PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

The Human Gene Map [and Discussion]

E. B. Robson and Kay E. Davies

Phil. Trans. R. Soc. Lond. B 1988 **319**, 229-237
doi: 10.1098/rstb.1988.0045

Phil. Trans. R. Soc. Lond. B **319**, 229–237 (1988)
Printed in Great Britain

The human gene map

By E. B. Robson
Galton Laboratory, University College London, Gower Street, London WC1E 6BT, U.K.

Mapping started exactly 50 years ago when Bell & Haldane (*Proc. R. Soc. Lond.* **123**, 119 (1937)) measured the genetic distance between colour blindness and haemophilia. In their Discussion they wrote 'if. . . an equally close linkage were found between the genes determining blood group membership and that determining Huntington's chorea, we should be able, in many cases, to predict which children of an affected person would develop this disease, and to advise on the desirability or otherwise of their marriage'.

Progress in this direction has proceeded through the discovery of autosomal linkages by family studies, and the assignment of genes to particular chromosomes by somatic-cell hybridization techniques. Recombinant DNA technology has been successfully used in both approaches, with the result that many chromosomes are now roughly mapped. In practice, the map can already be used for prenatal diagnosis of several diseases, and may provide 'take-off' points for some molecular approaches to poorly defined genes. More fundamentally, it is beginning to provide insights into the nature of the meiotic process and the organization of the genome.

Introduction

To be asked to cover a topic such as the human gene map is today an impossible task, even within the general context of a particular meeting, and in attempting to do so I have taken my starting point as Bell & Haldane (1937), for not only did they make the first academic contribution to the map by measuring the genetic distance between colour blindness and haemophilia, but they pointed out the implications of the use of linked markers with respect to this symposium's subject, the prevention and avoidance of genetic disease, in very precise terms:

> The present case has no prognostic application, since haemophilia can be detected before colour-blindness. If, however, to take a possible example, an equally close linkage were found between the genes determining blood group membership and that determining Huntington's chorea, we should be able, in many cases, to predict which children of an affected person would develop this disease, and to advise on the desirability or otherwise of their marriage.

The pitifully slow development of the human gene map meant that although the logic of this proposition was irrefutable, and although technical advances in medicine meant that the more socially realistic approach via antenatal diagnosis became possible, the use of linked markers in diagnosing genetic disease remained for a long time little more than a pious hope included in the final paragraph of a grant application. Even the substantial advances in mapping made after the development of the somatic-cell hybridization technique failed to change the situation very much as most of the loci mapped were not polymorphic. It was only the vast reserves of variation unleashed by recombinant DNA technology that realised Bell & Haldane's anticipated result, and have brought the jargon of the trade, 'the lod score', from the very appendixes of textbooks to the market place of *The Lancet*. The last report of the Human Gene

Mapping Workshop 8 (1985) gave information on almost 1500 loci, and if the speed of advance continues only on the curve established between 1981 and 1985, we might expect to find approximately 3000 by the time of the next Workshop (September 1987). Views on the numbers of basic genetic functions, or gene clusters, are falling all the time: current estimates are of the order of 10000, and it is really remarkable that we are possibly within an order of magnitude of this number in mapping information, approximate as it is in most cases.

So what is a speaker charged with a general review to do? Summaries are already too long, and lists too dull, and so I shall attempt to examine present methods as a basis for the detailed examples to be given by other speakers. I shall not discuss purely molecular approaches because these are dealt with by Southern (this symposium); I shall try to concentrate on how the gene map has provided a starting point in the pursuit of disease genes of unknown structure and function, and in the context of this meeting not for general interest, but with the objective of their elimination, correction or confusion of these genes.

Classical methods

There are basically only two approaches to gene mapping: classical pedigree analysis which estimates genetic distance by measuring recombination; and experimental techniques which aim to locate genes by isolating individual chromosomes either in cells, or free of cells. Table 1 presents one way of looking at the relative merits of the different procedures.

Table 1. Comparison of different methods of gene mapping in relation to polymorphism

	families	individuals		
	meiosis = linkage groups	somatic cell hybrids = localization (bands)	chromosomes = localization *in situ* (bands)	
polymorphic	'characters'	—	—	
	biochemical markers (blood)	biochemical (cultured cell)	—	
	DNA markers	DNA	DNA	
non-polymorphic	—	biochemical and DNA	DNA	

Classical pedigree analysis requires essentially two things: genetical variation and a reasonably sized sample of the products of meiosis. Variation was first studied at the level of the phenotype ('characters'). In the case of Bell and Haldane these were colour-vision tested by a functional test, and an observable tendency to bleed. As our understanding of the 'character' increases it may pass through the category of biochemical marker (e.g. when the mental defect of Folling syndrome became phenylketonuria) and ultimately into the DNA category: or the more recalcitrant may remain as 'characters'. I have used the category of biochemical markers to represent all those polymorphisms revealed by immunology and enzymology, etc., which arise from technical developments such as electrophoresis or the production of monoclonal antibodies.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

The practical limitations on the markers useful for linkage used to be the availability of a suitable tissue source. This was generally blood, because saliva, urine and hair follicles, though non-invasive, are very tedious for both collectors and collected. Given testable markers, the remaining requirement is informative meioses and these depend on the level of heterozygosity and family size. The vagaries of human pedigree structure made methods of analysis less straightforward than in *Drosophila* but general agreement emerged on the analysis of pair-wise data, and linkage groups began to emerge, involving all types of marker.

It is difficult to assign linkage groups to particular chromosomes by using this method because the only approach is to substitute a morphological chromosome marker as one of the 'characters', and such markers are too uncommon for the method to be generally useful. The other absolute defect of the method is the need for variation and it is in both these respects that somatic cell hybridization proved to be so valuable. The data obtained relate to the correlation between gene products and the presence in the hybrid cells of particular human chromosomes, and are not dependent on variation within the species, but only on a difference between man and mouse, or hamster, a much more common event. The results are expressed in different units: not now in centimorgans† (cM) but in descriptive terms related to the banding patterns of the chromosomes. The technique of *in situ* hybridization gives essentially the same type of information but is restricted to cloned genes. More recent methods permit the manipulation of individual chromosomes outside cells and the general usefulness of 'flow-sorting' is discussed by Ferguson-Smith (this symposium), but in this context cloned sequences can be localized by dot–blot analysis to sorted chromosomes and the sorted chromosomes can be used for the manufacture of chromosome-specific probes. In essence, the information in all these categories is the same and gives localization at various levels of precision about metaphase bands.

It is obvious that the information obtained by the different methods must be pooled in a way that uses the various strengths and obviates the weaknesses. This was done first by achieving the assignment of linkage groups by the localization of one member, maybe an enzyme that could be studied by somatic cell hybrids so that eventually the movement of information between the cells of table 1 resulted in the first real genetic maps, which were surprisingly good. The map of chromosome 1 produced by P. Cook is still recognizable as the basis of the present map. But there are considerable deficiencies and I now look on how these are being handled. Space does not permit a discussion of the current technical improvements in relation to somatic cell hybridization but they do not affect the essential nature of the results produced. Refinements of analysis can be obtained using the vast storehouse of information collected by cytogeneticists about translocations. By selecting individuals with chromosomes rearranged in a defined manner we can obtain regional assignments much narrower than can be obtained on normal chromosomes, given the physical limitation of the method. Even so, the resolution of such methods is in terms of metaphase bands, at least an order of magnitude greater than 1 cM and so for monomorphic loci there appears to be a serious limitation on fine scale mapping. Efforts must therefore be directed towards reducing this class of loci and in fact this has been achieved at the DNA level with the discovery of the enormous amount of naturally occurring variation in both coding and non-coding sequences, and so the only unfortunate genes left in this category are those for proteins where the gene has not yet been cloned.

† The morgan is the unit of relative distance between genes on a chromosome. One centimorgan represents a crossover value of 1%.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Recombinant DNA methods

The first shift from non-polymorphic to polymorphic used the restriction fragment length polymorphisms (RFLPs) which at their simplest reflect base pair changes which create or destroy a cleavage site for a specific restriction enzyme, causing a change in the length of a DNA fragment. A marker locus based on a single such variant will have only two alleles (a site being either present or absent) and so limiting heterozygosity to 50% at best. Several restriction site polymorphisms may be detected within, or very close to, a gene, so increasing the heterozygosity, but the additional information will be limited by the amount of linkage disequilibrium encountered.

In contrast, a few DNA marker systems reveal a DNA restriction fragment of extremely variable length. The first report of a polymorphic locus discovered with an arbitrary DNA probe by Wyman & White (1980) described just such a system, with 15 alleles. Subsequently, similar systems were observed that all consisted of a set of tandem repeats of a short (11–60 base pair (b.p.)) oligonucleotide sequence, and where the length of the restriction fragment is a function of the number of copies of the repeat present in the fragment. Such a system is highly informative, with heterozygosities much higher than 50%. The first success obtained in a random linkage search with such a probe was that linking polycystic kidney disease with the α-globin locus by Reeders *et al.* (1985). Using a sequence 8 kilobases downstream (3′) from the α-globin gene cluster they found significant evidence for close linkage. This initial assignment focused attention on other markers on the short arm of chromosome 16 and even closer linkage was demonstrated with the red-cell enzyme locus *PGP* (phosphoglycollate phosphatase), a much less informative locus with only 30% heterozygosity, where, of course, many more families were required to obtain a meaningful answer (Reeders *et al.* 1986). During our earlier studies on *PGP* we were unable to test the predicted relation between *HBA* and *PGP* because of the lack of variability in α-haemoglobin in this country, so this clinical study was the occasion for completing a logical piece of the map, though not part of the primary aim. This illustrates the opportunistic nature of much of the work which has in fact proved to be quite successful in human gene mapping.

The deliberate development of highly informative probes has arisen out of Jeffreys' myoglobin sequence which has come to be associated with the idea of 'DNA finger-printing' (Jeffreys *et al.* 1985). The 'finger-print' is the sum of the variation at a large number of loci, all of which show a high level of heterozygosity, so the exact interpretation of any one pattern is extremely difficult. However, Wong *et al.* (1987) have now cloned and characterized DNA from six of the most variable sites which can serve as locus-specific probes under high stringency conditions. The probes have been assigned by using a somatic-cell hybrid panel. Fortunately, they turn out to be well dispersed, on 1p, 1q, 5, 7p, 7q and 12. As these are fairly representative of the larger chromosomes it suggests that such loci are randomly distributed over the whole genome, which is useful, although in the absence of regional assignments it is still possible that they may be preferentially located in regions known to be rich in simple-sequence satellite DNA, the centromeres and telomeres, which would severely limit their value. These new locus-specific mini-satellites act as very sensitive hybridization probes and can be pooled to detect several loci simultaneously. One reservation remains to be assessed and this arises from the nature of the sequences, because it seems likely that the tandem repeats increase and decrease in number through a high frequency of unequal crossing over and the sequences may indeed be hot spots for recombination.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

If there are no other unforeseen snags in these ideal probes, it seems that there are certainly sufficient of them to provide the whole set of mapping points needed to define the genome. R. White and his group have now described 77 new loci which were developed by probing human genomic libraries with a set of synthetic oligonucleotides corresponding to the consensus sequences of the tandem repeats (Nakamura *et al.* 1987). Apparently, despite the similarity of sequence, there is sufficient variation at different loci for hybridization at high stringency to permit unique identification of a single locus, only a small proportion of clones showing overlap. The heterozygosities vary between 33 and 95%, the level of heterozygosity of the most informative group agreeing well with the criteria selected for the six loci characterized by Wong *et al.* (1987). Seventeen of these new loci have been tested in the CEPH families, and although the full details have not yet been made public they are said to be distributed over at least 10 chromosomes.

I cannot leave the subject of hypervariability without stressing its importance in making possible the identification of all four parental chromosomes in tracing the origin of aneuploidy, for instance, and in deriving mapping information from chromosomal rearrangements either occurring in germ cells or malignant somatic cells. Again, a logical but previously limited approach has become a generally informative technique because of the increased levels of heterozygosity revealed by each new generation of markers.

The consequences of increasing heterozygosity have not only increased the amount of information that can be extracted from most families, but they have stimulated a complete re-examination of the general strategy. Given an unlimited supply of markers, one can choose the most informative set for both degree of heterozygosity and location, and so produce a network of fixed points against which to test any unknown locus. It is easy enough to say, but rather a lot of work, and perhaps because of this, ideas of collaboration have begun to be heard. The best families for this type of work are quite different from the usual source material, which normally consists of families selected for a special locus such as disease, for if all families are informative one can attempt to maximize the other limiting variables, family size and knowledge of phase, and two groups in particular have taken this road. In Paris, J. Dausset built up a collection of families for his HLA work where there were three complete generations, i.e. four grandparents, two parents and about eight children, and the Utah group had used the extensive Mormon families in a similar way. Obviously the idea was not new: such an approach could have been used for all the previous common polymorphisms. However, other factors leading to the change of attitude were that DNA can be stored much more readily than cells, or even haemolysates for enzyme or blood-group determinations, and the improving efficiency of making immortalized cell lines has meant that a permanent resource can be established, and shared, so that each laboratory can limit its work load to its particular interests while contributing to the general picture. This collaboration was established two years ago by Dausset under the title of CEPH, and has distributed DNA from 40 families to over 20 collaborating groups. The first round of data exchange has just been completed and has provided information on 172 loci. The sharing of probes is also proceeding apace and apart from exchanges between individuals, the documentation is undertaken by the Human Gene Map Workshop and the Yale Gene Library, and the American Type Culture Collection has been commissioned by the National Institutes of Health to act as a Repository of DNA probes and libraries. The holdings at the end of March consisted of over 300 probes, half of which are already available on request, and most of which have fairly good regional assignments.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Problems and future developments

Although the skeleton of the map of reference points can be constructed from these selected families, the location of genes with low heterozygosity will still depend upon relatively uncommon families with their own problems such as small family size, death before maturity, delayed age of onset and so on, but significant answers will be more readily obtained on such limited material because more exact hypotheses can be constructed by using the established framework.

Are there any other ways of maximizing information from special families? Occasionally, information on phase would be extremely valuable, but the parents are dead or uninformative. In such cases the two homologous chromosomes can be isolated in somatic-cell hybrids and phase established in this way. The source of all our problems is that in higher plants and animals meiotic products cannot be recovered in their original tetrads and so one is attempting to investigate linkage by sampling random meiotic products. In the absence of a free-living haploid phase, the gametic genotype must be determined retrospectively by the examination of the succeeding diploid generation – after many years of amplification – but perhaps we would be better employed learning how to convert back to a robust haploid phase rather than maintaining diploid cell banks for linkage studies.

If things look fairly straight forward at the level of DNA, and rather more difficult at the level of expressed products, we have a collection of maximum difficulty in our 'characters'. For diseases remaining in this category the nature of the disease has not yielded to rational biochemical approaches, so that linkage has been hailed as the only technique which could point to the gene, by the back door approach. This is where much of the effort of the last few years has been applied, and success obtained, but this is the business of later papers. Success has been largely in those cases where there was clearly a Mendelizing gene, even if of unknown function, but this success has persuaded people into adopting the same approach for conditions where the genetics is not so clear: heart disease, mental disease, and even cancer. Sophisticated analytical programmes have been designed to tackle these problems, but recent reports of positive findings have been largely based on selecting, from the range of non-Mendelizing families, those families that seem to fit into a simple mode of inheritance. The work on Alzheimer's disease and chromosome 21 provides a remarkable example (St George-Hyslop *et al.* 1987). Although such bias presents major problems in general mapping, the clues offered may provide valuable pointers for the molecular approach.

Complex phenotypes include many of the morphological variants that the public constantly enquire about, such as eye and hair colour, or shape of nose, all so obviously inherited but which defy exact definition. Penrose (1950) reported an analysis of what he called 'striking' red hair and, by using the sib pair analysis to get round the problem of darkening with age, he made an estimate of 10% recombination with the ABO blood-group locus. Curiously enough, the latest report on the topic is by J. Mohr, who was a student of Penrose at that very time. Mohr now has evidence that green eyes and brown hair are linked to *lutheran* and *secretor*, and weaker evidence for brown eyes linked to *ESD* (esterase D) (Eiberg & Mohr 1987). So chromosomes 9, 13 and 19 are involved in appearance! At the beginning I suggested that we may have information on a non-trivial proportion of all human genes in numerical terms, but it now begins to look as though we are sampling a broader range of genes than might have

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

seemed possible only a few years ago, and moving very rapidly from the anonymity of DNA into characters long neglected.

In discussing the material most relevant to today's topic I have not touched upon the real meat of gene mapping: the combination of the locations and linked pairs into maps. The aim of most of the work described is to go ever inwards, whereas genuine mappers want always to extend their frontiers. There is as yet no real agreement on how best this type of mapping can be achieved because the statistical problems of ordering are not very clear, and the amount of computation heavy. Although there are plenty of theoretical papers to show that the joint analysis of three or four markers is better than a simple pair-wise approach, and that this should be possible in the ideal families of CEPH for DNA markers, it is not quite so straightforward in a complete chromosome analysis where there is a need to fit in less variant loci that will only rarely be segregating with more than one other marker. For the moment, maps are being constructed on largely intuitive lines combining three or four loci multipoint analysis orders, with constraints imposed by regional assignments, and distances estimated by 'lod' scores for other loci. Many of the known biological problems are ignored in bringing the scale of the analysis into the realms of the possible, and indeed the sensible, in terms of the relatively small amount of data as yet available. Interference is often neglected, as is the sex differential in recombination. The latter is almost certainly serious because there is generally more crossing over in females than in males, but as tolerably reliable estimates vary up to a factor of 5 it is not yet clear whether or not there are consistent local differences. It is also of course accepted that there is one correct order. However, man is not an inbred line and because large inversions occur with recognizable frequency perhaps small inversions are sufficiently common to be a nuisance. And finally there is the assumption that recombination occurs at random. At the cytological level there is little unambiguous evidence in man about the distribution of chiasmata, whereas most of the molecular evidence points to hot spots, regions of the order of 1000 base pairs where recombination is at least 5–10 times more frequent than elsewhere. It could be that there are about 5000 hot spots in the human genome, and so perhaps a distance between them of 600000 b.p. Fortunately, if this is the right order of magnitude and the hot spots are at random, there would be reasonable agreement between observed recombination fractions and physical distance measured in nucleotide pairs as long as distances are large compared with the average distances between hot spots, ideally over 2–3 cM. That this is so is perhaps supported by the fact that family studies of recombination within the HLA locus suggested a length of about 2 cM whereas the estimate of size obtained by pulsed-field electrophoresis is of the order 3×10^6 b.p.

I have attempted to show that logical strategies for mapping the human genome are now available but for those of us who perhaps feel that the excitement of the chase has been lessened, let me end by drawing attention to a recent result on tuberous sclerosis. Instead of sitting back and awaiting the day when the gene could be neatly slotted into a map consisting only of anonymous reference points, Fryer *et al.* (1987), by using classical markers only, have assigned the tuberous sclerosis gene to the tip of 9q by linkage to the ABO blood-group locus, the oldest polymorphic locus of all, even antedating Bell and Haldane.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

References

Bell, J. & Haldane, J. B. S. 1937 The linkage between the genes for colour-blindness and haemophilia in man. *Proc. R. Soc. Lond.* B **123**, 119–150.

Cook, P. J. L., Robson, E. B., Buckton, K. E., Jacobs, P. A. & Polani, P. E. 1974 Segregation of genetic markers in families with chromosome polymorphisms and structural rearrangements involving chromosome 1. *Ann. hum. Genet.* **37**, 261–274.

Eiberg, H. & Mohr, J. 1987 Major genes of eye colour and hair color linked to LV and SE. *Clin. Genet.* **31**, 186–191.

Fryer, A. E., Chalmers, A., Connor, J. M., Fraser, I., Povey, S., Yates, A. D., Yates, J. R. W. & Osborne, J. P. 1987 Evidence that the gene for tuberous sclerosis is on chromosome 9. *Lancet* i, 659–661.

Human Gene Mapping Workshop 8 1985 *Cytogenet. Cell Genet.* **40**.

Jeffreys, A. J., Wilson, V. & Thein, S. L. 1985 Individual-specific 'finger prints' of human DNA. *Nature, Lond.* **316**, 76–79.

Nakamura, Y., Leppert, M., O'Connell, P., Wolff, R., Holm, T., Culver, M., Martin, C., Fujimoto, E., Hoff, M., Kumlin, E. & White, R. 1987 Variable number of tandem repeats (VNTR) markers for human gene mapping. *Science, Wash.* **235**, 1616–1622.

Penrose, L. S. 1950 Data for the study of linkage in man: red hair and the ABO locus. *Ann. hum. Genet.* **15**, 243–247.

Reeders, S. T., Breuning, M. H., Davies, K. E., Nicholls, R. D., Jarman, A. P., Higgs, D. R., Pearson, P. L. & Weatherall, D. J. 1985 A highly polymorphic DNA marker linked to adult polycystic kidney disease on chromosome 16. *Nature, Lond.* **317**, 542–544.

Reeders, S. T., Breuning, M. H., Corney, G., Jeremiah, S. J., Meera Khan, P., Davies, K. E., Hopkinson, D. A., Pearson, P. L. & Weatherall, D. J. 1986 Two genetic markers closely linked to adult polycystic kidney disease on chromosome 16. *Br. med. J.* **292**, 851–853.

St George-Hyslop, P., Tanzi, R. E., Polinsky, R. J., Haines, J. L., Nee, L., Watkins, P. C., Myers, R. H., Feldman, R. G., Pollen, D., Drachman, D., Growdon, J., Bruni, A., Foncin, J.-F., Salmon, D., Frommelt, P., Amaducci, L., Sorbi, S., Piacentini, S., Stewart, G. D., Hobbs, W. J., Conneally, P. M. & Gusella, J. F. 1987 The genetic defect causing Alzheimer's disease maps on chromosome 21. *Science, Wash.* **235**, 885–890.

Wong, Z., Wilson, V., Patel, I., Povey, S. & Jeffreys, A. J. 1987 Characterization of a panel of highly variable mini satellites cloned from human DNA. *Ann. hum. Genet.* **51**, 269–288.

Wyman, A. & White, R. 1980 A highly polymorphic locus in human DNA. *Proc. natn. Acad. Sci. U.S.A.* **77**, 6754–6758.

Discussion

Kay E. Davies (*Nuffield Department of Clinical Medicine, John Radcliffe Hospital, Oxford, U.K.*). What degree of resolution of the gene map should be aimed for?

E. B. Robson. This topic is considered in Professor Bobrow's summing up (this symposium), when Professor Southern's detailed account of the present state of molecular analysis is considered in relation to the achievements of family studies and cell biology that I have described.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Addendum. Total numbers of genes (*a*) and cloned DNA segments (*b*) assigned to individual chromosomes

(Based on the data submitted to Human Gene Mapping Workshop 9 (Paris, September 1987)). The entire X-chromosome and approximately 60 % of the autosomal length (estimated by using the physical assignments of the outer loci of linkage groups) have now been included in 22 linkage groups by using just over 450 marker loci, mostly in category (*b*).)

Total numbers of genes (*a*) and cloned DNA segments (*b*) assigned to individual chromosomes

chromosome	*a*	*b*	chromosome	*a*	*b*
1	128	73	13	22	49
2	82	56	14	40	19
3	50	34	15	37	33
4	51	122	16	45	90
5	48	74	17	52	71
6	92	38	18	19	22
7	68	395	19	57	25
8	41	37	20	22	19
9	50	27	21	28	113
10	40	21	22	35	20
11	102	285	X	137	286
12	60	31	Y	12	62
			totals	1318	2002

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES