GENE MAP OF THE EXTENDED HUMAN MHC

Roger Horton, Laurens Wilming*, Vikki Rand*, Ruth C. Lovering‡, Elspeth A. Bruford‡, Varsha K. Khodiyar‡, Michael J. Lush‡, Sue Povey‡, C. Conover Talbot Jr§, Mathew W. Wright‡, Hester M. Wain‡, John Trowsdale||, Andreas Ziegler¶ and Stephan Beck*.*

Abstract | The major histocompatibility complex (MHC) is the most important region in the vertebrate genome with respect to infection and autoimmunity, and is crucial in adaptive and innate immunity. Decades of biomedical research have revealed many MHC genes that are duplicated, polymorphic and associated with more diseases than any other region of the human genome. The recent completion of several large-scale studies offers the opportunity to assimilate the latest data into an integrated gene map of the extended human MHC. Here, we present this map and review its content in relation to paralogy, polymorphism, immune function and disease.

**Wellcome Trust Sanger Institute, Genome Campus, Hinxton, Cambridge CB10 1SA, UK.
‡HUGO Gene Nomenclature Committee, Department of Biology, University College London, London NW1 2HE, UK.
§HUGO Gene Nomenclature Committee, The Johns Hopkins School of Medicine, Institute of Genetic Medicine, The Johns Hopkins University, Baltimore, Maryland 21205, USA.
||University of Cambridge, Department of Pathology, Immunology Division, Tennis Court Road, Cambridge CB2 1QP, UK.
¶Institut für Immungenetik, Charité-Universitätsmedizin Berlin, Campus Virchow-Klinikum, Humboldt-Universität zu Berlin, 14050 Berlin, Germany.
Correspondence to S.B.
e-mail: beck@sanger.ac.uk*

doi:10.1038/nrg1489

Since its discovery in the mouse in 1936 (REFS 1,2), the major histocompatibility complex (MHC) has become one of the most intensely studied regions in vertebrate genomes[3]. Discovered on the surface of white blood cells (leukocytes), the first MHC gene products became known as leukocyte antigens, which is why the human MHC is also referred to as the human leukocyte antigen (HLA) complex. Although MHC molecules were originally studied for their ability to confer tolerance (histocompatibility) following tissue grafts or later, organ transplants, their primary function is to provide protection against pathogens. This is achieved through sophisticated pathways in which MHC class I molecules present endogenous antigens to CD8$^+$ T CELLS and class II molecules present exogenous antigens to CD4$^+$ T cells. An increasing number of other proteins are being found that support these two pathways; many of these proteins also map to the MHC and, together with the CLASS III COMPLEMENT PROTEINS and inflammatory CYTOKINE GENES, they are discussed in this paper.

Driven by the necessity and desire to improve the survival rate of transplant patients and to tackle infectious and autoimmune disease, successive and improved MHC gene maps were generated for various vertebrate species over several decades, culminating in the 1993 foldout map of the human MHC[4]. The first sequence-based map, which followed in 1999, described 224 gene loci, of which 128 (57%) were thought to be expressed[5]. Although the 3.6 megabase-pair (Mb) long sequence was contiguous, it had been derived from many individuals of unknown HLA type, therefore producing a virtual, or mosaic, MHC HAPLOTYPE. Around the same time, there was confirmation of high LINKAGE DISEQUILIBRIUM[6], and evidence of CONSERVED SYNTENY[7] and the presence of MHC-relevant genes (*HFE*, *TRIM38* and BTNs) extending far beyond the boundaries defining the human MHC at that time. Analysis of this evidence led to the idea of an extended MHC in humans (xMHC), as compared with the classical MHC[8]. A sequence of the newly defined xMHC was completed in 2003 as part of the sequencing of the entire human chromosome 6 and now covers a total of 7.6 Mb on the short arm of this chromosome[9]. Furthermore, 4.6 Mb of the virtual MHC haplotype of 1999 has been replaced by a single reference sequence of a homozygous haplotype derived from sequencing the consanguineous PGF CELL LINE[10]. The comparative analysis of other MHCs, in particular of mammals, contributed to the validation of the MHC gene content in human, and identified a mosaic structure formed from stretches of conserved and non-conserved genes[11]. Throughout this period, the names of genes within the human MHC changed many times, resulting in confusion, bemusement and irritation. Except for the eponymous RING finger gene, *RING1* (REF. 12), the time has come, therefore, to bid farewell to graduate-student inspired gene names such as RING (really interesting new gene) and BING (bloody interesting new gene) and agree on official gene symbols that are assigned by the HUGO Gene Nomenclature Committee (HGNC) and IMGT/HLA Sequence Databases[13,14] (see Online links box). The former has now also instigated the use of the root symbol 'HCG' for 'HLA complex group' rather than the defunct, class I 'haemochromatosis candidate gene', to designate loci throughout the xMHC. By merging and integrating all the publicly available data, we derived the xMHC gene map that is reported here and we discuss it in the context of the main hallmarks of MHC biology and disease.

Gene map

FIGURE 1 shows the gene map of the human xMHC from telomere to centromere on the short arm of chromosome 6. Its orientation is in line with the chromosome in the ENSEMBL and VEGA databases, where full annotation is available. The five subregions that make up the xMHC span about 7.6 Mb and are colour-coded in the figure. The nomenclature used in the map is that of the HGNC and its database[13]. The annotation of gene loci used is that currently available for the chromosome 6 reference sequence[9], with updates to make it consistent with the MHC haplotype project[10]. All annotation discussed here is to the standards set by the Human Annotation Workshop (HAWK). Of the 421 loci, 252 (60%) are classified as being expressed genes on the basis of cDNA and/or EST evidence; 30 loci (7%) are classified as transcripts on the basis of EST evidence, but without ORFs; and 139 loci (33%) are classified as pseudogenes on the basis of similarity to known proteins, while containing frameshifts and/or stop codons. Of the above loci, there are at least four pseudogenes (*PPP1R2P1* and three olfactory-receptor loci) that seem to occur as genes (or at least ORFs) in other haplotypes[10,15]. With 58 (23%) of the expressed genes, class III is the most gene dense subregion of the xMHC and of the human genome[16].

Compared with protein-coding genes, non-coding RNA genes (except for tRNA genes) are much more difficult to predict, and at the moment they still lack experimental verification[17]. They have therefore been excluded from the gene map in FIG. 1. However, the positions of the tRNA genes are indicated and will be discussed in the later outlook section. Also indicated are the positions of two hypervariable regions (the RCCX region[18] (class III) and the DRB region[19] (class II)) that vary in gene number depending on the haplotype.

When applied to the xMHC, the current standards of annotation might fail to replicate the efforts of previous investigators, either owing to lack of evidence of homology or because loci are now annotated as repeat elements[20]. For this reason, several of the loci that have been previously reported[5] are not represented in the map in FIG. 1. These include: in the class I subregion, the *NOB* pseudogenes, many or all of the *HCG2P* (*HCGII*), *HCG4P* (*HCGIV*), *HCG8* (*HCGVIII*), *HCG9P* (*HCGIX*), *HCG26P* (*3.8–1*) and *HCP5P* (*P5*) pseudogenes, and the genes *CAT75X* and *C6orf215* (*GT257*); and in the class II subregion, the pseudogenes *B30.2-L*, *RING13*, *RING14* and *BING5*.

Gene clusters

Recent studies indicate that around 5% of the human genome can be attributed to segmental duplications[21], and that both large- and small-scale duplications are required to explain the distribution of human gene families[22]. Duplication that results in the formation of gene clusters is a particular hallmark of the xMHC and is immediately apparent from FIG. 1. There might be explanations for clustering immune-system genes in the MHC. If the products of genes physically associate, for example, in *HLA-DQA* and *HLA-DQB*, linkage can ensure that the protein components will be co-expressed in quantities appropriate for the formation of heterodimers. Similar observations have also been made for sequence-unrelated genes with related functions, such as those involved in inflammation and antigen processing, including peptide transporters (*TAP1/2*), IMMUNO-PROTEASOME components (*PSMB8/9*) and the peptide chaperone *TAPBP*. Enhanced coordination of gene expression could be a possible advantage of this type of clustering. Duplication of genes in *cis* could also confer a selective advantage. MHC class I and class II genes seem to undergo periodic expansion and contraction within and between species, presumably in response to demands of the changing range of pathogens. Clustering of related sequences might also facilitate increased sequence exchange[23].

The two largest (and partially overlapping) gene clusters in the xMHC, and indeed the two largest clusters of their kind in the human genome, are those of the histone and tRNA genes. Transcripts of both gene classes are required in enormous quantities. RNA genes (including tRNAs) constitute about 80% of cellular transcription in eukaryotes[24], and about 10^8 molecules of each core histone are required during the brief S-phase of the cell cycle[25]. Histone and tRNA genes might therefore be under selection pressure to cluster in order to maximize transcription levels and/or co-localize in chromosomal regions such as the MHC, which for other reasons, have higher than average transcriptional activity (transcription hotspots). The odds of this hypothesis being correct have recently increased with the finding of a significant non-random correlation ($p<0.001$) between transcription hotspots (including the xMHC) and the location of tRNA clusters[9]. Taken together, these findings open intriguing new possibilities for explaining the observed linkage disequilibrium between regions within the classical and the extended MHC[6]. It now seems equally plausible that the MHC could be hitchhiking with the histone and/or tRNA cluster(s), rather than the other way round. It remains to be seen if this interesting feature is evolutionarily conserved.

Clusters are defined here as three or more paralogous genes or pseudogenes that are present within a 1 Mb stretch, and superclusters are defined as clusters with

T CELLS
These are lymphocytes that have an important role in the primary immune response, so named because their final stages of development occur in the thymus. CD8$^+$ killer or cytotoxic T cells destroy infected cells, whereas CD4$^+$ or helper T-cells regulate the function of other lymphocytes.

CLASS III COMPLEMENT PROTEINS
Proteins that are involved in a cascade of proteolytic cleavage of glycoproteins, ultimately leading to the induction of inflammatory responses and damage to pathogens.

CYTOKINE GENES
These encode a wide array of proteins that mediate signalling between cells either at close range or at a distance.

HAPLOTYPE
The combination of alleles at several loci on a single chromosome of a given individual. For example, for a marker with alleles M and m that is linked to another locus with alleles Q and q, possible haplotypes are MQ, Mq, mQ and mq.

LINKAGE DISEQUILIBRIUM
The non-random association of alleles at adjacent loci.

CONSERVED SYNTENY
The occurrence of genomic collinearity between homologous genes in different organisms.

PGF CELL LINE
A consanguineous B-lymphoblastoid cell line, derived from a European caucasoid male, with a homozygous HLA haplotype.

IMMUNO-PROTEASOME
Large protease complexes that degrade proteins into peptides for association with MHC class I molecules. The immunoproteasome contains some subunits that are induced by γ interferon, two of which are encoded in the MHC.

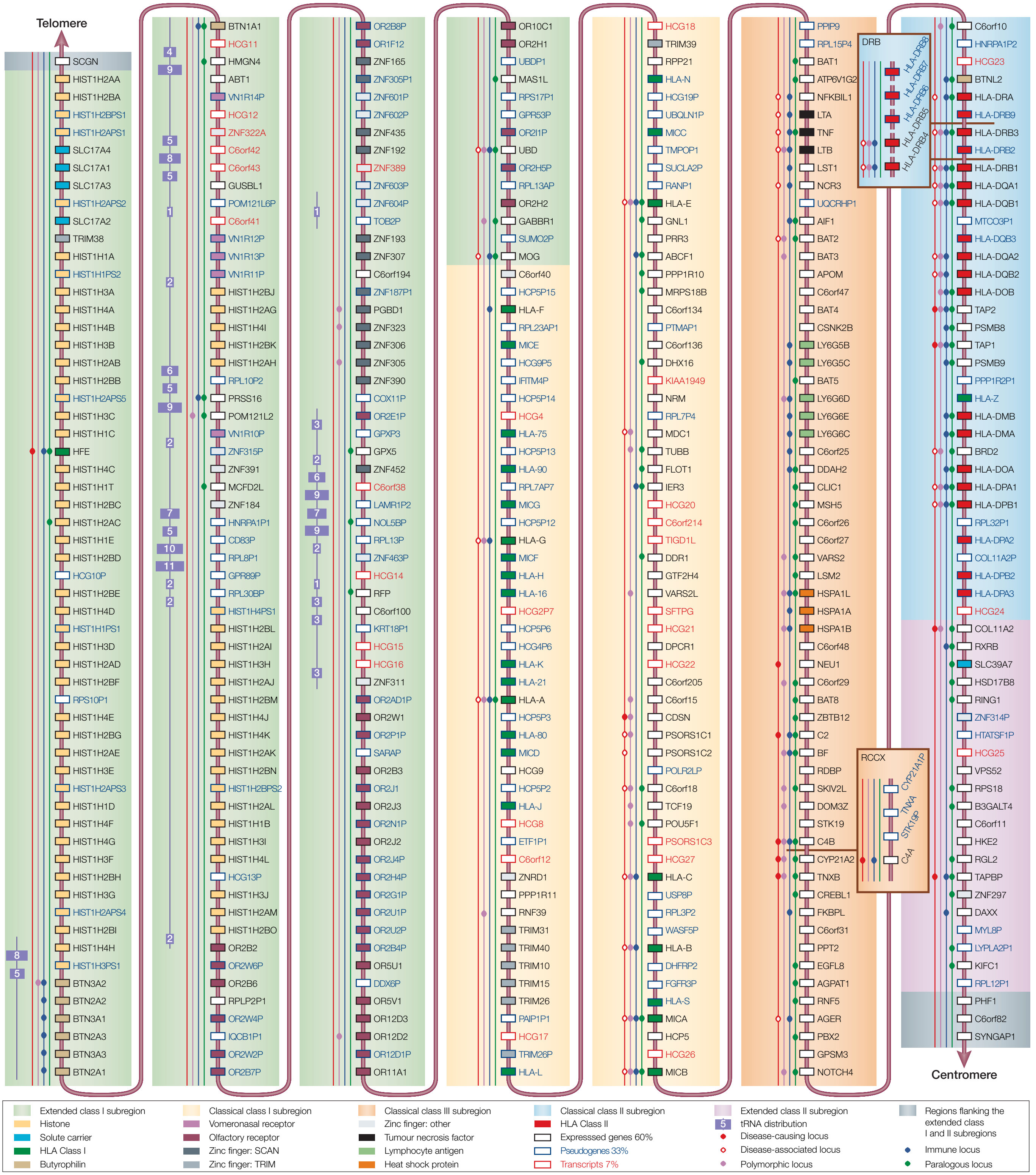

Figure 1 | **Gene map of the extended major histocompatibility complex (xMHC).** The gene map is shown from telomere (left) to centromere (right) on the short arm of chromosome 6. The five colour-coded subregions making up the xMHC span about 7.6 Mb and are defined as: the extended class I subregion (green block; *HIST1H2AA* to *MOG*; 3.9 Mb), the classical class I subregion (yellow block; *C6orf40* to *MICB*; 1.9 Mb), the classical class III subregion (orange block; *PPIP9* to *NOTCH4*; 0.7 Mb), the classical class II subregion (blue block; *C6orf10* to *HCG24*; 0.9 Mb) and the extended class II subregion (pink block; *COL11A2* to *RPL12P1*; 0.2 Mb). Regions that flank the extended class I and II subregions are shown as grey blocks. Insets denote the hypervariable RCCX and DRB regions. Numbers and positions of tRNA genes are represented by indigo bars, the length of which is proportional to the gene number between other loci. Vertical lines connect the two main groupings of tRNA genes of 1.6 Mb and 0.5 Mb of the sequence (separated by 0.6 Mb). Circles to the left of each locus indicate disease status, polymorphism, immune status and paralogy as described in the text. The gene map of the xMHC is also available as a poster, which accompanies this issue. The poster is available online (http://www.nature.com/nrg/journal/v5/n12/poster/MHCmap).

Table 1 | **Gene (super) clusters within the xMHC**

Cluster type	Total number of loci	Number of protein-coding loci	Number of pseudo-gene/transcript loci
Gene superclusters			
Histone	66	55	11
HLA class I	26	9	17
tRNA	157	151	6
Butyrophilin	8	8	0
Olfactory receptor*	34	14	20
Zinc finger protein	36	26	10
Gene clusters			
Solute carrier 17A	4	4	0
Vomeronasal receptor	5	0	5
Tumour necrosis factor	3	3	0
Lymphocyte antigen-6	5	5	0
Heat shock protein	3	3	0
HLA class II‡	24	15	9

*The distribution of olfactory loci between the gene and pseudogene categories is dependent on haplotype. ‡The number of loci in the HLA class II supercluster varies between different haplotypes. Please see text for details on each individual cluster. xMHC, extended major histocompatibility complex.

NATURAL KILLER CELL
Large granular non-T, non-B-type lymphocytes. Natural killer cells are important for the early response to viruses. They produce cytokines, kill certain tumour cells and have appropriate receptors for antibody-dependent cell-mediated cytotoxicity.

LEUKOCYTE RECEPTOR COMPLEX
A cluster of genes on chromosome 19q13.4. The products of some of the IgSF genes in the LRC are expressed on natural killer cells and serve as receptors for MHC class I molecules.

NATURAL KILLER COMPLEX
A cluster of genes on chromosome 12. The products of some of the lectin-related genes in the NKC are expressed on natural killer cells and serve as receptors for MHC class I ligands.

additional related gene(s) outside the core cluster, but within the xMHC. Following this definition, the xMHC harbours six clusters and six superclusters (TABLE 1), which are by no means restricted to only the immune genes. These clusters and superclusters are described according to their position on the chromosome from telomere to centromere.

Histone supercluster. Histones are basic proteins involved in nucleosome formation. They occur in five classes, H1 (linker histone), and H2A, H2B, H3 and H4 (core histones). With a total of 66 loci (55 expressed genes, 11 pseudogenes), they present the largest histone cluster in the human genome and the largest protein-encoding supercluster within the xMHC[25]. Individually there are 6 H1, 12 H2A, 15 H2B, 10 H3 and 12 H4 expressed genes.

Solute carrier cluster. Solute carrier (*SLC*) genes are part of a diverse family with poorly understood, but crucial physiological roles in solute and nutrient transport[26]. The *SLC17A* genes within the xMHC[27,28] have been shown to co-transport sodium and phosphate.

HLA class I supercluster. The HLA class I supercluster comprises the classical class I genes (*HLA-A, -B* and *-C*), the non-classical class I genes (*HLA-E, -F, -G, HFE* and 12 pseudogenes) and the class I-like genes (*MICA, MICB,* and 5 pseudogenes). Although the product of the gene *HFE* (in the telomeric region of the xMHC) resembles a classical class I protein in many respects, including the association of its heavy chain with β-2-microglobulin (*B2M*), its function is in iron metabolism rather than in antigen processing and presentation. There are many class I-like genes in the human genome[29]. In conjunction with *B2M*, the classical class I gene products present antigens to CD8$^+$ T-cells and are involved in the NATURAL KILLER CELL mediated immune response, as are the non-classical class I genes *HLA-E* and *-G*. This response involves interaction with receptors encoded in the LEUKOCYTE RECEPTOR COMPLEX (LRC) and the NATURAL KILLER COMPLEX (NKC). The expression profile of the class I-like *MIC* genes[30], the products of which are ligands for the activating receptor NKG2D (encoded by *KLRK1*), indicates a possible role in the mucosal immune system of the gut in addition to other tissues.

tRNA supercluster. tRNA genes are only 75–90 bases long and are crucial as the molecular adaptors in mRNA-mediated protein synthesis[31]. The presence of a tRNA cluster on the short arm of chromosome 6 at 6p21.2–p22.3 was first reported in 1996 (REF. 32). Later, as part of the chromosome 6 sequencing effort[9], it was found to be the largest tRNA cluster in the human genome, comprising 157 tRNA genes including all principal species except for Asn- and Cys-tRNAs (TABLE 2). Numbers and positions of tRNA genes are represented in FIG. 1.

Butyrophilin supercluster. Butyrophilin (BTN) genes are members of the immunoglobulin superfamily (IgSF) and display notable sequence similarity to other xMHC loci, particularly to mammalian *MOG* and the chicken MHC gene *B-G*[33]. The *BTNL2* locus in the classical class II subregion is highly variable in different vertebrate MHCs[34]. Comparative analysis of chicken sequences indicates that the BTN genes might be the result of exon shuffling events that occurred between the *B-G* and *B30* genes about 300 million years ago during the divergence of birds and mammals[34]. The function of the BTN genes is still unknown. One BTN gene (*BTN1A1*) is expressed exclusively and abundantly in milk[35].

Vomeronasal-receptor cluster. Vomeronasal-receptor (VNR) genes are members of the pheromone receptor family, which is involved in the subconscious perception of volatile substances such as pheromones. Humans seem to have lost this particular sensory pathway[36], which is concordant with the exclusive presence of pseudogenes in the VNR cluster.

Olfactory-receptor supercluster. With about 400 potentially functional members, the family of olfactory-receptor genes provides the basis for odour perception. The olfactory-receptor supercluster contains 34 olfactory-receptor loci, 14 of which are potentially functional. At least three additional members might be functional, depending on the haplotype. Similar to immune genes that provide protection from pathogens, they provide an essential survival tool in behavioural processes, including reproduction and predation. Also in common with some immune genes, olfactory-receptor genes are polymorphic[15] and subject to unorthodox expression and splicing, at least in the case of some MHC-linked olfactory-receptor genes in testicular tissue[37]. Together with polymorphic MHC antigens, sperm-expressed olfactory-receptor genes might be functionally involved in the selection of spermatozoa by the female, as proposed by the SPERM-RECEPTOR SELECTION HYPOTHESIS[38].

Zinc-finger supercluster. Genes that encode zinc-finger proteins are grouped according to the presence of particular zinc-finger domains rather than overall sequence similarity. After iron, zinc is the most abundant trace metal in the human body and many proteins have evolved that bind one or more zinc ions. Zinc-finger gene products have diverse functions and can act as enzymes, storage proteins, replication proteins and transcription factors[39]. The 36 loci making up this supercluster have been further subdivided (indicated by different shades of grey in FIG. 1) into gene products containing SCAN domains (13 loci)[40], tripartite motifs (*TRIM*, 8 loci) consisting of a ring-finger domain, B-box and a SPRY domain[41], or other zinc-finger domains (15 loci)[42].

Tumour necrosis factor cluster. The tumour necrosis factor (TNF) cluster contains genes for three cytokines (*TNF*, *LTA* and *LTB*). All three genes belong to the TNF superfamily and are involved in various inflammatory pathways[43].

Lymphocyte antigen cluster. Lymphocyte antigen 6 (LY6) genes encode glycosyl-phosphatidyl-inositol (GPI) anchored cell-surface proteins with putative immune function[44]. They are conserved in the mouse and present the largest gene cluster within the MHC class III subregion.

Heat shock cluster. Heat shock protein (HSP) genes are upregulated by cellular stress such as heat shock and act as chaperones in the synthesis, folding, assembly, transport and degradation of proteins[45]. The cluster of three HSP genes in the MHC class III subregion is involved in stress-induced signalling for immune system mediated elimination of damaged, infected or malignant cells[46].

HLA class II cluster. The HLA class II cluster comprises the classical class II genes (*HLA-DP, -DQ, -DR* and pseudogenes) and the non-classical class II genes (*HLA-DM* and *-DO)*. The classical class II genes are expressed on the cell surface as heterodimers consisting of corresponding α and β chains that present antigens to CD4$^+$ T cells. The non-classical class II genes are not expressed on the cell surface, but form heterotetrameric complexes involved in peptide exchange and loading onto classical class II molecules[47]. No class II-like gene has yet been found elsewhere in the human genome.

Paralogy

Gene duplication resulting in paralogous copies is an important mechanism driving genome evolution[48]. Immune genes, including MHC genes, might be under strong selection pressure to diversify in order to cope with ever changing pathogenic epitopes, and duplication followed by diversification provides one of the means to achieve this. According to the 'functionalization model'[49], a paralogous gene will undergo one of three fates: non-functionalization, whereby one gene copy degenerates after duplication; neo-functionalization, whereby one or both copies acquire new function(s); and sub-functionalization, whereby duplicate genes with multiple functions diverge by reciprocal loss of function. The xMHC has been a model for such studies for a long time and three hypotheses have been proposed to explain MHC paralogy[50]. The first, known as the '2R-hypothesis', proposes that two rounds of genome duplication occurred early in the vertebrate lineage[51]. Although this hypothesis explains why there are the three observed MHC paralogous regions on chromosomes 1, 9 and 19, it fails to maintain significance when applied on a genome-wide scale[52]. The second, 'selection-based clustering' hypothesis, allows for individual gene duplications, but needs to invoke a currently unproven selection-based mechanism to explain the clustering pattern[53]. The third, 'block duplication' hypothesis[54], solves the problem associated with the 2R-hypothesis, but remains discordant with the diverse duplication dates determined for some paralogues within the same block by phylogenetic analysis[53].

Although the debate will undoubtedly continue, we have compiled a comprehensive list of xMHC genes that have one or more paralogue(s) in the human genome but outside the xMHC. As the identification of 'true' paralogues is not straightforward, the analysis was conducted in a way that allowed assignment of

Table 2 | **tRNA genes within the xMHC**

tRNA species	Copy number
Ala	28
Arg	11
Asn	0
Asp	3
Cys	0
Gln	9
Glu	1
Gly	1
His	1
Ile	13
Leu	10
Lys	6
Met	13
Phe	7
Pro	2
Ser	17
Thr	10
Trp	2
Tyr	4
Val	12
Undetermined	1
Pseudogene	6
TOTAL	157

xMHC, extended major histocompatibility complex.

SPERM-RECEPTOR SELECTION HYPOTHESIS
This theory proposes that olfactory receptors that are expressed on spermatozoa and polymorphic antigens (for example, MHC class I molecules) might be functionally connected, ensuring that spermatozoa have a higher chance to fertilize a genetically different oocyte than spermatozoa that share alleles with the female, in particular on the MHC.

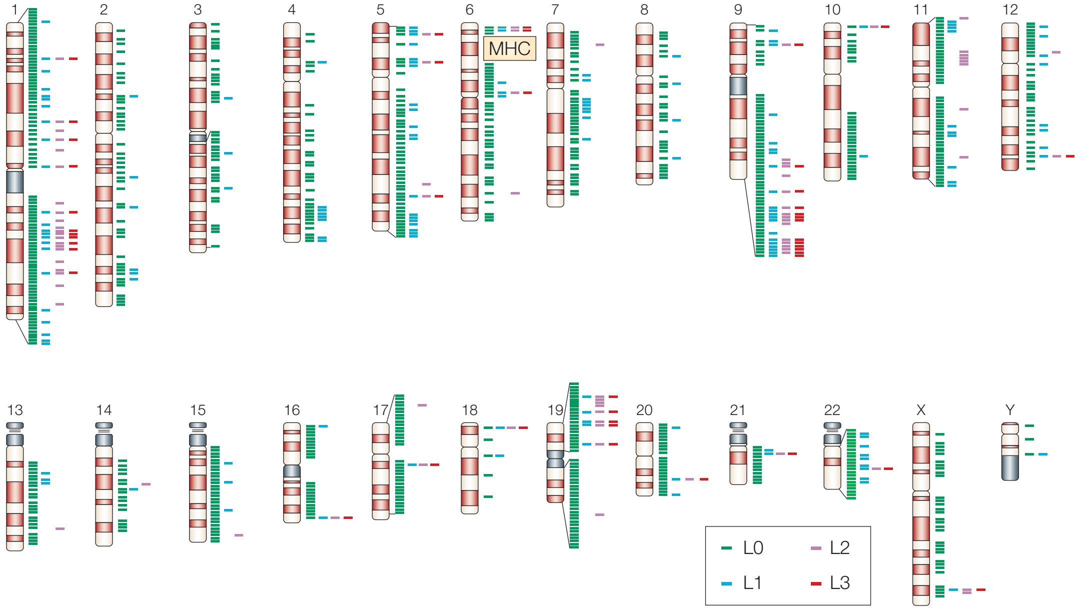

Figure 2 | **Distribution of major histocompatibility complex (MHC) paralogues in the human genome.** The approximate positions of the putative paralogues are colour-coded according to confidence level: L0 column represents BLAST similarity matches with a *p*-value of less than 10^{-5} (green); L1 column represents BLAST matches after filtering out domain-only matches (blue); L2 column represents BLAST matches after filtering for conserved gene structure[100] (purple); L3 column represents BLAST matches that passed both filtering steps (red).

confidence levels (L0, L1, L2, L3 in ascending order of confidence) to each paralogue. According to this analysis, 88 genes were identified within the xMHC that have a combined total of 791 putative paralogues (618 L0s, 91 L1s, 38 L2s and 44 L3s) throughout the human genome[55]. The 88 xMHC genes are shown next to their corresponding gene boxes in FIG. 1, and the distribution of the 791 putative paralogues is shown in FIG. 2. These data demonstrate that MHC paralogues cluster in, but are not restricted to the three previously known MHC-paralogous regions on chromosomes 1, 9 and 19, and that there is the potential for considerable functional redundancy for these 88 MHC genes elsewhere in the genome. A recent study further indicates that the NKC on chromosome 12 and the MHC might share a common ancestral region. On the basis of comparative analysis of the chicken MHC[56], which encodes two C-type lectin receptors (*BNK* and *Blec*), the study identified conservation in sequence, position and functional features of two of the human C-type lectin receptor genes (*KLRB1* and *CLEC2E*)[57], in addition to the *TAPBP*-like gene (*TAPBPL*)[58] in the NKC.

Polymorphism

HLA class I and class II genes are highly polymorphic[14] (see the IMGT/HLA Sequence Database in the Online links box) and *HLA-B* was recently confirmed to be the most polymorphic gene in the human genome[9]. SNPs are the most common type of variation and of these, coding SNPs (cSNPs) are the most informative sub-type. The ratio of non-synonymous (Ka) to synonymous (Ks) substitutions is indicative of whether a gene is under PURIFYING SELECTION (Ka/Ks <1), POSITIVE SELECTION (Ka/Ks >1) or neutral selection (Ka/Ks ~1)[59]. A systematic calculation of Ka/Ks ratios has not yet been carried out for the genes encoded within the xMHC. The most dramatic form of cSNPs are those that are non-synonymous or amino-acid changing. Those genes with greater than the average number of non-synonymous cSNPs (4.0 for the xMHC) as recorded in the ENSEMBL database[60], in the IMGT/HLA database[14] or by the MHC haplotype project[10], are shown in FIG. 1. By this definition, 56 genes (22%; excluding pseudogenes and tRNAs) can be classified as 'polymorphic', including 29 members of the 12 (super) clusters. To these could be added *PPP1R2P1*, *OR2J1*, *OR10C1* and *OR12D1P*, which can be genes or pseudogenes in different haplotypes[10,15]. Polymorphism is, however, not restricted to SNPs. The MHC is also rich in deletion/insertion polymorphisms (DIPs), which were recently catalogued by Stewart and colleagues[10]. In addition, the xMHC contains two large regions (FIG. 1) of gene polymorphism and duplication: the RCCX might contain a modular duplication of a region with a *C4* gene and *TNXA*, *CYP21A1P* and *STK19P* pseudogenes[18], whereas the DRB locus between *HLA-DRB1* and *HLA-DRB9* shows haplotype-specific arrangements of the *HLA-DRB3*, *4* and *5* genes and the *HLA-DRB2*, *6*, *7* and *8* pseudogenes[61]. Most of these and other MHC polymorphisms are freely available from the NCBI Single Nucleotide Polymorphism database and several efforts are now in progress to generate fully informative polymorphism and linkage disequilibrium maps that are a prerequisite for population-based mapping of MHC-associated disease[10,62–64].

PURIFYING SELECTION
Alternatively known as negative selection. A process in which more non-synonymous (amino-acid changing) than synonymous substitutions have been eliminated. It is observed, for example, when a substitution is deleterious and therefore has been eliminated from a population.

POSITIVE SELECTION
A process in which more non-synonymous (amino-acid changing) than synonymous substitutions have been preserved. It is observed when non-synonymous substitutions in a gene are selectively advantageous, for example, increasing the fitness of the species.

Table 3 | **A minimal set of immune-system genes in the human xMHC**

Category	Genes
Antigen processing/ presentation	*HLA-A*, *-B*, *-C*, *-DMA*, *-DMB*, *-DOA*, *-DOB*, *-DPA1*, *-DPB1*, *-DQA1*, *-DQA2*, *-DQB1*, *-DQB2*, *-DRA*, *-DRB1*, *-DRB3*, *-DRB4*, *-DRB5*; *PRSS16*; *PSMB8*, *PSMB9*; *TAP1*, *TAP2*, *TAPBP*; *UBD*
Immunoglobulin superfamily	*AGER*; *BTN1A1*, *BTN2A1*, *BTN2A2*, *BTN2A3*, *BTN3A1*, *BTN3A2*, *BTN3A3*, *BTNL2*; *C6orf25*; *MOG*
Inflammation	*ABCF1*; *AIF1*; *DAXX*; *IER3*; *LST1*; *LTA*, *LTB*; *NCR3*; *TNF*
Leukocyte maturation	*DDAH2*; *LY6G5B*, *LY6G5C*, *LY6G6D*, *LY6G6E*, *LY6G6C*
Complement cascade	*BF*; *C2*, *C4A*, *C4B*
Non-classical MHC class I	*HLA-E*, *HLA-F*, *HLA-G*; *HFE*
Immune regulation	*NFKBIL1*, *RXRB*, *FKBPL*
Stress response	*HSPA1A*, *HSPA1B*, *HSPA1L*; *MICA*, *MICB*

Most of these genes have established functions for innate or adaptive immunity; genes with remote links have been excluded. Some genes have been included because they are related by sequence to a known immune gene family but the precise function of these is still to be determined. The largest class of immune system genes is involved with antigen processing and presentation, and includes classical class I and II molecules, as well as some of the antigen processing machinery for loading peptides onto class I molecules. xMHC, extended major histocompatibility complex.

Immune function

Survival from attack by pathogens requires a large investment in many defence strategies. This is because some pathogens provide a 'moving target' by altering external antigens (for example, *Trypanosoma cruzi*) or proteins that could provide T-cell epitopes (for example, HIV). The scope of this investment is reflected at the genome level; in humans, more than 5% of expressed genes are involved in immune defence[65]. Some immune-system genes have features that reflect this crucial role. POLYGENY, polymorphism and rapid evolution are features that are exemplified by some MHC genes, including class I and class II. Indeed, comparisons of immune genes in different species indicate that they tend to differ in number, consistent with a drive for novelty, potentially in response to pathogen immune evasion[66,67].

Many MHC genes, including HLA class I and class II genes, are part of the adaptive immune system. This system, which is characterized by immunological memory, has developed to provide the flexibility to respond to numerous and highly variable antigenic targets. By contrast, the innate system recognizes microbial pathogens through a limited repertoire of germline-encoded receptors, such as the Toll family[68]. The division between adaptive and innate immunity, although convenient for textbooks, is not clear-cut and in fact the two systems are inextricably intertwined. The MHC helps to provide the link between the innate and adaptive immune systems. In addition to the role of class I molecules in presenting antigens to T-cell receptors, they are also used by natural killer cells, which are thought of as part of the innate system, to indicate immune evasion by a pathogen that has downregulated class I antigen expression. A wide range of viruses, particularly the large herpes viruses, but also smaller retroviruses such as HIV, broadly or specifically downregulate class I expression to avoid immune recognition[69]. Variation in both innate and adaptive immune-system genes influences their association with disease[68]. The MHC is linked more directly with innate immunity by encoding molecules such as MICA and MICB, which are upregulated by stress and which activate cells by the NKG2D receptor (encoded by *KLRK1*)[70].

In earlier MHC maps, it was noticed that immune-system genes tended to be overrepresented and comprised up to 40% of expressed genes in the classical MHC. In FIG. 1 and TABLE 3 we have listed the immune loci that are identifiable within the xMHC. Over this region, about 28% of the expressed transcripts are potentially associated with immunity and fall into one or more of the following categories: antigen processing/presentation, immunoglobulin superfamily, inflammation, leukocyte maturation, complement cascade, non-classical MHC class I receptor family, immune regulation and stress response. This variety of different immune functions might help to explain some of the complex associations of the region with disease. There are, however, still some genes that, on the basis of sequence similarity, are members of the immunoglobulin superfamily but have an unknown immune function. A notable example of such genes is the butyrophilin supercluster (*BTNL2* and the *BTN* loci), which are also part of the extended co-receptor superfamily including *CD80/CD86*.

Disease association

The MHC is associated with hundreds of diseases, including most if not all, autoimmune diseases[71]. Several disease candidate genes have been reported in the literature for which the aetiological mutation/variant has not yet been identified. We have therefore indicated whether the disease-related genes of the xMHC are disease-causing or disease-associated genes (TABLE 4, see also FIG. 1). In several cases, which could turn out to be the majority, a mutation in a given gene (for example, *HFE*) might both cause a disease (haemochromatosis[72]) and be associated with others (arthropathy, multiple sclerosis and hepatocellular carcinoma[73–75]).

Establishing a direct correlation between a disease and an MHC gene is made particularly difficult by two factors. Nearly all MHC-associated diseases are MULTIFACTORIAL DISEASES and, even when an association with the MHC has been established, the linkage disequilibrium characterizing many MHC haplotypes often prevents an unambiguous identification of the disease-causing or disease-associated locus. Despite these problems, the progress made in recent years is notable and has been achieved on different levels. For example, following about two decades of research, a promoter mutation of the *BRD2* gene has finally been linked to a common form of the idiopathic generalized epilepsies, juvenile myoclonic epilepsy[76], and the *NFKBIL1* locus has been identified as the second gene (after *HLA-DRB1* alleles) within the MHC that is associated with rheumatoid arthritis[77]. Another gene — the product of which is probably involved in autoimmunity and inflammation — is *AGER*, which

POLYGENY
The presence of several different but related genes with similar function. Polygeny of MHC class I genes ensures that each individual produces different MHC molecules.

MULTIFACTORIAL DISEASE
A disease that is influenced by multiple genetic, epigenetic or environmental factors.

ADVANCED GLYCATION
Non-enzymatic glycoxidation process involving sugars and basic amino acids of various proteins. The end products (AGEs) are chemically diverse, stable and implicated in various diseases in which deposits are formed, including amyloidosis, atherosclerosis and rheumatoid arthritis.

Table 4 | **Genes in the xMHC in which variation has a relationship to disease**

Gene symbol	Relationship to disease*
xMHC extended class I region	
HFE	Causes haemochromatosis; associated with arthropathy, multiple sclerosis, hepatocellular carcinoma
UBD	Associated with gastrointestinal and gynaecological malignancies
MOG	Associated with multiple sclerosis
xMHC class I region	
HLA-G	Associated with *Pemphigus vulgaris* in Jewish patients
HLA-A	Associated with autoimmune diseases; for example, birdshot chorioretinopathy
HLA-E	Associated with type 1 *Diabetes mellitus*; also influences age of onset of disease
MDC1	Associated with inadequate DNA damage responses owing to MDC1-deficiency
CDSN	Causes hypotrichosis simplex of the scalp
PSORS1C1	Associated with psoriasis
PSORS1C2	Associated with psoriasis
C6orf18	Associated with psoriasis
HLA-C	Associated with autoimmune diseases; for example, psoriasis
HLA-B	Associated with autoimmune diseases; for example, ankylosing spondylitis or Behcet disease
MICA	Associated with autoimmune diseases; for example, rheumatoid arthritis and coeliac disease
MICB	Associated with coeliac disease
xMHC class III region	
NFKBIL1	Associated with rheumatoid arthritis
LTA	Associated with myocardial infarction
TNF	Associated with septic shock, cerebral malaria
LTB	Associated with infective/inflammatory diseases
NCR3	Associated with impairment of NK cell function in HIV-1 infected patients
BAT2	Associated with influence on age at onset of type 1 *Diabetes mellitus*
NEU1	Causes type I and II sialidosis
C2	Causes C2 deficiency
C4B	Causes C4 deficiency
C4A	Causes C4 deficiency
CYP21A2	Causes several disorders owing to 21-hydroxylase deficiency
TNXB	Causes Ehlers–Danlos syndrome (hypermobility type) owing to tenascin X deficiency
AGER	Associated with amplification of inflammatory responses in rheumatoid arthritis
xMHC class II region	
HLA-DR loci	Associated with autoimmune diseases; for example, rheumatoid arthritis, type 1 and type 2 *Diabetes mellitus*
HLA-DQ loci	Associated with autoimmune diseases; for example, narcolepsy
TAP2	Causes bare lymphocyte syndrome type I owing to TAP2-deficiency; associated with various diseases; for example, rheumatoid arthritis
TAP1	Causes bare lymphocyte syndrome type I owing to TAP1-deficiency; associated with various diseases; for example, vitiligo in Caucasian patients that are young in age at onset
BRD2	Associated with juvenile myoclonic epilepsy
HLA-DP loci	Associated with autoimmune diseases; for example, chronic beryllium disease
xMHC extended class II region	
COL11A2	Causes autosomal dominant deafness (DFNA13) and several other diseases
TAPBP	Causes bare lymphocyte syndrome type I owing to TAPBP-deficiency

*It is only where a causal relationship has been established that it is stated that a mutated gene 'causes' disease, otherwise it is referred to as being 'associated with' a disease, even where the relationship is nearly obligatory (as is the case between narcolepsy and *HLA-DQB1*0602*). Some alleles might also confer resistance to particular diseases (see main text for an example). Occasionally, there is controversy regarding the type to which a particular disease belongs. For example, narcolepsy[98] has not always been considered to be an autoimmune disorder. By contrast, coeliac disease is now thought not to belong to this group, as it might result from an inappropriate inflammatory response to the modified dietary component gluten[99]. Further insight is provided by Hansen and Dupont[88]. NK, natural killer; xMHC, extended major histocompatibility complex.

microRNAs
A class of small (approximately 22-bp) non-coding RNAs that have an important role in gene regulation.

encodes the receptor for ADVANCED GLYCATION end products[78]. A mutation in *AGER* has been pinpointed that enhances inflammatory responses in patients with rheumatoid arthritis[79].

Remarkably, although the principle structural features of class I and II molecules have been established for a long time[80,81], the structures of additional molecules elucidated by X-ray crystallography might still hold surprises that have an influence on disease. First, the structure of the HLA-DQA1*0102/DQB1*0602 molecule indicates that the volume of the P4 pocket of the peptide binding groove is associated with susceptibility to narcolepsy. By contrast, the volume of the P6 and the specificity of the P9 pockets are linked with the dominant protection against type 1 diabetes that *DQB1*0602* confers[82]. Second, the ankylosing spondylitis associated HLA-B*2705 subtype can display a self-peptide in two extremely different conformations, whereas the B*2709 subtype that is not associated with ankylosing spondylitis and differs in only a single amino acid from B*2705 does not have this unorthodox property[83]. Individuals with these *HLA-B27* subtypes generate differential T-cell repertoires against this self-peptide, linking a gene polymorphism with peptide presentation, T-cell selection and ankylosing spondylitis development. It might also be that other pathological conditions are influenced by such subtle differences in antigen presentation[84,85]. It must also be pointed out that the enormous degree of MHC polymorphism, although highly desirable in healthy individuals for dealing with infections, poses great problems in a transplant setting, as the search for a suitable bone marrow or organ donor may be extremely difficult. Research in this area seeks to gain a deeper insight into the structure–function relationships of MHC class I and II molecules[82,83], define 'taboo' combinations of donor/recipient HLA alleles[86] and intervene selectively to establish tolerance in the recipient, without destroying essential components of the immune system[87].

Regarding genes outside the classical MHC, a locus in strong linkage disequilibrium with certain *HLA-A* alleles has been known for many years to be intimately associated with haemochromatosis. A frequent molecular defect in the *HFE* gene has finally been identified that prevents its product from associating with *B2M*, rendering the molecule non-functional[72]. Readers interested in more diseases and their association with HLA genes should consult the proceedings of the Histocompatibility Workshop and Congress, which are soon to be published[88].

Outlook

The MHC is the most important genetic region in the human genome in relation to infection, autoimmunity and transplantation. Over the past 50 years, the study of MHC genes and gene products has resulted in important contributions to immunology, population genetics, genotyping and transplant medicine[2]. The challenge for the future will be to study the function and interaction of these genes and their products within the immune system to better understand and treat infection, autoimmunity and MHC-linked disease in general. Despite best efforts, infectious disease remains a chief contributor to global mortality, causing more than 13 million deaths each year, particularly in developing countries[89]. Autoimmune diseases are more common in industrialized countries, affecting about 4% of the population. Their aetiologies are often complex (for example, polygenic and non-Mendelian) and involve genetic, epigenetic and environmental factors[90]. From monozygotic twin studies it is known that genetic factors make the largest contribution (ranging from 50–90%); this still leaves 10–50% that is attributable to non-genetic factors, and epigenetic factors such as altered DNA methylation are among the main contenders[91]. In the immune system, DNA methylation has a prominent role in the development of haematopoietic cells, immune competence, antigen receptor repertoire, autoimmunity and other immune processes[92]. The development of MHC-specific gene and peptide[93] microarrays already allows the monitoring and characterization of cellular immune responses on the molecular level (see the Human Genome Mapping Project Resource Centre web site in the Online links box). Another promising development is the recent discovery that endogenous microRNAs (miRNAs) have important regulatory functions, including the destruction or repression of mRNA targets, highlighting exciting new possibilities for potential future treatments[94]. Although the jury is still out as to whether RNA-mediated silencing has a defensive role against virus infection in higher vertebrates, miRNAs are known to mediate antiviral defence processes in plants and animals[95,96]. Much progress has also been made on the computational side in the development of immune-specific databases (for example, see the International ImMunoGeneTics Information System web site) and programs using neural networks and artificial intelligence. In fact, a new field termed 'immunoinformatics' has recently been created with the ultimate aim to create a computer-generated virtual immune system[97]. All these and future efforts stand to benefit from definitive genetic maps such as the gene map of the xMHC presented here.

1. Gorer, P. A. The detection of a hereditary antigenic difference in the blood of mice by means of human group A serum. *J. Genet.* **32**, 17–31 (1936).
2. Klein, J. Seeds of time: fifty years ago Peter A. Gorer discovered the H-2 complex. *Immunogenetics* **24**, 331–338 (1986).
3. Klein, J. *Natural History of the Major Histocompatibility Complex* (Wiley & Sons, New York, 1986).
4. Campbell, R. D. & Trowsdale, J. Map of the human MHC. *Immunol. Today* **14**, 349–352 (1993).
5. The MHC sequencing consortium. Complete sequence and gene map of a human major histocompatibility complex. *Nature* **401**, 921–923 (1999).
Describes the first sequence-based gene map of a (mosaic) human MHC.
6. Malfroy, L. *et al.* Heterogeneity in rates of recombination in the 6-Mb region telomeric to the human major histocompatibility complex. *Genomics* **43**, 226–231 (1997).
This paper describes the first evidence (linkage disequilibrium) for the existence of an extended MHC.
7. Yoshino, M. *et al.* Genomic evolution of the distal Mhc class I region on mouse Chr 17. *Hereditas* **127**, 141–148 (1997).
8. Stephens, R. *et al.* Gene organisation, sequence variation and isochore structure at the centromeric boundary of the human MHC. *J. Mol. Biol.* **291**, 789–799 (1999).
9. Mungall, A. J. *et al.* The DNA sequence and analysis of human chromosome 6. *Nature* **425**, 805–811 (2003).

10. Stewart, C. A. *et al.* Complete MHC haplotype sequencing for common disease gene mapping. *Genome Res.* **14**, 1176–1187 (2004).
The authors describe the sequences in the MHC from single haplotypes.
11. Kumanovics, A., Takada, T. & Lindahl, K. F. Genomic organization of the mammalian MHC. *Annu. Rev. Immunol.* **21**, 629–657 (2003).
12. Freemont, P. S., Hanson, I. M. & Trowsdale, J. A novel cysteine-rich sequence motif. *Cell* **64**, 483–484 (1991).
13. Wain, H. M., Lush, M. J., Ducluzeau, F., Khodiyar, V. K. & Povey, S. Genew: the human gene nomenclature database, 2004 updates. *Nucleic Acids Res.* **32**, D255–257 (2004).
14. Robinson, J. *et al.* IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. *Nucleic Acids Res.* **31**, 311–314 (2003).
15. Ehlers, A. *et al.* MHC-linked olfactory receptor loci exhibit polymorphism and contribute to extended HLA/OR-haplotypes. *Genome Res.* **10**, 1968–1978 (2000).
16. Xie, T. *et al.* Analysis of the gene-dense major histocompatibility complex class III region and its comparison to mouse. *Genome Res.* **13**, 2621–2636 (2003).
17. Eddy, S. R. Non-coding RNA genes and the modern RNA world. *Nature Rev. Genet.* **2**, 919–929 (2001).
18. Yang, Z., Mendoza, A. R., Welch, T. R., Zipf, W. B. & Yu, C. Y. Modular variations of the human major histocompatibility complex class III genes for serine/threonine kinase RP, complement component C4, steroid 21-hydroxylase CYP21, and tenascin TNX (the RCCX module). A mechanism for gene deletions and disease associations. *J. Biol. Chem.* **274**, 12147–12156 (1999).
19. Bergstrom, T. F. *et al.* Phylogenetic history of hominoid DRB loci and alleles inferred from intron sequences. *Immunol. Rev.* **167**, 351–365 (1999).
20. Kulski, J. K. & Dawkins, R. L. The P5 multicopy gene family in the MHC is related in sequence to human endogenous retroviruses HERV-L and HERV-16. *Immunogenetics* **49**, 404–412 (1999).
21. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
This paper shows that duplication is a common feature of the human genome and that it is not restricted to known (for example, immune) multigene families.
22. Gu, X., Wang, Y. & Gu, J. Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nature Genet.* **31**, 205–209 (2002).
23. Trowsdale, J. The gentle art of gene arrangement: the meaning of gene clusters. *Genome Biol.* **3**, COMMENT2002 (2002).
24. Paule, M. R. & White, R. J. Survey and summary: transcription by RNA polymerases I and III. *Nucleic Acids Res.* **28**, 1283–1298 (2000).
25. Marzluff, W. F., Gongidi, P., Woods, K. R., Jin, J. & Maltais, L. J. The human and mouse replication-dependent histone genes. *Genomics* **80**, 487–498 (2002).
26. Zhang, E. Y., Knipp, G. T., Ekins, S. & Swaan, P. W. Structural biology and function of solute transporters: implications for identifying and designing substrates. *Drug Metab. Rev.* **34**, 709–750 (2002).
27. Ruddy, D. A. *et al.* A 1.1-Mb transcript map of the hereditary hemochromatosis locus. *Genome Res.* **7**, 441–456 (1997).
28. Shibui, A. *et al.* Isolation and chromosomal mapping of a novel human gene showing homology to Na+/PO4 cotransporter. *J. Hum. Genet.* **44**, 190–192 (1999).
29. Radosavljevic, M. & Bahram, S. *In vivo* immunogenetics: from *MIC* to *RAET1* loci. *Immunogenetics* **55**, 1–9 (2003).
30. Bahram, S. *MIC* genes: from genetics to biology. *Adv. Immunol.* **76**, 1–60 (2000).
31. Hopper, A. K. & Phizicky, E. M. tRNA transfers to the limelight. *Genes Dev.* **17**, 162–180 (2003).
32. Buckland, R. A., Maule, J. C. & Sealey, P. G. A cluster of transfer RNA genes (*TRM1*, *TRR3*, and *TRAN*) on the short arm of human chromosome 6. *Genomics* **35**, 164–171 (1996).
33. Rhodes, D. A., Stammers, M., Malcherek, G., Beck, S. & Trowsdale, J. The cluster of *BTN* genes in the extended major histocompatibility complex. *Genomics* **71**, 351–362 (2001).
34. Stammers, M., Rowen, L., Rhodes, D., Trowsdale, J. & Beck, S. BTL-II: a polymorphic locus with homology to the butyrophilin gene family, located at the border of the major histocompatibility complex class II and class III regions in human and mouse. *Immunogenetics* **51**, 373–382 (2000).
35. Jack, L. J. & Mather, I. H. Cloning and analysis of cDNA encoding bovine butyrophilin, an apical glycoprotein expressed in mammary tissue and secreted in association with the milk-fat globule membrane during lactation. *J. Biol. Chem.* **265**, 14481–14486 (1990).
36. Giorgi, D., Friedman, C., Trask, B. J. & Rouquier, S. Characterization of nonfunctional V1R-like pheromone receptor sequences in human. *Genome Res.* **10**, 1979–1985 (2000).
37. Volz, A. *et al.* Complex transcription and splicing of odorant receptor genes. *J. Biol. Chem.* **278**, 19691–19701 (2003).
38. Ziegler, A., Dohr, G. & Uchanska-Ziegler, B. Possible roles for products of polymorphic MHC and linked olfactory receptor genes during selection processes in reproduction. *Am. J. Reprod. Immunol.* **48**, 34–42 (2002).
39. Coleman, J. E. Zinc proteins: enzymes, storage proteins, transcription factors, and replication proteins. *Annu. Rev. Biochem.* **61**, 897–946 (1992).
40. Lee, P. L. *et al.* Three genes encoding zinc finger proteins on human chromosome 6p21.3: members of a new subclass of the Kruppel gene family containing the conserved SCAN box domain. *Genomics* **43**, 191–201 (1997).
41. Meyer, M., Gaudieri, S., Rhodes, D. A. & Trowsdale, J. Cluster of *TRIM* genes in the human MHC class I region sharing the B30.2 domain. *Tissue Antigens* **61**, 63–71 (2003).
42. Matthews, J. M. & Sunde, M. Zinc fingers — folds for many occasions. *IUBMB Life* **54**, 351–355 (2002).
43. Gruss, H. J. & Dower, S. K. The TNF ligand superfamily and its relevance for human diseases. *Cytokines Mol. Ther.* **1**, 75–105 (1995).
44. Mallya, M., Campbell, R. D. & Aguado, B. Transcriptional analysis of a novel cluster of LY-6 family members in the human and mouse major histocompatibility complex: five genes with many splice forms. *Genomics* **80**, 113–123 (2002).
45. Milner, C. M. & Campbell, R. D. Structure and expression of the three MHC-linked *HSP70* genes. *Immunogenetics* **32**, 242–251 (1990).
46. Gleimer, M. & Parham, P. Stress management: MHC class I and class I-like molecules as reporters of cellular stress. *Immunity* **19**, 469–477 (2003).
47. Alfonso, C. & Karlsson, L. Nonclassical MHC class II molecules. *Annu. Rev. Immunol.* **18**, 113–142 (2000).
48. Ohno, S. *Evolution by Gene Duplication* (Springer, New York, 1970).
49. Mazet, F. & Shimeld, S. M. Gene duplication and divergence in the early evolution of vertebrates. *Curr. Opin. Genet. Dev.* **12**, 393–396 (2002).
50. Flajnik, M. F. & Kasahara, M. Comparative genomics of the MHC: glimpses into the evolution of the adaptive immune system. *Immunity* **15**, 351–362 (2001).
51. Sidow, A. Gen(om)e duplications in the evolution of early vertebrates. *Curr. Opin. Genet. Dev.* **6**, 715–722 (1996).
52. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
53. Hughes, A. L. Phylogenetic tests of the hypothesis of block duplication of homologous genes on human chromosomes 6, 9, and 1. *Mol. Biol. Evol.* **15**, 854–870 (1998).
54. Abi-Rached, L., Gilles, A., Shiina, T., Pontarotti, P. & Inoko, H. Evidence of *en bloc* duplication in vertebrate genomes. *Nature Genet.* **31**, 100–105 (2002).
The latest hypothesis to explain the MHC paralogy.
55. Rand, V. Genome evolution: a study of MHC paralogous genes in the human genome. Ph.D. Thesis, University of Cambridge, UK (2003). Access through http://www.sanger.ac.uk/Info/theses/
56. Kaufman, J. *et al.* The chicken B locus is a minimal essential major histocompatibility complex. *Nature* **401**, 923–925 (1999).
57. Rogers, S. & Kaufman, J. (personal communication).
58. Teng, M. S. *et al.* A human *TAPBP* (TAPASIN)-related gene, *TAPBP-R*. *Eur. J. Immunol.* **32**, 1059–1068 (2002).
59. Wagner, A. Selection and gene duplication: a view from the genome. *Genome Biol.* **3**, R1012 (2002).
60. Clamp, M. *et al.* Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.* **31**, 38–42 (2003).
61. Marsh, S. G. E., Parham, P. & Barber, L. D. *The HLA Factsbook* (Academic Press, San Diego, California, 2000).
62. Ahmad, T. *et al.* Haplotype-specific linkage disequilibrium patterns define the genetic topography of the human MHC. *Hum. Mol. Genet.* **12**, 647–656 (2003).
63. Walsh, E. C. *et al.* An integrated haplotype map of the human major histocompatibility complex. *Am. J. Hum. Genet.* **73**, 580–590 (2003).
64. Stenzel, A. *et al.* Patterns of linkage disequilibrium in the MHC region on human chromosome 6p. *Hum. Genet.* **114**, 377–385 (2004).
65. Trowsdale, J. & Parham, P. Mini-review: defense strategies and immunity-related genes. *Eur. J. Immunol.* **34**, 7–17 (2004).
66. Zdobnov, E. M. *et al.* Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science* **298**, 149–159 (2002).
67. Christophides, G. K. *et al.* Immunity-related genes and gene families in *Anopheles gambiae*. *Science* **298**, 159–165 (2002).
68. Lazarus, R. *et al.* Single nucleotide polymorphisms in innate immunity genes: abundant variation and potential role in complex human disease. *Immunol. Rev.* **190**, 9–25 (2002).
69. Hewitt, E. W. The MHC class I antigen presentation pathway: strategies for viral immune evasion. *Immunology* **110**, 163–169 (2003).
70. Vivier, E., Tomasello, E. & Paul, P. Lymphocyte activation via NKG2D: towards a new paradigm in immune recognition? *Curr. Opin. Immunol.* **14**, 306–311 (2002).
71. Lechler, R. & Warrens, A. *HLA in Health and Disease* (Academic Press, London, 2000).
72. Feder, J. N. *et al.* A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nature Genet.* **13**, 399–408 (1996).
The authors describe a novel role for an MHC class I molecule. This is a landmark paper that illustrates the problems posed by linkage disequilibrium in disease mapping in the MHC.
73. von Kempis, J. Arthropathy in hereditary hemochromatosis. *Curr. Opin. Rheumatol.* **13**, 80–83 (2001).
74. Rubio, J. P. *et al.* Extended haplotype analysis in the HLA complex reveals an increased frequency of the HFE-C282Y mutation in individuals with multiple sclerosis. *Hum. Genet.* **114**, 573–580 (2004).
75. Hellerbrand, C., Poppl, A., Hartmann, A., Scholmerich, J. & Lock, G. HFE C282Y heterozygosity in hepatocellular carcinoma: evidence for an increased prevalence. *Clin. Gastroenterol. Hepatol.* **1**, 279–284 (2003).
76. Pal, D. K. *et al.* *BRD2* (*RING3*) is a probable major susceptibility gene for common juvenile myoclonic epilepsy. *Am. J. Hum. Genet.* **73**, 261–270 (2003).
Identification of a promoter mutation within the MHC that is strongly linked to a common form of epilepsy.
77. Okamoto, K. *et al.* Identification of I_{κ}BL as the second major histocompatibility complex-linked susceptibility locus for rheumatoid arthritis. *Am. J. Hum. Genet.* **72**, 303–312 (2003).
78. Schmidt, A. M., Yan, S. D., Yan, S. F. & Stern, D. M. The biology of the receptor for advanced glycation end products and its ligands. *Biochim. Biophys. Acta* **1498**, 99–111 (2000).
79. Hofmann, M. A. *et al.* RAGE and arthritis: the G82S polymorphism amplifies the inflammatory response. *Genes Immun.* **3**, 123–135 (2002).
80. Bjorkman, P. J. *et al.* Structure of the human class I histocompatibility antigen, *HLA-A2*. *Nature* **329**, 506–512 (1987).
81. Brown, J. H. *et al.* Three-dimensional structure of the human class II histocompatibility antigen *HLA-DR1*. *Nature* **364**, 33–39 (1993).
82. Siebold, C. *et al.* Crystal structure of *HLA-DQ0602* that protects against type 1 diabetes and confers strong susceptibility to narcolepsy. *Proc. Natl Acad. Sci. USA* **101**, 1999–2004 (2004).
83. Hulsmeyer, M. *et al.* Dual, *HLA-B27* subtype-dependent conformation of a self-peptide. *J. Exp. Med.* **199**, 271–281 (2004).
84. Gao, X. *et al.* Effect of a single amino acid change in MHC class I molecules on the rate of progression to AIDS. *N. Engl. J. Med.* **344**, 1668-1675 (2001).
Along with reference 83, this paper shows how a single amino-acid change can influence structural and functional properties of *HLA-B27* antigens, and in *HLA-B35* molecules is associated with rate of progression from HIV infection to AIDS.
85. Ferrara, G. B. *et al.* Bone marrow transplantation from unrelated donors: the impact of mismatches with substitutions at position 116 of the human leukocyte antigen class I heavy chain. *Blood* **98**, 3150–3155 (2001).
86. Doxiadis, I. I. *et al.* Association between specific HLA combinations and probability of kidney allograft loss: the taboo concept. *Lancet* **348**, 850–853 (1996).
87. Chen, T. C., Waldmann, H. & Fairchild, P. J. Induction of dominant transplantation tolerance by an altered peptide ligand of the male antigen Dby. *J. Clin. Invest.* **113**, 1754–1762 (2004).
This paper shows that the balance of destructive and protective T cells in a transplant setting can be changed by using altered peptides for presentation by MHC molecules
88. Hansen, J. A. & Dupont, B. in *Proceedings of the 13th International Histocompatibility Workshop and Congress* (IHWG Press, Seattle, in the press).
89. Cohen, M. L. Changing patterns of infectious disease. *Nature* **406**, 762–767 (2000).
90. Vyse, T. J. & Todd, J. A. Genetic analysis of autoimmune disease. *Cell* **85**, 311–318 (1996).

91. Novik, K. L. *et al.* Epigenomics: genome-wide study of methylation phenomena. *Curr. Issues Mol. Biol.* **4**, 111–128 (2002).
92. Teitell, M. & Richardson, B. DNA methylation in the immune system. *Clin. Immunol.* **109**, 2–5 (2003).
93. Soen, Y., Chen, D. S., Kraft, D. L., Davis, M. M. & Brown, P. O. Detection and characterizationof cellular immune responses using peptide-MHC microarrays. *PLoS Biol.* **1**, E65 (2003).
94. Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297 (2004).
95. Vance, V. & Vaucheret, H. RNA silencing in plants — defense and counterdefense. *Science* **292**, 2277–2280 (2001).
96. Lecellier, C. H. & Voinnet, O. RNA silencing: no mercy for viruses? *Immunol. Rev.* **198**, 285–303 (2004).
97. Bock, G. & Goode, J. *Immunoinformatics — Strategies for Better Understanding of Immune Function* (Novartis Foundation, Chichester, UK, 2003).
98. Chabas, D., Taheri, S., Renier, C. & Mignot, E. The genetics of narcolepsy. *Annu. Rev. Genomics Hum. Genet.* **4**, 459–483 (2003).
99. Sollid, L. M. Coeliac disease: dissecting a complex inflammatory disorder. *Nature Rev. Immunol.* **2**, 647–655 (2002).
100. Brown, N. P., Whittaker, A. J., Newell, W. R., Rawlings, C. J. & Beck, S. Identification and analysis of multigene families by comparison of exon fingerprints. *J. Mol. Biol.* **249**, 342–359 (1995).

Acknowledgements

We wish to thank S. Rogers and J. Kaufman for sharing prepublication data. The work of the HUGO Gene Nomenclature Committee is supported by the US National Institutes of Health, the UK Medical Research Council and the Wellcome Trust.

Competing interests statement

The authors declare no competing financial interests.

Online links

DATABASES

The following terms in this article are linked online to:

Entrez: http://www.ncbi.nih.gov/Entrez/

HFE | *MICA* | *MICB* | *B2M* | *TRIM* | *TNF* | *LTA* | *LTB*

OMIM: http://www.ncbi.nlm.nih.gov/Omim/

Ankylosing spondylitis

FURTHER INFORMATION

Anthony Nolan Research Institute: http://www.anthonynolan.org.uk/research/

Beck's laboratory: http://www.sanger.ac.uk/Team/Teams50/

ENSEMBL: http://www.ensembl.org

HUGO Gene Nomenclature Committee (HGNC): http://www.gene.ucl.ac.uk/nomenclature/

Human Annotation Workshop (HAWK): http://www.sanger.ac.uk/HGP/havana/hawk.shtml

Human Genome Mapping Project resource centre: http://www.hgmp.mrc.ac.uk/Research/Microarray/HGMP-RC_Microarrays/description_of_arrays.jsp

IMGT/HLA Sequence Database: http://www.ebi.ac.uk/imgt/hla/

International Histocompatiblity Working Group: http://www.ihwg.org/

International ImMunoGeneTics Information System: http://imgt.cines.fr

MHC Haplotype Project: http://www.sanger.ac.uk/HGP/Chr6/MHC/

MHCPEP database of MHC-binding peptides: http://wehih.wehi.edu.au/mhcpep/

NCBI MHC database: http://www.ncbi.nlm.nih.gov/MHC

NCBI Single Nucleotide Polymorphism database: http://www.ncbi.nlm.nih.gov/SNP

Poster of the xMHC map: http://www.nature.com/nrg/journal/v5/n12/poster/MHCmap

Trowsdale's laboratory: http://www.path.cam.ac.uk/immuno/resint.html

UCSC Genome Browser: http://genome.ucsc.edu/

VEGA: http://vega.sanger.ac.uk

Ziegler's laboratory: http://www.charite.de/immungenetik

Access to this interactive links box is free online.