PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Population Genetics of C4 with the Use of Complementary DNA Probes

J. H. Edwards

Phil. Trans. R. Soc. Lond. B 1984 **306**, 405-417
doi: 10.1098/rstb.1984.0101

Phil. Trans. R. Soc. Lond. B **306**, 405–417 (1984)
Printed in Great Britain

Population genetics of C4 with the use of complementary DNA probes

By J. H. Edwards, F.R.S.

Genetics Laboratory, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, U.K.

The C4 system is unusually variant both in the number of expressed genes and in the variety of those expressed. It is also closely linked to two other complement loci (*C2* and *Bf*), which are structurally distinct but not linked to the two complement loci (*C3* and *C5*) which are structurally similar. The *C2*, *Bf*, *C4* segment lies within the MHC system, and their closeness to one another, and to the surrounding recognition loci (*HLA-A*, *-B*, *-C*, *-DR*) makes it difficult to attribute benefit or handicap, in relation to any disease, to any single locus by simple comparisons.

The variation at the protein level, expressed by specificity, or charge, or size, or haemolytic activity, can now be supplemented by variation at restriction sites, or in the length of DNA between sites. The latter may be either causally or coincidentally related to the variant proteins.

These data provide information on the evolutionary relations of the variants, both within and between species, and on the determinants related to inadequate or inappropriate defence against disease. They also provide evidence relevant to the sufficiency of the classical explanations of mutation, recombination and misalignment with recombination, and selection as explanations for what exists now.

The nature of population genetics

Population genetics is variously defined, but basically covers the development of model systems that are both simple enough to be mathematically tractable and faithful enough to be of some descriptive and predictive value. As with all models they are bounded by the rival demands of simplicity and fidelity, and different models can usefully coexist for different purposes. Such models are essential in the development of any science to bridge the gap between the basic vernacular and the new observations, to provide some framework by which neologisms can be rooted in the past, and to allow the model to be tested by prediction, or, in the case of biology, by consistency with a single evolutionary tree.

The models of population genetics differ from most models in that indeterminacy has a central place, ignorance being formalized within the essential operations of segregation and cosegregation. That is, if parentage occurs, the gene conveyed is allocated at random, and if pairs of genes are conveyed on the same chromosome, their chance of joint transmission is proportional to their closeness. These chaotic events are formalized by the first and second laws of Mendel; one consequence of the first is the equilibrium of Hardy and Weinberg. Given the further rare event of a gene changing by mutation, and the obvious requirement of parental survival if genes are to be transmitted, the basic Mendelian model is defined.

When technical advances allow either a major increase in the quality or the quantity of data, then models based on the coarser and simpler reality of the past may have to be modified, extended or rejected. Poincaré pointed out how fortunate the physicists were in that the advance

of instrumentation had not been so rapid as to outdistance concept formation. In genetics the new DNA techniques have unleashed a range of variability not easily accommodated within the standard models of classical genetics, and this has been accompanied by a failure of communication between these two subjects.

Shapiro (1983) has recently suggested the virtual rejection of classical population genetics on the grounds that little notice has been taken of new phenomena by theoretical biologists. While there is some justification for this view, the curiosity or attention of classical geneticists is hardly eased by the habitual misapplication of classical terms, or the development of transient nomenclatures which denote general phenomena in terms based on a technique or a species, and the indifference of population geneticists is partly due to this verbal obscurity. Shapiro's suggestion has been answered (Charlesworth & Smith 1983).

The Mendelian model has been adequate as a basis to explain both the main lines of evolution (Dobzhansky 1951), and the more detailed studies of our own species. Low but steady mutation rates occasionally changing otherwise immutable genes, whose neighbours occasionally change owing to the cut-and-join processes of recombination, and occasionally of mis-recombination, were sufficient to explain all that was known from the study of phenotypes down to the resolution of protein electrophoresis and amino acid sequencing. In our own species and its close relations the general synopses of Harris (1969), of Cavalli-Sforza & Bodmer (1972), of Goodman *et al.* (1975), and of Race & Sanger (1975) revealed no insufficiency in the classical Mendelian model of the usually immutable gene passive to the immediate mechanisms of its transmission.

In spite of such consistent sufficiency, the new DNA techniques soon provided evidence inconsistent with the unblending units postulated by Mendel, and the pure tree structures, undisturbed by any cross-connections, postulated by Darwin.

Highly repeated DNA sequences were first shown to be distributed so widely and consistently within (Southern 1970), but not between, allied species that some mechanism other than mutation and recombination had to be postulated: however such structures were not clearly related to transcribed segments, or genes, and had little impact on the assumed adequacy of the basic tree model to explain evolution at the level of the genes as usually defined.

The first three groups of loci to be examined in detail, those for the haemoglobins (see Jeffreys 1982), the gammaglobulins (Ollo & Roujeon 1983), and those within the MHC region (Steinmetz & Hood 1983), all revealed features only explicable in terms of 'blending' of the genetic elements during their passage through the diploid organism. Although the phenomenon was well known in simpler eukaryotes, and had been postulated (Winkler 1930) in the same year as Fisher's Genetical Theory of Natural Selection (Fisher 1930), this theory was explicitly based on the consequences of non-blending evolution, a matter occupying the first two pages. No clear-cut examples were defined in mammals until detailed sequencing revealed identities inconsistent with the expectations based on mutation alone. Unfortunately these examples of blending within genes acquired the name of 'conversion', a process originally defined in relation to homologous units at meiosis, to explain the blending behaviour of tandemly related duplications. This term is now widely used to cover blending of homologous, tandemly related, and distant segments.

The molecule of C4 provides the ultimate challenge to the population geneticist. It is one of the largest soluble proteins, and it takes part in numerous, diverse and highly specialized activities involving activation and inactivation, reversible and irreversible bonding, and

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

intimate associations with dissimilar proteins of great complexity with which it must have co-evolved (Porter 1983). Further, it must have the ability to work best under the unusual conditions of high fever. It is assembled after transcription into three chains, and its activities are largely related to distinct components liberated on its activation and dissolution. In man it is also unusually variable at the level of protein electrophoresis, its variability extending to both forms of its duplication, and apparently including null forms at either locus, but only rarely at both. Triplicate forms are also found as fairly common variants. While its main function, its central position in the complement pathways leading to the destruction of bacteria and larger parasites, and of virus-infected cells, is not in doubt, its protective functions are occasionally eclipsed by various disorders related to misdirected activities in which it has a key role. It would be surprising if its variant forms did not play a substantial role in varying susceptibilities to disease. Nothing is known of the genetic control of its glycosylation, which is extensive and unlikely to be inert or invariant.

These complexities in its form and function are largely protected from inferences based on population data by its position between the key recognition systems, the *HLA-DR* and *HLA-B*, *-C*, and *-A* loci, and, to a lesser extent, by close proximity to the rather less variant *Bf* and *C2* loci. Integrated systems in which genes with related functions are associated in a linkage group are unusual in eukaryotes although well recognized in some bacteria and most bacteriophages.

Compared with the haemoglobin molecule, which has a single basic function, and is known to be related to defence against only one family of diseases, and is not known to be linked to loci with distinct but related functions, the problems presented are formidable. We lack not only the extensive haplotype data relating to disease but also the means of analysing such data should we acquire it.

Models of inheritance of C4

The simplest genetic model is that in which a locus has two alleles that are randomly distributed to children, the various combinations possible having their proportions constrained by the demands of chance. The early observations of the fast and slow classes of C4 protein, now termed A and B respectively, were inconsistent with this model, Mendel's first law, with its consequence, the Hardy–Weinberg equilibrium. There was a gross excess of heterozygotes beyond the maximum possible proportion of a half.

This was resolved by increasing the number of allelic classes to include the complex '*AB*' allele, or alternatively to increase the number of loci to include the related duplication products *AO*, *AB*, and *OB*, in the powerful model which accommodated the Chido and Rodgers antigens and the C4 protein within a single system (O'Neill *et al.* 1978). This model provided an excellent fit to the data, but at a cost of introducing a system lacking any adequate precedents. The apparently similar variability in the haemoglobin alpha loci related only to phenotypes limited to a few populations and constrained by draconian selective pressures. Further, the gene products were not produced in amounts fully commensurate with their assumed number (table 1) (Hussain 1982; Mauff *et al.* 1984). While some inequality in the products of varying loci was well established in the haemoglobin loci, the correlation of amount and number was weak. The model was not fully satisfactory, suggesting the possibility of a model in which duplication with the pairs *AA*, *AB*, *BA* and *BB* would be maintained by some mechanism beyond the classical mechanisms of mutation, selection and rearrangement by recombination.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Table 1. Gene number, inferred from family studies, against C4 level in plasma defined by Mancini technique (Hussain 1982)

gene number	number	relative level
4	47	100 ± 16
3	57	99 ± 32
2	48	77 ± 28

Table 2. Models of C4 genotype–phenotype relations

loci	alleles	consequences
1	*A*, *B*	too many *AB*
1	*A*, *B*, *AB*	few precedents
2	*A*, *O*; *O*, *B*; *A*, *B*	too few *O*, *O*
2	*A*, *A*; *B*, *B*; *A*, *B*; *B*, *A*	conversion needed

Either the classical mechanism had to support the unlikely events of duplication with common, but almost never associated, null alleles, and a relative paucity of the product when both alleles were present; or the presence of exact duplicates has to be brought about by some novel mechanism that imposes an identity of structure on tandem repeats. Both haemoglobin clusters provided an adequate precedent for this. In the beta complex the gamma loci are similar to a degree inconsistent with their assumed antiquity (Slightom *et al.* 1980), while in the alpha complex with two alpha genes show the remarkable property of those on the same chromosome being more similar than their homologues (Zimmer *et al.* 1980). There is no adequate mechanism known for this, and no justification for assuming it to be meiotic, as mismatching can hardly be common with only two or three loci. It is only in mitosis and in the functioning gene complex that tandemly related genes are more closely related than homologues, and the bases are exposed. This phenomenon has been named, through its consequence, 'concerted evolution' (Zimmer *et al.* 1980) but lacks any primary name. It is clearly a major phenomenon in relation to the two *C4* tandem repeats, and may be sufficiently strong as a correcting mechanism to permit duplicated loci to be so faithfully corrected in tandem that they are usually identical on even the most exacting protein electrophoresis. That is, the basic *AO*, *OB*, *AB* model of O'Neill *et al.* (1978) may be, at the DNA level, represented by *AA*, *BB*, *AB* and *BA* loci. Any alternative explanation would be likely to impose frequencies of the double-null, represented by C4 deficiency, inconsistent with clinical observation. These various models are shown in table 2 and figure 1.

The Mendelian model of inheritance is based on inertness of the genes in relation to the mechanics of their transmission and on their independence. Although their consequences are usually related to their joint action they are undisturbed by any blending or interchange between one another. Consequently, every gene can be followed backwards from descendant to ancestor, passing unchanged rootwards and backwards in time, down any family tree. At every generation it comes from one and only one parent. Consequently, the tree which defines the ancestral lineage of any gene has fewer branches each generation, in distinction to the full ancestral tree, which doubles its branches every generation, eventually forming networks owing to coancestry.

This central feature of the Mendelian model, when interpreted in conjunction with the Darwinian model of speciation through common ancestral forms related by tree structures, provides a simple model in which all genes are, ultimately, derived from a single ancestral gene.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

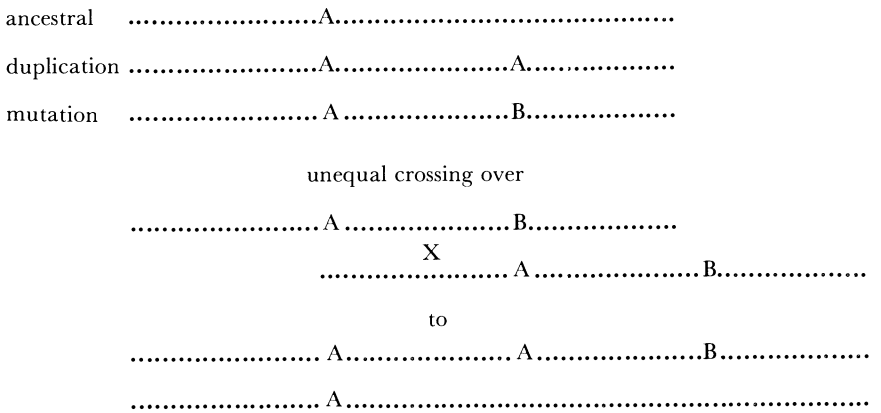

FIGURE 1. Mechanism postulated to explain duplication and variations in gene number.

The genes at a single locus represent merely a subset of the whole tree. Where various forms of theses genes can be defined then these variants, or alleles, have developed owing to mutations and, if each distinct mutational event was unique, then all alleles of the same form are simply related by being a subset of the tree. Two nearby genes, which may or may not have had a common origin in a duplication event, will share the same tree structure until they separate in a recombinational event (figure 2). On the classical model all genic variety can be interpreted in terms of a limited number of mutant events on a single tree, and all variety involving pairs of neighbouring genes in terms of recombination.

The approximate structure of such trees may be inferred from the rates of mutation and recombination. Numerous fairly consistent estimates on the mutation rate per base per year (see Kimura (1983) for a review) give a figure of about 10^{-9}, a figure curiously invariant to the generation time of the various eukaryotes on which it is based. In man, as in other mammals, there are about 3×10^9 base pairs, and in the human male about 60 chiasmata (Ford & Hamerton 1956), or recombinational events involving four strands, so that only half will be manifest as crossovers. The chance of recombination between any base pair is therefore about

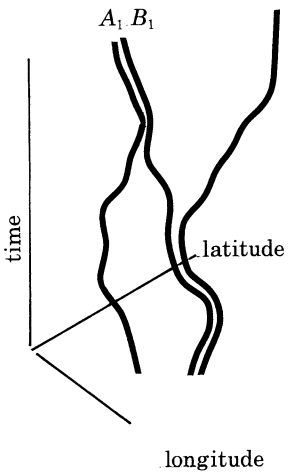

FIGURE 2. Effects of recombination on paired loci. On the classical model the half-life of the parallel lines is over 100000 years.

TABLE 3

(Expectations per megabase per thousand years or kilobase per million years.)

mutation	t, t	1
	d, d	< 1
recombination		1/3

Showing approximate rates.
t, t, Transitions and transversions.
d, d, Duplications and deficiencies.
One generation is taken as 30 years.

10^{-8} per generation, or less than 10^{-9} per year, and of the same order as the chance of a base changing into another base by a transition or transversion. Mutant events may also lead to duplications or deletions varying in size from one base to many kilobases. Small duplications or deficiencies are almost certainly rarer events than transitions or transversions. It is convenient to work in units of half life, as in mutant events involving atoms, and in units of kilobases and megabases, and time scales of thousands of years and millions of years (table 3). For rare events with a frequency u, the half life is equal to $u/\ln 2$ or about $0.7u$.

LINKAGE AND ALLELIC ASSOCIATION

If loci are linked then alleles are associated and the association of alleles in close relatives is the basis of classical linkage analysis. When loci are very closely linked, then this association will extend to distant relatives, that is, to randomly defined members of the same species. This is a necessary feature of close linkage in a finite species (Robbins 1918; Edwards 1980). It is also possible for pairs of neighbouring loci to maintain allelic associations owing to various pairs being advantageous. This positive association, based on selection rather than chance, was termed 'linkage equilibrium' by Fisher (1930). In the case of the *C4* loci and its neighbours there is clearly a case for the selective advantage of various allelic pairs, since C4 and C2 are functional partners needing mutually adapted binding sites. However, the anticipated scale of chance effects, and the confusion of secondary effects from the recognition loci flanking the effector segment (*C2*, *Bf*, *C4*), makes it impractical to distinguish joint selection, secondary selection – the 'hitch-hiker' effect, and chance – the 'founder' effect. If we ignore the obvious strong selective effects, then we may consider the implications of a closely linked set of loci rearranged by the random cut-and-join events of recombination.

In the segment that includes the *C2*, *Bf*, *C4* loci, which is about 100 kilobases long, a rearrangement has a 50% chance of happening in 700 generations, or about 20000 years. We are effectively sampling over half the haplotypes we study from a preglacial community, and from a necessarily rather small number of common ancestors. The various haplotypes that allowed our ancestors to survive to beget us are hardly likely to be related strongly to the same diseases manifest in dense populations on adequate and different diets. To attempt to explain the present variability in relation to contemporary disease is comparable to attempting to explain recent European history in terms of slings and arrows (figure 2).

Not only has our environment, and our nutrition, changed but all our parasites have the capacity to evolve faster than we do, and their evolution may include a transfer of preferred host species.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Since the *C4* loci are considerably larger than the gap between them, as are the *C2*, *Bf* pair, we may note that the chance of recombination within one of these pairs will exceed that between them, so that most apparent *Bf*, *C2* or *C4A*, *C4B* crossovers will involve gene cuts.

Conversion

This term, advanced by Winkler in 1930, was originally restricted to the asymmetrical interchange of genetic information between homologous loci. For example, two chromosomes with the alleles *ABC* and *abc* might lead to the products *AbC* or *aBc* with a frequency so inconsistent with expectations based on the recombination rates between any pair that some localized interchange had to be postulated. It is now clear that this event usually involves units of tens or hundreds of bases, rather than the tens of megabases which usually separate recombinational events (figure 3). Further, it is not a randomly defined event, in that the rates of conversion between any pair of alleles often differ. In the example above the expectations of the products *CdE* and *cDe* would be unequal. For a comprehensive review see Whitehouse (1983).

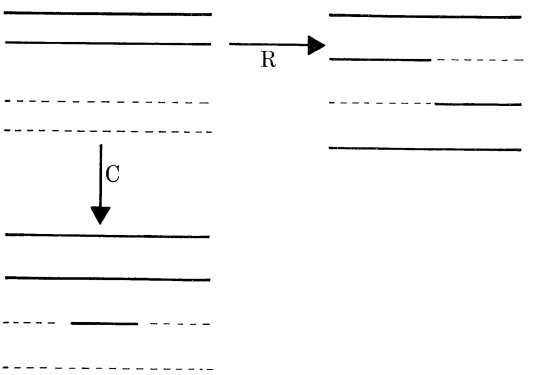

Figure 3. Consequences of recombination and conversion. The lines represent the chromosomes at meiosis.

This imposes serious difficulties in relation to the classical Mendelian model, since the tree structure is confused by localized networks. Although these events may be rare, they prevent extended inferences being based on unique observations. The potential for both dispersing and correcting mutational events gravely confounds inferences relating to gene flow and mutation rate based on evolutionary arguments.

The relation of recombination and classical conversion is shown in figure 3. Unfortunately the term has been extended to cover other distinct phenomena of correction, using this term to imply the replacing of one fragment of text by another without reference to which can be considered 'correct'. Given two chromosomes with the duplicated loci *D* and *E*,

$$\ldots\ldots DDDDDDD\ldots\ldots EEEEEEE\ldots$$
$$\ldots\ldots d\,d\,d\,d\,d\,d\,d\ldots\ldots e\,e\,e\,e\,e\,e\,e\ldots$$

then correcting mechanisms could act between elements *D* and *d* or *E* and *e*, as in classical conversion, and this would have to occur at meiosis in mammals as there is little other opportunity for the touching of homologues. Conversion by messenger, although possible, is unlikely in the sequence of cells connecting the germ cells unless the loci involved are active.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Correcting mechanisms could also act between elements of *D* and *d* or *E* and *e* and there is no reason to suspect meiosis, a relatively rare event compared to mitosis even in the germline. This seems the simplest type of correction to postulate, especially for small numbers of loci. Where one locus is transcribed with opposite polarity it is particularly easy to suggest possible mechanisms. There is a good precedent in *Drosophila* (Leigh Brown & Ish-Horowicz 1981).

In addition, correcting mechanisms could be between *D* and *e* or *d* and *E*, as postulated by Hood *et al.* (1975). When multiple copies, rather than duplicated copies, are present, recombination following misalignment is a possible mechanism. However this can only impose uniformity at an acceptable rate given numerous loci or a wide range of normal variation for numbers of loci. This mechanism is not classical conversion, although the consequences are similar; the event of misalignment, if common, would also give opportunities for classical conversion.

Finally we may consider the possibility of correction between loci on different chromosomes, which clearly occurs with small non-transcribed segments of DNA. The detection of such events is formally possible from the identity of variant segments in such similar but distant loci as *C3*, *C4* and *C5*.

These various mechanisms may all play some part in the inexplicable similarity of parts of the *C4A* and *C4B* sets of alleles. The absence of any agreed nomenclature makes any such discussion difficult. Any of them is sufficient to impose difficulties in any quantitative evolutionary inference.

Use of restriction enzymes

Both restriction enzymes and protein electrophoresis will reveal only a small proportion of the underlying variability, on the assumption that the variation is qualitative and related to transitions and transversions rather than to duplications and deficiencies. In practice both would be expected to reveal less than a tenth of the total variation. About one base in 4000 will be expected to initiate any specified sextet of bases, which would be recognized by a six-base restriction enzyme. Any such sextet could be inactivated by a change affecting any of its six bases: there is the same chance of a new restriction site being formed at any potential sextet, that is a sextet needing only a single base change to be complete. The chance of a point mutation's being recognized by a defined enzyme is therefore about 12/4000 or about 1/300. If 15 enzymes are used the chance of a variant being detected would be about 1/20. A randomly defined base in a gene has a chance of 1/5 or so of being in a coding sequence; if it is, the chance of a change being reflected in a different amino acid is about a half, and the chance of this being reflected in a sufficient charge difference to ensure recognition is about a third. Once again the chance of detection is about 1/20. These are very rough estimates, but they are of the same order. Restriction enzymes recognizing quartets will have a higher chance of defining a variant, but this is to some extent offset by the technical difficulties in recognizing very short fragments.

The assumption that the palindromic sites that can be recognized by restriction enzymes are randomly distributed is known to be false: they appear to be more densely distributed outside the genic segments (Steinmetz & Hood 1983). In some organisms segments of many kilobases may be uncuttable with a wide range of enzymes. In spite of these difficulties, it is probably reasonable to assert that a restriction site has only a small chance, of the order of one in ten

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

or so, of detection by the batteries of enzymes in current use, and the chance of a DNA variant being reflected by a charge difference in an amino acid is even less. Consequently the chance of any variant being defined as both a DNA and a protein variant by standard techniques will be of the order of 10% or less. Studies of DNA variants will usually be unrelated to functional variation.

If we postulate a tree connecting two protein variants through their common ancestor (figure 3), then we can expect several other distinct mutations to be present in the two genes, some of which may be manifest by restriction enzymes. We can also anticipate the unlikely possibility of a variant manifest in both ways, as in the base change related to sickle cell disease (figure 4). If we postulate the tree underlying the protein variants in C4, which can be split

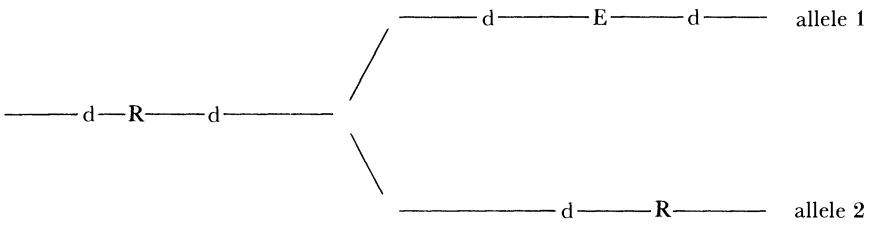

FIGURE 4. Evolutionary relations of base changes (d) some of which are represented by restriction enzyme variants (R) and some by electrophoretic variants (E).

into two main branches by both the fast and slow clusters of variants on electrophoresis and also by serological means, we can consider using restriction enzymes to increase definition. In principle (figure 5), the DNA variants may precede branching when they will define a group, or follow branching when they will define a split, or they may show a one-to-one relation consistent with a common cause or at least with two mutational events likely to be relatively close in time, or they may be inconsistent. Since most DNA variants will not cause differences in the protein, their study will be of limited value in studies relating to variants primarily related to disease.

DNA VARIANTS

A systematic search for DNA variants in a series of protein variants soon revealed an apparent one-to-one relation of a DNA variant with the fast non-haemolytic protein variant (A6) (Palsdottir *et al.* 1983). As this was consistent within and between families both in England and Iceland, it is at least plausible that it defines the base change related to the electrophoretic variant and that this is related to loss of haemolytic activity. While rewarding as an initial finding, a consistent relation is of little value unless it assists in locating the base involved. Grouping variants would be of interest in defining the tree structure, and might assist studies relating to disease by defining a wider group of C4 molecules with common biological features needlessly split by irrelevant features revealed on electrophoresis. However the recognition of such hidden functional variation would involve very large numbers of patients.

Such attempts assume the preservation of the basic tree structure with unblending genes. Observations have been made (M. C. Carroll, personal communication) which are inconsistent with a simple tree structure branching into the two major serological and electrophoretic classes. These findings can only be explained by some blending phenomenon involving the assumed

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

tandem loci *A* and *B*. The extreme conservation of bases, including third bases, over the substantial segments of the two tandem loci so far sequenced (Carroll *et al.* this symposium), is also inconsistent with classical mechanisms restricted to mutation and recombination.

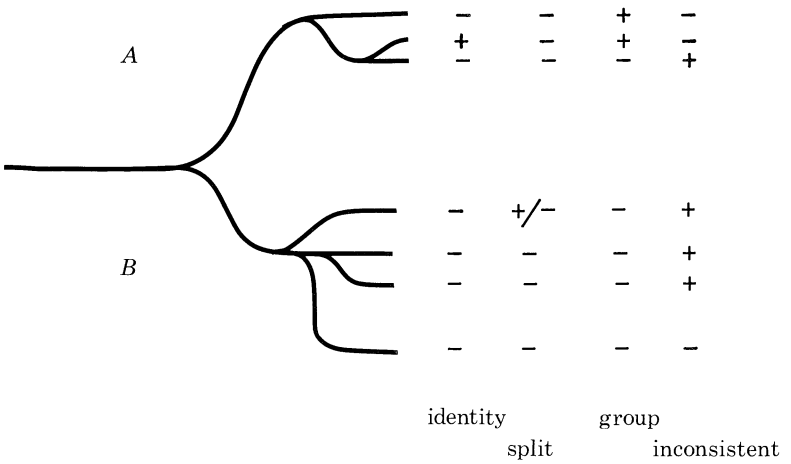

FIGURE 5. Possible relations of protein and DNA variants.

While the expected restriction of the DNA and protein variants to grouping and splitting has not been realized, the impossibility of using these techniques to define the ancestral tree is itself of far greater interest than the mere documentation of an expectation. Taken together with the evidence of extreme similarity of the sequenced segments, and the paradox of variant gene number not being reflected by serum levels of protein, this observation is sufficient to reject the classical model. At least there is a firm precedent for so doing in haemoglobins in our own species, in immunoglobulins in mice, and in numerous other loci in various invertebrate species.

CLINICAL APPLICATIONS

Probes, which can be applied to DNA from any cell, widen the scope of C4 typing to the dead and the unborn neither of whom can easily provide blood in good condition. This is unlikely to be of much value in practice, except in pursuing anthropological or disease-susceptibility studies under conditions in which fresh blood from the living is not available, or freezing or transport is difficult. While the direct application of DNA probes to clinical practice would seem somewhat limited, their application in the elucidation of HLA-related disease could be definitive. In view of the key position of the recognition loci in defence against bacterial and viral disease, it would be surprising if susceptibility to any such disorder were not to some extent HLA-related, and therefore secondarily related to complement variants. In view of the persisting variety at these loci it would seem reasonable to assume that there was a similar general 'fitness' for each haplotype, but that each was variously associated with susceptibility or resistance to various disorders, and that the elucidation of the relative influence of the *HLA-DR* and the *HLA-B*, *C*, *A* neighbours would be crucial to understanding the nature of the infection, or the nature of any excessive or inappropriate response. The *C4* locus, being between the two, could provide decisive evidence by assisting in the enumeration of the limited number of haplotypes likely to be strongly related to most cases of disease in any strongly associated

disorders in any homogeneous population. These variants must, however, be considered as a class distinct from the protein variants. They are less likely to affect protein function than a variant affecting behaviour in solution, including electrophoretic, immunological, and haemolytic activity, and the wealth of data on protein variation is already beyond the resources of formal analysis. It is in the detailed study of a small number of well-defined 'hot' haplotypes, and, in the case of diabetes, the 'cold' or protective *DR2*-containing haplotype, that future progress is likely.

We may consider the major problems presented by graft rejection; HLA-related acquired disease; disease caused by closely linked loci; and disease caused by complement deficiencies.

Graft rejection

In transplantation there is no likelihood that the various alleles at the complement loci would have any effect commensurate with that related to allelic variants at the recognition loci (*HLA-DR* and *HLA-B*) which flank them, even if techniques of sufficient rapidity were available.

HLA-related acquired disease

Most of the HLA-related acquired diseases are plausibly related to a delayed, inadequate, or excessive response to an infective episode. Examples (for review see Moller 1983) include ankylosing spondylitis, rheumatoid arthritis, multiple sclerosis, and insulin-dependent diabetes. Of these four well-recognized associations the former is clearly related to the *HLA-B* locus in view of its intense association with the *HLA-B27* allele. Insulin-dependent diabetes can be related to the *HLA-DR* loci on grounds of an interaction involving an increased risk of the 3/4 heterozygote compared with either homozygote, and a protective effect of another allele, *HLA-DR2*. Complement variants will necessarily show some differential effect owing to the allelic associations often termed linkage disequilibria (Kojima & Lewontin 1970) necessarily involved in such closely linked loci; there is no reason to postulate any primary association. In rheumatoid arthritis and multiple sclerosis the direct data are weaker, but there seems no reason to relate the varying course of these diseases, with their long remissions, with the likely consequences of variant forms of complement.

In all these disorders the risk of affliction, even of those with the 'hottest' haplotypes, does not exceed 10%. Data based on sib-incidence, and a quarter of sibs are HLA identical, suggest that more exacting definition of haplotypes, which will clearly be possible from complement typing at the DNA level, will not greatly increase the estimated risk related to the 'hot' haplotypes defined by present methods.

There is hardly likely to be an informed demand for the termination of pregnancies in relation to such uncertainties of affliction, especially as the consequence of the recognition of such associations is likely to lead to other methods of control before manifestation: most of the disorders involved, including multiple sclerosis and insulin-dependent diabetes, usually present after puberty.

There are also disorders that are associated with various HLA haplotypes and also with a primary disturbance to complement metabolism. Lupus erythematosus, often termed disseminated lupus erythematosus, or DLE, is the commonest of such conditions, and has recently been shown to be strongly related to the number of apparent null-loci by typing patients and their relatives for both HLA and complement markers, and attempting to unravel the

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

information by a detailed argument based on enumeration (Fielder *et al.* 1982). As there is evidence implicating *C4* as central to the disease process, it seems likely that the resolution of the null-locus problem, and the development of *A* and *B* specific probes, will define which *C4* locus is primarily related.

Finally, and so far uniquely, there is one known HLA-associated disorder, coeliac disease, related to a specific plant glycoprotein which damages the gut mucosa and submucosa in individuals with certain HLA-related haplotypes. While the exact relation is confusing, since both the human and wheat populations vary, it seems unlikely, from the strength of the *HLA–DR* relation, that complement variants would be directly involved. Once the diagnosis is made the offending factor can be avoided without undue difficulty; it could, in principle, be outbred from wheat and rye.

The distinctions between primary effects, secondary associations and linkage have to be by stratification based on large numbers of haplotypes. Elaborate statistical inferences are unlikely to compete with either simple numeration or mechanistic interpretations in reliability. The firmest direct inference relating to a primary association with *C4* (Fielder *et al.* 1982) required extensive family studies even to define an association with gene number. Very much larger series will usually be needed to define susceptibility by allelic variety. At present the only formal analytical procedure is based on the pair-by-pair associations of alleles with disorders, and with each other, and such pair-by-pair associations cannot extract the information provided by a multiple-locus haplotype. Hypothetical susceptibility loci can be postulated, and their position defined: however, the ability to locate a hypothetical locus does not confer reality.

Disease caused by closely linked loci

There are two disorders related to linked loci, excluding the hypothetical loci which can always be invented. The first, 21-hydroxylase deficiency, is a recessive with an incidence of about 1/4000 in caucasians. It is usually amenable to simple therapy once diagnosed and since it is a recessive most cases are the first in any family. A minority, in girls, involve serious genital abnormalities. Foetal diagnosis is possible following a defined case of known HLA phenotype to the same parents through HLA typing of amniotic cells, and in principle this could be replaced by the use of complement or other HLA related probes earlier in pregnancy. However, as the locus, which probably lies between the *C4* and *HLA-B* loci, can hardly fail to be identified and sequenced shortly, especially as rich sources of RNA are easily available, there are limited advantages in developing a novel technique for rare and ephemeral use. The other locus, whose nature and location is less certain, is that for haemochromatosis, a disorder of adult life easily controlled by becoming a blood donor once recognized, and easily recognized once suspected: this is hardly a candidate for foetal curiosity.

Disease caused by complement deficiencies

Primary deficiencies of a complement factor (Lachmann, this symposium), could in principle be diagnosed directly in foetal life if owing to deficiency of the locus, or indirectly owing to haplotyping. This would be an extremely rare situation in practice.

The main clinical value of the probes is likely to be secondary to an understanding of the detailed structure of the loci involved, and of their products, giving a sound understanding of the disorders owing to inadequate, excessive of inappropriate defence.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

References

Carroll, M. C., Porter, R. R. Campbell, R. D. & Bentley, D. R. 1984 A molecular map of the human major histocompatibility Class III region linking complement genes *C4*, *C2* and *factor B*. *Nature, Lond.* **307**, 237–241.

Charlesworth, B. & Smith, J. M. 1983 Population genetics. *Nature, Lond.* **303**, 748.

Dobzhansky, T. 1951 *Genetics and the origin of species*, 3rd edn, revised. New York, London: Columbia University Press.

Edwards, J. H. 1980 In *Population structure and genetic disorders* (ed. A. W. Eriksson). London, New York, Toronto, Sydney, San Francisco: Academic Press.

Fielder, A. H. L., Walport, M. J., Batchelor, J. R., Rynes, R. I., Black, C. M., Doch, I. A. & Hughes, G. R. V. 1983 *Br. med. J.* **286**, 425–428.

Fisher, R. A. 1930 *The genetic theory of natural selection*. Republished Dover Books 1958.

Ford, C. E. & Hamerton, J. L. 1956 The chromosomes of man. *Nature, Lond.* **178**, 1020.

Goodman, M., Tashian, R. E. & Tashian, J. H. (ed.) 1976 *Molecular anthropology. Genes and proteins in the evolutionary ascent of the primates*. New York and London: Plenum Press.

Harris, H. 1969 Enzyme and protein polymorphism in human populations. *Br. med. Bull.* **25**, 5–13.

Hood, H., Campbell, J. H. & Elgin, S. C. R. 1975 The organisation, expression and evolution of antibody genes and other multigene families. *A. Rev. Genet.* **9**, 305–353.

Hussain, R. 1982 Antibodies to complement. D.Phil. thesis, University of Oxford.

Jeffreys, A. 1982 In *Genome evolution*. London: Academic Press.

Kimura, M. (ed.) 1982 *Molecular evolution, protein polymorphism and the neutral theory*. Berlin, Heidelberg and New York: Springer-Verlag. Tokyo: Japan Scientific Societies Press.

Kojima, K. & Lewontin, R. C. 1970 Evolutionary significance of linkage and epistasis. In *Mathematical topics in population genetics* (ed. Ken-ichi Kojima), vol. 1 (*Biomathematics*), pp. 367–388. Berlin, Heidelberg, New York: Springer Verlag.

Leigh Brown, A. J. & Ish-Horowicz, D. 1981 Evolution of the *87A* and *87C* heat-shock loci in *Drosophila*. *Nature, Lond.* **290**, 677–682.

Mauff, G., Bender, K., Giles, C. M., Goldmann, S., Opferkuch, W. & Wachauf, B. 1984 Human C4 polymorphism: Pedigree analysis of qualitative, quantitative, and functional parameters as a basis for phenotype interpretations. *Hum. Genet.* **696**, 1–11.

Moller, G. (ed.) 1983 HLA and disease susceptibility. *Immunol. Rev.* **70**, Munksgaard.

Ollo, R. & Rougeon, F. 1983 Gene conversion and polymorphism: generation of mouse immunoglobulin *2a* chain alleles by differential gene conversion by *2b* chain gene. *Cell* **32**, 515–523.

O'Neill, G. J., Yang, S. Y., Tigoli, J., Berger, R. & Dupont, B. 1978 Chido and Rodgers blood groups are distinct antigenic components of human complement C4. *Nature, Lond.* **273**, 668–670.

Palsdottir, A., Cross, S. J., Edwards, J. H. & Carroll, M. C. 1983 Correlation between a DNA restriction fragment length polymorphism and C4A6 protein. *Nature, Lond.* **306**, 615–616.

Porter, R. R. 1983 Complement polymorphism, the major histocompatibility complex and associated diseases: a speculation. *Mol. Biol. Med.* **1**, 161–168.

Race, R. R. & Sanger, R. 1975 *Blood Groups in Man* (6th edn). Oxford, London, Edinburgh, Melbourne: Blackwells Scientific Publications.

Robbins, R. B. 1918 Some applications of mathematics to breeding problems. III. *Genetics* **3**, 375–389.

Shapiro, J. A. 1983 Evolution by numbers. *Nature, Lond.* **303**, 748.

Slightom, J. L., Blechl, A. E. & Smithies, O. 1980 Human fetal y^G- and y^A-globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. *Cell* **21**, 627–638.

Southern, E. M. 1970 Base sequence and evolution of guinea-pig alpha satellite DNA. *Nature, Lond.* **227**, 794–798.

Steinmetz, M. & Hood, L. 1983 Genes of the major histocompatibility complex in mouse and man. *Science, Wash.* **222**, 727–733.

Whitehouse, H. L. K. 1982 *Genetic recombination understanding the mechanisms*. New York, Brisbane, Toronto, Singapore: John Wiley.

Winkler, H. 1930 *Die Konversion der Gene*. Jena.

Zimmer, E. A., Martin, S. L., Beverley, S. M., Kan, Y. W. & Wilson, A. C. 1980 Rapid duplication and loss of genes coding for the alpha chains of hemoglobin. *Proc. natn. Acad. Sci. U.S.A.* **77**, 2154–2162.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES