Genetic Epidemiology 2

Genetic linkage studies

M Dawn Teare, Jennifer H Barrett

Lancet 2005; 366: 1036–44

This is the second in a Series of seven papers on genetic epidemiology.

Mathematical Modelling and Genetic Epidemiology Division of Genomic Medicine, Henry Wellcome Laboratories for Medical Research, University of Sheffield, Beech Hill Road, Sheffield S10 2RX, UK (M D Teare PhD); and Genetic Epidemiology Division, University of Leeds, Leeds, UK (J H Barrett PhD)

Correspondence to: Dr M Dawn Teare
m.d.teare@sheffield.ac.uk

See http://research.marshfieldclinic.org/genetics/

Linkage analysis is used to map genetic loci by use of observations of related individuals. We provide an introduction to methods commonly used to map loci that predispose to disease. Linkage analysis methods can be applied to both major gene disorders (parametric linkage) and complex diseases (model-free or non-parametric linkage). Evidence for linkage is most commonly expressed as a logarithm of the odds score. We provide a framework for interpretation of these scores and discuss the role of simulation in assessment of statistical significance and estimation of power. Genetic and phenotypic heterogeneity can also affect the success of a study, and several methods exist to address such problems.

Genetic linkage analysis can be used to identify regions of the genome that contain genes that predispose to disease. We begin by explaining the principles of such analysis in the context of major gene disorders, then consider methods more suited to complex diseases that do not require the specification of a disease model (model-free or non-parametric linkage). Finally we consider the power and interpretation of linkage studies and the choice of phenotype.

Linkage and linkage disequilibrium are two key concepts in genetic epidemiology (see first paper in this series[1]). Two genetic loci are linked if they are transmitted together from parent to offspring more often than expected under independent inheritance. They are in linkage disequilibrium if, across the population as a whole, they are found together on the same haplotype more often than expected (see a later paper in this series[2]). In general, two loci in linkage disequilibrium will also be linked, but the reverse is not necessarily true.

Linkage extends over much longer regions of the genome than does linkage disequilibrium. Two loci are linked if, during meiosis, recombination occurs between them with a probability of less than 50%.[1] By contrast, every time recombination occurs between the loci in the population, the linkage disequilibrium between them is weakened, and is maintained only if the two loci are very close together. Linkage analysis is often the first stage in the genetic investigation of a trait, since it can be used to identify broad genomic regions that might contain a disease gene, even in the absence of previous biologically driven hypotheses.

Parametric linkage analysis

Parametric or model-based linkage analysis is the analysis of the cosegregation of genetic loci in pedigrees. Loci that are close enough together on the same chromosome segregate together more often than do loci on different chromosomes. Loci on different chromosomes segregate together purely by chance. Each genotype for one genetic marker or locus is made up of two alleles, one inherited from each parent. Specific alleles are in gametic phase when they are coinherited from the same parent—ie, they were present together in the single transmitted gamete originating from that parent. The further apart two loci are on the same chromosome, the more likely it is that a recombination event at meiosis will break up the cosegregation. The main quantity of interest in parametric linkage analysis is the recombination fraction θ (the probability of recombination between two loci at meiosis). By genotyping genetic markers and studying their segregation through pedigrees, it is possible to infer their position relative to each other on the genome. This process can be done to map genetic markers or to map disease or trait loci. There now exist many sets of linkage-mapping markers, in which the markers have been selected to be regularly spaced across the genome (for example, the Marshfield Clinic resource).

Ehlers-Danlos disease

As an example, figure 1 shows a pedigree segregating a form of the Ehlers-Danlos disease (EDS-VIII [MIM 130080]). We will use the reported linkage analysis of this pedigree[3] to illustrate parametric linkage analysis. EDS-VIII is a very rare autosomal dominant disorder. 72 individuals from five generations were clinically examined in this family, and DNA samples were available for genetic analysis from 19 of them. Figure 1 shows only those parts of the pedigree segregating the disease (ie, many unaffected individuals are not shown). Figure 1 also shows genotypes for 17 selected genetic markers spanning 30 centimorgans (cM) on chromosome 12. For example, individual six is homozygous for allele 1 (denoted 1 1) for marker D12S352, whereas no genotype (denoted – –) is available for D12S356. This 30-cM region contains many more markers than those indicated here. The bold black vertical line indicates a haplotype that is shared between affected individuals. In the third generation, affected people have coinherited the same haplotype at all 17 loci, except for individuals 17 and seven; for individual 17 a recombination has occurred at some stage between markers D12S100 and D12S1615. In the fourth generation, although three affected people still share the same full haplotype, there has been one recombination in individual three and evidence of two ancestral recombinations in individual 18. By ancestral, we mean recombinations that have occurred in ancestors

that we have been unable to directly discern through genotyping. However all affected people have coinherited the same segment of chromosome 12, which is about 7 cM long (flanked by markers D12S314 and D12S1695). Figure 1 shows that everyone who has inherited the haplotype 6–10–5-3–7-6–3 has EDS type VIII, whereas no unaffected people have inherited this haplotype. So this region of chromosome 12 could contain the dominant gene causing the disease.

LOD scores

Linkage is usually reported as a logarithm of the odds (LOD) score (panel 1). This score was first proposed by Morton in 1955.[5] It is a function of the recombination fraction (θ) or chromosomal position measured in cM. This means that the LOD score is different depending upon which value of θ is being considered. Large positive scores are evidence for linkage (or cosegregation), and negative scores are evidence against. To calculate a LOD score a model for disease expression must be specified. This model includes the frequency of the disease allele and mode of inheritance (eg, dominant or recessive), marker allele frequencies, and a full marker map for each chromosome. The ultimate objective of the analysis is to estimate the recombination fraction between individual markers and the disease locus (two-point) or position of the disease locus relative to a fixed map of markers where the location of each marker is assumed to be known (multipoint). The best (maximum likelihood) estimate of θ or position is that which maximises the lod score function: the maximum LOD score.

The higher the LOD score, the greater the evidence for linkage. Traditionally, a score of 3 was regarded as significant evidence of linkage. This is equivalent to p=0·0001.[6] This seemingly stringent level of significance is because of the low prior probability of linkage to any particular marker and was originally set to allow for the sequential testing that Morton envisaged would follow. Morton assumed that groups would collaborate and genotype the same markers in more and more families until the total LOD score, for a predetermined θ, reached 3 (linkage accepted at that value of θ) or –2 (linkage

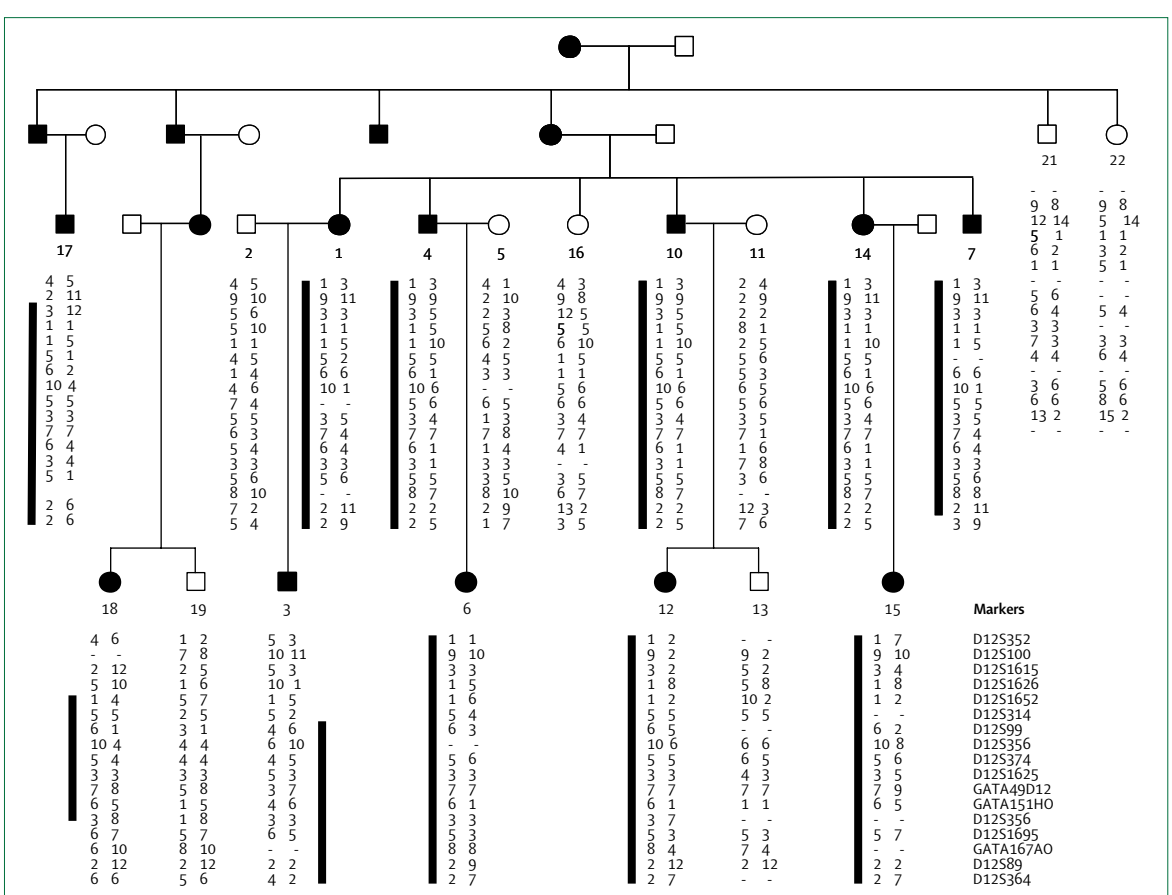


Figure 1: **Genotypes and haplotypes for family EDS-VIII for markers from chromosome 12p13**
Square=male. Circle=female. Filled=affected individual. Clear=unaffected individual. Position of each marker relative to D12S352 as follows: D12S100 (3·3cM), D12S1615 (4·6cM), D12S1626 (7·1cM), D12S1652 (7·6cM)(, D12S314 (11·4cM), D12S99 (12·6cM), D12S356 (14·2cM), D12S374 (14·2cM), D12S1625 (16·4cM), GATA49D12 (17·7cM), GATA151HO (17·7cM), D12S336 (19·0cM), D12S1695 (19·6cM), GATA167AO (20·3cM), D12S89 (23·2cM), D12S364 (29·4cM).
With permission of University of Chicago Press.[3]

Panel 1: LOD scores and likelihood ratios

Results of genetic linkage studies are often reported in the form of LOD scores. These scores are actually based on likelihood ratios. The use of maximum likelihood in genetic linkage analysis was originally proposed in 1947.[4] Its use became widespread once Morton[5] published his log-odd (LOD) tables, which enabled the sequential analysis of family-based linkage studies.

Maximum likelihood

Maximum likelihood provides a statistical framework to compare various hierarchical models and compute estimates of the various model parameters. The likelihood of the model, conditional on the data (represented as like[model]), is defined as the probability of the observations occurring, calculated according to the model. Hypotheses are tested by comparing two likelihoods (likelihood ratio test), the likelihood of an alternative model versus the likelihood of the null (or reduced) model. Under the null model, twice the natural logarithm of the ratio of the likelihoods is distributed as a χ^2. Extreme values of this test statistic are interpreted as evidence against the null hypothesis.

LOD scores

LOD score analysis is equivalent to likelihood ratio testing, but for historical reasons, instead of natural logarithms, logs to the base 10 are used. In the linkage analysis framework, the only parameter of interest is the recombination fraction (θ) between marker and disease locus or the map position of the disease locus with respect to a fixed map of markers. The null hypothesis represents no linkage between disease and marker locus (θ=0·5), and the alternative hypothesis assumes linkage exists (θ<0·5).
The LOD score function is then defined as:

$$LOD(\theta)=\log_{10}\left[\frac{Like(\theta)}{Like(\theta=\frac{1}{2})}\right]$$

The LOD score function is maximised with respect to θ—the recombination fraction in two-point analysis (a single marker and disease locus), or map position in multipoint analysis (disease locus and at least two markers at fixed relative positions).
The value of θ which gives the maximum LOD score is the maximum likelihood estimate of θ.

Heterogeneity LOD score

When linkage analysis is done allowing for more than one disease locus, the LOD score is maximised with respect to two parameters, θ and α (the proportion of families linked to this locus). The heterogeneity LOD score is defined as:

$$HLoD(\alpha,\theta)=\log_{10}\left[\frac{Like(\alpha,\theta)}{Like(\alpha=1,\theta=\frac{1}{2})}\right]$$

LOD score for non-parametric sib pair linkage analysis

The classic likelihood ratio test statistic obtained from a non-parametric sib pair linkage analysis is found by maximising the following ratio with respect to z_0 and z_1 (with $z_2=1-z_0-z_1$):

$$LR(z_0,z_1)=2\log_e\left[\frac{Like(z_0,z_1)}{Like(z_0=\frac{1}{4},\ z_1=\frac{1}{2})}\right]$$

This ratio can be converted into a LOD score for comparability with parametric analyses by dividing by 4·6 (ie, $2\times\log_e 10$), which changes the ratio to base 10 logarithms.

rejected). However, the common practice now is to maximise the LOD score over the recombination fraction. This increases the power to detect linkage, and linkage is viewed as being excluded for all values of θ at which LOD is less than –2. More recent work shows that a LOD score of 3 is equivalent to a genome-wide significance level of about 0·09.[7] In this theoretical work, it was assumed that researchers could genotype markers at very high density over the whole genome (ten markers per cM). A higher threshold of at least 3·3 would be necessary to ensure that the genome-wide type I error rate was in fact 0·05. Although the number of genetic markers used in a genome scan can be very large, once a certain density of markers is achieved, each new marker does not represent another independent statistical test and the threshold does not need to be increased.

Figure 2 is a plot of the LOD score function obtained when the disease gene is assumed to be at one of a series of regularly spaced positions. The maximum score is obtained when the disease gene is placed close to marker D12S356 (at 14·2 cM from D12S352). As we move away from this position, the LOD score decreases and at some points becomes very negative. Such large negative values are common in multipoint linkage analysis and this generally means that there is evidence that recombination(s) have arisen between marker(s) and the

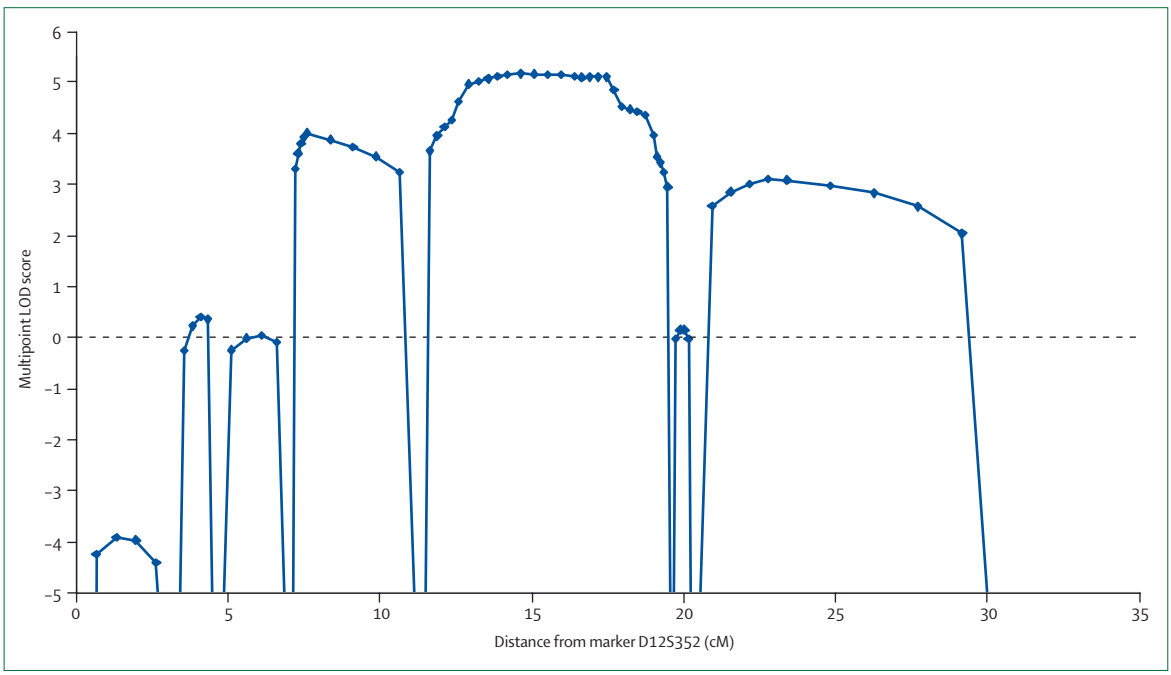


Figure 2: **Multipoint LOD score for EDS-VIII family by position (in cM) relative to marker closest to tip of chromosome 12 (D12S352 in this study)**
Order and cumulative distance between markers obtained from Marshfield Genetic Database and presented telomeric to centromeric, as follows: D12S352–3·3cM–D12S100–4·6cM–D12S1615–7·1cM–D12S1626–7·6cM–D12S1652–11·4cM–D12S314–12·6cM–D12S99–14·2cM–D12S356–14·2cM–D12S374–16·4cM–D12S1625–17·7cM–GATA49d12–17·7cM–GATA151h0–19·0cM–D12S336–19·6cM–D12S1695–20·3cM–GATA167a0–23·2cM–D12S89–29·4cM–D12S364. With permission of University of Chicago Press.[3]

disease locus. For example, such a recombination has occurred between D12S314 and D12S99 for individual three in figure 1. This person has inherited allele 2 (D12S314) from the affected mother and must have inherited the disease allele also. This situation can only happen if there is recombination between the marker and disease. When the LOD score is calculated assuming the disease gene is coincident with marker D12S314 (ie, these two locations are one and the same), we are assuming zero probability of recombination. So in calculating the LOD score at this point, we are fixing the recombination rate between the marker and disease locus to be zero. Joint consideration of data and model show this to be impossible or extremely unlikely, resulting in very strong evidence against linkage at that specific location.

Specifying the genetic model

For any parametric linkage analysis, the genetic model for the disease of interest must be specified. For a simple mendelian disease, this model amounts to mode of inheritance and frequency of disease allele. For some diseases, carrying the risk genotype does not always result in the individual being affected (incomplete penetrance). In more complex models, only a proportion of disease cases are due to a specific major gene, resulting in some risk of disease for individuals with any disease genotype (inclusion of a sporadic rate). Model parameters must be chosen before the linkage analysis. These model parameter estimates are preferably taken from population-based studies of the disease. Segregation analysis and estimation of familial relative risks can be used to ensure that appropriate models are used in the linkage analysis.

Genetic heterogeneity

The fact that the pattern of disease in families is consistent with a strong major gene component does not necessarily imply that only one gene is involved. There are many examples of diseases caused by inherited mutations in distinct genes. Some mutations give rise to the same disease but with a different mode of inheritance—for example, Charcot-Marie-Tooth disease has autosomal recessive, dominant, and X-linked forms, and mutations in up to ten genes are responsible for the different forms.[8] The EDS-VIII family illustrated yielded strong evidence of a disease gene on chromosome 12, but when four further smaller families were examined, two were not consistent with linkage to this region.

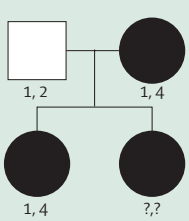
Heterogeneity LOD scores

Locus heterogeneity such as that with Charcot-Marie-Tooth disease can seriously affect the power of parametric linkage analysis. The most common solution is to assume that mutations in the disease genes will be so rare that each family will be linked to only one such gene. The genome scan is then done maximising a heterogeneity LOD score (panel 1). At each genomic position, the heterogeneity LOD score is maximised with respect to another parameter, α: the proportion of families linked to

this locus. If the genetic component of a disease is due to a few major genes, then power to detect linkage is reduced, but it might still be possible to detect the locus in this way. For example, the breast cancer genes *BRCA1* and *BRCA2* were detected by parametric linkage analysis despite heterogeneity.[9,10] If the genetic component is made up of a large number of distinct genes, parametric linkage analysis can be severely compromised and model-free alternatives become necessary, as discussed below.

Methods for reducing locus heterogeneity include limiting the analysis to strictly defined subtypes of disease (where there might be reason to suspect a stronger genetic component or where subtypes might themselves be diseases with distinct causes, each with a distinct genetic component) or targeting families in an isolated population where the number of original founder mutations could be low. Standard methods for accounting for heterogeneity assume that the genetic model for disease will be the same in linked and unlinked families. If genetic heterogeneity exists, then the estimates of model parameters from population genetics might no longer be appropriate when trying to identify individual loci contributing to the overall genetic component. It is therefore common for genome scans to be done for a range of parametric models allowing for heterogeneity. However, if the LOD score has been maximised over several models, then to maintain a low false positive rate the threshold (eg, 3·3) will need to be raised. The significance of the resulting maximum LOD score can be estimated by simulation.

***Panel 2:* Allele-sharing in sibling pairs**



In this affected sibling pair, the first sister has inherited allele 1 from her father and allele 4 from her mother. If the marker is unlinked to the disease, there are four equally likely combinations of alleles the second sister can inherit: (1, 1), (1, 4), (2, 1), or (2, 4), where the first number indicates the paternally inherited allele and the second number the maternally inherited allele. As illustrated in the table below, the numbers of alleles shared by the sisters that are IBD are 1, 2, 0 and 1 respectively. IBD sharing must be distinguished from the numbers of alleles shared that are identical by state (IBS). When the second sister has genotype (2, 1), both sisters have a type 1 allele; these alleles are IBS, but they are not IBD (assuming the parents are not inbred), since one is inherited from the father and one from the mother.

Genotype		Number of alleles shared	
Sibling 1	Sibling 2	IBD	IBS
1, 4	1, 1	1	1
	1, 4	2	2
	2, 1	0	1
	2, 4	1	1

Model-free (non-parametric) linkage analysis

For multifactorial diseases, where several genes (and environmental factors) might contribute to disease risk, there is no clear mode of inheritance. Methods to investigate linkage have therefore been developed that do not require specification of a disease model. Such methods are referred to as non-parametric, or model-free. The rationale is that, between affected relatives excess sharing of haplotypes that are identical by descent (IBD) in the region of a disease-causing gene would be expected, irrespective of the mode of inheritance. Various methods test whether IBD sharing at a locus is greater than expected under the null hypothesis of no linkage.

Sibling pairs

The simplest approach is to study sibling pairs, both of whom are affected. At any locus, according to the null hypothesis of no linkage, the number of IBD alleles shared by a pair of siblings is none with probability 0·25, one with probability 0·5, or two with probability 0·25 (panel 2). If IBD sharing in the families is known, the observed proportions of pairs sharing no, one, and two alleles at a candidate locus can be compared with these expectations. Linkage would be suggested if the pairs of siblings, both of whom are affected by a disease, share significantly more alleles IBD than expected by chance. The best test for linkage to use depends on the true mode of inheritance but in a wide range of situations the most powerful test is the so-called mean test, in which the mean number of alleles shared IBD is compared with the expected value of 1.[11]

In practice, IBD sharing between a pair of siblings is rarely known with complete certainty because the parents may not have been genotyped and the markers might not be sufficiently polymorphic to distinguish between sharing IBD or IBS (panel 2). In such cases, the proportions of IBD sharing can only be estimated. A general algorithm for calculating these proportions considers all possible parental genotypes that are consistent with the data.[12] More recently, maximum likelihood methods have been used.[13–17]

Other groups of relatives

Pairwise comparisons between relatives can easily be modified for types of relative pair other than siblings. However, in studies that set out to examine affected sibling pairs, additional affected siblings are often recruited. Various methods have been proposed to extend the pairwise approach to sibships larger than two. Selecting one pair at random or using only independent pairs means discarding information, so using all possible pairs is preferred. Should larger sibships be down-weighted to account for non-independence between pairs

with this approach, and if so, how?[18,19] Weighting might improve power, but since the type I error rate depends on factors such as the informativeness of the markers, it is recommended that significance levels be estimated by simulation,[19,20] a point we expand on below.

Alternative methods have been developed to analyse families with larger numbers of affected relatives of differing relationship, also based on the degree of IBD sharing. Each pedigree can be assigned a score that measures IBD sharing, and the test for linkage is based on comparing this score with the expected score according to the null hypothesis (combining over pedigrees). The score can be based on pairwise comparisons, but a more powerful alternative score has been proposed[21] that increases sharply as the number of affected members sharing the same allele IBD increases. This score is part of the program Genehunter.[22] In the absence of complete information, the score is replaced by its estimated value, leading to a conservative test. These methods have been modified to provide accurate likelihood-based tests,[23] all implemented in the much faster program Allegro.[24]

In linkage studies, genotyping is usually done at a set of linked markers (in some cases covering the whole genome). Here, IBD sharing can be estimated more accurately by multipoint analysis, by use of information from all markers on the chromosome. At any point along a chromosome, the pattern of inheritance within a pedigree can be described by an inheritance vector. This vector records, for each non-founding member of the family, whether they have inherited the grandpaternal or grandmaternal allele from each of their parents. The full inheritance vector might not be uniquely determined by the marker data, but its probability distribution conditional on the marker data can be calculated.[17] Calculation of the full multipoint IBD distribution for a pedigree is a non-trivial problem. Most available methods are based on algorithms that are limited either in the number of markers or in the complexity of the pedigrees that can be analysed. To cope with problems of large size in both dimensions, methods have been developed based on Markov chain Monte Carlo estimation.[25] Currently, whole genome screens are being done with several thousand single nucleotide polymorphisms,[26] increasing the number of markers by an order of magnitude over previous studies. An analytical method based on gene flow trees has been developed to handle such data, implemented in the program Merlin.[27]

There have been corresponding methodological developments in quantitative trait linkage analysis. In 1972, Haseman and Elston[12] suggested using sibling pairs to investigate linkage by regressing the squared difference in the siblings' trait values on the (estimated) proportion of alleles shared IBD. Two siblings that share more alleles IBD would be expected to have more similar trait values if the marker is linked to a gene influencing the trait. Consequently, in the presence of such linkage, there should be a negative relation between the squared trait differences and the estimated IBD sharing. More powerful variants of this method have since been developed that also incorporate information from the sum of the siblings' trait values, for example using the mean-corrected cross-product (sibling covariance) as the dependent variable[28] (panel 3). Variance components methods have been developed to analyse quantitative trait linkage in general pedigrees. These methods model the trait covariance between relatives, partitioning this value into components due to a specific chromosomal region (on the basis of estimated IBD sharing) and unlinked genes (on the basis of degree of kinship). These methods can now be done with multipoint methods to estimate IBD sharing (for example with the program SOLAR[29]).

Numerous whole genome screens in a wide range of complex diseases have now been done with model-free linkage analysis for both qualitative and quantitative traits. An early example was the affected sibling pair study of type 1 diabetes,[30] which successfully identified linkage to the HLA region with a LOD score of more than 7. Many quantitative traits related to cardiovascular disease have been investigated with some interesting results; for example in a whole genome screen of high-density lipoprotein-cholesterol, evidence of linkage to a locus on chromosome 9p was identified.[31] A review of genome screens of complex diseases[32] showed that disappointingly few studies published before 2001 were able to demonstrate significant linkage according to the criteria of Lander and Kruglyak[7] described below. With the larger study sizes and more focused sampling strategies often employed more recently, this situation may be improving.

Issues of power and interpretation

A fundamental issue in understanding the results of a linkage analysis is the interpretation of statistical significance. Whenever statistical tests are done, a balance must be struck between making claims many of which fail to be substantiated and adopting criteria so stringent that true findings are missed. For the parametric analysis of single gene disorders, it was suggested early that a threshold of 3 for the LOD score indicated a significant result at the genome-wide level. This approach has been instrumental in avoiding the reporting of large numbers of false-positive results, but at the same time allowing linkage analyses of single-gene disorders to lead successfully to the identification and cloning of disease genes.

The threshold issue is more contentious with complex traits. In 1995, Lander and Kruglyak[7] made proposals that have proved highly influential. Assuming a dense map of fully informative markers, they used mathematical theory to derive the threshold required for a LOD score to achieve genome-wide significance of 5%. Since the LOD scores used in different approaches have slightly different properties, the thresholds vary slightly from method to method. For parametric linkage analysis, a threshold of 3·3 is necessary, whereas in an affected sibling pair study

Panel 3: Quantitative trait linkage analysis using sibling pairs

Suppose we have a set of sibling pairs, and let the trait values for the jth sibling pair be x_{1j} and x_{2j}. In the original method proposed,[12] the squared difference in trait values $(x_{1j}-x_{2j})^2$ is regressed on the number of alleles shared IBD. This dependent variable ignores the information in the sum of the siblings' trait values, which would also be related to the number of alleles shared IBD under the alternative hypothesis. Combining the sum and the difference in trait values (corrected by the overall mean trait value μ), the following dependent variable is suggested:

$([x_{1j}-\mu]+[x_{2j}-\mu])^2-([x_{1j}-\mu]-[x_{2j}-\mu])^2$

which simplifies to give a multiple of the mean-corrected cross-product:

$4(x_{1j}-\mu)(x_{2j}-\mu)$

The expected value of the mean-corrected cross-product is just the sibling covariance.

significant linkage needs a LOD score of 3·6 (or 4·0 if the so-called possible triangle constraints[16] are used). From a whole genome scan Lander and Kruglyak suggested that areas of suggestive linkage (evidence expected to occur once overall by chance) and nominal linkage (p=0·05 from a single test without adjustment for multiple testing) should also be reported, although in the latter case no claims for linkage should be made.

The stringency of the criteria for genome-wide significance has been questioned, since they are based on the assumption of a dense marker map with no missing data. An alternative and flexible approach that takes into account the particular features of the study is to use simulation. Datasets can be simulated according to the null hypothesis of no linkage across the whole genome with the same family structures, marker map, allele frequencies, and patterns of missing data as in the study itself. With advances in computing, simulation is becoming the standard method of assessment in large studies.[31,33–36] This approach also has the advantage that the correct significance level can be identified for any method of analysis (including the different weighting methods for large families discussed above).

The threshold for statistical significance from simulation is likely to be lower than that from theoretical results; how much lower will depend on the features of the study. For a study of whole genome scans of siblings with multiple sclerosis that used simulation,[20] it was estimated that a LOD score of 3·2 would be achieved without linkage in only one in 20 studies—ie, a LOD of only 3·2 would be significant at the genome-wide level at 5%. Further investigation[37] suggested that map density and the extent of missing data have a substantial effect on significance levels. Even if no locus were to show significant evidence of linkage, a genome-wide study could contain independent peaks of linkage exceeding some lower threshold (eg, 2·0). In the locus counting approach, a series of such thresholds is considered, and the number of independent regions expected to exceed these thresholds under the null hypothesis of no linkage is estimated by simulation. This null distribution is then used to interpret the results of the genome screen, by determining whether or not more peaks were evident than would be expected by chance.

Considerations of genome-wide significance apply to whole genome scans, but it can be argued that the situation is not so much different in candidate gene linkage studies. Here, instead of whole genome scanning, regions containing genes selected on biological grounds are investigated. In practice, since for most diseases a good case can be made for a very large number of candidates, the significance of results should not be interpreted very differently from those of genome-wide scans. Researchers also often undertake numerous subgroup analyses, which can again inflate the false-positive rate if not interpreted correctly.

Simulation has an equally important role in study design. Genetic linkage studies can be expensive and investigators will not want to begin a study with low power. Power calculations by simulation will inform decisions about the number and type of families required and the necessary marker density. Generally, the more affected individuals in a pedigree, the more informative the family is. However, some familial configurations will be more informative than others. The most common difficulty is missing data—there might be little available information about older family members and not everyone will consent to providing DNA. Simulation allows investigators to estimate the power of the family collections at their disposal. If the power is judged adequate, an initial genome screen would be done with markers no more than 20 cM apart. Then, depending on the results from this first stage, further markers would be genotyped in promising candidate regions or further sets of genome-wide markers would be used to increase the density to every 5–10 cM.

The frequency of many diseases varies widely between populations. This differential incidence can be due to variations in both environmental and genetic background. For example, the autosomal recessive Tay-Sachs disease is 100 times more common in Ashkenazi Jews than non-Jews.[38] Site-specific cancer shows a very high degree of variation in incidence even within Europe.[39] Many forms of genetic differences have been identified between populations,[40] and studies might need to consider these additional sources of heterogeneity and if possible allow for them in the analysis. Such population differences usually reduce statistical power.

The limited success of linkage analysis for complex diseases so far is at least in part due to studies being too small to detect genes of modest effect. The interpretation of apparently negative findings depends crucially on power. The sample size necessary to detect linkage to genes with a genotype relative risk of less than 2 could be

unachievable.[41] Genotype error also affects power.[42] With large pedigrees, genotype error is easy to detect because such errors often lead to mendelian inconsistencies within the pedigree, but where only affected sibling pairs are genotyped, with no other family data, incorrect genotypes will probably not be spotted.

Choice of phenotype

Some traits or diseases have a clear phenotype definition. For simple mendelian traits, it is straightforward to identify affected and unaffected individuals and even in a disease such as cancer, once symptoms are experienced the diagnosis is based on pathological findings. However, other illnesses such as psychiatric disorders are more problematic because the diagnosis often depends upon several distinct symptoms, and there is often disagreement as to what constitutes a definitive diagnosis.[43] The absence of a clear definition of phenotype will lead to uncertainty about the classification of affected and unaffected individuals, and to potential inconsistency between studies. Sometimes use of a quantitative trait can circumvent this difficulty; for example, the number of distinct symptoms could be used as a measure of disease severity. Conversely, there might sometimes be good reason to transform a quantitative trait into a binary one. For instance, an individual may be classed as obese if his or her body-mass index is above a defined threshold and non-obese if it is below it. However, simplifying a quantitative trait to a binary phenotype can result in loss of power if an inappropriate threshold is used.[44]

Genetic linkage studies are very rarely done with population-based family datasets. Usually some other selection criteria are applied to the phenotype before the families are selected. These criteria are often driven by the need to maximise power and to reduce heterogeneity. A disease is aetiologically heterogeneous if it can result from more than one distinct pathway. Families are usually selected because of segregation of the disease of interest, and might only be studied if many members are affected. There might also be a focus on severely affected individuals such as those with early age at onset or those with other critical symptoms. Some diseases, such as Charcot-Marie-Tooth disease, might be classified into clear clinical subtypes. Sometimes diseases can be associated with other phenotypes, and families can be categorised as to whether or not the other phenotype is present—eg, families with breast cancer with or without ovarian cancer within the pedigree. Eligibility criteria such as these can reduce heterogeneity, and it can be helpful at the analysis stage if datasets can be split into meaningful subgroups. However, these devices might cause problems later when trying to interpret the linkage finding in terms of the general population. For example, a study that finds that half of highly selected families are linked to a specific locus might not be able to predict the proportion of the disease in the general population that is due to this locus.

If the phenotype of interest is a diagnosis requiring treatment or registration, the eligible families will often be ascertained via specialist clinics. In other cases, the phenotype itself might not be a treatable disease but a risk factor for disease. A good example is obesity. Obesity can be measured in different ways and studies can be difficult to compare. There is also the problem of ascertainment. Most linkage studies of obesity have arisen through datasets designed to study other primary endpoints such as heart disease, osteoporosis, and diabetes. There have been over 30 published linkage studies of obesity-related phenotypes.[45] Many of these individually large studies have reported significant linkage, but few such findings have been replicated, indicating not only genetic heterogeneity and low power, but also the heterogeneity of study designs and choice of phenotype.

One way of tackling replication in complex diseases is meta-analysis of pooled linkage study results.[46,47] However, meta-analysis works best when studies have been done under homogeneous conditions, when phenotypes have been measured with the same criteria, and when statistical analyses have been similar. Datasets might need to be recoded before a meta-analysis is done. For complex phenotypes such as obesity, where ascertainment is so variable, collaborative analyses on raw data are essential.

Linkage analysis: what next?

A linkage analysis of the whole genome can identify regions that show evidence of containing a disease gene. In the study of mendelian traits, crossover events often narrow down the region sufficiently to define a small interval of interest. Linkage analysis of complex diseases can only identify large regions (typically tens of cM). Location estimates indicated by the linkage peak are highly variable, and increasing the density of the marker map only somewhat improves the resolution.[48] Although a very strong candidate gene might exist within the linkage region, such regions often contain hundreds of genes, many of which are biologically plausible candidates. One way to narrow the region in studies of cancer is examination of loss of heterozygosity in tumours.[49] When markers that are heterozygous in germline DNA exhibit loss of heterozygosity in tumour cells, this can indicate deletion of a region of the chromosome, and the pattern of such loss can be used to narrow down the location of a tumour suppressor gene. Other approaches are based on linkage disequilibrium, which extends over much smaller distances than linkage, and is the subject of the next paper in this series.[2]

References

1 Burton PR, Tobin MD, Hopper JL. Key concepts in genetic epidemiology. *Lancet* 2005; **366:** 941–51.

2 Cordell HJ, Clayton DG. Genetic association studies. *Lancet* (in press).

3 Rahman N, Dunstan M, Teare MD, et al. Ehlers-Danlos syndrome with severe early-onset periodontal disease (EDS-VIII) is a distinct, heterogeneous disorder with one predisposition gene at chromosome 12p13. *Am J Hum Genet* 2003; **73:** 198–204.

4 Haldane JBS, Smith CAB. A new estimate of the linkage between the genes for colour-blindness and haemophilia in man. *Ann Eugen* 1947; **14:** 10–31.

5 Morton NE. Sequential tests for the detection of linkage. *Am J Hum Genet* 1955; **7:** 277–318.

6 Chotai J. On the LOD score method in linkage analysis. *Ann Hum Genet* 1984; **48:** 359–78.

7 Lander ES, Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* 1995; **11:** 241–47.

8 Berger P, Young P, Suter U. Molecular cell biology of Charcot-Marie-Tooth disease. *Neurogenetics* 2002; **4:** 1–15.

9 Hall JM, Lee MK, Newman B, et al. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* 1990; **250:** 1684–89.

10 Wooster R, Bignell G, Lancaster J.et al. Identification of the breast cancer susceptibility gene BRCA2. *Nature* 1995; **378:** 789–92.

11 Blackwelder WC, Elston RC. A comparison of sib-pair linkage tests for disease susceptibility loci. *Genet Epidemiol* 1985; **2:** 85–97.

12 Hasemen JK, Elston RC. The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* 1972; **2:** 3–19.

13 Risch N. Genetics of IDDM: evidence for complex inheritance with HLA. *Genet Epidemiol* 1989; **6:** 143–48.

14 Risch N. Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. *Am J Hum Genet* 1990; **46:** 242–53.

15 Risch N. Linkage strategies for genetically complex traits. II. The power of affected relative pairs. *Am J Hum Genet* 1990; **46:** 229–41.

16 Holmans P. Asymptotic properties of affected-sib-pair linkage analysis. *Am J Hum Genet* 1993; **52:** 362–74.

17 Kruglyak L, Lander ES. Complete multipoint sib-pair analysis of qualitative and quantitative traits. *Am J Hum Genet* 1995; **57:** 439–54.

18 Suarez BK, Van Eerdewegh P. A comparison of three affected-sib-pair scoring methods to detect HLA-linked disease susceptibility genes. *Am J Med Genet* 1984; **18:** 135–46.

19 Holmans P. Likelihood-ratio affected sib-pair tests applied to multiply affected sibships: issues of power and type I error rate. *Genet Epidemiol* 2001; **20:** 44–56.

20 Sawcer S, Jones HB, Judge D, et al. Empirical genomewide significance levels established by whole genome simulations. *Genet Epidemiol* 1997; **14:** 223–29.

21 Whittemore AS, Halpern J. A class of tests for linkage using affected pedigree members. *Biometrics* 1994; **50:** 118–27.

22 Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. Parametric and nonparametric linkage analysis: a unified approach. *Am J Hum Genet* 1996; **58:** 1347–63.

23 Kong A, Cox NJ. Allele-sharing models: LOD scores and accurate linkage tests. *Am J Hum Genet* 1997; **61:** 1179–88.

24 Gudbjartsson DF, Jonasson K, Frigge ML, Kong A. Allegro, a new computer program for multipoint linkage analysis. *Nat Genet* 2000; **25:** 12–13.

25 Sobel E, Lange K. Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *Am J Hum Genet* 1996; **58:** 1323–37.

26 Kennedy G, Shephard N, Cao M, et al. Whole genome scan in a complex disease using 11,245 SNPs. *Am J Hum Genet* 2003; **73** (suppl): 186.

27 Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin: rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 2002; **30:** 97–101.

28 Elston RC, Buxbaum S, Jacobs KB, Olson JM. Haseman and Elston revisited. *Genet Epidemiol* 2000; **19:** 1–17.

29 Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet* 1998; **62:** 1198–211.

30 Davies JL, Kawaguchi Y, Bennett ST, et al. A genome-wide search for human type 1 diabetes susceptibility genes. *Nature* 1994; **371:** 130–136.

31 Arya R, Duggirala R, Almasy L, et al. Linkage of high-density lipoprotein-cholesterol concentrations to a locus on chromosome 9p in Mexican Americans. *Nat Genet* 2002; **30:** 102–05.

32 Altmüller J, Palmer LJ, Fischer G, et al. Genomewide scans of complex human diseases: true linkage is hard to find. *Am J Hum Genet* 2001; **69:** 936–50.

33 Broeckel U, Hengstenberg C, Mayer B, et al. A comprehensive linkage analysis for myocardial infarction and its related risk factors. *Nat Genet* 2002; **30:** 210–14.

34 Caulfield M, Munroe P, Pembroke J, et al. Genome-wide mapping of human loci for essential hypertension. *Lancet* 2003; **361:** 2118–23.

35 Williams NM, Norton N, Williams H, et al. A systematic genomewide linkage study in 353 sib pairs with schizophrenia. *Am J Hum Genet* 2003; **73:** 1355–67.

36 Xu J, Meyers DA, Ober C, et al. Genomewide screen and identification of gene-gene interactions for asthma-susceptibility loci in three US populations: collaborative study on the genetics of asthma. *Am J Hum Genet* 2001; **68:** 1437–46.

37 Wiltshire S, Cardon LR, McCarthy MI. Evaluating the results of genomewide linkage scans of complex traits by locus counting. *Am J Hum Genet* 2002; **71:** 1175–82.

38 Kaback MM, Rimoin DL, O'Brien JS. Tay-Sachs Disease: Screening and Prevention. New York: Alan R Liss, 1977.

39 Bray F, Sankila R, Ferlay J, Parkin DM. Estimates of cancer incidence and mortality in Europe in 1995. *Eur J Can* 2002; **38:** 99–166.

40 Cavalli-Sforza LL, Menozzi P, Piazza A. The history and geography of human genes. Princeton University Press, Princeton, New Jersey, 1996.

41 Risch N. Searching for genetic determinants in the new millennium. *Nature* 200; **405:** 847–56.

42 Douglas JA, Boehnke M, Lange K. A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data. *Am J Hum Genet* 2000; **66:** 1287–97.

43 Kennedy JL, Farrer LA, Andreasen NC, Mayeux R, St George Hyslop P. The genetics of adult-onset neuropsychiatric disease: complexities and conundra? *Science* 2003; **302:** 822–26.

44 Duggirala R, Williams JT, Williams-Blangero S, Blangero J. A variance component approach to dichotomous trait linkage analysis using a threshold model. *Genet Epidemiol* 1997; **14:** 987–92.

45 Chagnon YC, Rankinen T, Snyder EE, Weisnagal SJ, Perusse L, Bouchard C . The human obesity gene map: the 2002 update. *Obesity Res* 2003; **11:** 313–67.

46 Wise LH, Lanchbury JS, Lewis CM. Meta-analysis of genome searches. *Ann Hum Genet* 1999; **63:** 263–72.

47 Levinson DF, Levinson MD, Segurado R, Lewis CM. Genome scan meta-analysis of schizophrenia and bipolar disorder, part I: methods and power analysis. *Am J Hum Genet* 2003; **73:** 17–33.

48 Roberts SB, MacLean CJ, Neale MC, Eaves LJ, Kendler KS. Replication of linkage studies of complex traits: an examination of variation in location estimates. *Am J Hum Genet* 1999; **65:** 876–84.

49 Parrella P, Fazio VM, Gallo AP, et al. Fine mapping of chromosome 3 in uveal melanoma: identification of a minimal region of deletion on chromosomal arm 3p25·1-p25·2. *Cancer Res* 2003; **63:** 8507–10.