

THE HUMAN MAJOR HISTOCOMPATIBILITY COMPLEX: Lessons from the DNA Sequence

Stephan Beck¹ and John Trowsdale²

¹*The Sanger Centre, Wellcome Trust Genome Campus, University of Cambridge, Cambridge CB10 1SA United Kingdom; e-mail: beck@sanger.ac.uk*

²*Immunology Division, Pathology Department, Cambridge CB2 1QP, United Kingdom; e-mail: jt233@mole.bio.cam.ac.uk*

Key Words MHC, HLA, DNA sequence, polymorphism, autoimmunity

■ **Abstract** The entire 3.6-MbpDNA sequence of a human major histocompatibility complex, derived from a composite of DNA clones from different haplotypes, was completed in 1999, primarily through the work of four main groups. At that time, it was the longest contiguous human DNA sequence to have been determined. The sequence is of extremely high quality and accuracy. In this review, we discuss how the DNA sequence has facilitated our understanding of the biology and genetics of the major histocompatibility complex. We suggest some ways in which the sequence may be exploited in the future to explore the relationship between the extraordinary polymorphism of the region and its association with both autoimmune and infectious diseases.

INTRODUCTION

The major histocompatibility complex (MHC) at chromosome 6p21.31 was discovered over 50 years ago (38). This region of the chromosome was known to specify histocompatibility genes, but their nature has been resolved only in the last 20 years, with the advent of DNA cloning and determination of the structures of several class I and class II molecules. The drive to complete the MHC sequence was stimulated by the complex biology and genetics of the mouse MHC (H-2) and human leukocyte antigen (HLA) regions (42, 47). The MHC contains 224 identified loci (44). A map of the extended MHC region is shown in Figure 1, encompassing an additional 40 loci. It is the most gene-dense region of the human genome that has been discovered so far. It encodes the most polymorphic human proteins known to date, the class I and class II molecules.

Historically, the MHC has been divided into three regions: class II (centromeric), class III, and class I (telomeric). Recent and ongoing analyses of the immediate flanking regions reveal that the classical class I and class II regions extend much

EXTENDED MHC

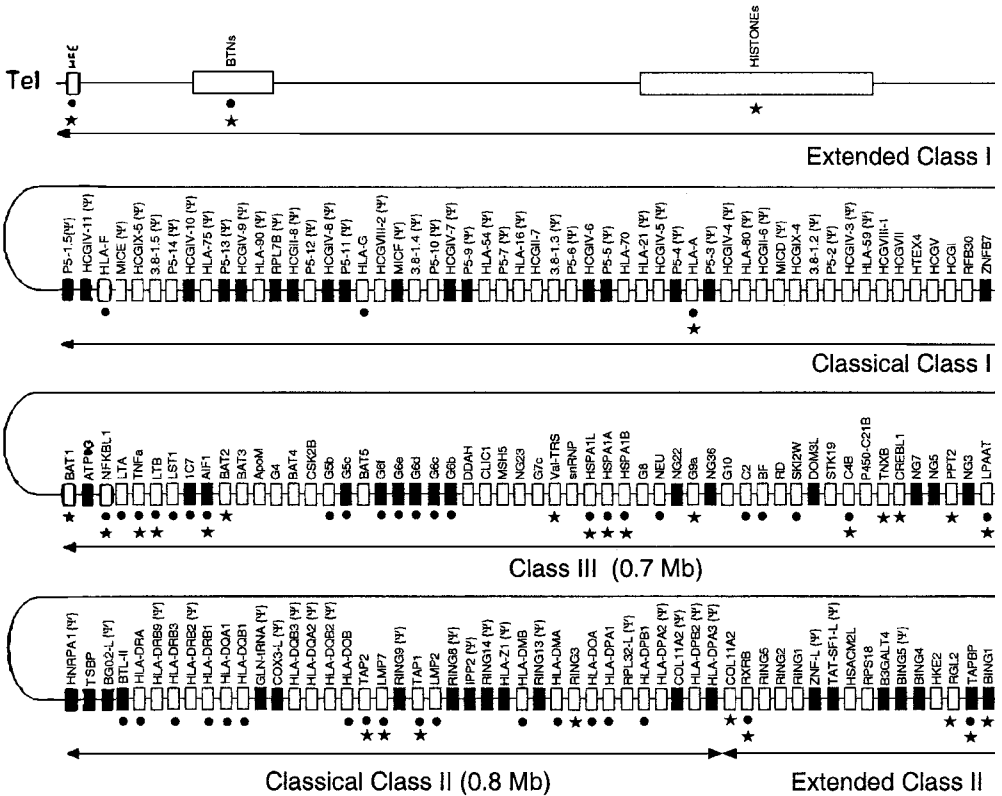


Figure 1 Gene map of the extended major histocompatibility complex. Shown in order but not to scale are 264 loci (genes/pseudogenes) from telomere to centromere (modified from 44; A Ehlers, S Beck, SA Forbes, J Trowsdale, A Voltz, submitted for publication; R Younger et al, manuscript in preparation). Genes ('New Loci') that were discovered or located to the extended MHC based on the genomic sequence are highlighted by filled boxes and pseudogenes are indicated by Ψ . Genes presenting "Immune Loci" are marked by filled circles under the corresponding loci and fall into at least one of the following categories: homology to the immunoglobulin domain or other immune superfamilies, expression that is specific to immune tissues, involvement in antigen processing and presentation (histocompatibility) or inflammation, implication in regulation of expression of immune loci, and induction by immune mediators such as interferon. Genes marked by asterisks represent loci for which paralogs have already been identified on chromosome 1, 9, or 19. *HLA-A* and *hs6M1-20* are marked as representatives for all paralogous class I and olfactory receptor loci, respectively.

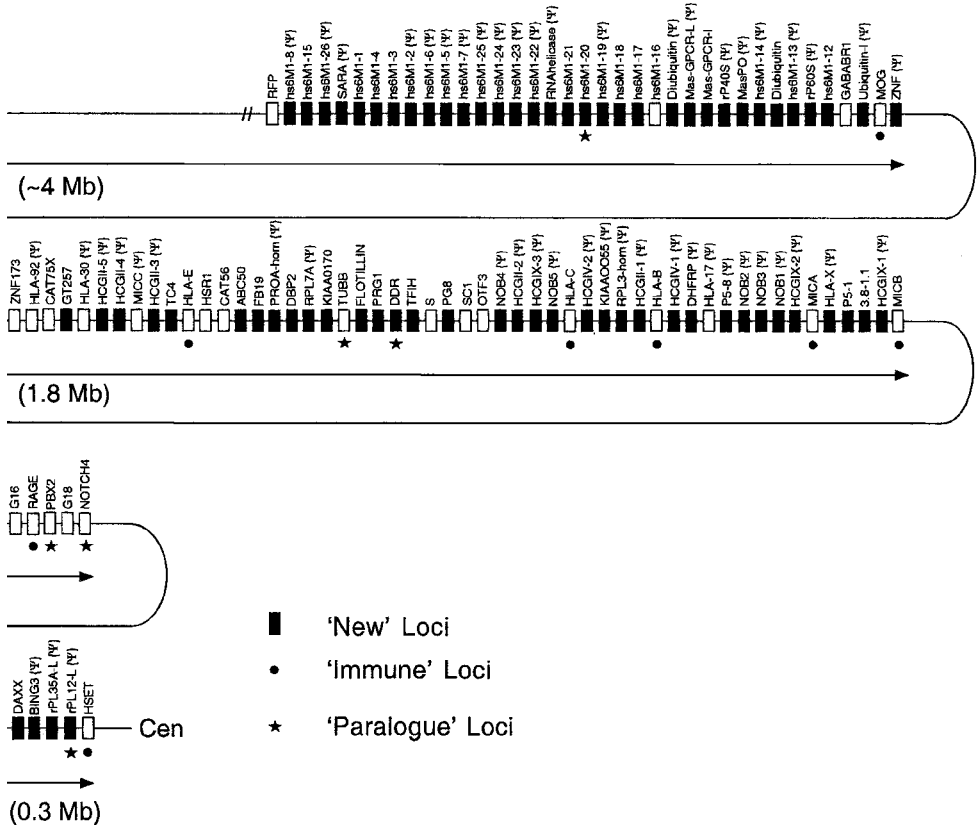


Figure 1 (Continued)

further than previously thought. These sections have been termed extended class I and class II regions (56).

THE DNA SEQUENCE AND GENE CONTENT

The human MHC sequence was assembled from a tile path of overlapping genomic clones (phage lambda, cosmid, PACs, BACs, and YACs) that had been carefully mapped before random sequencing by Sanger dideoxy technology. Any remaining gaps were closed by polymerase chain reaction (PCR). The 3.6-Mbp reference sequence covers ~0.12% of the human genome. Some of the MHC has been sequenced in different haplotypes, giving the first insights into the extraordinary polymorphism.

Many of the 224 identified MHC gene loci, of which 128 are predicted to be expressed, are of unknown function. Loci in the extended class I region are not included in this count because their exact number still needs to be determined. Of the 224 MHC loci, 93 (42%) were discovered or located at the MHC solely as a result of genomic sequencing. Including pseudogenes, the average gene density over 3.6 Mb is one gene per 16 kb. The class I and class II regions contain many pseudogenes. Up to half of the genes in the class I region are nonfunctional. Both regions appear to have duplicated multiple times, generating novel gene family members, which have then diverged (9, 53). The high numbers of pseudogenes may not be totally redundant because they could in theory play a role in generating new alleles by gene conversion. The gene density differs markedly in the three regions. The class III region contains an expressed gene for every ≤ 15 kbp and is extremely gene dense. In some cases (e.g. *TNXB* and *P450-C21B*), mRNA transcripts overlap (5). This region is also unique in that, except in certain haplotypes in which the *C4* regions have been duplicated, there are no pseudogenes for the entire region spanning 800 kbp (44).

Apart from the immune system genes (described below), the MHC contingent includes genes involved in a variety of processes. These include a large set of olfactory-receptor genes (73; A Ehlers, S Beck, SA Forbes, J Trowsdale, A Voltz, et al, submitted for publication) in the extended class I region, as well as some members of the ubiquitous zinc-finger, RING-finger, and transcription factor gene families (19). In the class I region, there is a set of loci potentially involved with DNA repair or cell growth, including *TFIIH* (transcription factor), *DDR* (receptor tyrosine kinase), *PRG1* (expressed in pancreatic carcinoma), *DBP2* (RNA helicase), and *TC4* (Ras-related) (53). There is also a large representation of genes involved in other cellular control processes, including *NOTCH4*, *RXRb*, *SC1*, *FB19*, and *HSR1*.

CLUSTERING OF IMMUNE SYSTEM GENES

Considering expressed loci only, 40% of the total contingent of loci in the MHC are immune related (Figure 1). This figure includes at least 10 novel genes that were identified from the genomic sequence. The class III region contains several such novel genes, which are members of the immunoglobulin (Ig) or Ly6 super families (*C5b*, *C5c*, *G6f*, *G6b*, *G6c*, *G6d*, *G6e*, and *IC7*). In the extended class I region is a set of butyrophilin-related loci (61), another member of which is located between the class II and class III regions (54b). The clustering of immune-related genes in the MHC region may not be coincidental (4, 27, 67). All of the genes in the class II region, with one exception (*RING3*, which is still of unknown function), have immune functions. This includes class II A and B genes, *LMPs*, *TAPs*, and *TAPBP* (22, 23), which are in the extended class II region. Over seven genes involved in inflammation, including three members of the tumour necrosis factor (TNF) superfamily, within the class III region have been referred to as the class IV region

(19). This clustering of immunity genes may relate to coevolution of functions or coexpression of genes with related functions.

POLYMORPHISM

The extreme polymorphism that is the hallmark of the MHC is most likely driven by resistance to infectious pathogens, although the identification of these agents has been difficult. The polymorphism is not homogenous throughout the 3.6-Mbp region. In the noncoding sequences, variation appears to increase, flanking the most polymorphic gene loci (26). "Hitchhiking" (4, 54a), along with natural selection of expressed, polymorphic class I and class II genes, has been suggested to explain this.

Although direct protein sequence of MHC molecules were available first, it was the easy availability of DNA sequence information, particularly that of cDNAs, that proved vital in understanding the variation of MHC class I and class II molecules. These sequences also had an impact on determination of class I and class II structures and structure-function relationships. Variation levels of $\leq 17\%$ have been reported at some of the loci, the most variable being *HLA-DP*, *DQ*, *B*, and *C* (44). In fact, the highest level of polymorphism in the expressed genome so far is in the MHC loci. The key finding made from comparison of sequences was that variation between alleles is predominantly localized within exons 2 and 3 for class I molecules and within exon 2 for *HLA*, *DRB*, *DQA*, *DQB*, *DPA*, and *DPB*. There is a predominance of nonsynonymous substitutions in these regions. Comparison of structures revealed how the amino acid differences occur in regions influencing the peptide-binding site. This pattern of variation in HLA molecules is different from that in most other protein-coding genes, in which allelic variation tends to occur more in introns than in exons. In classical HLA class I genes, there is relatively more variation in the exons, at the synonymous (silent or noncoding) positions, than in adjacent introns. This suggests that, although diversity is selected in some exons, introns may be subjected to opposing homogenizing forces.

In addition to the variation in specific exons that encode the peptide-binding domain, *DRB* genes are variable in number and position in different haplotypes. Comparisons of intron sequences (51) and repeat elements (58) can reveal the putative evolutionary history of the different gene arrangements. In the hypothetical scheme, there were two ancestral arrangements that diverged from an initial haplotype. One arrangement, containing the *DRB4* gene, may have arisen early because related haplotypes have been identified in other primates. The other lineage, containing the *DRB3* locus, has a characteristic *ERV9* long terminal repeat (LTR) insertion in intron 5, which is also found in *DRB1*, suggesting that the *DRB3* gene was duplicated from *DRB1*. All of the haplotypes that were studied presumably arose by duplication and recombination.

Duplication must have played a prominent role in the evolution of the MHCs, and it is therefore not surprising that the region is littered with class I and class II

pseudogenes, which may be the remnants of previous configurations that were left to decay. Sequencing identified a fragment of a class I pseudogene (*HLA-Z1*) in the class II region, the origin of which is difficult to determine (2).

The generation of the unique pattern of polymorphism in MHC molecules is the subject of an ongoing debate, and there is a role for both point mutations and conventional recombination, as well as gene conversion mechanisms (41).

EVOLUTIONARY ORIGIN OF THE SEQUENCE

Sequence comparisons revealed that several MHC genes (*TUBB*, *TNXB*, *PBX2*, *NOTCH4*, *RXR*, and *RPS18*) are syntenic in invertebrate genomes such as those of *Drosophila melanogaster* and *Caenorhabditis elegans*, indicating that the origin of the locus that is now known as MHC predates the emergence of the adaptive immune system (66; Figure 1). *RING3*, a class II gene of yet unknown function, is similarly conserved all the way down to yeasts, where it is found right next to *SKI2*, a gene in the class III region in humans (65). In yeasts, the *SKI2* gene is involved in viral cytopathology, in some respects a functional equivalent of class I antigen presentation. These data suggest that the class III region may be the oldest and that the three regions of the MHC had separate origins. Analysis of other species is not consistent with this view. The chicken and fish maps are informative in this regard. The MHC may have evolved within a group of genes cutting across the traditional class boundaries. In teleosts, class I and class II genes are not linked, but classical class I loci are closely associated with the immune proteasome components *LMP2* and *LMP7* and to *TAP* transporter loci (3, 59). Several other genes at the centromeric end of the human MHC, such as *RXR* and *RING3*, are also clustered in bony fish (18). In some sharks, there is evidence for linkage between class I and class II regions, suggesting that the association was lost at a later stage in the teleosts.

Dot matrix analysis of the sequence with a novel similarity-matching algorithm (J Mullikin, personal communication) shows how the class I and class II regions have evolved by multiple duplications (Figure 2). The regions identified in this way (e.g. *HLA-DP* to *DR*, *MICB* to *HLA-C*, and *HLA-E* to *HLA-F*) clearly carry the hallmarks of fairly recent duplications from shared ancestral sequences. Interspersed within these regions are several nonsimilar islands (e.g. *BTL-II*, *LMP/TAP*, and *RING3*), which therefore did not originate by duplication but were recruited from outside the MHC. Although the class III region is now considered to be the oldest region of the MHC, it appears not to have been involved in these massive duplication events that gave rise to the class I and class II regions, except for the *C4* locus. It is interesting that, so far, *C4* is the only class III locus found in the chicken MHC alongside class I and class II loci, indicating that this arrangement may reflect the primordial MHC organization (36).

The class I region appears to have been subjected to the most duplication events, in both humans and mice, such that various haplotypes in both species can differ

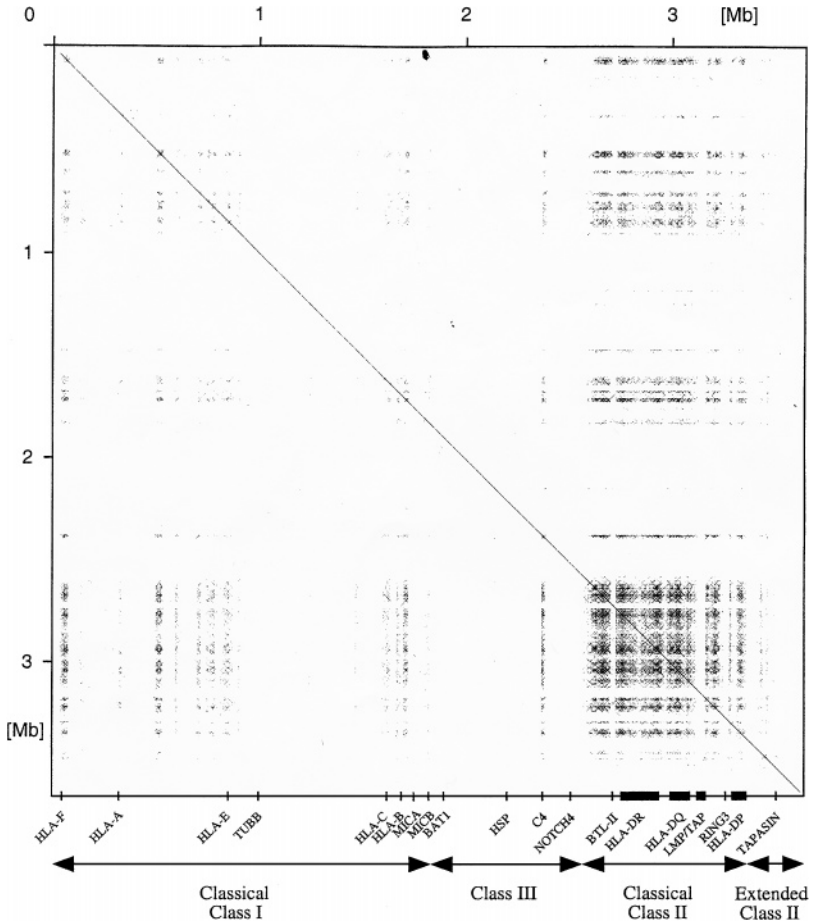


Figure 2 Dot matrix plot of the extended major histocompatibility complex (excluding the extended class I region). The 3.6-Mb-long sequence is compared with itself by using a novel similarity-matching algorithm (J Mullikin, personal communication). The signal footprint clearly identifies regions of similarity or shared origin (e.g. *HLA-DP* to *DR*, *MICB* to *HLA-C*, and *HLA-E* to *HLA-F*) interspersed with islands of dissimilarity (e.g. *BTL-II*, *LMP/TAP*, and *RING3*).

markedly in the number of loci (48). Some species of mice are estimated to contain hundreds of class I sequences (10). The expansions and contractions of sequences appear to have taken place against a background of anchor or framework loci that are invariant in number, based on the so-called framework hypothesis (1). One mechanism that has been proposed to explain the generation of the reiterated duplications is that of capture by retroelements such as retroviruses (9). Detailed dot-matrix analysis of the 1.7-Mbp class I region showed how all of the *MIC* genes

(*MIC-A, B, C, D, and E*) share upstream homology that extends over 15 kbp (53). These homologous sections contain a unique mix of genes that are all members of a multigene family, respectively called *HCGIX*, *3.8-1*, *P5*, *HCGIV*, *class I*, and *HCGII* (48). These elements were usually found in a similar orientation and order. Particularly striking is the 300-kbp telomeric portion of the class I region, linking *HLA-J* to *HLA-F* and encompassing *MICD*, *MICE*, and *MICF*. This region comprises over >30 homologous segments that range from 8 to 20 kbp. From these data, Shiina et al developed a model based on seven rounds of successive segmental duplications to shape the class I region. Of the 18 class I and 6 *MIC* genes, 15 class I and 5 *MIC* genes are associated with the shared elements. The proposed sequence of events began with the emergence of *HLA-F* and *MICE* genes, as the ancestral class I units (53, 54). Thereafter a series of over seven duplications gave rise to the current class I order. It should be stressed that this order and arrangement are not necessarily identical in all haplotypes. The sequence of events proposed by Shiina et al is supported by the use of *Alu* and other repeat sequences as a molecular clock (see below).

Analysis of repeats, such as *Alu* repeats, can help to date a chromosomal region. This type of analysis has been performed for a section of the class II region. *Alu* sequences can be divided into two main classes, *J-Alu* (old) and *S-Alu* (new). The ratio of *S/J Alus* in some representative regions of the human genome is generally around three. In the class II region, this ratio is more like 1.4, that is, two fold lower. This has been interpreted as an indication that the class II region became sensitive to further repeat expansion after the emergence of *S-Alus* >30 mya (2).

PARALOGY

Regions that are paralogous to the MHC on chromosomes 1, 9, and 19 have been proposed to result from ancient chromosomal duplications, although this has been disputed based on phylogenetic analysis (34, 71). The first clues to MHC paralogy came from studies of class III genes (*NOTCH4* and *PBX2*) and class II proteasome genes (17, 37, 57). Their paralog, the *PSMB7* gene encoding a constitutive subunit that is replaced by another interferon-inducible subunit, was mapped to a region on chromosome 9q33-34, which contains at least 10 genes with duplicated copies in the MHC (32). These include an ABC transporter, similar to *TAP*. Other chromosomal regions, namely 1q21-q25 and 19p13.1-p13.3, were found to harbor similar sets of genes (35). The recent increase in MHC sequence information allowed a detailed comparison of sets of genes in the four regions, although, at this stage, the MHC-encoded genes are by far the best documented. Some related genes are found at all four locations, such as the *NOTCH* family (*NOTCH1*, 2, 3, and 4) and the *RING3* family (*RING3*, *HUNK1*, *BRDT*, and *ORFX*). The list of gene families found in two, three, or four of the locations is expanding and already exceeds 30 (33; Figure 1). In some cases in which a large family of genes exists, such as

histones or olfactory receptor loci, it is difficult to rule out chance linkage, but the evidence is nevertheless compelling.

The distributions of copies on particular chromosomes provides clues as to the sequence of duplications that gave rise to them. For example, some pairs of genes, such as those related to *RXR*, are more highly related to each other on chromosomes 1 and 9, arguing that these copies shared an immediate common ancestor. It could also be argued that all of the linked genes should be equidistant from their paralog on the other chromosomes if they are duplicated en bloc. This is not always observed, indicating problems with construction of trees to analyze sequence relationships or more complex rearrangements. It is easy to imagine sequence exchange taking place after duplication, for example, or tandem gene duplication before chromosomal segment duplication, followed by loss of different copies. The sequence comparisons are consistent with duplications taking place in early vertebrate evolution, in a common ancestor of jawed vertebrates after separation from the jawless fish, possibly as a result of genome-wide duplication (33).

It seems likely that the adaptive immune system came to prominence only in species that arose subsequent to the development of jaws. Because genome duplication is an attractive way to develop a novel network of interacting molecules, it has been speculated that tetraploidization of the early vertebrate genome might have been one of the factors that enabled the emergence of an adaptive immune system. The recent discovery of the first living tetraploid mammal, the red viscacha rat (16), helps to make this theory attractive. If indeed the cluster of MHC genes predates the emergence of the adaptive immune system, it could be argued that the linkage of class I with *TAPASIN*, *TAP*, and *LMP* genes is simply fortuitous. Where and when did class II genes emerge? It has been argued that they predated class I (14). This seems less likely given the lack of any class II-paralogous genes on other chromosomes. Another issue in the evolution of the MHC, which is still without a satisfactory explanation, is the origin of class I or class II molecules. At one time it was proposed that they may have developed from *HSP70*-like ancestral molecules, especially because some *HSP70* relatives reside in the MHC class III region (50). Sequence and structures of *HSP70* make this hypothesis unlikely (72; for a synthesis of the origin of Ig domains in relation to MHC structures, see 11).

ORTHOLOGY

The MHC sequence analysis has not been restricted to humans. The chicken sequence has already been published, and fish sequences are also well under way (36). The sequences of other species will no doubt follow, including those of primates, cattle, pigs, sheep, and cats, as well as, of course, mice and rats (68). Some of the main features of the differing gene arrangements in these species have already been established. These include, for example, the class I genes at the centromeric end in rodents, *H-2K* in mice and *RT1.A* in rats, and the inversion of the class I

and III regions in chickens. There are also some interesting class II genes outside the classical MHC region in cattle (39).

The main features of the rodent MHC region recapitulate those of the human map, although it is obvious that some regions show greater flexibility than others. These include the class I region, where the genes are not orthologous to those in humans, indicating that rederivation of the repertoire of loci took place subsequent to speciation between the ancestors of the two species some 80 mya (46). A model has been proposed that the H-2 region consists of stable sections or framework loci interspersed with regions of considerable plasticity (1). It is not difficult to see how this scheme could be applied to MHCs in general, and it need not be confined to H-2.

The chicken MHC is informative because its B-F/B-L region contains a compact group of genes that have been interpreted to be a minimal set, referred to as a "minimal essential MHC" (36). Thus, classical, polymorphic class I and class II genes are found alongside *TAP* and *TAPASIN*, as well as *C4*. It appears that many of the other genes associated with the mammalian MHCs are deleted or moved. It is interesting that *LMP* genes are absent. This may relate to the specificity of the peptide-binding site in chicken class I molecules, which, unusually, accepts peptides with negatively charged COOH termini. The *DO* gene pair is absent, but *DM* equivalents are found. Lectin-like genes were identified, and these may be related to natural killer receptor loci. Such loci are on chromosome 12, not linked to the MHC, in humans, but, because the ligands of several of the natural killer receptor molecules are class I, a possible early configuration includes a potential genetic linkage, at least in some species.

Two fish have been studied in detail at the genomic level, zebrafish and pufferfish (3, 18, 40). These model organisms have specific advantages as biological systems, namely speed of reproduction in zebrafish and the compactness of the genome in pufferfish (12).

REPLICATION TIMING AND ISOCHORE STRUCTURE

The sequence provides an opportunity to examine chromosomal features such as isochores, that is, long-range regions of homogenous G+C content. The low G+C isochore characterizing the classical class II region is a good example of an isochore with identifiable margins (56). Its boundaries correlate with switching of replication timing from "later" in the classical class II region to "earlier" replication at the centromeric boundary (30) and at the telomeric boundary (63). There may be a link between isochores and the replicon structure of the human genome.

HUMAN LEUKOCYTE ANTIGEN TYPING

The genomic sequence has been of some value in permitting alleles to be unequivocally assigned to different loci, especially at the variable *DRB* loci. cDNA

sequences have fulfilled a major role in determination of allelic variation for tissue typing, most of which is now done by DNA techniques (21). As more refined sequence data becomes available, typing techniques evolve alongside, but the value of typing in different transplant situations is not generally accepted. An accurate match is clearly essential to bone marrow transplantation, but it is often ignored for solid-organ transplants, at least in some centers, despite the evidence of a relationship between graft survival and level of matching. A discussion of these issues was recently published (43).

DISEASE ASSOCIATION AND MAPPING OF CANDIDATE DISEASE GENES

The MHC is associated with more diseases than any other region of the human genome (49). It is linked to most, if not all, autoimmune conditions. Other nonimmune-disease phenotypes have also been linked to the region, ranging from cancer to narcolepsy. The availability of both cDNA and genomic DNA sequence has been essential in identifying candidate MHC loci that predispose to these diseases, although in few cases has the precise locus been identified. As predicted, the highly polymorphic class I and class II loci are the major determinants of MHC-associated disease, but the strong linkage disequilibrium across the complex makes it difficult to rule out a contribution from other linked genes, in the class III region for example.

There are multiple different explanations for association of autoimmune conditions with antigen-presenting molecules. It is interesting first of all that there is no wild type for class I and class II. "Disease" alleles are common in the normal, unaffected population, consistent with the notion that autoimmune conditions are influenced by multiple contributory factors including multiple other genes and environmental effects. In no case can it be argued that a particular class I or class II allele is necessary or sufficient to cause disease. In the most dramatic models, such as *HLA-B27* and ankylosing spondylitis, although 95% of patients express B27, only 3% of Caucasians with the allele develop the condition. The association between narcolepsy and *HLA-DQB1*0602* is another example in which the sequence is present in nearly all patients and is a useful diagnostic indicator, but the frequency of the same allele in the normal, unaffected population approaches 25% (65a).

Hemochromatosis is one of the few diseases that have been picked up by MHC variation, but it was subsequently identified as being some distance away from the MHC, at a locus that is paradoxically related to classical class I genes (13). After years of effort in the vicinity of the *HLA-A* locus (*HLA-A3* was one of the early indicators of the disease), the *HLA-HFE* locus was eventually found, by brute-force cloning and sequencing, to be several megabase pairs away. In this disease, the strong linkage disequilibrium was crucial for mapping at an early stage but became a mixed blessing for identifying the locus. Even with current statistical

approaches, the precise identification of genes predisposing to autoimmune conditions within the MHC remains a complex process, even when we are armed with the complete sequence over the region. Typing of new microsatellites that are derived from the genomic sequence has narrowed down the candidate region for psoriasis vulgaris (an inflammatory skin disorder) to a critical segment of 111 kb, containing four previously known genes—*S*, *PG8*, *SCI*, and *OTF*—and several novel genes (45, 60).

Other complications arise in analysis of the relationship between MHC markers and infectious diseases. Although resistance to infection is believed to drive the extreme polymorphism, it has proved difficult to pinpoint relationships between such infections and specific alleles. The idea that it may be possible to uncover a simple cause and effect relationship between diseases of this type and alleles of class I and class II has been shown to be limited. In some cases, such as *HLA-B*5301*, a believable relationship with resistance to malaria is observed in some populations (24). On the whole, studies of other infectious diseases have been less informative. One explanation for this could be that measurement of mortality or morbidity is too crude to pick up subtle effects of the advantages afforded by different HLA alleles. These may be only contributory, affecting, for example, viral load or longevity after infection or other specific features such as responses to certain epitopes or antibody levels (7, 29), advantages that work in concert with other effects. In the study of HIV, HLA alleles appeared to support the concept of heterozygous advantage (7). This concept has been known in the field of population genetics for decades, but there are only a small number of convincing examples. With a chronic infection such as HIV, in which the virus sequence is unstable, it is reasonable to propose that heterozygotes have the advantage of having an increased chance of possessing an HLA allotype that can present an appropriate pathogen-derived peptide for T-cell recognition.

Candidate gene mapping will be facilitated by having large numbers of readily scorable polymorphic markers that span the HLA region, a goal that is now in sight. Novel high-throughput technologies can be used on both population and family material to gather the data sets. Phase-known patient MHCs can be analyzed for “haplotype decay” in relation to phenotype. In other words, the region associated with a disease can be refined by tracking the extent of conserved blocks of sequence. The question remains as to how many samples have to be included to identify a susceptibility gene in a region like the MHC, where there may be weak relationships between the presence of alleles and the disease. In multiple sclerosis or diabetes, for example, sib-pair analysis and other linkage approaches have had limited impact. Novel experimental approaches based on linkage disequilibrium may be necessary (62). Population-based association (case/control) studies are powerful, but transmission-disequilibrium testing (TDT) on families in which nonrandom transmission of alleles has occurred has the advantage of being resistant to population stratification effects (31). Whatever approach is adopted, closure requires a functional correlate, and once an association has been found with a particular

region, intelligent candidate screening is feasible if the complete sequence is available.

RECOMBINATION AND LINKAGE DISEQUILIBRIUM

Recombination is an area of research that is now dependent on DNA sequences. Before sequences were available, the MHC became a useful model for recombination because of the extreme variation of the constituent loci, which were scored by tissue typing. Recombination has been studied systematically in the mouse H-2 system, using inbred strains (55). Studies in both humans and mice have suggested that different haplotypes affect the frequency and location of crossovers. This has been difficult to approach in out-bred human populations, but the explosion of single nucleotide polymorphism (SNP) and variable microsatellite data as sequences accumulate, in combination with sperm typing, may help to alleviate this problem (25).

It is generally assumed but not proven that variation in the MHC is in response to pathogens. The two models that have been put forward to explain maintenance of the variation are (a) heterozygote advantage, in which there may be a more diverse immune response to a pathogen from sets of alleles on the two chromosomes, and (b) frequency-dependent selection, in which rare alleles are at an advantage in a population in which a pathogen has evolved to evade elimination. In reality, both mechanisms may overlap. Either one could lead to the preservation of haplotypes, with combinations of alleles at two or more loci that could work well in concert to eliminate pathogens. One could imagine that the alleles on a common haplotype such as *A1B8DR3* are in linkage disequilibrium (at a greater frequency together than predicted by the individual frequencies of the single alleles) because of such selective mechanisms. Another explanation is that recombination differences in different haplotypes, especially absence of recombination, have helped to promote certain combinations of alleles. Another possibility is historical; human populations have grown from a restricted number of families in the last few thousand years, and sufficient time has not elapsed for combinatorial equilibrium.

Systematic studies of MHC recombination provide evidence in support of its nonrandom nature. Two regions are well known to be devoid of crossovers, namely the regions between *HLA-B* and *-C* and between *DQA1* and *DRB1*. The telomeric region of *HLA-A* linkage disequilibrium is also marked, for example between *HLA-A* and *HLA-HFE*, a distance of ~ 4 Mbp. Again, the explanation for this has not been established, but variations in genomic arrangement that disturb homology upon pairing could play a role.

Recombination hotspots have also been identified, for example in the class II region between *HLA-DNA* and *RING3*, *DQB3-DQB1*, and *TAP1* and *TAP2* (6, 8). Crossovers have been located in these regions to within a few hundred base pairs,

and comparison with equivalent rodent hotspots has identified recombinogenic chi-like sequences in some of these locations, the significance of which remains to be determined.

As with the disease association, the future exploration of this topic depends on the availability of the closely spaced, highly variable markers that DNA sequences will provide, in concert with novel techniques for studying recombination, such as sperm typing (for a comprehensive review of recombination, see 6).

SYNTHESIS

The availability of multimegabase genomic sequences such as the MHC reference sequence allows the study of chromosomal features and function in unprecedented detail. As described above, the initial analysis of the MHC reference sequence has already revealed several intriguing new features, in addition to the gene content. Variation analysis of loci (>50 kb of noncoding sequences) in the classical class I, classical class II, extended class II, and class III regions has shown dramatic local differences (5- to 50-fold) in variation levels (20, 26, 56; L Rowen, personal communication). G+C content analysis has revealed the presence of low and high G+C isochores, regions of long-range, uniform base composition (15, 44, 56). Replication timing studies uncovered distinct local and regional differences (particularly between the class II and class III regions), which appeared to correlate with predicted isochore boundaries (30, 63).

Figure 3 shows a simplified summary of these findings. The extrapolated correlation between replication timing and isochores (represented by the G+C content) is quite obvious and striking. Less obvious is a possible correlation between variation levels and G+C content and possibly replication timing as well. In this figure, the levels of variation that are characteristic of MHC class I and class II genes occur in regions of relatively low G+C content. More data are needed to establish this point, especially because early work showed clustering of variation at areas that are rich in CpG dinucleotides (62). The high CpG dinucleotide levels correlate with gene conversion events (25). Replication timing data are available only for the centromeric part of the MHC. If this correlation is real, it raises two immediate questions: (a) Are there regions in the genome that, independent of functional selection, have higher than average levels of variation; and (b) what other mechanisms exist to generate such nonrandom variation, apart from functional selection? The answer to the first part of question *a* is, clearly, yes, in view of the highly variable noncoding regions. However, whether these are independent of functional selection and what then might be causing them are difficult to answer. In some cases (Figure 3, locus *c*), the region of high variation extends far (~20 kb) beyond the known transcriptional unit of the nearest gene, making functional selection in this region less likely. Hitchhiking has been suggested as a possible explanation in this case (26). In an indirect way, the lack or repression of recombination (which

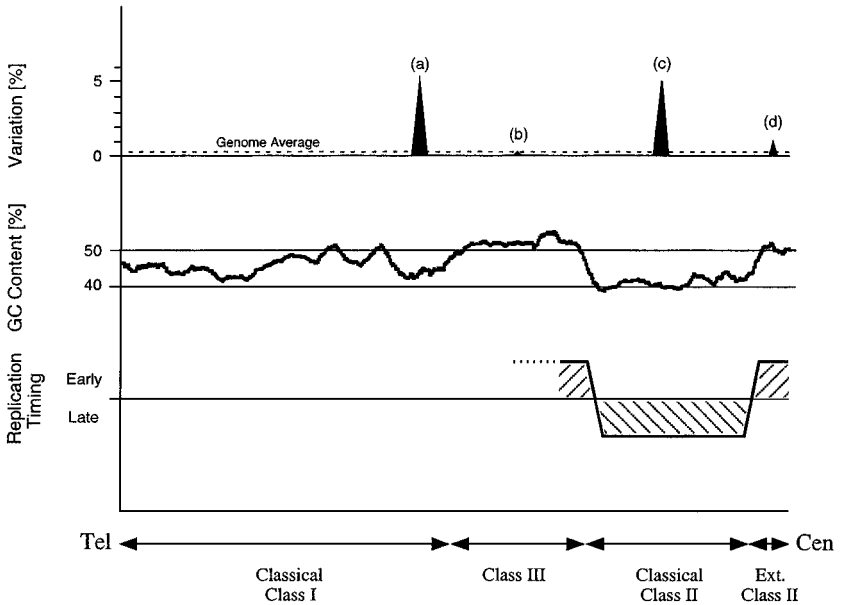


Figure 3 Simplified summary of major histocompatibility complex feature data, illustrating the observed correlation between G+C content, replication timing, and variation levels. For further details see Synthesis section in text.

has been observed in regions of high variation) may even contribute to the high variation in such regions, by not subjecting them to the effects of homogenization, which is a result of recombination. Further evidence that variation rates can be independent of natural selection comes from a study of closely linked loci that were found to exhibit significantly more similar variation rates than unlinked loci, suggesting that variation is influenced by genomic location (52). The involvement of replication timing as a possible influence on this phenomenon was first proposed >10 years ago, when it was shown that the mutation rate varies between different regions in mammalian genomes (70). Replication of the entire human genome takes several hours, and it would be surprising if the variation rate did not change during this process as a result of changes in nucleotide/enzyme concentrations that are involved in replication. In fact, it has been proposed that isochores arose as a direct result of such replication-dependent variant conditions (70). Taken together, these issues suggest that some chromosomal locations are more likely to randomly accumulate higher rates of variation, and it is tempting to speculate that such regions (which could be low in G+C content and recombination rate and/or could be late replicating) have been exploited by genes under natural selection pressure to generate high variation.

SEQUENCES OF COMPLETE HAPLOTYPES

The MHC reference sequence is now the starting point for numerous studies, and various databases of MHC-associated data are available, including the following: (a) a genomic database, <http://www.sanger.ac.uk/HGP/Chr6/>; (b) allele databases, http://www.swmed.edu/home_pages/ASHI/ashi.htm and <http://www.anthonynolan.com/HIG/index.html>; and (c) peptide ligand databases, <http://wehih.wehi.edu.au/mhcpep> and http://bimas.dcrn.nih.gov/molbio/hla_bind/.

A systematic nomenclature has been devised to accommodate the growing number of new alleles, with the unique locus designation followed by an asterisk (*) and a four-digit allele identifier. For example, *HLA-DRB1**0201 refers to the *DRB1* locus, allele 2. The 01 refers to a subdivision or minor variation of the second allele. Further optional numbering can be added for synonymous nucleotide changes and noncoding allelic variation. This nomenclature system is in force for all MHC class I and II loci, as well as other genes such as *TAP* and *LMP*.

The biological importance of the MHC justifies the resequencing and epigenetic analysis of several common haplotypes, which may differ in sequence and gene content. Efforts towards this goal are already in progress. Such studies will help to facilitate the precise identification of disease loci, and haplotypic differences in gene organization may help to better understand features such as recombination and polymorphism. Determination of the DNA sequence is coincident with international efforts to harvest and identify the functional SNPs or quantitative trait nucleotides (QTNs) that will be used for scoring quantitative trait loci (QTL). The correlation of the presence of different alleles with disease phenotypes provides a way of establishing causality between a pathway in which the gene containing the SNP acts and the disease. As discussed above, the MHC is one of the most important regions of the human genome, regarding autoimmune conditions and infection. The extreme polymorphism extends from class I and class II loci to intergenic regions that differ by insertion/deletion variation (indels). Despite the intense focus on the region, very few QTNs have been identified. The complete sequences of several common haplotypes, such as *DR1-10*, would be an efficient way of picking up the total variation content distinguishing these haplotypes. Knowledge of the variation would lead to the subset of QTNs that determines the disease. Thus the sequence of additional haplotypes will provide the basis for association studies of all MHC-linked diseases. The completeness of the catalog of QTNs will, for the first time, facilitate a survey of variation in all parts of genes, no matter what the nature of a regulatory sequence or the size of an intron. Incomplete surveys of gene polymorphism, which are difficult to interpret when evaluating the association of a gene cluster with disease, will be relegated to the past. The contiguous map of SNPs will permit the determination of which parts of the haplotypes—or ancestral segments thereof—are identical by descent. Linkage disequilibrium of the SNPs with each other and their ancestral haplotype associations will be known in intimate detail. High-throughput automation of SNP determination based on DNA sequence is an emerging technology that is rapidly developing in sophistication.

Not all diseases are caused by changes in the primary DNA sequence. Complex diseases, including cancer, are likely to have other contributing factors, including epigenetic factors such as changes in methylation patterns. For the same reasons as already outlined above, the MHC presents an ideal model as a pilot study for the recently proposed human epigenome project (69).

The full MHC gene content deduced from the genomic sequence will also allow the use of DNA chip technology to explore tissue-specific and disease-specific expression profiles. Complementary to the physical order of genes along chromosomes provided by the genomic sequence, this approach addresses the order and logic of genetic programs and biological pathways. It has already proved successful in studying several such programs in yeasts and humans (64). Similar studies are essential to elucidate the full interaction and genetic program of the particularly tight linkage group of human MHC genes and gene products.

ACKNOWLEDGMENTS

We thank our many colleagues from the MHC community for comments and unpublished results and Jim Mullikin for the dot matrix plot (Figure 2). The authors are supported by the Wellcome Foundation.

Visit the Annual Reviews home page at www.AnualReviews.org

LITERATURE CITED

1. Amadou C. 1999. Evolution of the Mhc class I region: the framework hypothesis. *Immunogenetics* 49:362–67
2. Beck S, Abdulla S, Alderton RP, Glynne RJ, Gut IG, et al. 1996. Evolutionary dynamics of non-coding sequences of the human MHC. *J. Mol. Biol.* 255:1–13
3. Bingulac-Popovic J, Figueroa F, Sato A, Talbot WS, Johnson SL, et al. 1997. Mapping of the Mhc class I and class II regions to different linkage groups in the zebrafish, *Danio rerio*. *Immunogenetics* 46:129–134
4. Bodmer WF. 1972. Evolutionary significance of the HL-A system. *Nature* 237:139–45
5. Bristow J, Tee MK, Gitelmann SE, Mellon SH, Miller WL. 1993. Tenascin-X: a novel extracellular matrix protein encoded by the human XB gene overlapping P450c21B. *J. Cell Biol.* 122:265–78
6. Carrington M. 1999. Recombination within the human MHC. *Immunol. Rev.* 167:245–56
7. Carrington M, Nelson GW, Martin MP, Kissner T, Viahov D, et al. 1999. HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage. *Science* 283:1748–52
8. Cullen M, Noble J, Erlich H, Thorpe K, Beck S, et al. 1997. Characterization of recombination in the HLA class II region. *Am. J. Hum. Genet.* 60:397–407
9. Dawkins R, Leelayuwat C, Gaudieri S, Tay G, Hui J, et al. 1999. Genomics of the major histocompatibility complex: haplotypes, duplication, retroviruses and disease. *Immunol. Rev.* 167:275–304
10. Delarbre C, Jaulin C, Kourilsky P, Gachelin G. 1992. Evolution of the major histocompatibility complex: a hundred-fold amplification of the MHC class I genes in the African pigmy mouse *Nannomys setulosus*. *Immunogenetics* 37:29–38

11. Du Pasquier L. 2000. The phylogenetic origin of antigen-specific receptors. In *Origin and Evolution of Vertebrate Immune System*, ed. L Du Pasquier, GW Litman, pp. 159–90. Berlin: Springer.
12. Elgar G, Sandford R, Aparicio S, Macrae A, Venkatesh B, et al. 1996. Small is beautiful: comparative genomics with the pufferfish (*Fugu rubripes*). *Trends Genet.* 12:145–50
13. Feder JN, Gnirke A, Thomas W, Tsuchihashi Z, Ruddy DA, et al. 1996. A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nat. Genet.* 13:399–403
14. Flajnik MF. 1991. Which came first, MHC class I or class II? *Immunogenetics* 33:295–300
15. Fukagawa T, Sugaya K, Matsumoto KI, Okukura K, Ando A, et al. 1995. A boundary of long-range G+C% mosaic domains in the human MHC locus: pseudoautosomal boundary-like sequence exists near the boundary. *Genomics* 25:184–91
16. Gallardo MH, Bickham JW, Honeycutt RL, Ojeda RL, Kohler N. 1999. Discovery of tetraploidy in a mammal. *Nature* 401:341–43
17. Glynne R, Powis SH, Beck S, Kelly A, Kerr LA, et al. 1991. A proteasome-related gene between the two ABC transporter loci in the class II region of the human MHC. *Nature* 353:357–60
18. Gongora R, Zaleska-Rutczynska Z, Takami K, Figueroa F, Klein J. 1998. Linkage of RXRB-like genes to class I and not to class II Mhc genes in the zebrafish. *Immunogenetics* 48:141–43
19. Gruen JR, Weissman SM. 1997. Evolving views of the MHC. *Blood* 90:4252–65
20. Guillaudeux T, Janer M, Wong GK, Spies T, Geraughty DE. 1998. The complete genomic sequence of 424,015 bp at the centromeric end of the HLA class I region: gene content and polymorphism. *Proc. Natl. Acad. Sci. USA* 95:9494–99
21. Hansen J. 2000. The detection and application of DNA polymorphisms. *Rev. Immunogenetics*. In press
22. Herberg JA, Beck S, Trowsdale J. 1998. Tapasin, Daxx, RgL2, KE2 and four new genes (BING1,3–5) form a dense cluster at the centromeric end of the MHC. *J. Mol. Biol.* 277:839–57
23. Herberg JA, Sgouros J, Jones T, Copeman J, Humphray SJ, et al. 1998. Genomic analysis of the Tapasin gene, located close to the TAP loci in the MHC. *Eur. J. Immunol.* 28:459–67
24. Hill AVS, Allsop CEM, Kwiatkowski D, Anstey NM, Twumasi P, et al. 1991. Common West African HLA antigens are associated with protection from severe malaria. *Nature* 352:595–600
25. Hogstrand K, Bohme J. 1999. Gene conversion can create new MHC alleles. *Immunol. Rev.* 167:305–17
26. Horton R, Niblett D, Milne S, Palmer S, Tubby B, et al. 1998. Large-scale sequence comparisons reveal unusually high levels of variation in the HLA-DQB1 locus in the class II region of the human MHC. *J. Mol. Biol.* 282:71–97
27. Howard JC. 1993. Restrictions on the use of antigenic peptides by the immune system. *Proc. Natl. Acad. Sci. USA* 90:3777–79
28. Jaulin C, Perrin A, Abastado JP, Dumas B, Papamatheakis J, et al. 1985. Polymorphism in mouse and human class II H-2 and HLA genes not the result of independent point mutations. *Immunogenetics* 22:453–70
29. Jeffery KJM, Usuku K, Hall SE, Matsumoto W, Taylor GP, et al. 1999. HLA alleles determine human T-lymphotropic virus-I (HTLV-I) proviral load and the risk of HTLV-I-associated myelopathy. *Proc. Natl. Acad. Sci. USA* 96:3848–53
30. Jonhonn P, Williamson J, Beck S, Sheer D. 2000. Analysis of replication timing in the human MHC. *Cytogenet. Cell Genet.* 88:188
31. Jorde LB. 1995. Linkage disequilibrium as

- a gene-mapping tool. *Am. J. Hum. Genet.* 56:11–14
32. Kasahara M. 1997. New insights into the genomic organisation and origin of the MHC: role of chromosomal (genome) duplication in the emergence of the adaptive immune system. *Hereditas* 127:59–65
 33. Kasahara M. 1999. The chromosomal duplication model of the major histocompatibility complex. *Immunol. Rev.* 167:17–29
 34. Kasahara M, Hayashi M, Tanaka K, Inoko H, Sugaya K, et al. 1996. Chromosomal localisation of the proteasome Z subunit gene reveals an ancient chromosomal duplication involving the major histocompatibility complex. *Proc. Natl. Acad. Sci. USA* 93:9096–101
 35. Katsanis N, Fitzgibbon J, Fisher EMC. 1996. Paralogy mapping: identification of a region in the human MHC triplicated onto human chromosomes 1 and 9 allows the prediction and isolation of novel PBX and NOTCH loci. *Genomics* 35:101–8
 36. Kaufman J, Milne S, Gobel TWF, Walker BA, Jacob JP, et al. 1999. The chicken B locus is a minimal essential MHC. *Nature* 401:923–25
 37. Kelly A, Powis SH, Glynn R, Radley E, Beck S, et al. 1991. Second proteasome-related gene in the human MHC class II region. *Nature* 353:667–68
 38. Klein J. 1986. *Natural History of the Major Histocompatibility Complex*. New York: Wiley. 775 pp.
 39. Lewin HA, Russell GC, Glass EJ. 1999. Comparative organization and function of the MHC of domesticated cattle. *Immunol. Rev.* 167:145–58
 40. Lim EH, Brenner S. 1995. Sequence analysis of Mhc class II β -like fragments in the pufferfish *Fugu rubripes*. *Immunogenetics* 42:432–33
 41. Little AM, Parham P. 1999. Polymorphism and evolution of HLA genes and molecules. *Rev. Immunogenet.* 1:105–23
 42. McCluskey J. 1999. Immunobiology of the MHC. *Rev. Immunogenet.* 1:1–123
 43. McCluskey J, Peh CA. 1999. The human leucocyte antigens and clinical medicine: an overview. *Rev. Immunogenet.* 1:3–20
 44. MHC Sequencing Consortium. 1999. Complete sequence and gene map of a human major histocompatibility complex (MHC). *Nature* 401:921–23
 45. Oka A, Tamiya G, Tomizawa M, Ota M, Katsuyama Y, et al. 1999. Association analysis using refined microsatellite markers localizes a susceptibility locus for psoriasis vulgaris within a 111Kb segment telomeric to the HLA-C gene. *Hum. Mol. Genet.* 12:2165–70
 46. Parham P. 1994. The rise and fall of great class I genes. *Semin. Immunol.* 6:373–82
 47. Parham P, ed. 1999. *Immunological Reviews. Genomic Organisation of the MHC: Structure, Origin and Function*. Copenhagen: Munksgaard. 379 pp.
 48. Pichon L, Carn G, Bouric P, Giffon T, Chauvel B, et al. 1996. Structural analysis of the HLA-A/HLA-F subregion: precise localization of two new multigene families closely associated with the HLA class I sequences. *Genomics* 32:236–44
 49. Price P, Witt C, Allcock R, Sayer D, Garlepp M, et al. 1999. The genetic basis for the association of the 8.1 ancestral haplotype (A1, B8, DR3) with multiple immunopathological diseases. *Immunol. Rev.* 167:257–74
 50. Rippmann F, Taylor WR, Rothbard JB, Green NM. 1991. A hypothetical model for the peptide binding domain of hsp70 based on the peptide binding domain of HLA. *EMBO J.* 10:1053–59
 51. Satta Y, Mayer W, Klein J. 1996. Evolutionary relationship of HLA-DRB genes inferred from intron sequences. *J. Mol. Evol.* 42:648–57
 52. Sharp PM, Matassi G. 1994. Codon usage and genome evolution. *Curr. Opin. Genet. Dev.* 4:851–60
 53. Shiina T, Tamiya G, Oka A, Takishima N, Yamagata T, et al. 1999. Molecular dynamics of MHC genesis unraveled by sequence

- analysis of the 1,796,938 bp HLA class I region. *Proc. Natl. Acad. Sci. USA*. 96:13282–87
54. Shiina T, Tamiya G, Oka A, Takishima N, Inoko H. 1999. Genome sequencing analysis of the 1.8 Mb entire human MHC class I region. *Immunol. Rev.* 167:193–99
 - 54a. Smith JM, Haigh J. 1974. The hitchhiking effect of a favorable gene. *Genet. Res.* 23:23–35
 - 54b. Stammers M, Rowen L, Rhodes D, Trowsdale J, Beck S. 2000. *BTL-II*: a polymorphic locus with homology to the butyrophilin gene family, located at the border of the MHC class II and class III regions in human and mouse. *Immunogenetics*. 51:373–82
 55. Steinmetz M, Stephan D, Fischer-Lindahl K. 1986. Gene organisation and recombination hotspots in the murine major histocompatibility complex. *Cell* 44:895–900
 56. Stephens R, Horton R, Humphray S, Rowen L, Trowsdale J, et al. 1999. Gene organisation, sequence variation and isochore structure at the centromeric boundary of the human MHC. *J. Mol. Biol.* 291:789–99
 57. Sugaya K, Fukagawa T, Matsumoto KI, Mita K, Takahashi EI, et al. 1994. Three genes in the human MHC class III region near the junction with the class II: gene for receptor of advanced glycosylation end products, PBX2 homeobox gene and a notch homolog, human counterpart of mouse mammary tumor gene int-3. *Genomics* 23:408–19
 58. Svensson AC, Andersson G. 1987. Presence of retroelements reveals the evolutionary history of the human DR haplotypes. *Hereditas* 127:113–24
 59. Takami K, Zaleska-Rutczynska Z, Figueroa F, Klein J. 1997. Linkage of LMP, TAP, and RING3 with Mhc class I rather than class II genes in the zebrafish. *J. Immunol.* 159:6052–60
 60. Tazi-Ahni R, Camp NJ, Cork MJ, Mee JB, Keohane SG, et al. 1999. Novel genetic association between the corneodesmosin (MHC-S) gene and susceptibility to psoriasis. *Hum. Mol. Genet.* 8:1135–40
 61. Tazi-Ahni R, Henry J, Offer C, Bouissou-Bouchouata C, Mather IH, et al. 1997. Cloning, localization, and structure of new members of the butyrophilin gene family in the juxta-telomeric region of the major histocompatibility complex. *Immunogenetics* 47:55–63
 62. te Meerman GJ, Nolte IM, Spijker GT, Boon GT, Buys CHCM. 2000. Systematic haplotype sharing analysis as an asymptotically efficient tool to find gene positions in regions with high haplotype conservation and in complex diseases, illustrated with fine mapping of genes on chromosome 6 in multiple sclerosis and hemochromatosis. *Cytogenet. Cell Genet.* In press
 63. Tenzen T, Yamagata T, Fukagawa T, Sugaya K, Ando A, et al. 1997. Precise switching of DNA replication timing in the GC content transition area in the human MHC. *Mol. Cell. Biol.* 17:4043–50
 64. The chipping forecast. 1999. *Nat. Genet.* 2(Suppl.): 1–60
 65. Thorpe KL, Abdulla S, Kaufman J, Trowsdale J, Beck S. 1996. Phylogeny and structure of the RING3 gene. *Immunogenetics* 44:391–96
 - 65a. Tiwari JL, Terasaki PI, eds. 1985. *HLA and Disease Association*. New York: Springer
 66. Trachtulec Z, Hamvas RMJ, Forejt J, Lehrach HR, Vincek V, et al. 1997. Linkage of TATA-binding protein and proteasome subunit C5 genes in mice and human reveals synteny conserved between mammals and invertebrates. *Genomics* 44:1–7
 67. Trowsdale J. 1993. Genomic structure and function in the MHC. *Trends Genet.* 9:117–22
 68. Trowsdale J. 1995. “Both man & bird and beast”: comparative organization of MHC genes. *Immunogenetics* 41:1–17

-
69. Walter J, Beck S, Olek A. 1999. From genomics to epigenomics: a loftier view of life. *Nat. Biotechnol.* 17:1144
70. Wolfe KH, Sharp PM, Wen-Hsiung L. 1989. Mutation rates differ among regions of the mammalian genome. *Nature* 337:283–85
71. Yeager M, Hughes AL. 1999. Evolution of the mammalian MHC: natural selection, recombination and convergent evolution. *Immunol. Rev.* 167:45–58
72. Zhu XT, Zhao X, Burkholder WF, Gragerov A, Ogata CM, et al. 1996. Structural analysis of substrate binding by the molecular chaperone DnaK. *Science* 272:1606–14
73. Ziegler G, Ehlers A, Forbes S, Trowsdale J, Uchanska-Ziegler B, et al. 2000. Polymorphic olfactory receptor genes and HLA loci constitute extended haplotypes. In *Major Histocompatibility Complex: Evolution, Structure and Function*, ed. M Kasahara, pp. 110–30. Tokyo: Springer-Verlag