PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Prospects for a Complete Molecular Map of the Human Genome [and Discussion]

E. M. Southern and H. Sharma

Phil. Trans. R. Soc. Lond. B 1988 **319**, 299-307
doi: 10.1098/rstb.1988.0051

Phil. Trans. R. Soc. Lond. B **319**, 299–307 (1988)
Printed in Great Britain

Prospects for a complete molecular map of the human genome

By E. M. Southern, F.R.S.
Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, U.K.

Linkage maps are limited by the number of recombinations that can be scored in human pedigrees to a resolution of *ca.* 1 centimorgan (relative distance between genes on a chromosome having a crossover value of 1%) which is estimated to be about 10 megabases. Molecular maps can be formed at any resolution down to the base sequence. To complement the linkage approach, the most useful molecular map would be one that helped to locate disease loci, by using restriction fragment length polymorphisms (RFLPs) and accurate localization of recombinations, and which then helped to find candidate genes in this region, by providing the positions of coding sequences. This paper discusses the appropriate form and scale of such a map, how it can be produced with methods now available, and the most efficient strategy for building the map, based on present knowledge of the organization of the human genome.

Introduction

The potential value of a complete molecular map of the human genome to studies of human inherited disease is well illustrated by other papers in this symposium (Tsui *et al.*; Worton *et al.*; Caskey *et al.*; Weatherall *et al.*; Harper *et al.*). These studies illustrate several important aspects of the molecular approach: the power of linkage analysis using restriction fragment length polymorphisms (RFLPs) as markers to locate a disease locus to a limited region; the application of linked probes in genetic counselling; further refinement of the map at the molecular level to locate the affected gene; and finally, the characterization of defects in the gene that lead to the genetic disorder. At present, only a few diseases are amenable to these approaches, though these are the most prevalent diseases. Without help from chromosomal defects, the task of locating a gene with no characterized product is a large one. It is now clear that two steps in the process require a great deal of effort using present techniques and resources. The first is the initial localization when this is done by linkage analysis; indeed it may be impossible to achieve linkage by random probes if appropriate families are not available (Edwards, this volume). The second difficult step is the identification of the gene within the large region of the genome exposed by linkage. The idea of using a complete linkage map based on RFLPs to locate the position of disease loci was discussed by Botstein *et al.* (1980) At that time, the notion of building a molecular map of the genome was a daunting prospect, but recent technical advances, such as pulsed-field gel electrophoresis (Schwartz & Cantor 1984), cloning in yeast (Burke *et al.* 1987), and advances in our knowledge of the human genome, such as the discovery of the *Hpa*II tiny fragments (HTF) islands (Bird 1986), have made it possible to design parallel strategies for forming a physical map of the human genome. This paper discusses some such strategies which merge the two approaches, optimizing the search for genes, in particular those loci associated with inherited diseases.

Sequence types and the organization of the human genome

Some proposals for mapping the human genome are based on approaches that have been used successfully with less complex genomes. To take the most extreme example, sequence analysis, which has been successful in revealing the sequence organization of viral genomes, could be applied to the human genome if applied on a massive scale, and if ways could be found to overcome the problems presented by repeated sequences and polymorphism. However, although this approach would produce a wealth of fine detail, it is criticized because it is inefficient, as it is known that much of the human genome consists of sequences with no important genetic function, so before considering more efficient mapping strategies it is worthwhile examining our present knowledge of the human genome to help us to decide what more we need to know, and how to set about the necessary analysis.

Sequence types

It is convenient to divide the genome into three types which differ in their functional importance (table 1).

1. Coding sequences and their associated controls. There is a great deal of evidence to show that the majority of mutations in these sequences affect the fitness of the organism adversely. However, as we shall see below, estimates of the number of genes are much larger than the number of known genetic diseases, implying that most mutations are either lethal to the embryo, or irrelevant to genetic disease for some other reason.

2. Non-coding sequences, such as introns, the sequences that lie between genes and the simple-sequence or satellite DNAs that are found associated with constitutive heterochromatin. With few exceptions, changes in these sequences have no known effect. It is clear, however, that they diverge more rapidly than coding sequences, implying that they may be neutral (Cooper *et al.* 1985).

Table 1. Genetic importance of major sequence classes

sequence	approximate percentage	consequences of mutation
Genes	1–5	mainly deleterious
Introns and satellites	70	mainly neutral
Dispersed repeats (transposons?)	20	possibly advantageous

3. Mobile sequences, such as retroposons. Some of the dispersed repeated sequences in the human genome have properties that suggest that they derive from retrotransposons (Baltimore 1985; Deka *et al.* 1986). Such sequences are an important cause of mutation in *Drosophila melanogaster*, suggesting that the presence of such a sequence indicates absence of any important function for the region that contains it. Mutations that immobilize retrotransposons may increase the fitness of the organism.

Sequence composition and organization

Almost all coding sequences that have been studied are interrupted by introns. Genes such as γ-interferon, which has no intron, are in a minority. The size and number of introns vary

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

greatly from gene to gene. There are many genes in the size range 5–20 kilobases. However, there are examples of genes, such as that for clotting factor VIII (Gitschier *et al.* 1984) and for Duchenne muscular dystrophy (Monaco *et al.* 1985), well in excess of 100 kilobases in total length. Assuming an average of 30 kilobases for the total length of DNA occupied by a gene would give a total of 100000 genes in the human genome.

The only measurements of the amounts of DNA between genes come from analyses of gene families such as the globin genes discussed by Weatherall *et al.* (this symposium). These examples suggest that the length of DNA between genes may be longer than the total made up from introns and exons.

Simple-sequence DNAs are found in blocks that are often several megabases long. Most of these blocks are close to the centromeres, in the heterochromatin (see references in Beridze (1982)).

Dispersed repeats are present throughout the genome. There are two major families; the *Alu* family and the *Kpn* family which are present in high copy number (see references in Southern (1984)). Their significance for mapping is that they can be used as probes for human DNA sequences in contexts where other DNAs are present; for example, they can be used to detect human DNA in somatic cell hybrids.

Of greater importance for mapping, because they mark the position of sequence polymorphisms, are the so-called minisatellite sequences (Jeffreys *et al.* 1985). These sequences are tandemly repeated, and so able to expand and contract by unequal recombination. They belong to several related families. Used as probes, these sequences provide a DNA 'fingerprint'. To produce probes that detect alleles it is necessary to locate a flanking single copy sequence in a clone and to isolate this to make the probe. There are likely to be hundreds, if not thousands, of sites for minisatellites in the human genome, giving enormous potential for mapping the positions of recombinations in family studies.

The numbers used in this description of the organization of the human genome are very approximate, but even these make the important point that any mapping procedure that deals with sequences in units smaller than the size of the average gene will spend a lot of effort in the analysis of DNA that is probably unimportant. This is especially true of random procedures that treat all sequences as equal in importance. It follows that any method that focuses on the position of genes will be more efficient in producing the kind of map that is needed to locate genes involved in genetic disorders.

It is evident that probes made from copy DNAs (cDNAs) are useful for finding their corresponding gene. A less obvious way of locating the positions of a subset of genes has been suggested by Bird (1986), Brown & Bird (1986) and Lindsay & Bird (1987) from studies of DNA methylation.

DNA methylation and HTF islands

DNA methylation has become an area of intense study. It has importance for the control of gene expression, and the work of Bird and his colleagues has shown how it can be exploited for mapping the genomes of higher eukaryotes. They discovered a set of sequences that are degraded to small fragments by the restriction enzyme *Hpa*II; they called these sequences HTF islands (for *Hpa*II tiny fragments). These sequences are distinguished from the bulk of human DNA in that they are high in G+C (*ca.* 60% against *ca.* 40%), and they have the high content of the doublet CpG that is predicted from their base composition (table 2). The C of the CpGs in the islands is not methylated, and it is this that makes the islands sensitive to a class of restriction

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

TABLE 2. METHYLATED SEQUENCES AND HTF ISLANDS

	bulk of euchromatic DNA	HFT islands
G+C(%)	40	65
CpG	very low content C methylated	expected content C non-methylated
Length	50–100 kilobases	1 kilobase

Other properties of HTF islands: *ca.* 30000 in the genome; targets for 'CpG' restriction enzymes; mainly single copy; mark the 5′ ends of some genes.

enzymes that have CpG in their recognition site. The cutting action of most of these enzymes is blocked by methylation of the CpG. By contrast with the HTF islands, the bulk of human DNA is heavily methylated at CpG, and the CpG doublet is present at only one fifth the level predicted from the base composition. Most importantly, where the HTF islands have been thoroughly characterized, they have been shown to be close to the 5′ ends of coding sequences. From the size of the fragments produced from total human DNA, which are in the range from 50 to more than 1000 kilobases, it has been estimated that there are a few tens of thousands of HTF islands in the human genome. The distribution is not even. There are regions where HTF islands are clustered, for example the ribosomal genes have all the properties of HTF islands, and islands are frequent in the region of the α-globin genes and the major histocompatibility complex (MHC) locus; some regions, such as the Duchenne muscular dystrophy (DMD) locus, have very few HTF islands over several hundred kilobases (Burmeister & Lehrach 1986). Nevertheless, the HTF islands provide a means of cutting the DNA at infrequent sites, and in addition they point to the sites of important sequences, perhaps quite a high proportion of genes (Lindsay & Bird 1987).

MAPPING STRATEGIES: PROBES, CLONES, RESTRICTION FRAGMENTS AND GELS

Mapping requires a way of fragmenting the genome and a way of detecting overlaps between the fragments. For a restriction map the final product is a linear display of part of the genome, which is scaled by the positions of the restriction sites, and which includes features such as coding sequences determined by hybridization to restriction fragments. Maps can also be formed from a set of overlapping clones. These have the advantage of providing any sequence of interest in bulk and in pure form for further analysis or manipulation.

Fragmentation and separation

Restriction enzymes provide the main method of site-specific cleavage, and can be used to produce fragments in any size-range from 100 or so base pairs (b.p.) up to more than one megabase. The principal methods of separating fragments are gel electrophoresis, which gives an accurate measure of fragment size (Elder & Southern 1983; Schwartz & Cantor 1984), and cloning. Both methods span the range from a few hundred base pairs to more than one megabase. Scale is emphasized because it determines the amount of effort needed to produce the map and the amount of detail that the map provides.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Probes

Fragments may be linked together to form a map in a variety of ways. For very small fragments, sequence analysis is an appropriate method, but for the scale of map considered here, the choice of method is either to determine restriction-site maps for clones and then to overlap the clones by finding overlapping restriction-site maps, or to locate sequences in the fragments, separated either by gel electrophoresis or cloning, by molecular hybridization using cloned probes. The latter approach has the advantage that probes can be chosen that not only help to form the map, but also place important features in the map.

For mapping the human genome, three kinds of probe are important.

1. Probes for RFLPs, especially those which find high levels of polymorphism (Botstein *et al.* 1980) such as those associated with minisatellite sequences (described by Jeffreys *et al.* (1985)) are used in linkage mapping. Indeed, it is possible that a linkage map of probes derived from minisatellites will be produced before a physical map. Such a map would provide an ideal set of probes for linking up a coarse physical map, though recombination can never produce the number of break points needed for high-resolution mapping (Robson, this symposium).

2. HTF islands locate the positions of a major subset of genes, and also link together fragments produced by cutting the genome with enzymes that have sites in these islands (Brown & Bird 1986).

3. cDNA clones locate the positions of coding sequences.

The additional practical advantage of cDNAs and HTF island sequences is that they are mainly single-copy sequences, so there is no need to remove repeated sequences before they can be used as hybridization probes (Sealey *et al.* 1985).

What is the appropriate scale for a map of the human genome?

For the applications that are being discussed at this meeting, there are two important attributes required of a molecular map. First, it should help in the accurate location of the recombination points in families where recombination has been seen between a disease locus and a linked polymorphic probe. The more accurately this point can be located the easier it will be to move on to the next stage, which is location of candidate coding sequences that may be implicated in the disease. The second important attribute of the map is that it should help in the location of these candidate sequences.

As we have seen, the average distance between genes, taking account of introns and intergenic DNA, is likely to be tens of kilobases. Ideally, a map that will establish the position of all exons should therefore be on this scale. However, such a map on this scale would be made up of several hundred thousand components, and would take a great deal of work to assemble. A map at the 100–1000 kilobase level would be a more realistic first objective.

The potential of a megabase map

A physical map at the megabase level of resolution would help to locate the positions of recombinations ten times more accurately than it is possible to achieve by measuring genetic distance between markers and using statistical methods (figure 1). Once an area of interest had been established at this coarse level, it would be possible to look at the region in much greater detail. This is a generalization of the approach that has been taken with the genes for cystic

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

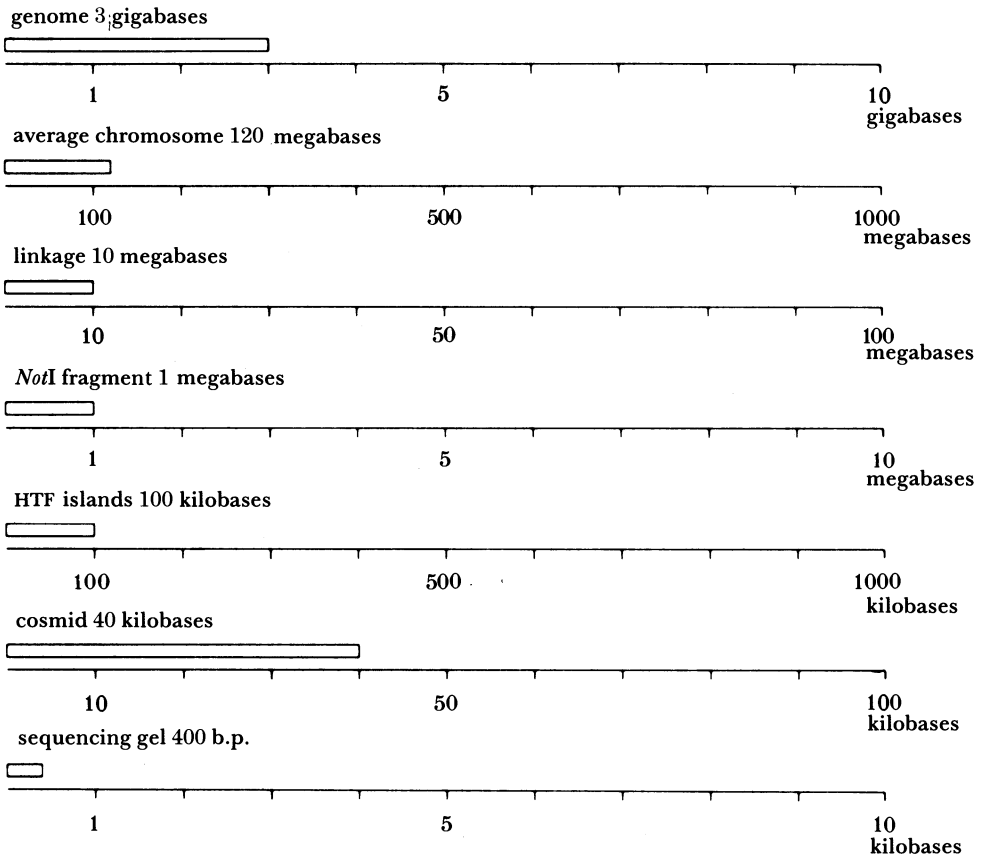

FIGURE 1. The sizes of the human genome, an average human chromosome, of length 1 cM†, an average *Not*I fragment, the average spacing of HTF islands, a typical cosmid clone, and the amount of sequence that can be derived from a single sequencing gel are represented by the horizontal bars. Note that each scale differs by an order of magnitude from the scale above and below it.

fibrosis (Tsui *et al.*, this symposium), and for Huntington's disease (Gusella *et al.* 1983; Harper *et al.*, this symposium). The work that was needed to establish linkage in those cases would have been much reduced if a map of the whole genome had been available, because information from all meioses could have been used to exclude regions as well as to point to potential locations of the gene. But perhaps more importantly, the work that remains to be done to find the precise positions of recombinations in the chromosomes would already have been done, and the positions of candidate genes would already have been established.

Strategy for producing a megabase map

The method of choice for producing fragments in the size range 100–1000 kilobase is digestion with infrequently cutting restriction enzymes. These fragments can be separated and measured by pulsed-field electrophoresis. As has already been argued, the most useful hybridization probes for putting the map together are those based on HTF islands, cDNA clones and minisatellites. The number of analyses that will be needed to produce a complete map of the human genome at this level is quite large, as can be seen from the recently published

† 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

mapping of the *Escherischia coli* genome, which is smaller by three orders of magnitude (Smith *et al.* 1987). Methods for cloning fragments in yeast are being developed that will permit cloning in this size-range. Such methods will undoubtedly make the task of mapping complex genomes much easier (Burke *et al.* 1987).

Searching for genes in local regions

Given a complete map of the human genome at the megabase level, the task of locating a disease locus within a region defined by a linked RFLP is still a large one. There are several approaches that can be taken. Mutations such as deletions and inversions, which would not have been picked up by cytogenetic analysis, may be revealed by restriction-site mapping. A map on the scale of 100–1000 kilobases would detect these if they were several kilobases long by using the relatively simple method of pulse-field gel electrophoresis, and though this approach is able to detect a large proportion of mutations in the DMD locus, it is unlikely that mutations in all disease loci will be detected in this way. It will then be necessary to isolate coding sequences in the region and examine them for mutations. However, as linkage may narrow the location of the gene to a region as big as 10 megabases, there could be hundreds of coding sequences to search. Several methods have been used to locate coding sequences. The first, and the easiest, is to examine all the HTF islands in the region. The second is to establish a set of overlapping clones, in phage or cosmids so far, but perhaps presently in yeast (Burke *et al.* 1987), and then to examine these clones for sequences that are conserved between species (Monaco *et al.* 1985), or for sequences that are transcribed. The task of locating coding sequences would obviously be greatly simplified if a global map were available that included all these features. Present methods for locating all coding sequences are unreliable and involve a great deal of effort; there is a need for new approaches. There is also a need for new methods of examining coding sequences for mutations.

Present state of the molecular map

Several regions of the human genome have already been mapped with some of the approaches in this paper because of interest in specific genes. These regions include several segments of the X- and Y-chromosomes, and chromosomes 4, 6, 7, 11, 13, 16, 19, 21 and 22. These mapping exercises have already produced detailed maps of a substantial portion of the human genome, and it is this experience that suggests the optimum strategy for extending the map to the whole of the genome will be one that helps with the accurate location of recombination points, and helps to locate the coding sequences in the region between linked probes.

Conclusion

The resources needed to produce a map of the type suggested in this paper, are:

(*a*) a complete set of cloned probes for HTF islands and their flanking sequences;

(*b*) a set of around 3000 polymorphic probes;

(*c*) well characterized cDNA libraries, to provide as complete a representation of the genome as possible.

Most of these are being produced in various laboratories, but coordination of the effort

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

would undoubtedly lead to more efficient use of these resources, as has happened with the use of other similar resources such as families, and chromosome-specific libraries. If effort is to be put into a global sequencing project, a good starting point would seem to be the coding sequences, as represented in cDNAs, because this information will be directly relevant to many aspects of biology, including the analysis of human disease.

Methods for producing various kinds of map are well established: the production of large fragments and their separation by pulsed-field gel electrophoresis, and the analysis of families by polymorphic probes could all be used without further development to produce a map on whatever scale is chosen. Future cloning of large fragments in yeast should increase the speed and the power of analysis as cloning in *E. coli* did the analysis of smaller sequences. But it should be borne in mind that even the relatively approximate level of resolution discussed here represents a major undertaking. There is no escaping the fact that the human genome carries several thousand genes, and many analyses will be needed to locate them all. The procedures used for mapping are not the kind that lend themselves to automation, and without automation, or radically new methods, there is a choice of strategies, a choice that has important implications for the way in which research in human biology is organized and funded. There is no doubt that coarse maps, at the 1 cM and 1–10 megabase levels, will be produced by linkage and physical methods of the kind discussed in this paper. From these coarse maps, should one proceed directly to a high resolution global map, which will make the analysis of each disease easier, or is it better to carry on building the detailed map around each disease locus as they are brought into view by linkage analysis? There is no doubt that starting with a global map is the more efficient strategy, but those working in the field are aware that the motivation for mapping comes from interest in individual loci.

References

Baltimore, D. 1985 Retroviruses and Retrotransposons. *Cell* **40**, 481–482.

Beridze, T. 1982 *Satellite DNA*. Berlin, Heidelberg, New York and Tokyo: Springer-Verlag.

Bird, A. P. 1986 CpG-rich islands and the function of DNA methylation. *Nature, Lond.* **321**, 209–213.

Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. 1980 Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. hum. Genet.* **32**, 314.

Brown, W. R. A. & Bird, A. P. 1986 Long-range restriction-site mapping of mammalian genomic DNA. *Nature, Lond.* **322**, 477–481.

Burke, D. T., Carle, G. F. & Olson, M. V. 1987 Cloning of large DNA segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science, Wash.* **236**, 806–812.

Burmeister, M. & Lehrach, H. 1986 Long-range restriction map around the Duchenne muscular dystrophy gene. *Nature, Lond.* **324**, 582–585.

Cooper, D. N., Smith, B. A., Cooke, H. J., Niemann, S. & Schmidtke, J. 1985 An estimate of unique sequence DNA heterozygosity in the human genome. *Hum. Genet.* **69**, 201–205.

Deka, N., Paulson, K. E., Willard, C. & Schmid, C. W. 1986 Properties of a transposon-like human element. *Cold Spring Harb. Symp. quant. Biol.* **51**, 471–477.

Elder, J. K. & Southern, E. M. 1983 Measurement of DNA length by gel electrophoresis II: comparison of methods for relating mobility to fragment length. *Analyt. biochem.* **128**, 227–231.

Gitschier, J., Wood, W. I., Goralka, T., Wion, K., Chen, E., Eaton, D., Vehar, G., Capon, D. & Lawn, R. 1984 Characterisation of the factor VIII gene. *Nature, Lond.* **312**, 326–330.

Gusella, J. F., Wexler, N. S., Conneally, P. M., Naylor, S., Anderson, M. A., Tanzi, R. E., Watkins, P. C., Ottina, K., Wallace, M., Sakaguchi, A., Young, A., Shoulson, I., Bonilla, E. & Martin, J. B. 1983 A polymorphic DNA probe genetically linked to Huntington's disease. *Nature, Lond.* **306**, 234–238.

Jeffreys, A., Wilson, V. & Thein, S. 1985 Hypervariable 'minisatellite' regions in human DNA. *Nature, Lond.* **314**, 67–71.

Lindsay, S. & Bird, A. P. 1987 Use of restriction enzymes to detect potential gene sequences in mammalian DNA. *Nature, Lond.* **327**, 336–338.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Monaco, A. P., Bertelson, C. J., Middlesworth, W., Colletti, C.-A., Aldridge J., Fishbeck, K. H., Bartlett, R., Pericak-Vance, M. A., Roses, A. D. & Kunkel, L. M. 1985 Detection of deletions spanning the Duchenne muscular dystrophy locus using a tightly linked DNA segment. *Nature, Lond.* **316**, 842–845.

Schwartz, D. C. & Cantor, C. R. 1984 Separation of yeast chromosome-sized DNA by pulsed field gradient electrophoresis. *Cell* **37**, 67–75.

Sealey, P. G., Whittaker, P. A. & Southern, E. M. 1985 Removal of repeated sequences from hybridisation probes. *Nucl. Acids Research* **13**, 1905–1921.

Smith, C. L., Econome, J. G., Schult, A., Kleo, S. & Cantor, C. R. 1987 A physical map of the *Escherichia coli* K12 genome. *Science, Wash.* **236**, 1448–1453.

Southern, E. M. 1984 DNA sequences and chromosome structure. In *Higher order structure in the nucleus* (ed. P. R. Cook & R. A. Laskey) (*J. cell Sci.* (suppl. 1)), pp. 31–42.

White, R., Leppert, M., O'Connell, P., Nakamura, Y., Julier, C., Woodward, S., Silva, A., Wolff, R., Lathrop, M. & Lalouel, J.-M. 1986 Construction of human genetic linkage maps: I. progress and perspectives. *Cold Spring Harb. Symp. quant. Biol.* **51**, 29–38.

Discussion

H. Sharma (71 *Barrack Road, Hounslow, U.K.*). Relying on data produced by various groups to study particular genes, rather than an organized programme to map human DNA, would provide data in early days where groups are motivated by problems of disease relevant to genes being studied. To generate data on genes not being actively studied, as part of a programme in particular disease, would require an organized effort.

E. M. Southern. Studies of local regions around disease loci have already made a significant contribution to the human gene map. A point I would like to emphasize is that the work that goes into the initial characterization of the chromosome carrying the gene provides resources that could be exploited to map the whole chromosome. Groups concentrating on a particular gene will of course narrow the focus of their work to the region of the gene. It would make efficient use of these resources if, in an organized programme, they could be handed over to those working towards a complete map.