by MHC molecules, raises intriguing possibilities as to mechanisms of chemodiscrimination. The most traditional explanation for MHC diversity, namely heterozygote advantage (because MHC heterozygotes present a wider range of foreign peptides to T-cells), is still the most popular, although the number of examples of predicted associations between MHC genotypes and resistance to infectious disease is still small. It is still under-appreciated that some of the best examples of strong associations between MHC loci and parasite resistance or between MHC loci and expression of sexually selected traits come from birds. For example, 95% of chickens with certain haplotypes survive infection by Marek’s disease (a type of viral leukemia), whereas individuals with other haplotypes may have nearly 0% survival. MHC-disease associations in chickens might be stronger than in mammals because chicken MHC is compact, possibly intensifying the strength of selection on fewer target genes. Whereas these and other examples suggest selection reduces diversity because of resistance conferred by a specific haplotype, a much-needed example of selection that enhances diversity, in the form of heterozygote advantage, has recently emerged from analyses of HLA loci and resistance to hepatitis B virus among West Africans Consistent relationships between MHC variants and resistance could be difficult to detect even in control populations for a variety of reasons, including improper statistical design, low resolution of serotypes or FLP markers versus sequences, geographically variable MHC adaptation owing to geographical differentiation of the para-site, adaptation to parasites no longer present, influence of other linked or epistatic loci, overlap in the peptide repertoires of individual MHC alleles, or importance to parasite resistance of loci outside the MHC. Because some of the most important infectious diseases of the past are epidemic, selection models for maintaining MHC variation that incorporate differential selection over time and space may be most appropriate.

Getting to ‘s’

There are various ways in which the intensity of selection has been estimated at MHC genes. In general, the selective difference (s) between heterozygotes and homozygotes is estimated (Box 4). Several approaches measure the combined selection intensity over all or several fitness components. For example, in some aboriginal populations (which have a limited number of alleles, are fairly isolated, and have little or no sequence variation within serotype) large deviations from Hardy-Weinberg expectations, giving s values of approximately 0.3 in the current generation, have been observed. The level of selection derived from Hardy-Weinberg
deviations in many nonaboriginal populations can be hard to estimate either because nearly all individuals are heterozygous for the variable HLA loci or because past selection pressures are no longer evident owing to better health care and sanitation. Deficiencies of MHC homozygotes have also been reported in some4, but not all48, natural populations investigated so far. Using another indirect method based on the unusual pattern of variation in MHC genes, Satta et al. estimated r in humans to be 0.013, with the highest value of 0.042 for the HLA-B gene.

The extent of selection for viability, mate choice or parasite resistance can in principle be examined separately. The extent of viability selection can be determined by examining deviations from expected Mendelian segregation in families, and selection based on mate choice by determining the deviation of progeny-genotypes from random-mating proportions or by preference experiments. Recent estimates of r using these approaches vary from 0.254 for viability, for a large number of American progeny47, to around 0.6 for mate choice, in laboratory mice48 and humans in general49. Hill et al.5 estimated r for particular HLA-A and DRB1 alleles as 0.029 and 0.038, respectively (based on the relative risk of genotypes contracting malaria and the mortality rate from malaria), and an estimate of r for hepatitis B for HLA-DQ haplotypes as 0.0804. The variation among estimates of r derived from such diverse approaches can reflect differences in r over time, space, species or experimental context.

The future
The amount and sophistication of theory and technology being thrown at the MHCs of diverse organisms are staggering. The information flowing from ongoing large-scale sequencing efforts will prove a goldmine for molecular evolution and systematics. However, more extensive molecular characterizations of MHCs in nonmammalian vertebrates will be needed to infer the directions and timing of major events in MHC evolution. More realistic models incorporating mutation, recombination and potentially high levels of convergence in estimating population genetic parameters for MHC alleles, or further study of the conditions under which these problems are inconsequential, are needed. Finally, more extensive investigations of mate choice and infection by disease, and estimates of their selective effects in natural populations, if detectable, would help determine why MHC genes are so variable – for many researchers the ‘holy grail’ of MHC biology49.

Acknowledgements
We thank D. Gerasszy, H. Erlich, M. Flajnik, M. Kaufman, M. Kasahara, C. Ober and Y. Satta for helpful discussions and sharing of unpublished manuscripts. J. Kingsolver, M. Nei, T. Otta, P. Parham, Y. Satta, J. Trowsdale and three anonymous reviewers provided helpful suggestions on the manuscript. S.V.E. was supported, in part, by grants from the National Science Foundation.

References
6. Mizuki, N. et al. (1997) Nucleotide sequence analysis of the HLA class I region spanning the 227-Kb segment around the HLA-B and -C genes, Genomics 42, 45–66
9. Flajnik, M. F. (1993) A novel type of class I gene organization in vertebrates: a large family of non-MHC-linked class I genes is expressed at the RHA level in the amphibian Xenopus, EMBO J. 12, 4380–4386
10. Flajnik, M. F. (1997) Mapping of MHC class I and class II regions to different linkage groups in the zebrafish, Danio rerio, Immunogenetics 46, 125–134
21. Otta, T. On the pattern of polymorphism at major histocompatibility complex class I loci, in The Human HLA System, in press
The future as the key to the past for palaeobotany?

D.J. Beerling

T
e two factors have provided a strong impetus for international experimen
tal and modelling programmes on plant responses to global environmental change (GEC): the threat of future increases in the concentration of atmospheric CO2 and associated climatic change; and the central role of terrestrial plants and vegetation in the global carbon cycle. Revised predictions by the Intergovernmental Panel on Climate Change indicate an atmospheric CO2 concentration by the year 2100 of between 450 ppm and 800 ppm. These upper and lower limits are based on assumptions of economic growth, population growth and fossil fuel supply that are either low or moderate, respectively. The upper limit is not new to terrestrial vegetation and ecosystems, which have evolved and survived under such CO2 concentrations (or higher) for at least the past 250 million years, with the exception of the short-lived fluctuations in the Quaternary period.

Continued increase in the concentration of atmospheric CO2 and its possible effects on global climate has generated intense research interest on the likely responses of terrestrial plants and vegetation. Results from this new research provide quantitative information on plant function and growth in an environment with a high CO2 concentration, but are also relevant to understanding plant function during episodes of high CO2 concentration in the geological past (e.g. the greenhouse world of the mid-Cretaceous and mid-Eocene); and they are beginning to provide new insights into the effects of CO2 on the relationship between plants and climate.

Past climates from plant fossils

Three aspects of the relationship between plants and their environment are used by palaeobotanists to give a signal of terrestrial climate from the fossils: leaf phosphonoxy, the tolerance of plants to low temperature and the relationship between tree rings, CO2 and climate; it therefore signals the need for caution in interpreting palaeoclimate from fossils.