The Turkish Journal of Pediatrics 33:1 - 12, 1991

GENE MAPPING USING LINKAGE ANALYSIS*

Özden Sanal MD**, Tatiana Foroud***, Richard A. Gatti MD****

SUMMARY: Sanal O, Foroud T, Gatti RA. (Department of Pediatrics, Hacettepe University Faculty of Medicine, Ankara, Turkey, and Departments of Biomathematics and Pathology, University of California, (UCLA), Los Angeles, California, USA). Gene mapping using linkage analysis. Turk J Pediatr 33: 1-12, 1991.

Linkage analysis is a recombinant technology used for gene mapping. If two genetic loci segregate together in a pedigree more often than by random chance, they are said to be linked, that is they lie close on the same chromosome. What makes the detection of linkage between two markers on a chromosome possible, are the recombination events which occurred during meiosis. Thus, information can be obtained for the location of a mutant gene by using markers with known locations.

Since the first published reports of linkage over thirty years ago, the use of the "lod score" method to determine the chromosomal location of a disease gene has become widespread. With the continual production of new RFLPs, the mapping of new loci will continue to become faster and more efficient. The power of the linkage method has spawned the development of several new techniques which utilize marker data to graphically represent the most likely location of a disease gene. Such advances, however, can best be utilized only after initial chromosomal localization of the particular gene through the lod score method. *Key words: linkage analysis, gene mapping, restriction fragment length polymorphism, southern blotting.*

Linkage analysis is an important tool for genetic mapping of polymorphism loci. By this approach, many disease genes have been localized to a specific chromosomal region. In at least one case, cystic fibrosis, this approach has played a direct role in the isolation of the gene. We have a similar interest in trying to "fine map" and eventually, to isolate the gene(s) for ataxia-telangiectasia[1-3].

A human cell contains 23 pairs of homologous chromosomes; one chromosome per pair is inherited from the mother and the other from the father. During the first phase of meiosis, the homologous chromosomes in a progenitor cell are duplicated and distributed to the germ cells, each of which contains 23 single chromosomes. The parental chromosomes, however, are not transmitted intact.

---

* From the Department of Pediatrics, Hacettepe University Faculty of Medicine, Ankara, and the Departments of Biomathematics and Pathology, University of California (UCLA), Los Angeles California, U.S.A.

** Professor of Pediatric Immunologist, Hacettepe University Faculty of Medicine.

*** Biomathematician, University of California, Los Angeles (UCLA).

**** Professor of Genetics Pathology, University of California, Los Angeles (UCLA).

This work was supported by the US. Department of Energy, the Ataxia-Telangiectasia Medical Research Foundation, the Joseph Drown Foundation and the Margaret E. Early Trust. Richard A. Gatti is a member of the Jonsson Comprehensive Cancer Center.

In the course of meiosis, homologous chromosomes exchange segments in a process termed recombination or crossing-over. Thus, each chromosome that is transmitted to a germ cell generally carries segments from both parental chromosomes. Chromosomes are covered with loci, which are like signposts along a road. Loci can be functional genes or RFLPs (Restriction Fragment Length Polymorphism). Any two loci from a given parent that appear to segregate together more than half the time are considered to be linked. Linked loci lie on the same chromosome. The concept of linked loci has been widely used in the study of inherited disease in order to establish linkage between a disease gene and another locus (marker).

Figure 1 illustrates the theory behind linkage with a disease gene. In this example, homologous chromosomes carry different alleles at two markers (A and B)[4]. One chromosome also bears a mutant allele causing disease (m) and the other chromosome bears the normal allele (+). In the precursor cell, the disease is associated with allele 1 of both marker A and marker B. After replication, the homologous chromosomes cross-over. Here crossing-over takes place between loci A and B. As a result of this exchange, two germ cells (a, d) carry the parental combinations of alleles, and two (b, c) contain recombinant chromosomes. In cell b, the mutant gene is still found with allele 1 at locus A but is now also on the same chromosome as allele 2 at locus B. After studying many instances of meiosis, if the frequency of cross-over is low between the disease gene and the marker A, then the disease gene and the genetic marker are closely linked. The probability that a cross-over will occur increases as the distance between the two genes increases.

One percent recombination corresponds to approximately one cento Morgan (cM) or 1×10^6 nucleotides. In other words, the genetic distance between two markers is defined as one cM when the markers have a probability of one in 100 of recombining at each generation. If a marker and a disease gene lie well apart on a chromosome, the recombination frequency will approach 50 percent. The marker and the gene are then unlinked. The same probability may be seen when a marker and a mutant allele are on different chromosomes since 50 percent approximates random chance.

An inheritable phenotype can serve as a marker (e.g HLA antigens or ABO blood groups). In recent years RFLPs have been used as markers for linkage analysis. A good marker for linkage studies should have high informativeness, codominant expression and simplicity of use and interpretation in the laboratory[5]. Here informativeness means the ability to determine if a recombination event has occurred between the two markers being studied. For example, if one of the alleles of a marker is present at a very high frequency in the population, the probabilty that the two parental chromosomes will have the same alleles at one locus is high. Since all of their offspring will have only that one allele at that locus

it will not be possible to follow the segregation of paternal and maternal chromosomes; the results will be uninformative. Thus, to detect a recombination event between a marker locus and the disease gene, it is necessary to be able to distinguish between the two chromosomes at the marker locus.

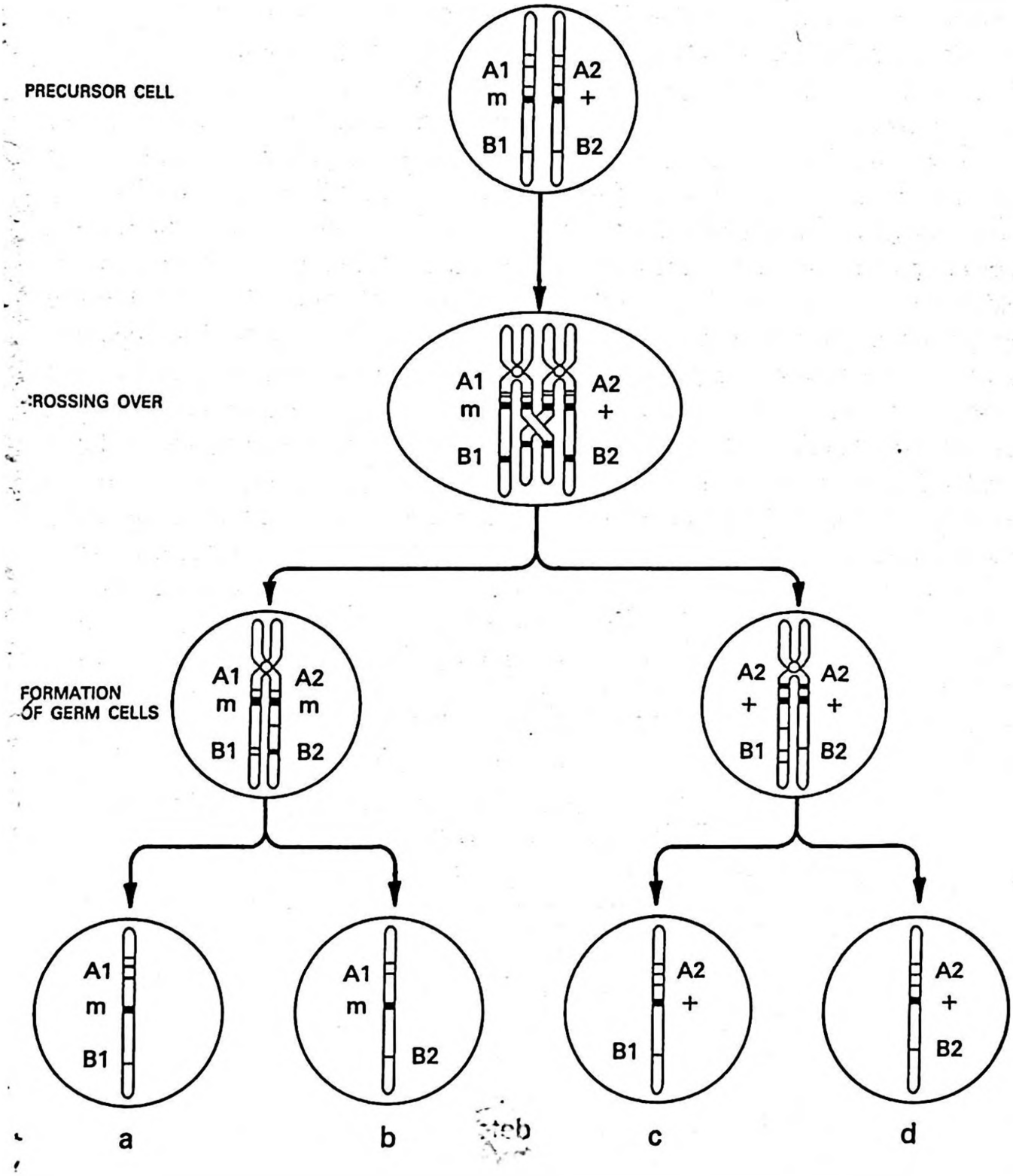

Fig. 1: The recombination event; exchange of segments of equal length between homologous chromosomes during meiosis (Reprinted with the permission of White R, Lalouel JM *Scientific American* 258:40, 1988).

The most informative markers are the ones that show many alleles in the population, none of which is present at a very high frequency. The following example illustrates the segregation of an autosomal dominant disease and a marker in a three generation pedigree (Fig. 2). Individual I-1 has both affected and unaffected children; therefore, she must be heterozygous for the disease gene. She is also heterozygous for the RFLP. If we follow the segregation of these two loci, it is clear that the affected offspring in the second generation (II-2, II-7 and II-8) do not have the same genotype at the marker locus. Since the father is homozygous (has the same two alleles at the marker locus), the affected mother has donated marker allele b along with the disease gene to two of her offspring; she has also donated the disease gene with marker allele a to the third offspring. If we consider the unaffected offspring in generation II, three children have received marker allele a with the mother's unaffected gene; however, one normal child received marker allele b. Therefore, either three of her children (II-4, II-5 and II-8) are recombinant and the other four nonrecombinant or these three children are nonrecombinant and the other four are recombinant. Since it is not known whether she carries RFLP allele a or allele b on the disease bearing chromosomes, we can not tell which is true. The situation is similar for the children of II-2. It would appear that individual III-1 and III-2 are recombinant. If we assume in this pedigree the minimal number of recombinant events, we can determine the recombination fraction as:

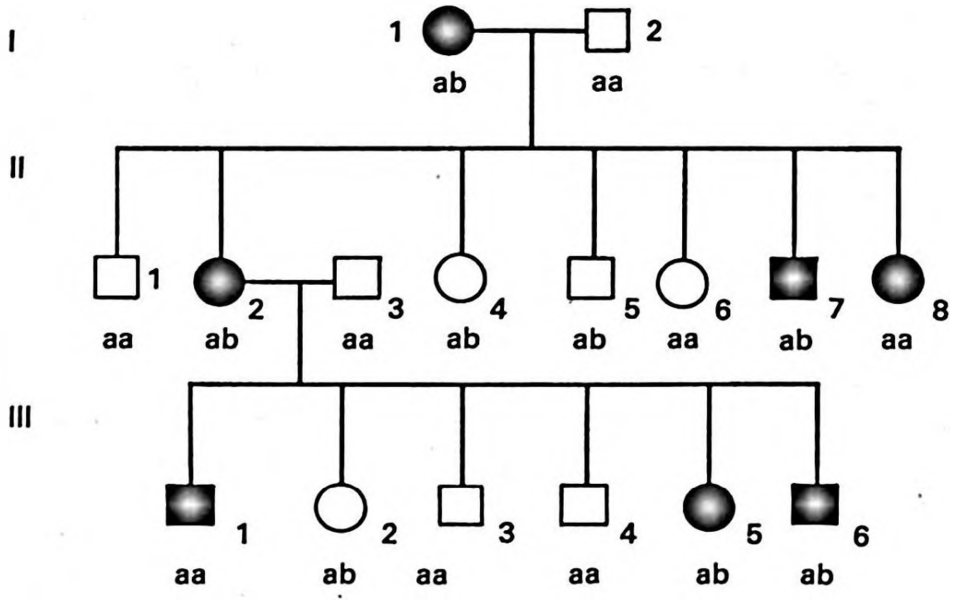

Fig 2: Segregation of an autosomal dominant disease and possibly linked marker.

$$\text{recombination fraction} = \frac{\#\text{ of recombinant chromosomes}}{\text{total number of chromosomes}} = \frac{3+2}{13} = \frac{5}{13} = < 0.50$$

It would appear that there is weak linkage between the marker and the disease gene.

Closely linked markers provide information about the location of a mutant gene and can be used for carrier detection and prenatal diagnosis[6]. Eventually, the information from linkage can be used to isolate the defective gene that causes the disease. However to be able to do all of this, markers linked to the disease of interest must be detected. Molecular biology has provided the polymorphic markers that are needed for linkage studies.

Very high levels of variation, i.e. polymorphism; found in the sequence of nucleotides of DNA allows linkage analysis to be possible. Approximately one in every 400 nucleotides is different between homologous chromosomes. These allelic variants provide a limitless number of markers scattered throughout the human genome.

RFLPs are recognized by detecting size differences between fragments obtained by cutting the DNA with restriction enzymes. RFLPs show Mendelian inheritance. The restriction endonucleases are bacterial enzymes that exist in nature to protect the host DNA from exogenous DNA. Each restriction enzyme cleaves DNA molecules whenever it meets a short, specific nucleotide sequence. The DNA fragments produced by this cleavage are very reproducible and are called restriction fragments. Any alteration in the primary sequence of the DNA will eliminate or create a cutting site. RFLPs are revealed by differences in the banding patterns shown by Southern blotting analysis between individuals for the same enzyme[7]. An example is given in Fig. 3.

A single restriction enzyme finds millions of cutting sites in the total human DNA. To detect one or two variant fragments among millions, the fragments are first separated by electrophoresis through a gel. The mobility of a fragment is inversely proportional to its length. Therefore, short fragments will travel more quickly through the gel, while large fragments will move proportionately shorter distances in the same time interval at the same voltage. Southern blotting is a sensitive technique used to detect these fragments. Double stranded DNA can be separated when heated or exposed to a high pH. The nucleotides along the single strands of DNA can then pair hybridize with a complementary DNA sequence. This feature is the essence of DNA hybridization techniques.

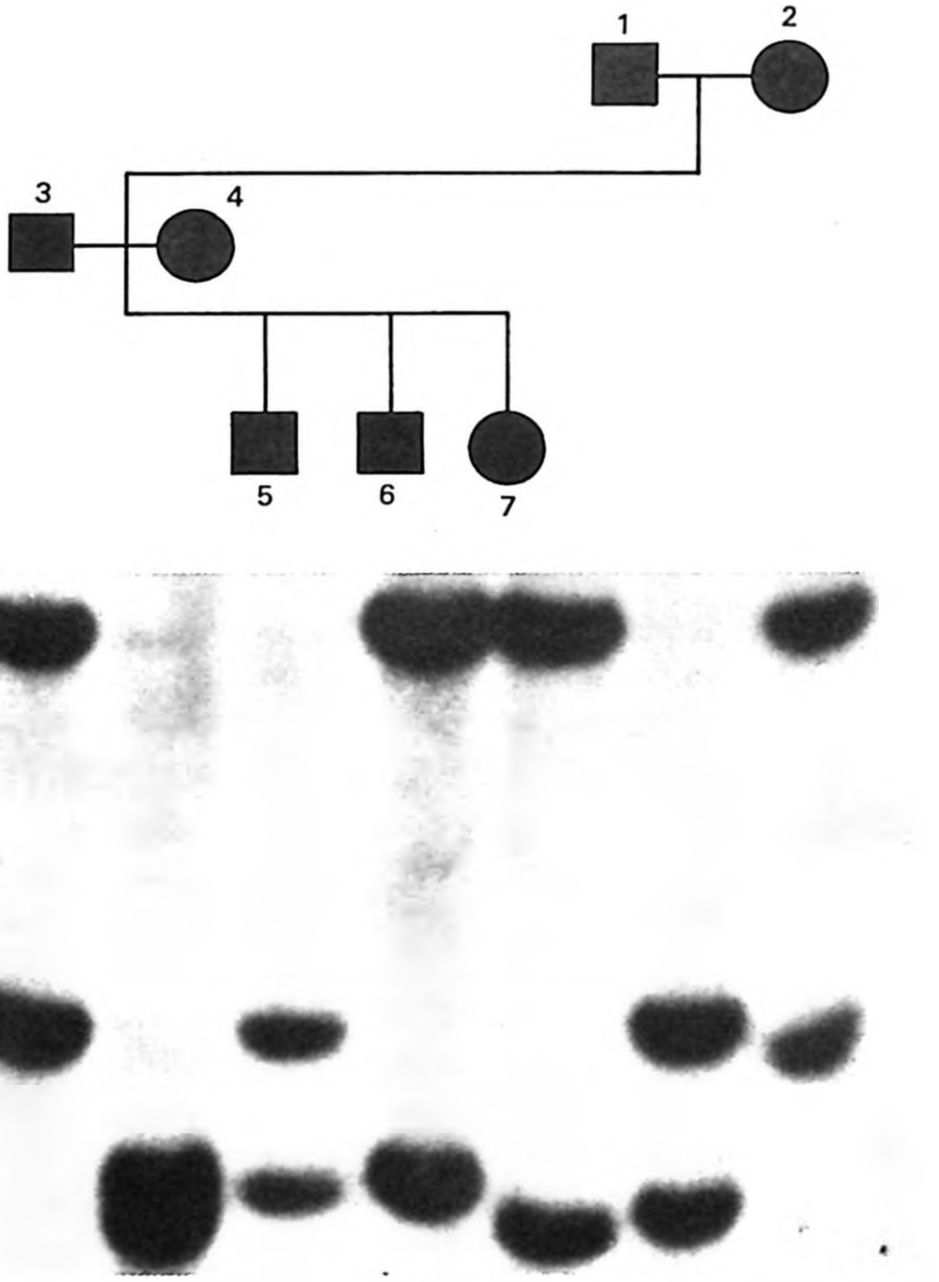

Fig. 3: A DNA marker displaying Restriction Fragment Length Polymorhism (RFLP).

In Southern blotting (Fig. 4), human genomic DNA is digested with a restriction enzyme and the fragments produced are separated by gel electrophoresis. The DNA is unwound and the single strands are blotted onto a nylon membrane. The probe, a single strand of DNA that is complementary to the piece of DNA of interest, is labelled by incorporating radioactive nucleotides and the nylon membrane is exposed to this probe. The probe will bind only to the DNA fragments complementary to this base sequence. A special film is exposed to this membrane to allow detection of the positions and patterns of bands that result from the DNA polymorphism. If the probe reveals a RFLP, then DNA from

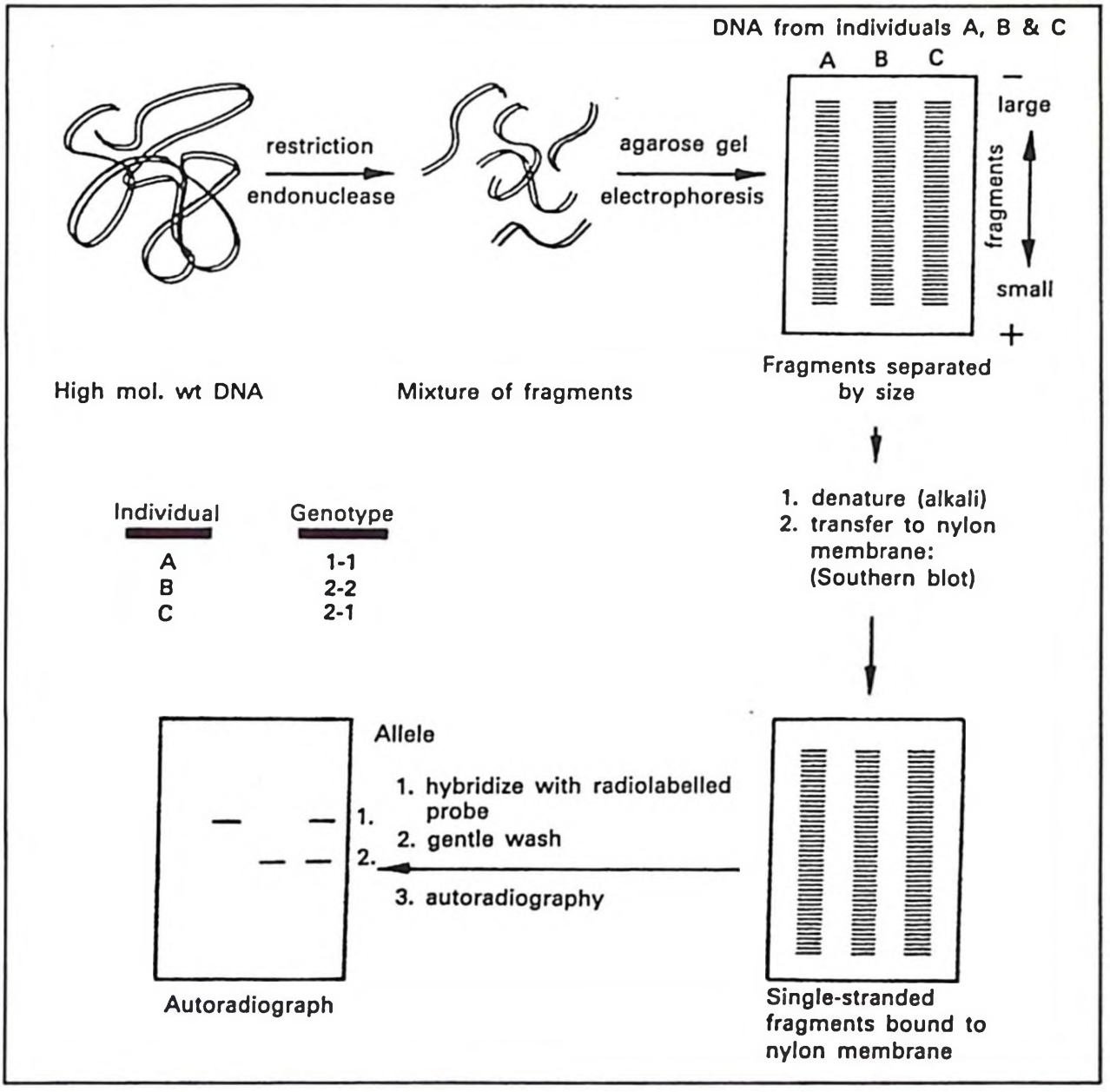

Fig. 4: Southern blotting technique for detecting RFLP.
(Reprinted with the permission of Davies KE, Read AP. *Molecular Basis of Inherited Disease*, 1988, p.8).

different individuals will show hybridization signals (bands) at different places on the blot. To detect a RFLP, a probe complementary to the DNA near the restriction enzyme cutting site should be found. These probes are chosen from a collection of cloned DNA fragments, representing the full human genome. The number of genes and anonymous DNA segments (uncoding or arbitrarily cloned DNA sequences) that detect RFLPs has increased continuously[8]. The total number of RFLPs described for each of the last four Human Gene Mapping Workshops (HGM) and the number of identified polymorphic loci per chromosome by June 1989, are shown in Fig. 5 and 6, respectively.

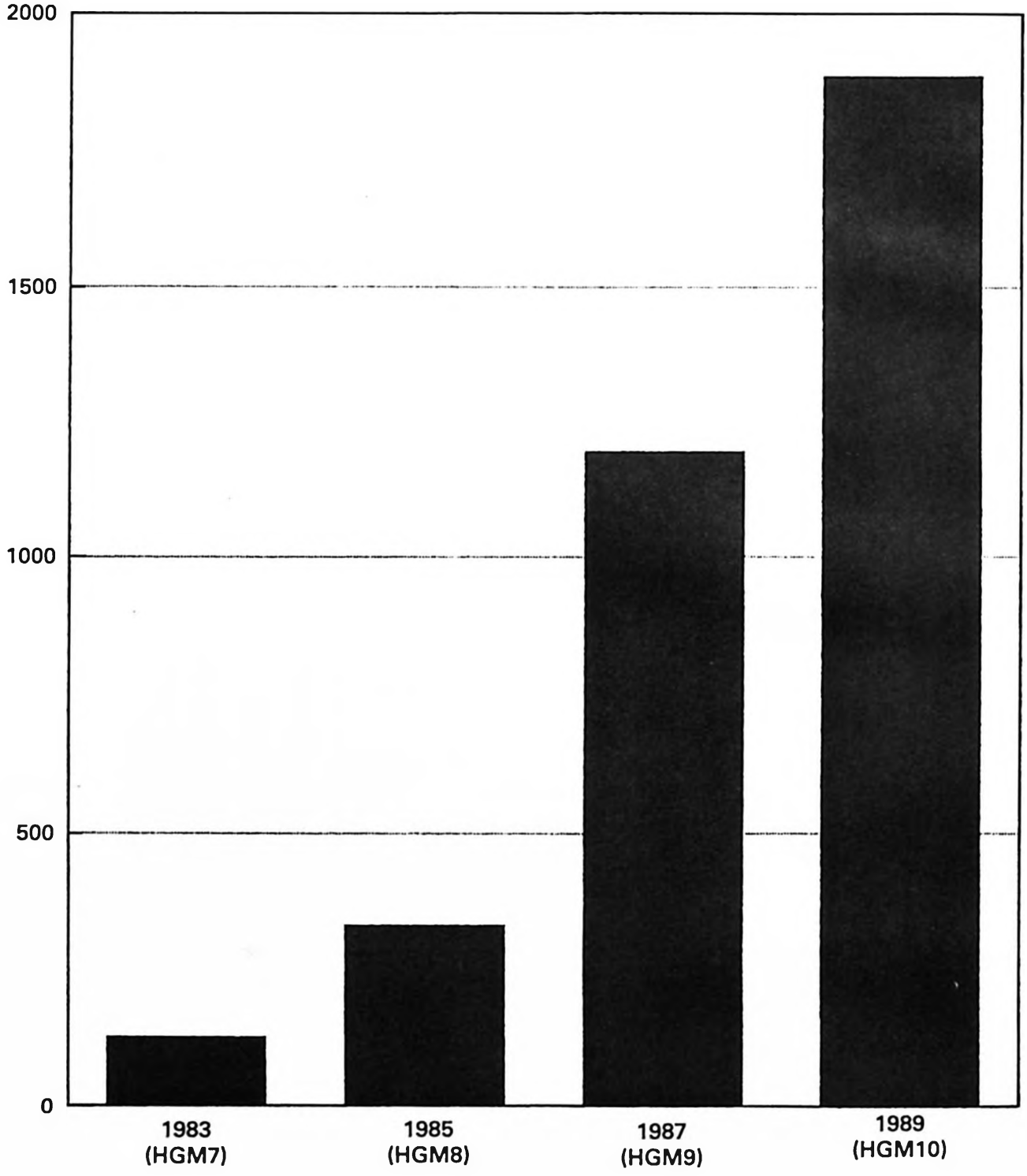

Fig. 5: The total number of RFLPs that have been assigned to specific chromosomes at each of the last four Human Gene Mapping Workshop (HGM).

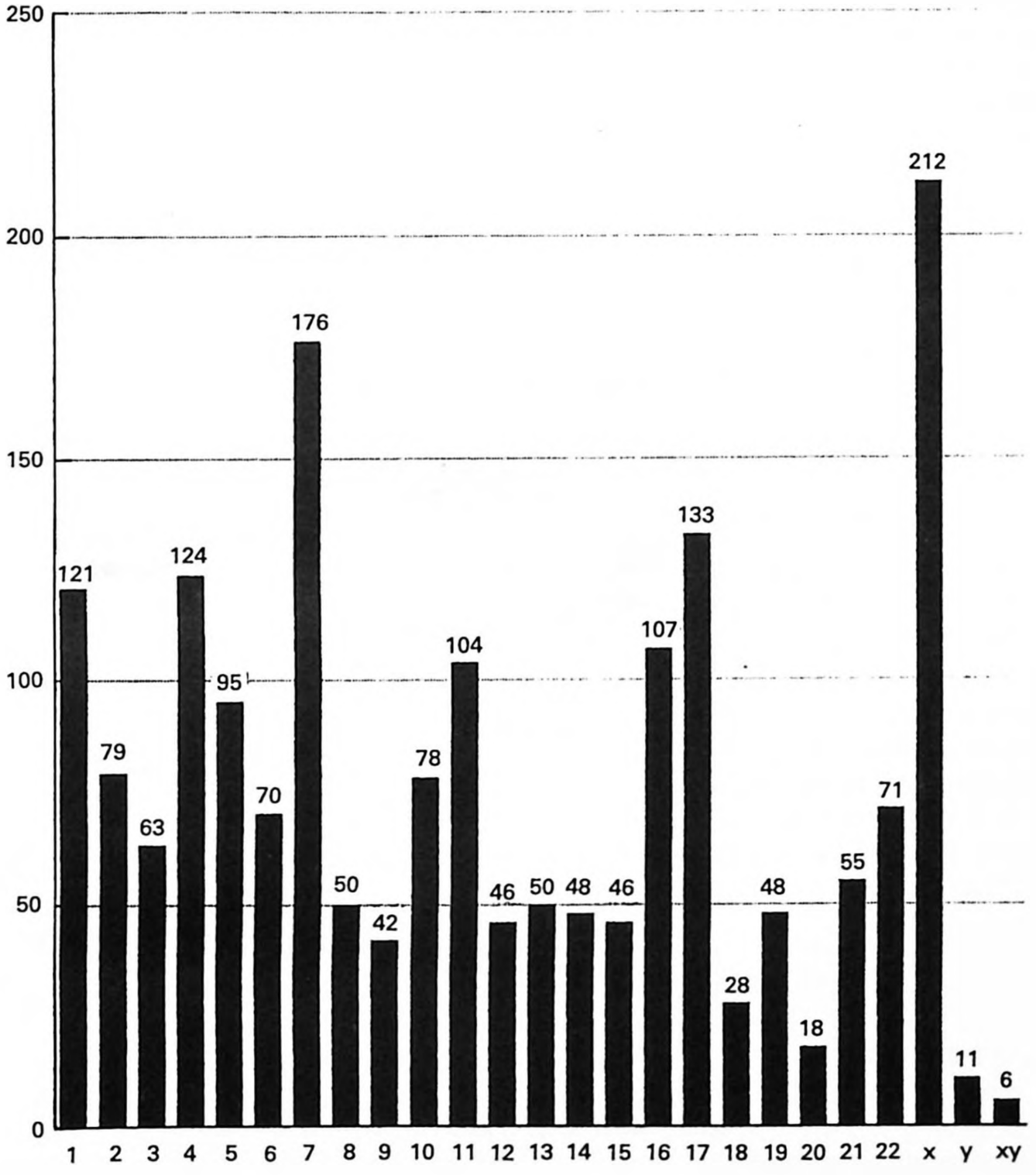

Fig. 6: The number of identified polymorphic loci per chromosome by June 1989 (HGM10).

Although the principle for linkage mapping is simple, it requires a great deal of work to both generate the data and to analyse it. A haploid genome contains about 3×10^9 nucleic acids. It has been estimated that in the human genome there are appoximately 50,000-100,000 genes. In a disease with a Mendelian pattern of inheritance, a single gene among the 100,000 genes will be defective. Where does one begin to search to localize the gene? To be able to locate any disease gene one needs to study families with 100-200 evenly-spaced genetic markers. DNA must be collected from hundreds of individuals, in dozens of large families and tested for the RFLPs characterizing each marker set. If one knows which chromosome to search, the number of markers that must be tested can be reduced from several hundreds to dozens. If, for example, a genetic disease shows an X-linked inheritance pattern, then finding the gene will only require that markers known to be carried on the X chromosome be tested. The "candidate gene" approach may also reduce the number of probes to be used. This approach is based on selecting the marker because of its position on the genetic map and its known function[9]. For example, by biochemical study, type IV Ehler's-Danlos syndrome was shown to have a deficiency of type III collagen, which is coded by a gene symbolized COL3A1. Absolute linkage of a COL3A1 RFLP and E-D IV indicates that at least in the families studied, the disorder is caused by a mutation in the structural gene for type III collagen.

For the analysis of linkage studies, the lod score method is used to quantify the degree of linkage. A lod score is the logarithm of the odds in favor of the linkage, calculated at a particular value of θ. To calculate the lod score at a particular value of θ, $\hat{\theta}$, it is necessary to calculate the likelihood of the pedigree with $\theta=\hat{\theta}$ and $\theta=0.5$. Then, the lod score at $\theta=\hat{\theta}$ is the logarithm, base 10 of the ration of two likelihoods:

$$\text{LOD} = \text{Log}_{10} = \frac{\text{likelihood of family given, 2 loci are linked with a recombination fraction, } \hat{\theta}}{\text{likelihood of family given, 2 loci are unlinked, } \theta = 0.5}$$

A positive lod score favours linkage and a negative lod score is evidence against linkage. The threshold for significance are +3 and −2. With a lod of 3, the odds in favor of linkage are 1000 to 1. A plot of the lod scores against recombination fractions is shown in Fig. 7. Curve 1 does not cross the +3 or −2 threshold and, therefore, is inconclusive. Curve 2 shows linkage, the peak is a lod score of 7 at zero, indicating an absence of recombination, suggesting that the marker localizes very close to the disease gene. Curve 3 crosses the −2 threshold at 0.1 recombination fraction, indicating that linkage closer than 10 cM can be excluded.

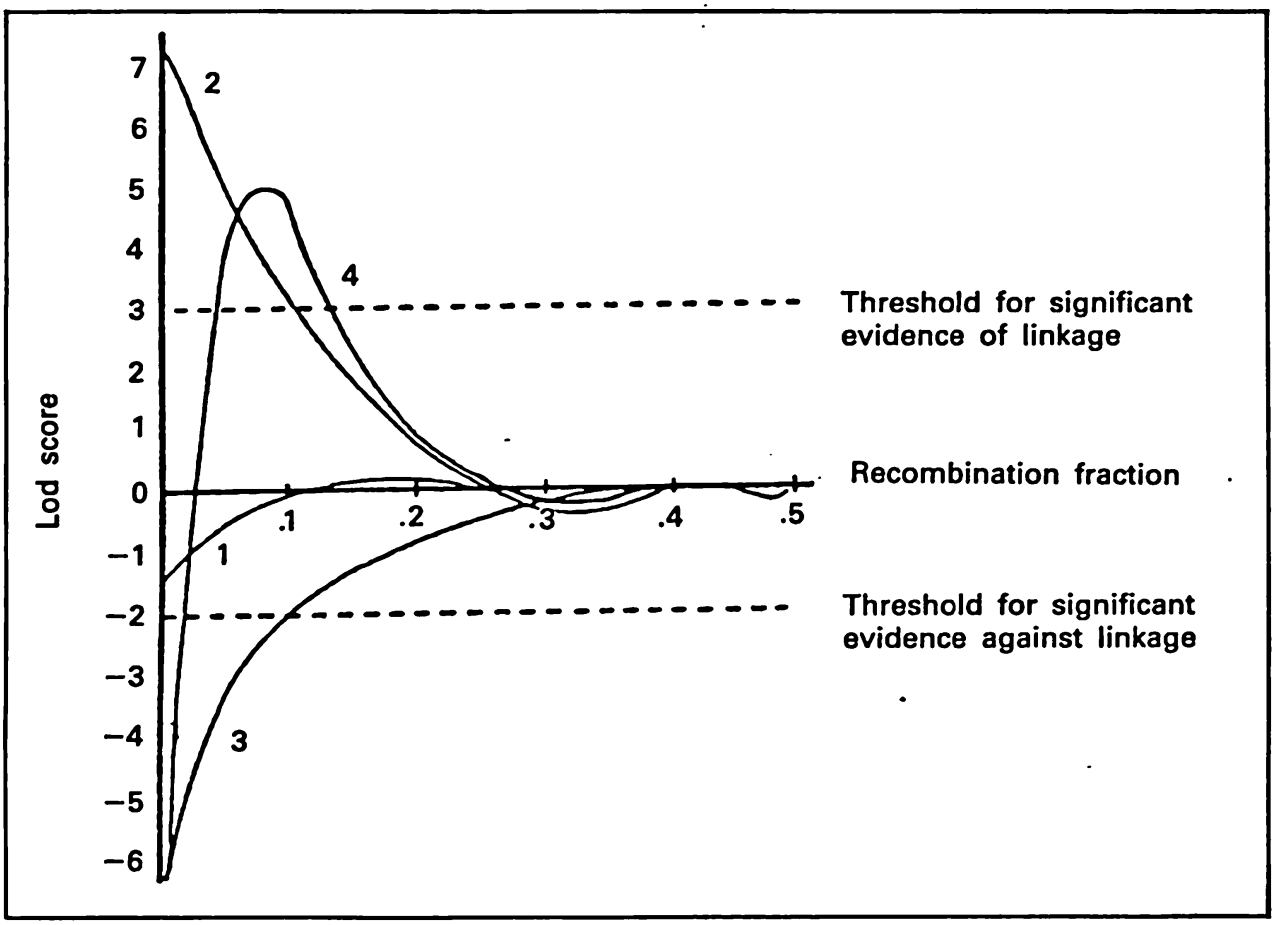

Fig. 7: Evaluation of the result of a linkage analysis (plotting of the lod scores against recombination fractions. Reprinted with the permission of Davies KE, Read AP. *Molecular Basis of Inherited Disease*, 1988 p. 19).

Curve 4 shows linkage, the peak lod score being around 5, at an estimated distance of 10 cM.

The lod score obtained from each pedigree are added together to obtain a cumulative lod score for the disease of interest and a particular marker. When several researchers are studyng the same disease, the lod scores obtained by each group can be added together provided that similar parameters are used in the calculations (e.g. allelic frequencies or disease frequency). This feature is very useful, particularly when working with a rare disease, where in each center only a few families can be studied. We have applied exactly this approach to the linkage analysis of a rare disease called "ataxia-telangiectasia"[1,3,10].

REFERENCES

1. Gatti RA, Berkel I, Boder E, et al. Localization of an ataxia-telangiectasia gene to chromosome 11q22-23. *Nature* 336:577, 1988.
2. Concannon P, Malhotra U, Charmley P, et al. The ataxia-telangiectasia gene (ATA) on chromosome 11 is disting from the ETS-1 gene. *Am J Hum Gen* 46:789, 1990.
3. Sanal O, Wei S, Foroud T, et al. Further mapping of an ataxia-telangiectasia locus to chromosome 11q23 region. *Am J Hum Genet* (in press).
4. White R, Lalouel JM. Chromosome mapping with DNA markers. *Sci Am* 258:40, 1988.
5. Drayna D. Fine structure linkage mapping of the human genome using RFLP analysis. *Bio Techniques* 4:412, 1986.
6. Davies KE, Read AP. Mapping and tracking genes. In Rickwood D (ed). *Molecular Basis of Inherited Disease.* Washington DC: IRL Press, 1988, pp. 11-29.
7. Gusella JF. Recombinant DNA techniques in the diagnosis of inherited disorders. *J Clin Invest* 77:1723, 1986.
8. Kidd KK, Bowcock AM, Schmidtke J, et al. Report of the DNA committee and mapped genes and DNA polymorphisms. *Cytogenet Cell Genet* 51:622, 1989.
9. McKusick VA. The morbid anatomy of the human genome: a review of gene mapping in clinical medicine. 4. *Medicine* (Baltimore) 67:1, 1988.
10. Foroud T, Sobel E, Ziv Y, et al. Localization of the AT locus to an 8 cM interval defined by STMY and S132. *Genomics* (submitted for publication).