PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Linkage Analysis: Progress and Problems

D. Timothy Bishop

Phil. Trans. R. Soc. Lond. B 1994 **344**, 337-343
doi: 10.1098/rstb.1994.0072

Linkage analysis: progress and problems

D. TIMOTHY BISHOP

ICRF Genetic Epidemiology Laboratory, St James's University Hospital, Beckett Street, Leeds LS9 7TF, U.K.

SUMMARY

Linkage analysis involves identifying the locations of genes responsible for disease by identifying other genes that co-segregate through families with the risk of disease. Subsequently, molecular approaches are applied to find the precise gene involved. Although these approaches have been successful to date, applications to diseases which do not show clear Mendelian inheritance have presented problems. The nature of these problems is discussed.

1. INTRODUCTION

Linkage analysis attempts to identify the co-segregation of pairs of loci by examining families. Specifically, the loci are recognized either indirectly by the occurrence of a trait (often a disease) in some family members or directly by an assay which indicates the variants that each family member carries at a particular genetic locus. Typically, we would study such a disease if there were evidence to show that the disease had a genetic basis (i.e. if the disease were transmitted through a family in a manner consistent with being due to a gene) and the intention of the study would be to identify the chromosomal location of the responsible gene or genes. Often the other loci are DNA sequences which are variant between individuals (called 'markers'). These loci are chosen because their precise sequence is known, it routinely shows variation from one copy of that chromosome to another and this variation can be routinely assayed; the two variants that each family member carries can then be assessed. Finally, the chromosomal location of the sequences for these markers must be known. Prior to DNA-based techniques, an example of such a marker would be the ABO blood system; this could be routinely assayed and in the general population there is considerable variation.

The observation of linkage is made by showing statistically that the two loci are not transmitted independently of each other through the families implying that the loci are physically close to each on the same chromosome. Identifying co-segregation will therefore imply that the location of the 'disease' locus is in this same chromosomal region. Although a precise limit is not possible to define, if two loci are within perhaps 25 000 000 DNA base pairs (b.p.) of each other then statistically it is often feasible to show linkage. Because there are of the order of 3×10^9 b.p. in a full complement of human chromosomes and the gene could be anywhere in this set of chromosomes this implies that the probability that a randomly selected marker will be within a distance commensurate with showing linkage is of the order of 0.02. There are many provisos to this simple calculation because, for instance, if the disease could be due to two or more genes acting separately then the confusion created by taking families with different causes would dramatically lower the probability of showing linkage. Some of these issues will be discussed later.

The analysis provides an estimate of the genetic distance between two loci, that is, the proportion of times that the two loci are separated by a recombination event during meiosis; this can be translated to a crude estimate of the distance between the loci in terms of DNA base pairs. The identification of co-segregation is then the initial step to cloning the disease gene through precise mapping of the disease locus by genetic methods and then subsequent recognition of the important mutation by molecular methods. The steps in the process involve: (i) identifying pedigrees segregating the disease of interest; (ii) typing various DNA markers on available family members to define broadly the location of the disease locus ('showing linkage'); (iii) examination of evidence for heterogeneity (i.e. multiple genes being responsible for the same disease); (iv) refinement of the location of the disease locus so that DNA markers which flank the disease locus are identified and an estimate of the distance between the two loci is obtained; and (v) cloning of the disease gene through molecular methods. This discussion will focus on steps (i)–(iv). Step (v) is discussed in detail in many different publications and is the focus for research into the most efficient techniques.

The recognition of the importance of linkage analysis in human genetic is due to R. A. Fisher (for instance, 1935) although Morton (1955) produced the statistical framework on which analyses are currently based. Morton (1955) introduced the logarithm of the odds (LOD) score defined by

$$\max \mathrm{Log}_{10}\{L(\theta)/L(0.5)\} \qquad 0 \leqslant \theta \leqslant 0.5,$$

where $L(\theta)$ is the likelihood of a specific recombina-

Phil. Trans. R. Soc. Lond. B (1994) **344,** 337–343
Printed in Great Britain

tion fraction, θ, between the disease locus and the marker locus compared to the likelihood ($L(0.5)$) when the two loci are segregating independently (i.e. $\theta = 0.5$). The situation $\theta = 0.5$ then corresponds to the two loci being on separate chromosomes or greatly separated on the same chromosome. When θ is small ($\theta \approx 0.0$), the mutation at the disease locus and the variant at the marker locus which are shown to be on the same physical copy of a chromosome will be transmitted together throughout the pedigree. θ then, can take values between 0 and 0.5. The likelihood ratio $L(\theta)/L(0.5)$ measures the evidence in favour of a hypothesis that the two loci are linked at the recombination fraction, θ, as compared with being unlinked. The data consist of observations on pedigrees involving the transmission of risk of disease and the transmission of the alleles at the marker locus (see, for example, Ott 1991).

Morton (1955) modelled the process of identifying linkage as selection of a marker and then typing enough individuals to either eliminate that locus from consideration because of accumulated statistical evidence against linkage or postulating linkage when evidence for co-segregation was identified (i.e. inferring that $\theta < 0.5$); his assumption was therefore based on the analogy with sequential analysis. Further, in the absence of a biological reason for examining one part of the genome as compared with any other for the disease gene, the a priori probability of linkage is low (of the order 0.02 for a large study as discussed above (Elston & Lange 1975)) so that the criterion for a critical LOD score should be strict. Morton (1955) suggested a LOD score of 3.0 as the critical limit; although the precise interpretation of the significance of this level varies with the study design, the limit has proved appropriate in practice as few studies of Mendelian traits achieving such LOD scores have been subsequently shown to be false (Morton 1990). For instance, Rao *et al.* (1978) estimated that less than one in a hundred results achieved with a LOD score of 3.0 or more were false.

Until the early 1980s, there were few loci that could be used as markers as the chromosomal locations were frequently unknown and most genes only rarely showed any variation between individuals (a notable exception was the ABO blood group). The scarcity of markers meant that the probability of finding linkage was negligable. Then in the 1980s, with the development of DNA technology, enormous effort has been placed on identifying DNA polymorphisms which can be used as DNA markers for disease studies and also as a basis for building more detailed maps of the chromosomes with the ultimate goal of providing a complete sequence of the human genome (Congress of the United States 1988). There are now thousands of DNA-based markers and the methodological improvements have meant that it is possible to generate more markers within a defined genetic region often in a matter of weeks.

When these DNA-based markers started becoming available, many authors recognized that in the initial search to show linkage between a disease gene and a marker by typing markers that were close to each other would duplicate linkage information and so would be inefficient (Solomon & Bodmer 1979; Botstein *et al.* 1980; Bishop & Skolnick 1980). Bishop & Skolnick (1980) examined the best 'spacing' between markers and suggested that to choose 150 equally spaced markers across the chromosomes would produce the most efficient study; such an arrangement would avoid duplication but keep the probability that a disease gene would fall too far from its closest marker (and hence make it impossible to show linkage) to a low level. Maps which are sufficently detailed for linkage analysis are now available (Weissenbach *et al.* 1992) but, in fact, even before these maps were completed, the success of the project in the localization of disease causing loci has been apparent. The startling success of the search for markers has meant that essentially all diseases with a clear Mendelian component and for which sufficient family information is available have been mapped. This class of diseases are those by which examination of the pattern of disease segregation within a family appears to identify a clear dominant or recessive gene. For instance, a rare disease for which males and females are equally likely to be affected and families expressing the disease, show disease expression in essentially every generation and for which affected parents produce on average an equal proportion of affected and unaffected children is likely to be due to a rare autosomal dominant gene. Of course, many of the loci remain to be cloned and so the precise genes are not yet known.

2. MAPPING 'COMPLEX' DISEASES

The success in mapping the rare Mendelian syndromes as well as the completion of the first generation of genomic-wide linkage maps has focused attention on diseases which do not show such a clear Mendelian pattern within families but are often considerably more common in the general population. These diseases represent significant health care problems and often show significant family aggregation (for instance, all of the common cancers (Easton & Peto 1990), schizophrenia and affective disorder, and diabetes). The application of these methods to diseases which have a clear familial (but which is not consistent with being due to a single gene) component offers the opportunity to understand the etiology directly from the genetic susceptibility rather than trying to reconstruct the natural history of the disease. We shall refer to such diseases as 'complex' in this discussion.

From a procedural viewpoint, the methods described above do not need to be modified. Statistically, the definition of the LOD score requires some modification to take into account the dependence on the mode of inheritance at the disease locus, both the 'true' model and the model assumed for analysis as we will not know the truth. Of course, the model required specification before complex diseases were studied but such definitions were trivial. We will write T to indicate the assumed parameters of such inheritance. T will involve specification of number

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

and type of (autosomal or sex-linked) loci involved in susceptibility, frequencies of the various alleles at the loci and the genotype-specific risks of disease (assuming that there is no selective effects on these loci).

The parameters of T may be defined in three distinct ways:

1. In some cases analysis of pedigrees is sufficent to define their values with accuracy (as for instance would be the case for a rare disease which clearly satisfied Mendelian segregation).
2. Analysis of an independent set of families (from those being analysed with linkage analysis) provides an estimate of the parameters (as for instance if segregation analysis has been performed prior to collection of this set of families).
3. No initial estimates are available so the investigator postulates values.

Difficulties in mapping are in part due to a lack of information concerning T and raises probably the most frequently voiced concern about the possibilities of mapping a specific disease.

3. MODEL FOR EXAMPLES

For the initial part of this discussion, and to introduce some of the difficulties, we will concentrate on diseases caused at least in part by a single autosomal locus with two alleles (D,d). D is assumed to be dominant to d for disease susceptibility and is known to be rare. Carriers of the allele labelled 'D' have a risk of disease (t) while non-carriers never develop the disease; the value of t is not known precisely. 't' is termed the penetrance of the disease. For simplicity, we assume that family information is sufficient to show convincingly that the father is a heterozygote at the disease locus and to provide the phase of the disease locus with the DNA marker so that the father has phase DA/da (i.e 'D' and 'A' on one copy of the paternal chromosome pair while 'd' and 'a' are on the other) where A and a are the two alleles at the marker locus. Because the disease allele is so rare, the mother is assumed an homozygote for the disease locus and, for convenience, also at the marker locus; her genotype is therefore da/da. Because the mother can only produce a single gamete for transmission to her children (i.e. da), there are four possible genotypic combinations for their children (see table 1). For comparisons, we will compute the expected LOD score, that is, the average information for comparing the hypotheses of no linkage under various misspecified modes of inheritance. The misspecification is due to a lack of knowledge concerning t, and is the only 'complexity' for this disease.

4. IDENTIFYING LINKAGE

The introduction of DNA technology has modified the way in which in experiments are conducted so that we must reconsider the statistical analysis of the methods. Current technology dictates that most studies are essentially fixed sample size (in terms of the number of samples typed) unlike the methods on which Morton (1955) based his formulation of the linkage problem. Of course, once linkage is identified usually as many samples as are available will by typed; we will consider that aspect of the study later. Under a fixed sample size assumption, the statistic defined by $Z(\theta;T) = \log_{10}\{(\theta; T)/L(0.5;T)\}$ when evaluated at the maximum likelihood estimate of the recombination fraction θ between the disease locus and a marker locus is the LOD score (T is included in the formula to make its involvement more apparent). Under this definition, at the maximum likelihood estimate, $2\ln(10)Z(\theta;T)$ should asymptotically have a distribution that is χ^2 with 1 d.f. under the hypothesis of no linkage assuming that the correct mode of inheritance is employed (i.e. choice of T) in the evaluation of $Z(\theta; T)$ (situation (1) in §2); the factor of $\ln(10)$ is required to transform the logarithms back to natural logarithms. For the situation where T cannot be known with certainty, MacLean *et al.* (1993) investigated situation (1) and showed that for even modest samples (25 offspring with $t=0.5$), the asymptotic distribution (χ^2 with 1 d.f.) was appropriate.

In the more complicated situations ((2) and (3) in §2 above), if t is incorrectly specified then the frequency of false assertions of linkage should not increase (Williamson & Amos 1990). In fact, the theorem of Williamson & Amos (1990) applies more generally so that the LOD (suitably scaled) has the appropriate χ^2 statistic under the assumption of no linkage whenever the mode of inheritance is incorrectly specified. Again, MacLean *et al.* (1993) confirmed the small sample properties of the test statistic. The method of analysis should therefore be robust to misspecification of the mode of inheritance under the hypothesis of no linkage.

When the mode of inheritance is misspecified and there really is linkage, we can assume that this will lead to biases in the estimated recombination fraction between the disease locus and the marker locus and a

Table 1. *The probabilities of the four distinct outcomes for a family with a rare partially penetrant dominant disease*

(The probabilities of the four distinct outcomes are for a family with a rare partly penetrant dominant disease (determined by a single locus with alleles D and d) when the father is affected and has the genotype DA/da (so that he produces the gametes DA, Da, dA, da with probabilities $(1-\theta)/2$, $\theta/2$, $\theta/2$, $(1-\theta)/2$ respectively). The alleles at the marker locus are A and a. Any child carrying a copy of D is affected with probability, t, while no child without D can be affected. The mother is assumed to have genotype da/da (so that the mother always produces a gamete 'da').)

child's disease status	paternally derived marker allele	probability of observation
affected	A	$t(1-\theta)/2$
unaffected	A	$(1-t+t\theta)/2$
affected	a	$t\theta/2$
unaffected	a	$(1-t\theta)/2$

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

resulting loss of power. The biases we might assume, could be so severe that all evidence of linkage is lost. However, two point analysis has been shown both analytically and anecdotally to be quite robust to this misspecification, that is there is little loss of power through this misspecification. In the most detailed analysis, Clerget-Darpoux *et al.* (1986) showed that assuming an incorrect penetrance (i.e. value of t) did not lead in general to a significant decrease in power or (more precisely, loss in expected LOD score). This can be seen for the example discussed above (figure 1) for the case of misspecified penetrance. In this example, analysis is performed assuming an (incorrect) penetrance, t, while the true penetrance is 0.8. The quantity calculated is the expected log likelihood difference between the hypothesis of linkage and no linkage when analysis is conducted assuming the incorrect penetrance. As a guide to the change in the results, this expected log likelihood difference is graphed as a function of the recombination fraction ($E[Z(\theta);T]$); these log likelihood differences are averaged over the distribution of samples that would result from the correct model. We label these results as expected LOD scores (ELOD) to follow others notation although precisely these are expected log likelihood differences. Because of the misspecification of the penetrance, asymptotic bias, of course, exists in the recombination fraction but the expected loss of information is often only marginal even for major differences between the true and postulated penetrances. The robustness can be intuitively explained by the recombination fraction being available to accommodate the misspecification in the penetrance, simply too few or too many recombination events between the disease and marker locus affects the estimated recombination fraction but the overall evidence for linkage is only marginally changed. The one situation to lead to complete loss of information is when dominantly and recessively determined diseases are confused, so to assume a recessive mode of inheritance for a truly dominant disease or vice versa does lead to inappropriate exclusion of linkage (Clerget-Darpoux *et al.* 1986).

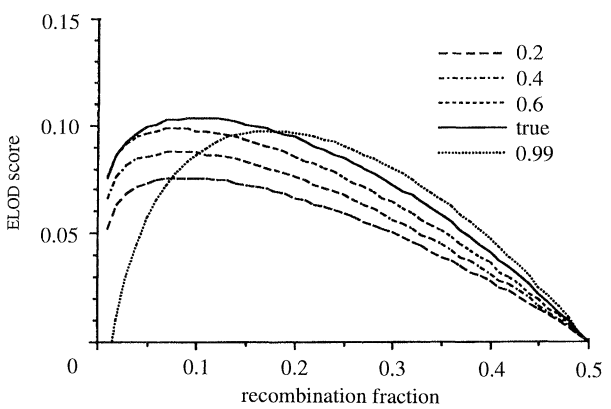

Figure 1. The ELOD score for an individual from a family with probabilities of disease as given in table 1 when the true recombination fraction between the two loci is 0.8 and the true recombination fraction between the loci is 0.1 but, in the analysis, the penetrance is misspecified. The ordinate refers to the recombination fraction assumed in the analysis. Overall, as expected, the maximum ELOD occurs when precisely the correct parameters are assumed but even for widely misspecified values of the penetrance, the loss in ELOD is minimal.

Although model misspecification does not invalidate the linkage analysis, it should be stressed that to date we have only considered specifications prior to analysis. Examining multiple modes of inheritance during linkage analysis can seriously exaggerate the significance of a result as essentially it entails multiple testing. For instance, assuming the model described above and maximizing the LOD score over penetrance and recombination fraction leads to a statistic which has a χ^2 distribution with 2 d.f. asymptotically rather than 1 d.f. as would be obtained if the maximization over penetrance had not occurred (MacLean *et al.* 1993). Optimization over more complex models can lead to test statistics with even more degrees of freedom; the precise degrees of freedom are usually impossible to determine in a straightforward manner although an upper limit is available (J. A. Williamson, personal communication).

In practice, the analysis of more complex diseases has already produced a number of 'false' reports of linkage. Examples include schizophrenia and affective disorder where LOD scores which exceed 3.0 were subsequently shown to be produced in the absence of linkage (see Risch (1990*d*) for a review and interesting commentary). Although few conclusions can be reached from individual cases, analysis of robustness shows that the most likely method to produce erroneous assertions of linkage is the examination of multiple assumptions concerning the mode of inheritance. In the extreme case of optimization over all possible modes of inheritance to produce the largest possible LOD score, this leads to an additional degree of freedom for resulting asymptotic distribution of the (scaled) likelihood ratio for the example described above. However, it should be stressed that this example contains only a single parameter and optimization over models containing more parameters could produce and even more exaggerated effect (MacLean *et al.* 1993; Weeks *et al.* 1990).

5. HETEROGENEITY

Once a gene causing a disease has been mapped, it is natural to investigate first whether all of the families with this disease are due to this gene or, if not, to estimate the proportion of families with this disease which are due to this precise gene as a first attempt to characterize such families. Specifically, we might investigate whether the families due to this gene have a specific clinical expression, perhaps representing a distinct natural history, pathological or prognostic subgroup. Heterogeneity analysis is a statistical approach to address this issue; basically, the assumed model considers two alternative forms of inheritance, one linked to the region of interest, the other unlinked to that region (Ott 1991). Within the analysis, the proportion of families due the candidate region is estimated. A basic assumption of the model is that the risk of disease is the same independent of the genetic

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

cause of the disease. If such an assumption is not valid then the estimated proportion of linked families is in most circumstances meaningless because the proportion is highly dependent on the way in which the families were identified. This information is usually unknown in the context of linkage analysis as families are identified because of a cluster of cases rather than as part of a systematic study in which all families of a particular structure are identified and sampled.

As an example of heterogeneity analysis, table 2 contains a brief summary of the results of a linkage study of breast cancer reported by Easton *et al.* (1993). A gene for hereditary breast cancer has been mapped to 17q (this locus is called BRCA1); analysis of 214 families showed that there were evident subgroups of families which showed wide variation in the proportion of linked families. For instance, all families with at least one case of ovarian cancer (breast-ovary families) were consistent with linkage while only 45% of families with breast cancer in the absence of any ovarian cancer cases and also any male breast cancer cases ('breast only' families) were estimated to be due to BRCA1. Among the 'breast only' families, families with average age of onset prior to 45 years were more likely to be linked than families with later age of onset ($p < 0.10$ for the comparison of equal proportions using these results while $p < 0.05$ for testing mean onset before 45 years versus after 45 years (Easton *et al.* 1993)). Finally, there is little evidence for linkage to families with at most three cases of breast cancer before the age of 60 years. In summary, heterogeneity analysis suggests that the majority of breast-ovary families are due to BRCA1, as are 45% of breast only families. There is also evidence that families with early onset (before age 45 years) and four or more cases before the age of 60 years are more likely to be linked than other families. Identifying families that are most probably linked also aids in identifying the location of the gene on 17q (Easton *et al.* 1993).

6. AFFECTED RELATIVE PAIR METHODS

The methods described above require that T is specified prior to analysis. For many diseases, investigators are often unwilling to specify T and so methods which are 'non-parametric' have been developed. These methods are based on the observation that two relatives both of whom are affected with the same disease are likely to share an inherited susceptibility. So, suppose that the mutation has a dominant effect and that, for instance, two brothers are affected; they are likely to have inherited the same copy of a mutated allele which increases the risk of disease from the same parental chromosome. Of course, the rarer the mutations, the more likely that the two share exactly the same copy. When the mutations are more common, then the possibility that both parents carry a copy of the mutated allele increases as does the probability that one parent carries two copies. For a mutation that has a recessive effect, both parents must, of course, be carriers of a mutated allele but the possibility still arises that for more common alleles, one or more parents carry two copies of the mutated allele. More generally for any genetic effect, the two affected relatives should share more frequently identical copies of part of a chromosomal region which contains a risk increasing mutation than would be expected for any two relatives of precisely that genetic relationship. For instance, consider the case of two brothers (table 3) who should, on average, inherited exactly the same allele from both parents a quarter of the time, the same allele from exactly one parent one half of the time and be entirely discordant a quarter of the time. Of course, instead of studying the segregation of this gene, we will only be studying a marker which is adjacent to the disease gene. If the two are in close proximity then the results for that marker will mimic the segregation of the disease gene. Any deviation between the observed and expected sharing in the direction of excess sharing can be taken as some evidence for linkage.

This method was first described for diseases in which the HLA system was thought to play a role (see, for instance, Bodmer 1981; Suarez *et al.* 1978). The antigens which form this system are located on white blood cells and determine histocompatibility. This system contains many different variants and for many different diseases, affected and unaffected individuals show quite distinct allelic combinations. This method of analysis can be extended to more distant relatives and to less polymorphic systems (Bishop & Williamson 1990; Risch 1990*a–c*, 1992; Holmans 1993).

The power of these methods has been examined for different modes of inheritance (for instance, Bishop & Williamson 1990; Risch 1990*a–c* 1992). For a disease determined by a single gene, then affected sibpairs represent the most efficient sampling unit if the disease is recessive, while the situation is more complicated if the disease is caused by a dominant gene. In such a situation, the most efficient sampling design depends upon, among other factors, the recombination distance between the disease gene and the marker, the genetic relationship of the affected relatives and the frequency of the high risk allele at the disease locus.

Table 2. *The estimated proportion of families with breast cancer due to BRCA1, adapted from Easton et al. (1993)*

type of families (number)	LOD score	estimated proportion linked
breast-ovary families (57)	20.79	1.00
breast only families (153)	6.01	0.45
breast only families with average age of onset:		
<45 years (54)	6.56	0.67
45–54 years (63)	0.40	0.19
$\geqslant 55$ years	0.08	0.38
breast only families by number of cases diagnosed before age 60 years:		
$\leqslant 3$ cases	0.36	0.26
4 or 5 cases	3.29	0.60
$\geqslant 6$ cases	2.72	0.45

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Table 3. *Possible inheritance for a single locus when the father has genotype A1A2 and the mother has genotype A3A4*

(There are 16 possible segregations to two children defined by (first child's genotype, second child's genotype) categorized by the number of alleles shared from the two parents. If there is no linkage between the disease and the marker then the probability of each of the 16 distinct possibilities is equal so that the probability of observing two affected children who share their two alleles is 4/16 (=0.25), 8/16 that they share exactly one and 4/16 that they share neither.)

share alleles from both parents:
{(A1A3, A1A3); (A2A3, A2A3); (A1A4, A1A4); (A2A4, A2A4)}
share alleles from one parent only:
{(A1A3, A1A4); (A1A4, A1A3); (A2A3, A2A4); (A2A4, A2A3); (A1A3, A2A3); (A2A3, A1A3); (A1A4, A2A4); (A2A4, A1A4)}
share alleles from neither parent:
{(A1A3, A2A4); (A1A4, A2A3); (A2A3, A1A4); (A2A4, A1A3)}

7. DISCUSSION

Identifying genes responsible for disease through linkage mapping has been one of the most startling successes of the genome project to date. Although simply mapping the genes is not an end in itself it does provide the opportunity to identify the gene responsible through the subsequent cloning and then to look at the product of that gene and ask about the abnormal function of that gene within the disease families. The success though is largely (although not totally) attributable to the identification of genes that segregate in families in a Mendelian fashion. Because these families are due to genes with a huge effect (in terms of disease risk) and population genetics tells us that any such disease which involves diminished fertility will be rare in the general population, that is 'few' individuals carry mutations in these genes. Although we should at no time belittle the importance of finding the genes for the rarest of syndromes on humanitarian and scientific grounds, the next stage of the process is then to move onto genes responsible for 'common' diseases (cancers, diabetes, coronary disease, etc.) which affect large proportions of the general public. Studying such disease introduce many complexities, partly statistical, partly logistical.

The success of the next stage depends on many factors and especially the true 'complexity' of the underlying processes. For instance, in all of the examples discussed above, we have assumed that the disease is due entirely to a single gene in each family; different families may have different genes but within a family the disease is caused by the same gene (this is the basis of the heterogeneity test described above). Suppose instead that the genetic etiology resulted from the interaction of two genes so that only specific combinations of genes produced affected individuals. The probability of dissecting this process will depend upon the precise way in which these genes interact. If, for instance, a copy of a specific mutation at each locus was required, then this would be accessible to linkage analysis as each of the two loci would separately be identifiable. More subtle interactions would not be so readily approached. The discussion of the number of genes involved and the way they interact is a topic of current research (see, for example, Risch 1990*a*,*b*). Until we understand more about the possible interactions, we are not in a position to predict the ultimate level of success in this process.

REFERENCES

Bishop D.T. & Skolnick M.H. 1980 Numerical considerations for linkage studies using polymorphic DNA markers in humans. In *Banbury Report 4: Cancer incidence in defined populations* (ed. J. Cairns, J. L. Lyon & M. Skolnick), pp. 421–433. Cold Spring Harbor Harbor Laboratory.

Bishop, D.T. & Williamson, J.A. 1990 The power of identity-by-state methods for linkage analysis. *Am. J. Hum. Genet.* **46**, 254–265.

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

Botstein, D., White, M., Skolnick, M. & Davis, R.W. 1980 Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32**, 314–321.

Clerget-Darpoux, F., Bonaïti-Pellié, C. & Hochez, J. 1986 Effects of misspecifying genetic parameters in lod score analysis. *Biometrics* **42**, 393–399.

Congress of the United States, Office of Technology Assessment 1988 *Mapping our genes. Genome projects: how big, how fast?* Baltimore: The Johns Hopkins University Press.

Easton, D.F., Bishop, D.T., Ford, D., Crockford, G.P. & the Breast Cancer Linkage Consortium 1993 Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. *Am. J. Hum. Genet.* **52**, 678–701.

Easton, D.F. & Peto, J. 1990 The contribution of inherited predisposition to cancer incidence. *Cancer Surv.* **9**, 395–416.

Elston, R.C. & Lange, K. 1975 The prior probability of autosomal linkage. *Ann. Hum. Genet.* **38**, 341–350.

Fisher, R.A. 1935 The detection of linkage with dominant abnormalities. *Ann. Eugen.* **6**, 187–201.

Holmans, P. 1993 Asymptotic properties of affected-sib-pair linkage analysis. *Am. J. Hum. Genet.* **52**, 362–374.

MacLean, C.J., Bishop, D.T., Sherman, S.L. & Diehl S.R. 1993 Distribution of lod scores under uncertain mode of inheritance. *Am. J. Hum. Genet.* **52**, 354–361.

Morton, N.E. 1955 Sequential tests for the detection of linkage. *Am. J. Hum. Genet.* **7**, 277–318.

Morton, N. E. 1990 Genetic linkage and complex diseases: a comment.*Genet. Epidemiol.* **7**, 33–34.

Ott, J. 1991 *Analysis of human genetic linkage*. Baltimore: The Johns Hopkins University Press.

Rao, D.C., Keats, B.J.B., Morton, N.E., Yee, S. & Lew, R. 1978 Variability of human linkage data. *Am. J. Hum. Genet.* **30**, 516–529.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Risch, N. 1990*a* Linkage strategies for genetically complex traits. I. Multilocus models. *Am. J. Hum. Genet.* **46**, 222–228.

Risch, N. 1990*b* Linkage strategies for genetically complex traits. II. The power of affected relative pairs. *Am. J. Hum. Genet.* **46**, 229–241.

Risch, N. 1990*c* Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. *Am. J. Hum. Genet.* **46**, 242–253.

Risch, N. 1990*d* Genetic linkage and complex diseases, with special reference to psychiatric disorders. *Genet. Epidem.* **7**, 3–16.

Risch, N. 1992 Corrections to 'Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs'. *Am. J. Hum. Genet.* **51**, 673–675.

Solomon, E. & Bodmer, W.F. 1979 Evolution of the sickle cell variant gene. *Lancet* **i,** 923.

Suarez, B.K., Rice, J. & Reich, T. 1978 The generalized sib pair IBD distribution: its use in the detection of linkage. *Ann. Hum. Genet.* **42,** 87–94.

Weeks, D.E., Lehner, T., Squires-Wheeler, E., Kaufmann, C. & Ott, J. 1990 Measuring the inflation of the LOD score over model parameter values in human linkage analysis. *Genet. Epidemiol.* **7**, 237–243.

Weissenbach, J., Gyapay, G., Dib, G. *et al.* 1992 A second-generation linkage map of the human genome. *Nature, Lond.* **359**, 794–801.

Williamson, J.A. & Amos, C.I. 1990 On the asymptotic behavior of the estimate of the recombination fraction under the null hypothesis of no linkage when the model is misspecified. *Genet. Epidemiol.* **7**, 309–318.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES