PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Gene Clusters and the Evolution of the Major Histocompatibility System

W. F. Bodmer, J. Trowsdale, J. Young and Julia Bodmer

Phil. Trans. R. Soc. Lond. B 1986 **312**, 303-315
doi: 10.1098/rstb.1986.0009

Phil. Trans. R. Soc. Lond. B **312**, 303–315 (1986)
Printed in Great Britain

Gene clusters and the evolution of the major histocompatibility system

By W. F. Bodmer, F.R.S., J. Trowsdale, J. Young and Julia Bodmer
Imperial Cancer Research Fund Laboratories, P.O. Box 123, 44 *Lincoln's Inn Fields, London WC2A* 3*PX, UK*

Gene clusters containing one or more sets of duplicated genes with related functions are probably the basic genetic functional units. The major histocompatibility systems, such as HLA and H2, are among the most complex gene clusters so far known and studied, and illustrate many of the features of their structure and evolution. They cover about one thousandth of the mammalian genome and include two major sets of cell surface products with different but related functions in the control of immune interactions, as well as genes for complement components and 21-hydroxylase. Molecular analysis has revealed an extraordinary complexity at the genetic level, reflecting a very long and involved evolutionary history. A description of the organization of the HLA system, especially the HLA-D region, and its function and polymorphism forms the basis for considering the evolution of such complex gene clusters.

Introduction: evolution of gene clusters

Gene clusters are sets of closely linked genes, usually evolutionarily related and with similar or interacting functions. The simple theory, which derives from the original description of duplications in *Drosophila*, is that once a gene has been duplicated the copies can diverge from the original to produce different, related or new functions. Close linkage may favour mutually related control of gene activity. A cluster may, therefore, be maintained by the interaction between selection and linkage, as originally postulated by R. A. Fisher in 1930, in the sense that a genotype in which the cluster was disrupted would be selected against. Fisher's theory also, of course, allowed for the possibility that genes that interact favourably in their function and are not closely linked could be brought together. Thus, a genotype harbouring a translocation, inversion or transposition which brought together favourably interacting genes would be selected for. Classical examples of gene clusters included the *Primula* incompatibility and human Rh blood group systems. But the first well established example of a gene cluster at the molecular level was that for the haemoglobin β chains. In this paper we shall first discuss some general ideas concerning the evolution of gene clusters, and then illustrate them by examples from the major histocompatibility system, in particular the human HLA system. For further background and references see Bodmer (1983).

Molecular studies show that a wide variety of genetic functions in higher eukaryotic organisms are controlled by gene clusters and suggest that the gene cluster is a basic genetic functional unit. Clusters range in size from the relatively simple haemoglobin systems, containing no more than six genes, to the much more complex immunoglobulin and major histocompatibility systems, which include up to a hundred or more genes. The immunoglobulins and the major histocompatibility system are part of a large family of evolutionarily related gene products with functions associated with the immune system. This family now includes the T-cell receptor, with at least three different gene clusters, the heavy and two light chain gene clusters

of immunoglobulins, $\beta_2$microglobulin, and Thy1 and certain other lymphocyte surface molecules (Hood *et al.* 1985). Some sets of products, such as the collagens and actins, are controlled by related genes that are not clustered, but widely dispersed throughout the genome. These differences must clearly be related to the strategy for control of expression of the products at different times, and in different tissues, and may also be associated with the frequency of variant production by mechanism such as non-homologous intrachromosomal gene conversion.

Gene clustering in prokaryotes is generally quite different (see, for example, Bodmer 1970). The transition from prokaryotic to eukaryotic organization was presumably accompanied by the development of a nuclear membrane, RNA processing and the intron–exon organization of genes. This was a critical evolutionary transition which favoured the evolution of gene clusters as now seen in higher eukaryotes. Within the eukaryotes there is evidence to suggest that the complexity of gene clustering increases with the complexity of the organism, certainly in the transition from single cell to multicellular eukaryotes, and probably again from invertebrates to vertebrates. The evolution of the intron–exon organization enormously increased the flexibility of the evolutionary process, but at the potential price of increased amounts of DNA. RNA processing could well have originated from the processing of ribosomal and transfer RNA. In this context, it is important to realise that transfer RNA itself has a three-dimensional structure which gives it a function, and that the RNA sequence derived from an intron, at least in a yeast tRNA gene, can function as an RNA splicing enzyme (see, for example, Flavell 1985). Perhaps this is a reflection of the fact that RNA may have been the original key molecule in prebiotic evolution, since it could serve both an informational and a functional role (Eigen 1983; Woese 1983). Now, we see the vestige of this in residual functions carried out by RNA sequences derived from introns and controlling regions, sometimes, perhaps, only inside the nucleus. This emphasizes the need to be cautious before dismissing the functional significance of changes in these parts of a gene. It may also be important to consider the significance of such sequences in terms of the structures they could give rise to, rather than their direct sequential informational content. These sequences adjacent to exons should also answer questions concerning the strategy for the control of expression within and between clusters of functionally related genes.

Genes controlling functions common to prokaryotes and eukaryotes may have an intron–exon organization less closely related to protein domains and so, overall, simpler genomic organization (Bodmer 1983). It is perhaps especially among comparatively recent gene systems, such as immunoglobulins and the major histocompatibility system of eukaryotes, that one should look for, and expect, the evolution of more complex gene clusters. These may have evolved over a very long period of time by a wide variety of mechanisms, including duplication, inversion, deletion, conversion, transposition and frame shifts, coupled with differential splicing of transcripts. This variety of mechanisms can yield a most complex organization within a gene cluster, that has to some extent an evolution of its own, almost as if it were a micro-organism within the overall genome. Diversification of a basic gene product can increase its flexibility and lead to more refined functional adaptation. A good example is the distinction between embryonic, foetal and adult haemoglobins, each adjusted to the needs of an oxygen carrier in the blood at different stages of life. In addition, by mechanisms such as gene conversion, transposition and double unequal intrachromosal crossing over, which can transfer sequences from one part of a gene cluster to another, increasing complexity of a gene cluster also increases the opportunity for producing functionally effective variants. Inserting a sequence from one gene into the equivalent position of another closely related gene is more likely to yield a

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

functionally effective variant than random mutation, since the transposed sequence has already been functionally tested in its original environment.

Many of these features of gene clusters and their evolution are illustrated by the major histocompatibility system, which is one of the most complex gene clusters so far described.

The HLA system

The HLA system is the major histocompatibility system of man, which has its counterparts in other species of mammals, birds and reptiles, including in particular the much-studied H2 system of the mouse. In mammals it encompasses about one thousandth of the genome, corresponding, probably, to about 3 million nucleotide pairs, and it codes for at least five known sets of gene products. The first are the HLA-ABC, or class I, determinants found on most nucleated cells, which are 43000 Da molecular mass glycoproteins associated with another protein, $\beta_2$microglobulin (β2m), which is coded for on a different chromosome. Second are the HLA-D or class II products, which are composed of two glycoprotein chains both coded for in the HLA region. These two sets of cell surface products mediate functional interactions between the cells of the immune system, and so play a major role in the control of the immune response. Two further sets of products in the HLA region are part of the cascade of complement components which also have important effects or functions in the immune system. C2 and C4 are the second and fourth components of the classical complement pathway, and factor B, closely related to C2, functions in the alternative pathway. There are two *C4* genes in the region and, immediately adjacent to each of these, lie copies of the fifth product, the *21-hydroxylase* gene, which, when deficient, causes congenital adrenal hyperplasia.

Figure 1 shows a schematic comparison of the maps of the mouse H2 and human HLA genetic regions. The H2 region sequence could have been derived from that for HLA by transposition

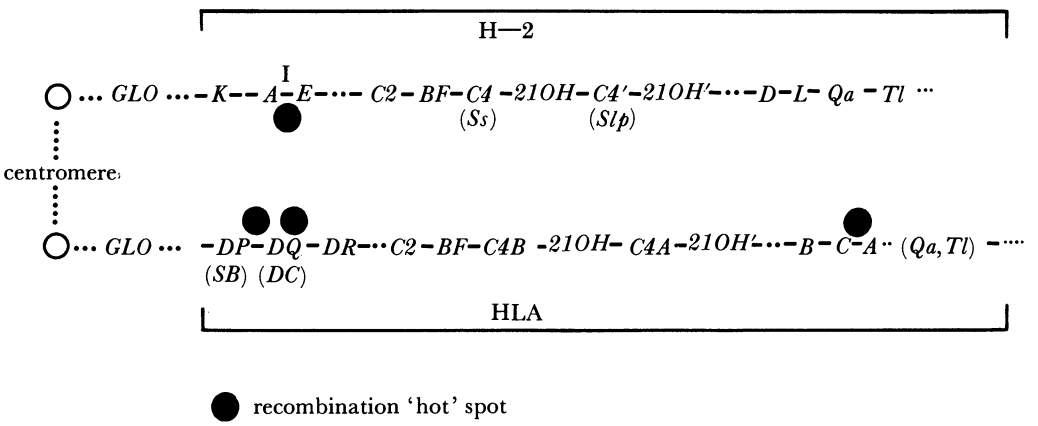

Figure 1. Schematic genetic maps of the HLA and H2 regions. The regions are aligned to give the maximum correspondence of the sequences, leaving only H2K out of line with the HLA sequence. *H2KDL*, *Qa* and *Tl*, and *HLA-A*, *B*, *C* are the class I determinants. *H2 I* regions *A* and *E* correspond, respectively, to *HLA-DQ* and *DR*. *C2* and *C4* are the second and fourth components respectively of the classical complement system, and *BF* is factor B closely related to C2 of the alternative pathway. *21-OH* stands for the 21-hydroxylase gene deficient in congenital adrenal hyperplasia. *GLO* stands for the gene for glyoxalase enzyme, linked to both H2 and HLA. *Ss* and *Slp* are the original names for the two *C4* genes in the H2 region. *SB* and *DC* are the former names for *HLA-DP* and *DQ* respectively. The basis for the assignment of the recombinational hot spots is discussed in the text.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

of the mouse equivalent of HLA-A, namely H2K, by, for example, an intrachromosomal double unequal cross over (Bodmer 1981). In all other respects the sequence of the genes in the two species is remarkably similar. The class I region of H2, and presumably also HLA, contains two sets of genes called *Qa* and *TL*, whose products, like HLA-ABC, are β2m associated but which have a much more restricted distribution, mainly on certain types of lymphocytes, and whose function is completely unknown. The work of Hood and others (see Hood *et al.* 1985) has shown that there may be anywhere from 25 to 35 genes in the class I region, all with a similar overall genomic organization, falling into the three subsets, *KDL*, *Qa* and *TL*.

Sequence comparisons within and between class I and class II genes clearly show them to have a common, if in some cases very distant, evolutionary origin. There is, however, no discernable evolutionary relationship between these sets of genes and the complement and *21 hydroxylase* (*21-OH*) genes. Neither is there any obvious relationship between the three sets of products C2/BF, C4 and 21-OH. C4 is thought to be evolutionarily related to C3, the third complement component, which maps not to the HLA region, but to chromosome 19. This suggests that the *C4* gene was transposed into the HLA region after the evolutionary establishment of HLA class I and class II products and complement. Presumably this transposition gave rise to a selective advantage for the association between *C4* and the pre-existing HLA region products. This may be an example of the mechanism that Fisher originally had in mind in suggesting how gene clusters could evolve because of the interaction between selection and linkage. Perhaps the *21-hydroxylase* gene was simply brought in as a hitch-hiking neighbour to the evolutionary precursor of the *C4* gene. It would be interesting, therefore, to know whether there are any P450 mixed function oxidases similar to 21-OH next to, for example, *C3* on chromosome 19. C4 and C2 are associated in their complement activities and this, perhaps, accounts for the juxtaposition of their genes. The similarity of the overall arrangement of the H2 and HLA systems indicates that their evolution pre-dates the major evolutionary divergence of the mammals, some 150 to 200 million years ago. But there is as yet no information as to when, before this, the events placing the complement and *21-OH* genes within the major histocompatibility regions might have taken place.

Linkage disequilibrium and recombinational hot spots

Linkage disequilibrium, which is the population association between alleles at different loci on a given haplotype and which leads to a phenotypic association, is commonly observed between alleles at different loci of the HLA system. (see, for example, Bodmer & Bodmer 1978.) The association between alleles on a given haplotype, measured by the linkage disequilibrium coefficient, decreases at a rate of $1 - r$ per generation, where r is the recombination fraction between the loci. Linkage disequilibrium between alleles at different loci is therefore very much dependent on the magnitude of the recombination fraction between the loci. When this is very small, associations are maintained for a long time and no special mechanisms other than close linkage are then needed to account for linkage disequilibrium. This is, for example, the probable explanation for the extensive linkage disequilibrium between alleles of the *HLA-B* and *C* loci, which is similar among all major human racial groups. There are, on the other hand, only a relatively small number of highly significant population associations between alleles of the *HLA-A* and *B* loci. Most notably, there is strong linkage disequilibrium between *A1* and *B8* and the pattern of distribution of this combination in European populations has suggested that in this

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

case the association is probably maintained by some form of selective interaction. The difference between *A* and *B* locus associations, compared with those between *B* and *C*, simply reflects the fact that the recombination fraction between *HLA-A* and *C* is at least four or five times, and possibly more, than between *HLA-C* and *B*. This initially suggested that there may be other loci between *HLA-A* and *C*, not yet identified. But the molecular data on class I antigens in the mouse suggests that this is unlikely, and that it is more probable that the relatively high recombination fraction between *HLA-A* and *C* reflects the existence of a 'recombinational hot spot' between these two loci. Such recombinational hot spots have been observed in the human β-globin gene cluster (Chakravarti *et al.* 1984) and, more directly by Steinmetz *et al.* (1982) and Kobori *et al.* (1984), who showed that all the recombinants that have been observed in that part of the mouse H2-I region, which has so far been cloned, could be localized to a 1.7 kilobase interval out of a region that is at least 150–200 kilobases long. Patterns of linkage disequilibrium between alleles of genes within a gene cluster can clearly identify the positions of putative recombinational hot spots. On this basis, because of the relatively loose linkage and weak linkage disequilibrium between alleles of the *HLA-DP* subregion and those for *HLA-DR* and *DQ*, and also because of weak linkage disequilibrium between alleles at *HLA-DX* and *HLA-DQ*, two putative recombinational hot spots can be identified within the *HLA-D* region as indicated in figure 2.

It seems likely that recombination, at least in mammals, is mainly localized to recombinational hot spots. In the human genome, for example, if these gave rise, on average, to recombination fractions of 0.5%, then there would be a total of approximately 5000 recombinational hot spots

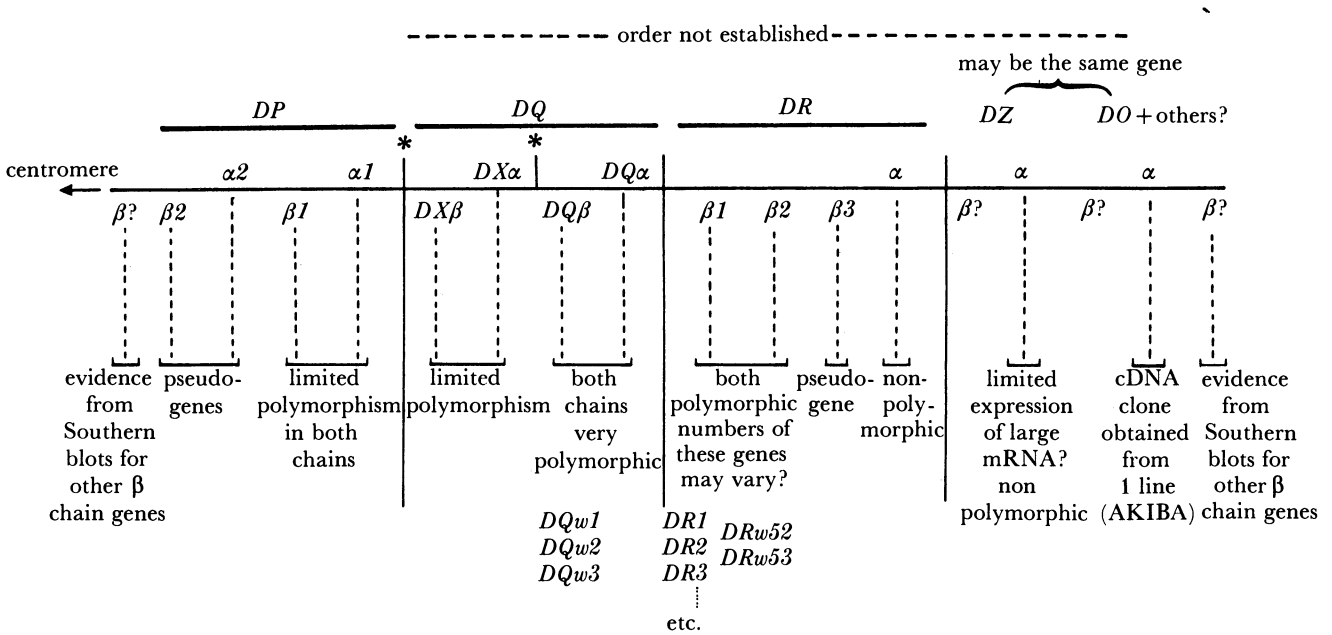

FIGURE 2. Schematic map of the *HLA-D* region (based on Bodmer (1984) and Trowsdale *et al.* (1985). α genes are placed above the line and β genes below them. The order is based on a combination of data from deletion mutants and patterns of linkage disequilibrium. The asterisks indicate more precisely the postulated positions of recombinational hot spots between *DP* and *DQ* and within the *DQ* subregion. *DZ* (and *DO*) may actually lie between the *DP* and *DQ* subregions. The question marks after β genes indicate the possible existence of as yet undetected β genes associated with *DZ* (and *DO*), and at least in some individuals an extra β gene in the DP subregion.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

in the genome overall. This would imply up to half a dozen within the HLA region. Recombination outside these hot spots may occur on average at least 5–10 times less frequently than within them. Alleles at loci not separated by such a recombinational hot spot would usually be expected to be in strong linkage disequilibrium, whereas those at loci separated by one or more hot spots would, in general, not show extreme linkage disequilibrium. Recombination fractions over distances that are long relative to the interval between recombinational hot spots would on average give a reasonable estimate of the physical distance between marker genes. But once one is looking at genes within a cluster at intervals which are not large compared with the average interval between recombinational hot spots, then the correlation between physical and recombination distance breaks down. These considerations have a major influence on the interpretation of patterns of population associations between alleles of the genes within a gene cluster such as the HLA system. The remainder of our discussion on the HLA system will be focused on the HLA-D region.

The *HLA-D* region

The HLA-D specificities were originally identified by using cellular techniques. Subsequently, a set of serologically identified specificities, HLA-DR, was described, mainly on B lymphocytes and monocytes, that corresponded closely to the cellular specificities. Further serological and biochemical analysis, especially with monoclonal antibodies, showed that there were at least three major sets of products, now called HLA-DR, DQ and DP, each formed from two transmembrane glycopolypeptides, α (the larger in overall molecular mass), and β. Molecular analysis of the *HLA-D* region at the DNA level has now shown that there are at least six α chain genes and seven β chain genes, which can be assigned to the three major subregions, *DR*, *DQ* and *DP*, and a newer, less well characterized, subregion, provisionally called *DZ*. A schematic summary of the current view of the genetic structure of the *HLA-D* region is shown in figure 2. The *DP* and *DQ* subregions consist of associated pairs of α and β genes, while in *DR* there appears to be just one α gene associated with up to three or more β genes. The situation for *DZ* is not yet clear. At least one α gene has been defined and there is now provisional evidence (E. Long, personal communication) for at least one unassigned β gene, *DQβ*, which could be the *DZα* partner. (For an up-to-date description of the HLA system, including the HLA-D region, see Albert 1984. The data on the HLA-D region discussed here are based on Trowsdale *et al.* (1985) and Trowsdale & Kelly (1985).)

Serological and biochemical data indicated that a major part of the polymorphism for the *HLA-D* region was associated with two of the *DRβ* genes and one of the *DQβ* genes, while limited polymorphism was seen for *DQα* by using these techniques. Restriction enzyme polymorphism studies, and direct DNA sequencing, have now shown that two of the *DRβ* genes, as well as *DQα* and *β* are very highly polymorphic, *DXα*, *DXβ* and *DPα1* and *DPβ1* show limited polymorphism, while the remaining genes show a very limited level of polymorphism, especially in contrast to *DRβ* and *DQ*.

The genomic organization of some of the *D* region β and α genes is shown in figure 3. As for the immunoglobulins and HLA class I products, there is a remarkable concordance between exons and postulated protein domains. The β genes tend to have larger introns and show more subdivision of the connecting peptide, transmembrane and cytoplasmic regions. Intron–exon organization is highly conserved for all the genes, and there is a clear pattern of similarity

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

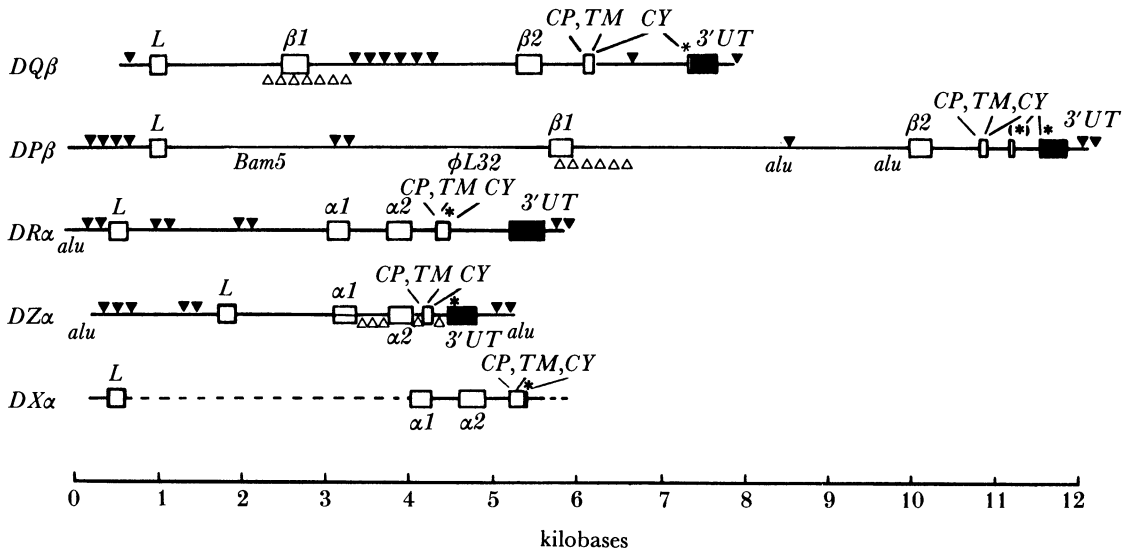

FIGURE 3. Exon–intron organization of human *HLA-D* region α and β genes. *β1*, *β2*, *α1* and *α2* refer to the exons corresponding to the external domains of the protein. *CP*, *TM* and *CY* stand, respectively, for connecting peptide, transmembrane and cytoplasmic regions. *L* for leader sequence and *3′UT* for the 3′ untranslated region. *Bam* 5, *φL32* and *alu* identify different human repetitive sequences, *φL32* being a pseudogene for the ribosomal L32 protein. Closed triangles denote regions rich in AT and open triangles rich in GC (from Trowsdale *et al.* 1985).

between the α chain, and between the β chain gene organizations. The highly polymorphic regions are predominantly associated with certain subsegments of the *α1* (for *DQ*) and *β1* (for *DR*, *DQ* and to a lesser extent *DP*) exons. The *DP β* chain introns, interestingly, contain presumably non-functional transposed gene sequences, as well as examples of the *alu* repeat family. The *φL32* sequence, for example, is the first example of a pseudogene version of a ribosomal protein with obvious transposon-like features to be found in an intron. It seems unlikely that this serves any functional role and so its presence in the *DPβ* intron either reflects a chance increase in frequency, which will occur of course from time to time especially if such transpositions are relatively common, or might have been helped into the population by asymmetrical gene conversion, which could favour an inserted looped out sequence in a heteroduplex (Bodmer 1981).

HLA-D POLYMORPHISM

There are very characteristic patterns of association, or linkage disequilibrium, for polymorphisms detected both serologically and by restriction enzymes. Thus, *DQβ*, *DQα*, and *DRβ* variation is very highly associated in all populations studied so far. Limited data also show a high degree of association between *DXα* and *DXβ* determined DNA polymorphisms, but relatively weak association between these and *DQα*, *DQβ* and *DRβ*. Similarly, while the *DPβ* DNA polymorphisms show strong associations with the *DP* types as established by cellular techniques, they show weak or negligible associations with variations for *DQ* and *DR*. These patterns justify the placing of recombinational hot spots between the *DP* and *DQ* subregions and also within the *DQ* subregion between *DX* and *DQ*. Recent data from the mouse (R. Flavell

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

and G. Widera, personal communication) suggest in contrast to what has been published so far, that there may be at least vestiges of a similar set of genes to *DP* and *DZ* in the mouse H2 region. Sequence data have clearly shown that *DR* corresponds to mouse *H2-IE*, while *DQ* corresponds to H2-*IA*.

There are a number of additional interesting features of the observed polymorphic variation, especially using restriction enzymes. The most striking feature of the patterns of polymorphism for the *HLA-D* region genes is the very high level of polymorphism for some of the genes, namely *DQα*, *DQβ* and *DRβ*, which is restricted to particular parts of these genes, and the very low level of polymorphism for the other genes. Differential variability of this sort is strong evidence for the action of natural selection, since population structure effects should affect all genes equally, and regional differences in mutation rates, which could account for these patterns, seem most implausible. The selective forces driving the polymorphism are likely to be differential susceptibility to disease associated with immune response differences. This leads to a form of frequency-dependent selection that tends to favour new alleles that give rise to protection against new, or newly varied, pathogenic organisms (Bodmer 1972). The difference in levels of polymorphism between the various genes suggests functional differentiation, namely that variation for the highly polymorphic genes must be more likely to confer differential resistance to pathogens than does variation for the less polymorphic genes.

The majority of the restriction enzyme variations lie outside the coding regions. It is really quite striking that probes from either end of a 40000 nucleotide pair cosmid *DQα* insert, will reveal polymorphism with virtually any restriction enzyme, while probes from an exactly analogous *DRα* containing cosmid of the same length reveal only very restricted polymorphism with one or two enzymes. Since most of the restriction site variation lies outside the coding regions, this suggests an overall higher degree of polymorphism surrounding the *DQα* than the *DRα* genes and, more generally, around the highly polymorphic than the minimally polymorphic genes. Similar differences in patterns of variability have been pointed out by Steinmetz *et al.* (1984). Their data specifically, and ours by implication, suggest that recombinational hot spots demarcate the boundaries of these polymorphic regions. It seems most likely that these regional differences in levels of polymorphism are associated with patterns of selection for polymorphic variants. Thus, when selection favours the increase of a new allele, say at the *DQα* locus, any closely linked differences in the DNA sequence will be pulled along into the population with the new allele so long as they confer no selective disadvantage (Bodmer & Parsons 1962). This phenomenon is sometimes called 'hitch-hiking' and is due to strong linkage disequilibrium between very closely linked alleles. Variations associated with a new *DQα* allele, on the same haplotype, could, for example, be produced by the mechanism that gave rise to the new allele, for instance by gene conversion or unequal double crossing over involving relatively extensive regions of DNA sequence. Variation could also be produced in association with a recombination event giving rise to a new allele, for example by sequence errors introduced during branch migration. Variation that happened to exist in the haplotype in which the newly selected polymorphic allele was produced will also be hitch-hiked into the population. The region of variability surrounding *DQα* and the other other highly polymorphic genes, over which associated DNA polymorphisms will be pushed into the population by the hitch-hiking effect, will clearly be bounded by recombinational hot spots, since across these linkage-disequilibrium will be markedly less than it is within the bounded region.

Average sequence differences among a few alleles for *DPα* and *DQα*, separating out the

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Table 1. Comparison of the level of polymorphism for selected *DPα* and *DQα* alleles

	nucleotides		amino acids	
	DPα1	*DQα*	*DPα1*	*DQα*
SP	2.2^{93}	3.7^{82}	3.3^{31}	8.7^{23}
α1	2.8^{252}	12.2^{246}	3.6^{84}	20.7^{87}
α2	2.1^{282}	2.8^{282}	3.2^{94}	5.3^{94}
CP, TM, CY	1.3^{156}	6.4^{155}	3.9^{51}	3.9^{51}

SP, CP, TM and CY refer, respectively, to signal peptide, connecting peptide, transmembrane and cytoplasmic regions. α1 and α2 are the two external domains of the polypeptide. The large numbers are the maximum observed percentage variation among subsets of three or four alleles. The superscript numbers give the length of the sequence in either nucleotides or amino acids for each of the respective domains (based on Trowsdale *et al.* 1985).

different exons, are given in table 1. The data show overall variation throughout each gene, but with a marked increase in the extent of sequence differences between *DQα* alleles in the *α1* exon or domain. This is consistent with the fact that most of DQα polymorphism is found in the *α1* domain. The data suggest that the background level of variation outside the *DQα1* exon is that expected from neutral substitutions and so, that these differences can be used to estimate the age of the alleles. Thus, Kimura (1983) has calculated that the average rate of substitution due to random drift of neutral base pair differences is about 4×10^{-9} per base, per year. This means that on average 1% of base pairs should be substituted due to neutral evolution about every 2.5 million years. An overall 2% difference, therefore, as among the *DPα* alleles suggests a separation time for these alleles of about 5 million years. This indicates that the *DPα* alleles arose well before the evolution of *Homo sapiens*, but after the divergence of the hominids from the great apes, which is thought to have taken place up to 10 million years ago. The background level of differences between *DQα* alleles outside the *α1* exon domain is slightly higher than that for the *DPα* alleles. This increase should reflect the increased level of variation surrounding the *DQ* alleles that is observed using restriction enzymes, and which is probably explained by hitch-hiking of neutral by selected variation. It is, therefore, hard to estimate the age of the *DQα* alleles. It may be similar to that for the *DPα* alleles and is at most about twice as long, and so still less than the time since the divergence of the hominids from the great apes.

Evolution of *HLA-D* region genes

Sequence comparisons are beginning to provide an overall picture of the pattern of evolution of the *HLA-D* region genes. The biggest differences are between α and β chains, which on average share only about 30% of their amino acid sequences even in their membrane proximal α2 or β2 domains. This is only slightly more than the percentage of shared sequences between class I and class II products which is around 25%, whereas the sharing between HLA products and immunoglobulin sequences is around 15–20%. The next level of comparison within the *D* region is between the products, either α or β chains, of the *DP*, *DQ* and *DR* subregions. These share on average about 55% of their amino acid sequences. The extent of sharing, however, varies for different parts of the molecules. Thus, in all cases, the membrane proximal domains, whether of class I or class II products, are the most conserved and these, for example, share approximately 70% of their sequence between DP, DQ and DR products. Also highly conserved, with an average similarity of about 70% (Travers *et al.* 1984), are the transmembrane

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

regions. Least conserved are the signal sequences, with only around 40% sharing of amino acid sequences. These differences presumably reflect differential functional constraints associated with different parts of the molecule, and so with greater or smaller probabilities of substitution of selectively advantageous replacements. In some cases, for example in the connecting peptide and cytoplasmic tail regions, the extent of differences could actually reflect positive selection for differentiation of the products to accommodate somewhat divergent functions.

The duplicates within a subregion, such as DP or DQ, are more similar to each other than they are to the products of other subregions. Thus DP1α and DP2α, the alpha chain duplicates within the DP subregion, share about 75% of their amino acid sequence while within the DQ subregion DXα and DQα share 93% of their sequence. The divergence between DXα and DQα, excluding the α1 domain, is however only 2.4% and the difference between the α1 domain divergence (16.1%) and the remainder is highly significant ($\chi_1^2 = 14$). This parallels the differences in levels of polymorphism between the different exons, which probably have a selective basis related to the function of the α1 exon. The difference between *DXα* and *DQα* in the α1 exon thus suggests that, like alleles, they have diverged as functional products and so implies that *DXα* has functioned for at least a substantial part of its evolutionary history.

TABLE 2. AMINO ACID SEQUENCE SIMILARITIES BETWEEN HLA-D PRODUCTS

comparison	approximate percentage amino acid sequence shared
Dα and β chains	30
DP, DQ and DR (α or β)	55
DP1α and DP2α	75
DXα and DQα	93

The overall pattern of sequence divergence, summarized in table 2, clearly suggests a series of successive gene duplications leading to the present set of diverged products. The earliest series of events were presumably those that gave rise to the precursors of the present-day HLA, immunoglobulin and other systems of the whole super family from a single ancestral gene. Next, within the HLA system, must have come the duplication that separated the class I and class II antigens. As pointed out by Travers *et al.* (1984) the common ancestor of the present class II, *HLA-D* region products was presumably a homodimeric molecule with a similar structure to the present αβ heterodimers, or for example, the constant region domains of immunoglobulins. Within the *HLA-D* region, the first duplication presumably gave rise to primordial α and β chain genes. Subsequently, duplication of a region containing an *α* and a *β* gene first of all gave rise to *DP*, *DQ*, *DR* and *DZ* and then within these to the duplicates within the *DP* and *DQ* subregions, the latter being by far the most recent event. By using the percentage similarity between sequences one can estimate an approximate divergence time for each of these events and so construct a presumed phylogenetic tree for the genes of the HLA system and, within them, the *HLA-D* region, as shown in figure 4.

Following Kimura (1983), we have taken an average figure of 10^{-9} for the rate of change of amino acid substitutions per year. Clearly the detailed time estimates may be subject to a considerable margin of error, especially for the more distant events. More precise estimates of divergence will come from comparisons of intronic and flanking region nucleotide sequences, which should more nearly correspond in their evolution to the rates expected for neutral

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

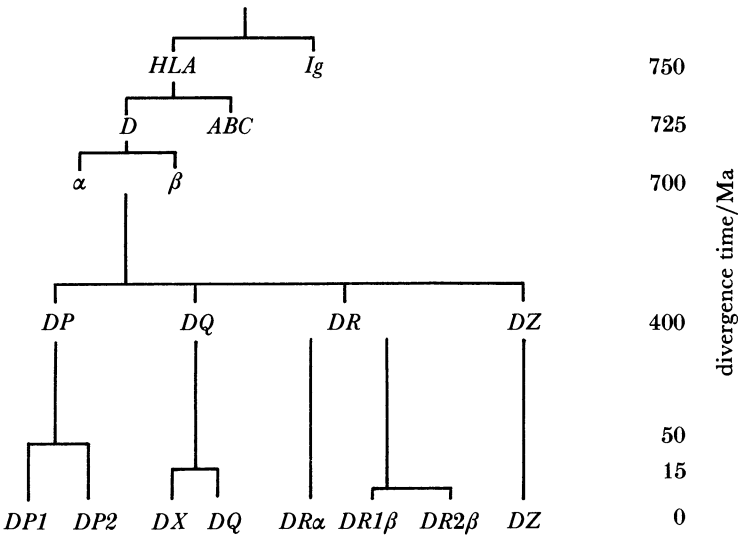

FIGURE 4. Evolutionary tree for the *HLA-D* region products. *HLA* and *Ig* refer to the primordial genes for the HLA and immunoglobulin systems. *D* and *ABC*, α and β refer similarly to the primordial genes, first for the class I and class II subregions and then separately for the class II α and β chains. Except where indicated, subsequent events involve pair-wise combinations of α and β genes. Approximate divergence time is given in millions of years.

substitution, but there are not yet enough data accumulated to justify making these more precise calculations. As pointed out by Travers *et al.* (1984), the initial period of divergence of the whole immunoglobulin and HLA system family, and the differentiation of the major products within the HLA system, may have occurred relatively rapidly over a period of perhaps 100 million years, some 700 or more million years ago, well before the emergence of the vertebrates. The major refinements of the HLA system had thus probably occurred, concordantly with the evolution of the other components of the immune system, by the time vertebrate species were diverging into fish, amphibians, reptiles, birds and mammals which started some 400 million years ago. The *DP* duplication appears to be relatively old, while the *DX*/*DQ* and the *DR* subregion duplications appear to have occurred more recently. The amino acid sequence differences between some of the alleles are not much less than those between the duplicates within the *DQ* and *DR* subregions and suggest that some alleles at least may be relatively old, certainly compared with the age of the human species. In agreement with this some monoclonal antibodies, for example those for the DQw1 determinant, and a DR7-specific monoclonal antibody produced from a mouse–mouse immunization, show that there are polymorphic determinants common to man and mouse. It was observations such as these which suggested the hypothesis that some of the polymorphism within the HLA region could be with respect to control of expression of different genes (Bodmer 1973). This is still a plausible explanation for some polymorphic differences, but certainly not for the majority.

Clearly, many of the mechanisms already mentioned, such as gene conversion, transposition and intrachromosomal unequal double-crossing over, will have contributed to the evolution of the *HLA-D* region as we see it now. Events that produce variants by transposing sequences from one gene to another can explain, for example, serological cross reaction between the products of alleles at different linked loci and can indeed also explain why such pairs of

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

cross-reacting alleles may be in strong linkage disequilibrium (Bodmer 1984). This phenomenon can make the separation and characterization of the different *HLA-D* region gene products with serological techniques very difficult. Polymorphism within the *DQ* subregion for both α and β genes creates the possibility of producing hybrid determinants formed by particular allelic combinations of $DQ\alpha$ and β genes (Spielman *et al.* 1984; Bodmer 1984). This is the explanation proposed, for example, for the particular association of juvenile onset insulin dependent diabetes with *DR3/DR4* heterozygotes. Bodmer (1984) has, furthermore, pointed out that such hybrid determinants may favour the selection of haplotypes containing particular $DQ\alpha$ and β alleles, because their products form combination determinants that are selectively advantageous. This would be a classical example of the potential for the interaction between selection and linkage and might explain why certain haplotype combinations are favoured over others. The predicted structures of α and β chains from different subregions make it unlikely that cross-locus combinations, for example involving $DQ\alpha$ with $DR\beta$ alleles, will form functionally effective combinations (Travers *et al.* 1984). However, within the *DQ* region it is clearly possible that, for example $DQ\alpha$ allelic products could be combined with $DX\beta$, since these sets of products have diverged only comparatively recently.

Conclusions

There is still a considerable amount of detailed sequence information needed, both to complete the picture of the genomic organization of the HLA system as a whole, and the *HLA-D* region in particular and also to describe the patterns of variation between duplicates, between loci, and between individuals, not only for the coding regions but also for introns and flanking regions. These patterns of variation defined, for example, as a function of the distance from intron–exon junctions, should provide much more precise information about the timing of the evolutionary divergence of the HLA region products. Clearly, fascinating data will come from a progressive delineation down the evolutionary scale, even to the most primitive chordates such as the tunicates, of the evolutionary equivalents of the HLA system and other components of the overall immune system. Within the human and mouse species further molecular and functional analysis should identify the control strategy of the region, the polymorphic sequences that are functionally effective in giving rise to immune response differences and disease associations, such as with diabetes, and so the specific nature of the differences that have been selected for during the creation of the extensive polymorphism for HLA-D region genes. The challenge is to identify the adaptive changes against the background noise of neutral evolution.

References

Albert, E. (ed.) 1984 *Histocompatibility testing 1984*. Berlin, Heidelberg: Springer-Verlag.

Bodmer, W. F. 1970 The evolutionary significance of recombination in prokaryotes. *Symp. Soc. gen. Microbiol.* **20**, 279–294.

Bodmer, W. F. 1972 Evolutionary significance of the HLA system. *Nature, Lond.* **237**, 139–145.

Bodmer, W. F. 1973 A new genetic model for allelism at histocompatibility and other complex loci: polymorphism for control of gene expression. *Transpl. Proc.* **5**, 1471–1476.

Bodmer, W. F. 1981 HLA structure and function: a contemporary view. *Tissue Antigens* **17**, 9–20.

Bodmer, W. F. 1983 Gene clusters and genome evolution. In *Evolution from molecules to men* (ed. D. S. Bendall), pp.197–208. Cambridge University Press.

Bodmer, W. F. 1984 The HLA system, 1984. In *Histocompatibility Testing 1984* (ed. E. Albert) pp. 11–22. Berlin, Heidelberg: Springer Verlag.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Bodmer, W. F. & Bodmer, J. G. 1978 Evolution and function of the HLA system. *Br. Med. Bull.* **34**, 309–316.
Bodmer, W. F. & Parsons, P. A. 1962 Linkage and recombination in evolution. *Adv. Genet.* **11**, 1–100.
Chakravarti, A., Buetow, K. H., Antonarakis, S. E., Waber, P. G., Boehm, C. D. & Kazazian, H. H. 1984 Nonuniform recombination within the human β-globin gene cluster. *Am. J. Hum. Genet.* **36**, 1239–1258.
Eigen, M. 1983 Self-replication and molecular evolution. In *Evolution from molecules to men* (ed. D. S. Bendall), pp. 105–130. Cambridge University Press.
Fisher, R. A. 1930 *The genetical theory of natural selection.* Oxford University Press.
Flavell, A. 1985 Introns continue to amaze. *Nature, Lond.* **316**, 574–575.
Kimura, M. 1983 *The neutral theory of molecular evolution.* Cambridge University Press.
Hood, L., Kronenberg, M. & Hunkapiller, T. 1985 T cell antigen receptors and the immunoglobulin supergene family. *Cell* **40**, 225–229.
Kobori, J. A., Winot, A., McNicholas, J. & Hood, L. 1984 Molecular characterisation of the recombination region of six murine major histocompatibility complex (MHC) 1 region recombinants. *J. molec. cell. Immunol.* **1**, 125–131.
Spielman, R. S., Lee, J., Bodmer, W. F., Bodmer, J. G. & Trowsdale, J. 1984 Six HLA-D α chain genes on human chromosome 6: polymorphisms and associations of DCα-related sequences with DR types. *Proc. natn. Acad. Sci. U.S.A.* **81**, 3461–3465.
Steinmetz, M., Minard, K., Horvath, S., McNicholas, J., Srelinger, J., Wake, C., Long, E., Mach, B. & Hood, L. 1982 A molecular map of the immune response region from the major histocompatibility complex of the mouse. *Nature, Lond.* **300**, 35–42.
Steinmetz, M., Malissen,M., Hood, L., Orn, A., Maki R., Dastoornikoo, G., Stephan, D., Gibb, E. & Romaniuk, R. 1984 Tracts of high or low sequence divergence in the mouse major histocompatibility complex. *EMBO J.* **3**, 2995–3003.
Travers, P., Blundell, T., Sternberg, M. J. E. & Bodmer, W. F. 1984 Structural and evolutionary analysis of HLA-D region products. *Nature, Lond.* **310**, 235–238.
Trowsdale, J., Young, J. A. T., Kelly, A. P., Austin, P., Carson, S., Meunier, H., So, A., Ehrlich, H. A., Spielman, R. S., Bodmer, J. & Bodmer, W. F. 1985 Structure, sequence and polymorphism in the HLA-D region. *Immunol. Rev.* **84**, 136–173.
Trowsdale, J. & Kelly, A. 1985 The human HLA class II α chain gene, *DZα*, is distinct from genes in the *DP*, *DQ* and *DR* subregions. *EMBO J.* **4**, 2231–2237.
Woese, C. R. 1983 The primary lines of descent and the universal ancestor. In *Evolution from molecules to men* (ed. D. S. Bendall), pp.209–233. Cambridge University Press.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES