STATISTICS IN MEDICINE
Statist. Med. **18**, 2961–2981 (1999)

TUTORIAL IN BIOSTATISTICS
GENETIC MAPPING OF COMPLEX TRAITS

JANE M. OLSON*, JOHN S. WITTE AND ROBERT C. ELSTON

Department of Epidemiology and Biostatistics, Rammelkamp Center for Education and Research, MetroHealth Campus, Case Western Reserve University, Cleveland, Ohio, U.S.A.

SUMMARY

Statistical genetic mapping methods are powerful tools for finding genes that contribute to complex human traits. Mapping methods combine knowledge of the biological mechanisms of inheritance and the randomness inherent in those mechanisms to locate, with increasing precision, trait genes on the human genome. We provide an overview of the two major classes of mapping methods, genetic linkage analysis and linkage disequilibrium analysis, and related concepts of genetic inheritance. Copyright © 1999 John Wiley & Sons, Ltd.

1. INTRODUCTION

In recent years, genetic study of complex human traits has increased dramatically. Most human diseases are now viewed as having some genetic component, and considerable effort is being made to find and study the genes involved. As a result, statistical methods used to find disease genes are receiving a great deal of attention, and improvements in methodology are continually being proposed. In this article we provide an overview of genetic mapping methods. We first explain concepts in genetic inheritance, focusing primarily on the genetic mechanisms that investigators exploit in genetic mapping, and introduce relevant terminology. We then introduce the reader to the two main areas of mapping methodology: genetic linkage analysis and linkage disequilibrium mapping.

* Correspondence to: Jane M. Olson, Department of Epidemiology and Biostatistics, Rammelkamp Center for Education and Research, 2500 MetroHealth Drive, Case Western Reserve University, Cleveland, Ohio 44109, U.S.A. E-mail: olson@darwin.cwru.edu

Contract/grant sponsor: National Center for Human Genome Research
Contract/grant number: HG01577

Contract/grant sponsor: National Center for Research Resources
Contract/grant number: PR03655

Contract/grant sponsor: National Cancer Institute
Contract/grant number: CA73270

Contract/grant sponsor: National Institute of General Medical Sciences
Contract/grant number: GM28345

CCC 0277–6715/99/212961–21$17.50

Received June 1998
Accepted January 1999

2. GENETIC TERMINOLOGY AND LINKAGE CONCEPTS

2.1. Genetic Models

Simple genetic models are derived from *Mendelian* laws of inheritance. Each individual has two sets of 23 *chromosomes*, one maternal and one paternal in origin. One of the 23 pairs of chromosomes are the *sex chromosomes*, and we shall concern ourselves with the remaining 22 pairs of *autosomal chromosomes* in this tutorial. Each chromosome consists of a long strand of *DNA*, a linear molecule with units known as *base pairs*. A chromosomal location (which may be a single base pair or a collection of consecutive base pairs) is termed a genetic *locus*. At each locus, there may be distinct variants, called *alleles*. In common parlance, the term *gene* is often used to denote both locus and allele, but the two should be regarded as distinct concepts by the statistician. For an individual, the pair of alleles (maternal and paternal) at a locus is called the *genotype*. A genotype is called *homozygous* if the two alleles are the same allelic variant and *heterozygous* if they are different allelic variants. If more than one locus is involved, the pattern of alleles for a single chromosome is called a *haplotype*; together, the two haplotypes for an individual is still called a (multilocus) genotype. Each offspring receives at each locus only one of the two alleles from a given parent; alleles are transmitted randomly (that is, each with probability 1/2), and offspring genotypes are independent conditional on the parental genotypes. The probability that a parental genotype transmits a particular allele or haplotype to an offspring is called the *transmission* probability, and is the first component of a genetic model.

The second component of a genetic model concerns the relationship between the (unobserved) genotypes and the observed characteristics, or *phenotype*, of an individual. The phenotype may be discrete or continuous. We define *penetrance* to be the probability (mass or density) of a phenotype given a genotype; a complete genetic model requires specification of the penetrances of all possible genotypes. The third component of a genetic model is the (distribution of) relative frequencies of the alleles in the population. These *allele frequencies* are used primarily to determine prior probabilities of genotypes when inferring genotype from phenotype.

These three components, taken together, fully describe the *genetic model* of a trait. Given a set of phenotypic data on pedigrees, one can estimate the genetic model using statistical techniques collectively known as *segregation analysis*.[1,2] While segregation analysis is beyond the scope of this paper, it is helpful to realize that in a segregation analysis, genotypes are latent variables inferred from trait phenotypes. For simple Mendelian traits, in which only one genetic locus is segregating, estimation of the genetic model is usually straightforward, as only one set of latent variables (genotypes) is involved. For complex traits, which are the emphasis of most genetic studies today and which are probably due to the effects of more than one locus, estimation of the genetic model is difficult to impossible, because each locus represents a different set of (possibly interacting) latent variables. As a result, two approaches to genetic linkage analysis have evolved: those that require prior specification of a genetic model for the trait under study (model-based methods), and those that do not (model-free methods). For a more detailed review of genetic models and genetic likelihoods, see Thompson.[3] We now discuss concepts specific to linkage analysis.

2.2. Recombination and Linkage

Two loci that are on the same chromosome are said to be *syntenic*. If they are close enough together, the alleles at the two loci that are paternal (maternal) in origin tend to pass to the same

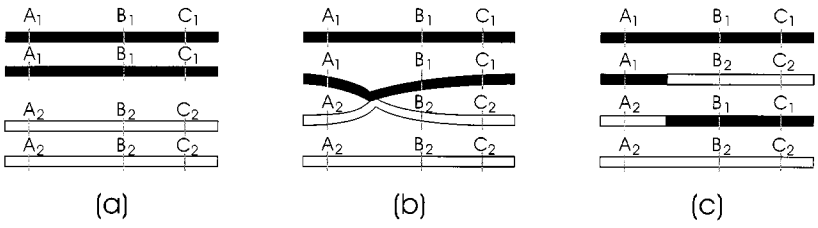

Figure 1. Diagram showing part of two homologous chromosomes at the time of gamete formation. In (*a*), the two chromosomes have paired up and each chromosome (parental solid, maternal open) has divided into two identical chromatids. Alleles at three loci (*A*, *B* and *C*) are indicated. In (*b*), a cross-over forms between loci *A* and *B*. In (*c*), the four resulting gametes are shown; two have recombined between the loci *A* and *B*, two between *A* and *C*, none between *B* and *C*

gamete (sperm or egg) and hence are transmitted together to an offspring; we thus have *cosegregation* at the two loci. However, when the chromosomes pair up together at the time of gamete formation (a process known as *meiosis*), portions of the paternal and maternal chromosomes interchange by a process known as *crossing over* (Figure 1). If an odd number of cross-overs occurs between two loci, then the alleles at the two loci that an offspring receives from one parent are no longer identical to those that occur in one of the parental chromosomes, but rather have *recombined*. Thus, at one locus the offspring receives from a parent an allele that comes from a grandmother, and at the other locus the allele from the same parent comes from a grandfather (Figure 1). The closer the loci are together, the smaller the probability of a recombination, and hence the larger the probability of cosegregation, a phenomenon known as genetic *linkage*.

The proportion of gametes in which recombination is expected to occur between two loci is the *recombination fraction* between them, usually denoted θ. If the two loci are far apart, segregation at one locus is independent of that at the other, and $\theta = 1/2$; all four types of gametes are produced in equal frequencies. When linkage occurs, $0 \leqslant \theta < 1/2$, and the 'parental-type' gametes are more frequent than the 'recombined-type' gametes. In linkage analysis, one locus is a measured locus called a genetic *marker*, for which the genotype is known with a high degree of certainty, and the other is a disease locus with unknown genotype, for which the genotype is inferred (with varying degrees of accuracy) only through the disease or trait phenotype. After typing marker loci at known locations in the genome, we can test each marker for linkage to a disease or trait and approximate the location of the disease or trait to the chromosomal region harbouring the linked marker.

If the parent is heterozygous at both loci, the recombination can be observed and the parent (or resulting offspring) is said to be *informative* for linkage. Otherwise, although an odd number of cross-overs may have occurred between the two loci, the recombination cannot be observed. The further apart two loci are, the higher the probability of recombination between them and this leads to the concept of *genetic distance*. Genetic distance is correlated with physical distance, but there is no simple function that relates them because the frequency of recombination changes as one moves along the chromosomes. The unit of recombination is the *Morgan*, defined as the distance in which exactly one cross-over is expected to occur. A Morgan is divided into *centiMorgans* (*cM*), where 1 Morgan = 100 cM.

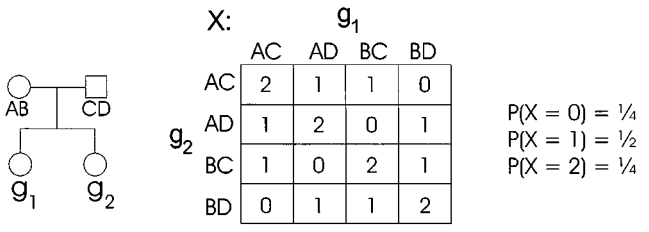

Figure 2. Prior distribution of the number of alleles shared identical-by-descent (X) by a sib pair

2.3. Identity-by-descent

A pair of related individuals shares an allele *identical-by-descent* (IBD) if that allele has a common ancestral source, that is, if the allele in each relative is from the same chromosome of the same ancestor. In the context of linkage analysis, the common ancestor is taken to be a recent ancestor, one within the sampled pedigree. For example, if the pair is a sib pair, the common ancestors are their parents, and the sibs may have inherited the same paternal allele and/or maternal allele at a particular locus. The concept of IBD plays an important role in linkage analysis, particularly model-free linkage analysis, because it forms the basis of a measure of genetic similarity of a pair of related individuals. If a pair of relatives shares alleles IBD at one genetic locus, they will also share alleles IBD at a second, linked locus with high probability, because linked loci tend to cosegregate. Generally, one correlates the extent of marker IBD sharing to some measure of disease or trait similarity to determine the genomic location of the disease or trait locus.

We now introduce notation that will be useful in later discussion. Consider a pair of siblings and a single genetic locus. Figure 2 shows the distribution of genotype pairs inherited from their parents according to Mendelian laws of segregation. The sibs are expected to share 0, 1 or 2 alleles IBD with probability 1/4, 1/2 and 1/4, respectively. In Figure 2, we have assumed that both parents are different heterozygotes and that IBD sharing can be determined with certainty. More generally, the number of alleles shared IBD by a particular sib pair, a, can be determined probabilistically. For notational use, let f_i, $i = 0, 1, 2$, be the prior (unconditonal) probability that the sib pair shares i alleles IBD at a locus, and $\hat{f}_i$ the estimated probability that the sib pair shares i alleles IBD conditional on available marker data, denoted I_{m}. A general form for computing IBD sharing probabilities is

$$\hat{f}_i = \frac{P(a = i, I_{\mathrm{m}})}{P(I_{\mathrm{m}})}$$

where the denominator is the probability, or likelihood, of the pedigree marker data and may be computed using an Elston–Stewart[4] ('peeling') algorithm, and the numerator can be written as a sum of the terms of the denominator consistent with sharing i alleles IBD, each term representing a *phase-known* (that is, the maternal or paternal origin is known for each allele) pedigree genotype.[5,6] Finally, let $\hat{\pi} = \frac{1}{2}\hat{f}_1 + \hat{f}_2$ be the estimated proportion of alleles shared IBD conditional on available marker data.

Algorithms for computing IBD sharing probabilities using data from multiple linked markers is an area of on-going research.[7–12] Use of multiple markers, generally termed *multipoint* linkage analysis, increases the power of a linkage study by increasing the overall marker informativity in

Table I. Haplotype frequencies for two diallelic loci under linkage equilibrium

Disease allele	Marker allele		Total
	B	b	
A	$p_A p_B$	$p_A(1-p_B)$	p_A
a	$(1-p_A)p_B$	$(1-p_A)(1-p_B)$	$1-p_A$
Total	p_B	$1-p_B$	1

a chromosomal region. Multipoint computational algorithms usually employ a hidden Markov model that assumes that IBD sharing at consecutive loci behaves in a first-order Markov manner.

2.4. Linkage Disequilibrium

The phenomenon of linkage results in the cosegregation of the alleles of two linked loci and thus in within-family association of specific alleles. Among families, it is common to assume in a linkage analysis that no association between the allelic variants of different loci is present, as different marker alleles may cosegregate with the disease allele in different families. If no population association exists between alleles at two loci, the loci are said to be in *linkage equilibrium*, and the population frequencies of each two-locus haplotype are the products of the single-locus allele frequencies (Table I). If a population association does exist, the loci are said to be in *disequilibrium*.

As an example of disequilibrium, assume we have two diallelic loci, one a disease locus with alleles A and a (with frequencies p_A and $1-p_A$, respectively), and the other a marker locus with alleles B and b (with frequencies p_B and $1-p_B$, respectively). Under random mating, the frequencies of the two-locus genotypes are determined by the products of the frequencies of the four haplotypes AB, Ab, aB and ab, which change over time (usually measured in generations).[13–15] Specifically, at a particular point in time, let the haplotype frequencies for AB, Ab, aB and ab be

$$h_{AB} = p_A p_B + D$$
$$h_{Ab} = p_A(1-p_B) - D$$
$$h_{aB} = (1-p_A)p_B - D$$
$$h_{ab} = (1-p_A)(1-p_B) + D$$

respectively, where $D = h_{AB}h_{ab} - h_{Ab}h_{aB}$ is the departure from equilibrium; that is, D is the disequilibrium between A and B.

The magnitude of disequilibrium between the disease and marker alleles dissipates as future generations of a population emanate and recombination occurs. How quickly equilibrium is reached depends on the number of new generations that have passed since the disease allele arose (mutation age), and the genetic distance between the disease and marker loci. Fewer recombinations imply that the disease and marker alleles are in stronger disequilibrium. This inverse relation implies that when there is very close linkage, disequilibrium may persist for long periods

of evolutionary time.[16] Other effects influence allele and haplotype frequencies in the population over evolutionary time, including *genetic drift*, the random component of allele frequency change over generations. The magnitude of genetic drift is inversely related to the size of the portion of the population relevant to mating and gene transmission, called the *effective population size*.

3. MODEL-BASED LINKAGE ANALYSIS

In a model-based linkage analysis we completely specify the mode of inheritance of the trait being studied: the number of loci involved; the number of alleles at each locus and their frequencies; and the penetrances of each genotype (which may further depend on age or other covariates). Typically, for computational reasons we assume that the trait is caused by segregation of just two alleles at a single locus, and that there is no other cause of familial aggregation for the trait. Thus one allele frequency and three penetrances need to be specified. The marker allele frequencies are also specified, but we shall see that these have no effect on the evidence for linkage if the marker genotypes of all the pedigree *founders* (those pedigree members from whom all other pedigree members are descended) are known or can be inferred with certainty.

Denoting the joint probability of all genotypes $P(g)$, and the joint probability of all observed data x (trait and marker phenotypes) conditional on genotypes $P(x|g)$, the likelihood for a set of data is

$$L = \sum_g P(g)P(x|g)$$

where the summation is over all the possible joint genotypes g (trait and marker) for all pedigree members. We assume that the only unknown parameter in this likelihood is the recombination fraction θ, on which $P(g)$ depends. (We shall assume in this account that θ is a scalar, though more generally it may be a vector if, for example, multiple marker loci are involved, or θ is made sex-dependent.) Thus the likelihood of interest is

$$L(\theta) = \sum_g P(g|\theta)P(x|g)$$

and we base inferences about θ on the likelihood ratio $\Lambda = L(\theta)/L(1/2)$ or, equivalently, its logarithm. In human genetics it is usual to take logarithms to base 10 and we define the lod score at θ to be $Z(\theta) = \log_{10}[L(\theta)/L(1/2)]$, with a maximum $Z(\hat{\theta})$ at the maximum likelihood estimate $\hat{\theta}$. Bernard[17] introduced the term lod, carefully distinguishing between the *forward* lod and the *backward* lod. The former is the logarithm of the odds for a hypothesis, that is, the probability that the hypothesis is true divided by the probability that it is false; the latter is now called the logarithm of the likelihood ratio. In the genetics literature lod is often mistakenly interpreted as the logarithm of the odds for linkage. Note that if $L(1/2) > L(\theta)$ for some value of θ, then the corresponding lod score is negative.

The vector of genotypes g can be partitioned into those that pertain to the pedigree founders, g_{f}, and those that pertain to the non-founders, g_{n}. We can thus write

$$\sum_g P(g|\theta) = \sum_{g_{\mathrm{f}}} \sum_{g_{\mathrm{n}}} P(g_{\mathrm{f}})P(g_{\mathrm{n}}|g_{\mathrm{f}}, \theta)$$

where $P(g_f)$ is determined by the marker and trait allele frequencies, while $P(g_n|g_f, \theta)$ is determined by the transmission probabilities. Furthermore, if the trait and marker genotypes are independently distributed in the population, an assumption often realistic when linkage analysis is performed, we can write $P(g_f) = P(g_{fm})P(g_{ft})$, where g_{fm} and g_{ft} are, respectively, the founders' marker and trait genotypes. It follows that, if the founders' marker genotypes are known, that is, there is only one possible marker genotype for each founder,

$$\sum_{g_{fm}} \sum_{g_{ft}} P(g_{fm})P(g_{ft}) = P(g_{fm}) \sum_{g_{ft}} P(g_{ft}),$$

so that $P(g_{fm})$ is a constant multiplier in the likelihood. It follows that the marker allele frequencies are then irrelevant when making inferences about θ.

Model-based linkage analysis has been described in detail by Ott.[18] Here we highlight some of the main statistical results when the trait is known to be *monogenic* (that is, caused by segregation of alleles at a single locus) and then briefly discuss the effects of analysing data under this assumption when it is false. Traditionally, $Z(\hat{\theta}) > 3$ has been taken as 'proof' of linkage.[19,20] From general likelihood theory, under the null hypothesis $\theta = 1/2$, $2[\log_e 10]Z(\hat{\theta})$ is asymptotically distributed as χ_1^2 if $Z(\hat{\theta})$ is a maximum, so that $Z(\hat{\theta}) > 3$ corresponds asymptotically to a χ_1^2 value $> 13{\cdot}8$, which translates to $p < 10^{-4}$ if we allow for the fact that we want a one-sided test of $\theta = 1/2$. Use of such an extremely small p-value was chosen in an attempt to limit to 0·05 the probability of making an error when concluding that linkage is present, using the fact that the prior probability of linkage between two random autosomal loci in the human genome is about 0·054. (If we assume all 22 pairs of autosomal chromosomes have equal length, the probability that two random loci are on the same chromosome is $1/22 = 0{\cdot}045$; the figure 0·054 allows for the different lengths of the chromosomes.) On the assumption that there is no appropriate prior probability of linkage in the case of complex traits, Lander and Kruglyak[21] proposed that the appropriate p-value should be based on the multiple testing performed when the whole genome is scanned for linkage, whether or not such a scan has been performed.[22,23]

Model-based linkage analysis is often used with guessed values of the disease allele frequencies and penetrances, and this will not inflate the significance of a result (that is, probability statements about the data on the assumption $\theta = 1/2$) provided the disease is in fact monogenic and there are no errors in the probability model assumed for the marker (it is not necessary for the marker to be error-free – only that the allele frequencies and penetrance functions for it to be correct).[24,25] Furthermore, under the assumptions underlying the likelihood, we can maximize the lod score over both θ and the parameters that describe the mode of inheritance of the trait, to obtain consistent estimates of these latter parameters, and, provided the pedigrees are ascertained on the basis of the trait only, under the assumptions the lods do not depend on the mode of ascertainment.[26,27]

Model-based linkage analysis has been used to show that more than one locus can cause a simple Mendelian disease. This was first done by determining if the recombination fraction is heterogeneous among pedigrees, indicating $\theta < 1/2$ in some and $\theta = 1/2$ in others, using the usual heterogeneity chi-square.[28] Later, Smith[29] proposed a model in which a proportion α of the families exhibit linkage with recombination fraction θ, while a proportion $1 - \alpha$ show no linkage. Thus the likelihood was formulated as $L(\alpha, \theta) = \alpha L(\theta) + (1 - \alpha)L(1/2)$ and maximized over $0 \leqslant \alpha \leqslant 1$ and $0 \leqslant \theta \leqslant 1/2$. We test the null hypothesis of no heterogeneity, $\alpha = 1$, by comparing the usual likelihood ratio statistic to a chi-square distribution with one d.f., taking half the

indicated probability because of the one-sided nature ($\alpha < 1$) of the alternative. In order to test for linkage in the presence of heterogeneity, we can use the same likelihood to define the model, but now the null hypothesis is $\theta = 1/2$, with α free to vary between 0 and 1. Because α is irrelevant when $\alpha = 1$, it is difficult to derive the null distribution of the likelihood ratio statistic.[30,31]

These tests for heterogeneity, or for linkage in the presence of heterogeneity, assume that $L(\theta)$ and $L(1/2)$ are appropriate for a particular mode of trait inheritance. The assumption of a monogenic mode of inheritance, and the reliance on $Z(\hat{\theta}) > 3$ in general as the criterion for accepting linkage, have led to anomalous statements being made in the genetics literature. For example, it is believed by some that a more powerful test for linkage, in the case of a complex disease, can be obtained by using trait model parameter estimates that overestimate the genetic effect of the locus being linked, because these estimates are more likely to result in $Z(\hat{\theta}) > 3$ in the presence of linkage. This argument ignores the fact that the significance level associated with such a test is affected by ignoring trait familial correlations that are due to loci other than the one that is modelled in the linkage analysis. However, it is generally recognized that estimates of θ made under a wrong model are usually biased upwards, and that a trait locus cannot be excluded from a genomic region on the basis of $Z(\theta)$ being less than, for example, -2; what is excluded by such a criterion is the existence, within a recombination fraction θ of the marker, of a trait locus with the particular mode of inheritance assumed in the calculation of $Z(\theta)$.

4. MODEL-FREE LINKAGE ANALYSIS

In contrast to model-based linkage methods, model-free linkage methods do not depend on prior specification of a model of inheritance for the disease or trait of interest. In other words, the frequencies and penetrances of disease genotypes need not be known in advance, and functions of these quantities and the recombination fraction may be estimated. It is important to recognize, however, that many of the methods do rely on assumptions about the underlying genetic model and some methods are in fact parametric or semi-parametric in nature. Generally, however, assumptions about the genetic model affect only parameter estimation and not the validity of the method for detecting linkage. In this section, we differentiate between two general types of model-free linkage analysis – those designed for qualitative traits and those designed for quantitative traits – although both theory and applications of these two groups of methods overlap. Model-free linkage methods typically evaluate marker locus IBD relationships among family members, often pairs of siblings, and thus are often referred to as relative-pair, or sib-pair, methods.

4.1. Qualitative Trait

Model-free linkage methods designed for qualitative traits usually consider samples of *affected sib pairs* or sibships with at least two affected members. If a trait and marker are linked, affected sib pairs should share more marker alleles IBD than expected by chance. Assuming that at least one genetic locus contributes to trait variability, under the null hypothesis of no linkage, sib pairs are expected to share exactly 0, 1 or 2 alleles IBD at a single marker locus with respective probabilities $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{1}{4}$. If the marker IBD state can be determined with certainty, so that a sample of n pairs can be partitioned into n_0, n_1 and n_2 pairs, corresponding to sharing 0, 1 or 2 alleles IBD, the data can be modelled using a multinomial distribution. If the marker IBD state cannot

be determined with certainty, then IBD probabilities can be used to model the data using a hidden multinomial framework.

In this context, estimated IBD probabilities can be used to compute non-parametric test statistics or parametric likelihoods. The most commonly used non-parametric test statistic, called the mean test,[32] has power close to optimal for most one-locus genetic models,[33–35] and compares the observed mean proportion of marker alleles shared IBD to its null value of $\frac{1}{2}$:

$$T_{\mathrm{m}} = \frac{[n_2 + n_1/2]/n - 1/2}{[1/(8n)]^{1/2}}.$$

A more general mean statistic substitutes $\bar{\hat{\pi}} = \sum_{j=1}^{n} \hat{\pi}_j/n$ for $[n_2 + n_1/2]/n$ and an empirical variance estimate for the denominator when IBD sharing cannot be determined with certainty.

Parametric modelling of affected sib pairs is based on the hidden multinomial distribution.[36] Let z_i be parameters defined as the probability that an affected sib pair shares i marker alleles IBD. The distribution of affected sib pair IBD sharing can be parameterized in terms of the *relative risk* of disease to siblings (λ_{s}) and offspring (λ_{o}) of affected individuals (that is, the risks to these individuals relative to the population prevalence of disease), and the recombination fraction between trait and marker loci (θ), assuming a single genetic locus underlies the disease:

$$z_0 = \frac{1}{4} - \frac{1}{4\lambda_{\mathrm{s}}}(2\psi - 1)[(\lambda_{\mathrm{s}} - 1) + 2(1 - \psi)(\lambda_{\mathrm{s}} - \lambda_{\mathrm{o}})]$$

$$z_1 = \frac{1}{2} - \frac{1}{2\lambda_{\mathrm{s}}}(2\psi - 1)^2[(\lambda_{\mathrm{s}} - \lambda_{\mathrm{o}})$$

$$z_2 = \frac{1}{4} + \frac{1}{4\lambda_{\mathrm{s}}}(2\psi - 1)[(\lambda_{\mathrm{s}} - 1) + 2(1 - \psi)(\lambda_{\mathrm{s}} - \lambda_{\mathrm{o}})]$$

where $\psi = \theta^2 + (1 - \theta)^2$. Note that, using the hidden multinomial framework, only two free parameters, which are functions of λ_{s}, λ_{o} and θ, can be estimated, as the z_i must sum to one. The relative risks are therefore identifiable only when θ is known, such as when there is complete linkage ($\theta = 0$). In terms of the relative risks and θ, the hypotheses of interest are in fact composite hypotheses. Under the null hypothesis, either $\theta = 1/2$ or $\lambda_{\mathrm{s}} = \lambda_{\mathrm{o}} = 1$, or both. The alternative hypothesis requires both linkage and a genetic effect: $H_1: \theta < 1/2$, $\lambda_{\mathrm{s}}, \lambda_{\mathrm{o}} > 1$.

For a single affected sib pair (ASP), the lod score for the pedigree marker data (MD) can be written

$$Z(\mathrm{MD}|\mathrm{ASP}) = \log_{10} \sum_{i=0,1,2} \frac{z_i \hat{f}_i}{f_i}. \tag{1}$$

Under the null hypothesis, $z_i = f_i$, for $i = 0, 1, 2$ and $Z(\mathrm{MD}|\mathrm{ASP}) = 0$. The lod score (1), summed over independent pairs, can then be maximized over the z_i. The likelihood ratio statistic equals $2\log_{\mathrm{e}} 10$ times the maximum lod score. The asymptotic distribution of the likelihood ratio statistic is non-standard, as constraints on the z_i consistent with a one-locus genetic model ($z_0 \geqslant 0$, $z_2 + z_0 \geqslant z_1$ and $z_1 \geqslant 2z_0$) are usually imposed, giving mixture of χ_0^2, χ_1^2 and χ_2^2 random

variables.[37] In a later data example, we will refer to this statistic as the ASP (affected sib pair) lod score.

Affected sib pair methods are particularly useful for detecting linkage to complex diseases, which are expected to the *oligogenic* (that is, have a few underlying loci). The single-locus lod score provides a valid test of linkage even if more than one locus contributes to a disease. Multilocus models can also be fitted.[38–40] A general multilocus model allows for *epistasis*, or interactions among loci. Specialized multilocus models are available that treat the loci as interacting multiplicatively or additively.

When small pedigrees are available for linkage analysis, affected sib pair analysis may not capture all the linkage information in the sample, and non-parametric model-free approaches have been proposed to analyse general pedigree structures. An example of a non-parametric approach can be found in Kruglyak *et al.*,[41] who propose calculating a scoring function $S(v, x)$ (first proposed by Whittemore and Halpern[42]) that depends on an inheritance pattern v and the observed disease phenotypes x in the pedigree. When the inheritance pattern is unknown, one computes its conditional expectation

$$\bar{S}(x) = \sum_{v} S(v, x) P(v)$$

where $P(v)$ is estimated using available marker data. The authors further discuss a model-free scoring function that considers IBD sharing among sets of affected family members. In a later data example, we will refer to this statistic as the NPL (non-parametric linkage) statistic.

4.2. Quantitative Trait

Many biomedical traits of interest are measured on a continuous or ordinal scale. Development of model-free methodology to study linkage between a marker locus and a locus underlying a quantitative trait has, for the most part, proceeded separately from methods for studying linkage to qualitative traits. It is useful to recognize, however, that both sets of methods employ similar concepts, most notably that of IBD sharing, and that the methods are often used in multiple settings. For example, if both affected and disordant sib pairs are collected, disease status can be treated as a quantitative trait using a method originally intended for continuous traits. Conversely, some investigators studying quantitative traits sample siblings from the extremes of a continuous distribution and treat the outcome as a qualitative variable.

We first consider the problem of linkage to a quantitative trait by assuming that a single genetic locus underlies a quantitative trait. For an observation X from the trait distribution, the genetic model may be written

$$X = \mu + g + e$$

where μ is an overall mean, g is effect of the genotype at the major locus and e is a residual effect with an unspecified distribution. The variance of g is a function of the parameters of the genetic model (allele frequencies and penetrances) and can be partitioned into *additive* (σ_a^2) and *dominance* (σ_d^2) components; the dominance genetic variance measures the variance due to the effect on penetrance of interaction between the individual's two alleles.

The squared difference $Y = (X_1 - X_2)^2$ between the measurements of a quantitative trait for a randomly sampled pair of siblings is a linear function of the Bayesian (taking the known allele frequencies as prior probabilities) estimate of the proportion of marker alleles shared IBD

between the members of the pair ($\hat{\pi}$) and the estimated probability that the pair share exactly 1 marker allele IBD ($\hat{f}_1$), that is

$$E(Y \mid I_m) = \alpha_s + \beta_s \hat{\pi} + \gamma_s \hat{f}_1$$

where

$$\beta_s = -2\sigma_g^2(1-2\theta)^2$$

$$\gamma_s = \sigma_d^2(1-2\theta)^4$$

and α is an intercept containing no linkage information. This model is the well-known Haseman–Elston model,[6] and is often implemented assuming $\gamma_s = 0$. As with the affected sib pair lod score model, only two free parameters containing linkage information can be estimated, and the genetic variance components are identifiable if θ is known. Under the null hypothesis of no linkage ($\theta = 1/2$) or no genetic effect ($\sigma_g^2 = \sigma_d^2 = 0$), the regression parameters equal zero. After fitting the regression model using least squares, an asymptotically normal one-sided Wald-type test of linkage may be constructed based on the estimate of β.[6]

As in the case of a qualitative trait, extensions to include families larger than sibships are of interest. Extensions to the Haseman–Elston regression model have been proposed that model all informative relative pairs in an extended family using regression relationships derived for specific relationships, such as half-sibling, grandparental and cousin.[43] Generalized estimating equations can be used to combine the information from multiple relative pairs in a sibship or extended family.[44] Likelihood-based approaches designed for extended families have also been proposed. For example, one might treat the pedigree quantitative trait values as multivariate normal, with a covariance matrix that is parameterized in terms of variance components, IBD sharing and the recombination fraction.[45,46] Parameters may be estimated using maximum-likelihood methods, if multivariate normality of errors is assumed, or by estimating-equation approaches. Modelling the trait covariance rather than the squared-pair trait difference gives an increase in power.

To this point, we have assumed that sib pairs, or pedigrees, are randomly sampled. Often, an investigator selects families based on the value of the trait of interest or some related characteristic. For example, an investigator interested in stroke might wish to study the genetics of blood pressure in families ascertained through stroke patients or through patients attending a blood pressure clinic. Such selected samples often contain more information about linkage by increasing the frequency of the 'interesting' alleles in the sample, relative to the population. For the purpose of detecting linkage (as opposed to estimating genetic variance), the methods we have discussed thus far can be used on families acquired through some type of specialized sampling scheme, provided the selection criteria depend only on the trait phenotype. On the other hand, by tailoring the statistical model to allow for the sampling scheme used, one can often maximize power as well.

For example, suppose we believe that high values of a particular trait are due in large part to a low frequency allele at some unknown locus. In this case, we might sample index cases from the upper tail of the trait distribution to increase the frequency of this allele in the sample of sib pairs, and thus also increase the potential information for linkage. A regression model developed specifically for this sampling scheme is more powerful than the Haseman–Elston model when applied to the resulting data.[47,48] Another design that has high power for most, but not all, genetic models is the extreme discordant sib pairs design, in which pairs comprising one sib with a trait value from the upper tail and the other from the lower tail of the trait distribution are

sampled.[49,50] For example, given a large sample of probands with an extreme value in one direction (usually that indicating disease), one might genotype only those pairs for which the sibling has a trait value in the opposite tail.

5. LINKAGE DISEQUILIBRIUM MAPPING

The genetic variants that one might be interested in mapping arise through, for example, novel mutations or immigration of carriers of mutant alleles into a population. When a mutation initially arises, it has a particular chromosomal location and specific neighbouring marker alleles. At this incipient point in time, the mutation is completely associated with the adjacent alleles; it is only observed when the marker alleles are also present.[51] Marker alleles that were in the neighbourhood of the disease gene when its mutation was introduced into the population will generally remain nearby over numerous generations (that is, in disequilibrium). One can estimate whether a particular marker locus appears to be in disequilibrium with a disease locus. In particular, if specific marker allele frequencies are higher among diseased versus normal chromosomes (for example, in unrelated unaffected subjects), this suggests linkage between that locus and a disease locus. The extent of this disequilibrium depends on the number of subsequent generations since the mutation was introduced into the population, the recombination between the disease and marker alleles, mutation rates and selective values. This allelic disequilibrium is commonly referred to as 'linkage disequilibrium', although linkage need not be present for disequilibrium to exist; allelic association is a better term to describe the general phenomenon.[16]

While model-based and model-free linkage analysis approaches have proved successful for mapping many disease and trait genes, in some gene mapping investigations the limited number of meioses occurring within available pedigrees limits one's ability to detect recombination events between closely spaced (<1 cM) loci.[52] One can instead use information on all recombinations occurring since the incipient mutation, and attempt to map disease genes more finely by disequilibrium.

Mapping by disequilibrium entails determining the relative location of a disease locus by comparing marker allele locations with estimators of the relation between the corresponding alleles and disease alleles.[53,54,55] These include measures of association and recombination between disease and marker alleles. As with linkage analysis, relevant properties of the disease locus are inferred based on phenotype.

The most basic linkage disequilibrium efforts contrast single markers with disease. Let B denote a single marker allele that is being evaluated in relation to a disease allele A. Standard linkage disequilibrium models assume that a randomly mating population is in a steady state; a constant size and in equilibrium between the effects of genetic drift and recombination. These steady-state assumptions imply that the number of generations that have passed since the mutation was introduced is of the same order as the constant effective population size, N_e. Under these conditions, the squared correlation between linked disease and marker alleles is

$$\rho^2 = D^2/[p_A(1-p_A)p_B(1-p_B)],$$

using the notation given in the above subsection on disequilibrium concepts. For a particular data set, one can estimate the squared correlation by substitution of observed frequencies. This observed squared correlation is equivalent to χ^2/n, where χ^2 is the standard test statistic from a 2×2 table of observed haplotype counts (Table II), and n is the total number of haplotypes.[60]

Table II. Observed haplotype counts for two diallelic loci

Disease allele	Marker allele		Total
	B	*b*	
A	n_{AB}	n_{Ab}	n_A
a	n_{aB}	n_{ab}	n_a
Total	n_B	n_b	n

Table III. Observed counts for transmission/disequilibrium test

Non-transmitted allele	Transmitted allele		Total
	A	*a*	
A	*a*	*b*	$n_{\cdot A}$
a	*c*	*d*	$n_{\cdot a}$
Total	$n_{A\cdot}$	$n_{a\cdot}$	n

One can also investigate disequilibrium using the transmission/disequilibrium test (TDT).[55–59] The TDT compares the frequencies of alleles transmitted from parents to diseased offspring with those of alleles that are not transmitted (Table III). Because each parent of an affected offspring contributes exactly one transmitted and one non-transmitted allele, the TDT is simply the McNemar test resulting from a matched case-control design. The TDT provides a joint test of linkage and association (that is, linkage in the presence of association or vice versa). By combining evidence of association between marker and disease alleles among families (disequilibrium) and within families (linkage), the TDT provides increased power to detect disease gene location.

The squared correlation between linked disease and marker alleles can also be written as

$$\rho^2 = 1/(4N_e\theta + 1) \tag{2}$$

where θ is the recombination fraction.[60] Note that (2) models disequilibrium as being inversely related to the recombination fraction. If one has estimates of N_e and ρ^2, then (2) provides an estimate of the recombination fraction θ. In large, stable populations, however, ρ^2 can be quite small.[61] Furthermore, due to large variances, ρ^2 can be relatively uninformative for estimating θ.[62,63] Note that θ and N_e are completely confounded (N_e must generally be estimated from external data). There exists in the literature numerous other formulae for the expected value of ρ^2, though there is no general exact explicit equation.[62] Devlin and Risch[64] and Guo[65] evaluate the properties of some of these disequilibrium measures.

For most populations the conventional linkage disequilibrium assumption of a steady-state (that is, equilibrium) population over evolutionary time is unreasonable. Furthermore, isolated or small populations that have recently undergone rapid expansion can have more linkage disequilibrium,[61] and can provide reasonable estimates of founding population size and founding date (to approximate the mutation's age) while addressing issues of admixture. When one does not have an equilibrium population, the Luria–Delbrück method for estimating mutation rates in

exponentially growing bacteria[66] can be applied to estimate recombination frequencies – using disequilibrium information – in populations.[52] To model this in a basic fashion, we assume that all disease alleles arise from the original mutation introduced into the population. Then, m successive generations after the ancestral mutation was introduced, under complete random mating, disequilibrium can be modelled as

$$D_m = (1 - \theta)^m D_0 \tag{3}$$

where D_i is the disequilibrium at generation i, $i = 0, \ldots, m$, and the two loci in disequilibrium are linked with recombination fraction θ. If one assumes that there was complete disequilibrium when the disease mutation was introduced into the population, then $D_0 = 1$. An estimate of D_m is given by the observed proportion of diseased individuals who have the marker allele. With this and an estimate of the number of generations m (that is, the age of the mutation), the corresponding recombination fraction θ can be estimated. This approach, however, may not provide very accurate confidence bounds for the recombination fraction.[67,51] Another approach for dealing with non-steady-state populations entails using a Poisson branching process to model disease progression in a growing population and simulations to estimate the likelihood for the recombination fraction and corresponding support intervals.[68]

Likelihood-based linkage disequilibrium approaches provide a unifying framework for fine-scale mapping. Hill and Weir[60] derive likelihoods for D conditional on the observed number of haplotypes, N_e and θ. One can also use a likelihood approach to estimate disequilibrium while accounting for demographic factors that might affect a population's steady state (that is, growth, sampling effects and genealogical associations).[69] When observed data appear compatible with a large number of potential ancestries, however, evaluating likelihoods may become cumbersome.[70] Ultimately, iterative approaches will be required to estimate disequilibrium.[71] For example, one can obtain linkage disequilibrium estimates from population-based data by using the EM algorithm[72] or Markov chain Monte Carlo simulation approaches.[54] When mapping more complex traits, linkage disequilibrium approaches may require extension to allow for multiple markers. For multiple alleles and/or loci, basic extension of the single marker disequilibrium measures presented above have been developed.[67,68,73,74] Like linkage analysis, multipoint disequilibrium can be more efficient than single-marker analysis.[54,64,75] Furthermore, using multiple markers can provide estimates for ancillary population ancestry parameters and possibly give more accurate translation from genetic to physical disease.[70] A likelihood-based multipoint approach to linkage disequilibrium mapping loci can be found in Terwilliger.[75] When a narrow region is being considered for linkage disequilibrium fine-scale mapping, conditioning on the distances between markers allows the use of a composite likelihood to extract information from multiple markers.[76] Xiong and Guo[54] give a general likelihood framework for linkage disequilibrium mapping that incorporates multiallelic markers, multiple loci and mutational processes at the disease and marker alleles. Finally, estimating linkage disequilibrium measures is easiest and most efficient if one has haplotype data, which will generally require familial information.[16] Nevertheless, linkage disequilibrium estimates can still be obtained from genotypic data.[77]

6. EXAMPLE

To illustrate the use of some of the methods described in this tutorial, we simulated a set of 25 small pedigrees. Each pedigree contains marker and disease phenotype data for both parents and two offspring. We assumed that, in the population, the marker locus has 20 equally frequent

Table IV. Example genetic data set

Family	Disease status*		Marker phenotype			
	Mother	Father	Mother	Father	Offspring	
1	1	2	13, 15	1, 17	1, 15	1, 15
2	1	2	6, 11	1, 12	1, 6	1, 6
3	2	1	1, 11	8, 14	1, 14	1, 14
4	2	2	1, 8	3, 5	1, 5	1, 3
5	1	2	11, 13	1, 4	1, 11	1, 11
6	1	1	3, 17	5, 7	5, 17	3, 7
7	1	2	16, 18	1, 10	1, 18	1, 16
8	2	1	1, 6	10, 13	1, 13	1, 13
9	1	2	16, 20	1, 19	1, 20	1, 16
10	2	1	1, 15	8, 10	1, 8	1, 10
11	2	1	1, 15	8, 16	15, 16	1, 16
12	2	1	1, 5	11, 17	1, 11	1, 11
13	2	1	1, 5	11, 19	1, 19	5, 19
14	1	2	10, 17	1, 5	1, 10	1, 10
15	2	1	1, 20	8, 12	1, 12	1, 8
16	2	1	1, 18	17, 20	1, 20	1, 20
17	1	1	3, 8	2, 7	3, 7	3, 7
18	2	1	1, 18	7, 16	1, 7	1, 16
19	2	1	1, 19	3, 12	1, 12	1, 3
20	1	1	10, 19	1, 18	1, 10	1, 19
21	2	1	1, 12	3, 15	1, 3	1, 3
22	2	1	1, 14	5, 7	1, 7	1, 5
23	2	1	1, 16	6, 10	1, 6	6, 16
24	2	1	1, 17	9, 19	1, 9	1, 9
25	1	1	6, 16	1, 12	1, 6	1, 6

* 1, unaffected; 2, affected. All offspring are affected

alleles, numbered from 1 to 20 that were measured without error, so that the marker phenotype and genotype are the same. Individuals with one or two copies of the disease allele are affected with probability 0·8, while individuals with no copy of the disease allele are affected with probability 0·05 (*sporadics* or *phenocopies*). (In such a situation, we say that the disease allele is *dominant* to the normal allele. If individuals with two copies of the disease allele are affected with high probability, while those with zero or one copy are affected with low probability, we say that the disease allele is *recessive* to the normal allele.) In our simulation, we further assumed that the disease and marker loci are tightly linked and perfectly correlated; chromosomes carrying marker allele 1 also carry the disease allele and chromosomes carrying other marker alleles do not carry the disease allele. The frequency of the disease allele in the population is therefore assumed to be 0·05.

We sampled only those families with two affected offspring and further required that the parents carry four distinct marker alleles. The first requirement increases the probability that the sampled families are segregating for the disease allele and is a common practice in genetic studies. The second requirement was imposed artificially so that the reader can observe segregation more directly. In practice, however, selection procedures that rely on both marker and disease phenotype may be biased, so care should be used in the design of such studies. The data are shown in Table IV.

Table V. Analysis of example data set

Method	Statistic	*p*-value
Model-based		
Dominant	2·40 (lod score)	0·00044
Recessive	0·87 (lod score)	0·02262
Model-free		
ASP	2·39 (lod score)	≈0·00090
NPL	3·32 (normal)	0·00018
TDT (2-allele)	34·78 (χ^2_1)	$1{\cdot}1 \times 10^{-7}$
TDT (20-allele)	69·62 (χ^2_{19})	$3{\cdot}7 \times 10^{-9}$

We then tested for linkage and/or association using several methods and report the test statistics and approximate *p*-values in Table V. We first estimated the recombination fraction assuming the (correct) dominant model. The recombination fraction was correctly estimated to equal 0·0. The lod score, when converted to a likelihood ratio statistic and compared to a χ^2_1 distribution (one-sided) showed considerable evidence for linkage. If the genetic model was incorrectly assumed to be recessive, so that the penetrance of the heterozygote was 0·05, evidence in favour of linkage decreased considerably.

We then computed two model-free linkage statistics, the affected sib pair (ASP) lod score, which uses no information about parental affected status, and the non-parametric linkage (NPL) statistic, which uses some information about parental affected status. Both statistics give considerable evidence for linkage. The NPL statistic follows a normal distribution for a one-sided test, while the ASP likelihood ratio statistic follows a distribution that is a mixture of χ^2_0, χ^2_1 and χ^2_2 and is roughly stochastically slightly smaller than a χ^2_1, so we report the *p*-value associated with χ^2_1 as a crude approximation.

To illustrate the presence of the evidence in favour of linkage, consider the ASP lod score. Because we have chosen families in such a way that the markers are *fully informative*, that is, identity-by-descent (IBD) sharing can be determined with certainty, one can easily show that, of the 25 ASPs, 1 shares zero alleles IBD, 12 share one allele IBD, and 12 share two alleles IBD. Using the multinomial distribution, estimates of the IBD sharing parameters (z_0, z_1, z_2) are (0·04, 0·48, 0·48), which are quite different from their null hypothesis (no linkage) values of (0·25, 0·5, 0·25). In particular, the estimated mean number of alleles shared IBD $(0{\cdot}48 + 2(0{\cdot}48) = 1{\cdot}44)$ is larger than the value of 1·0 expected if no linkage is present. The lod score is computed as follows:

$$\text{lod score} = 1 \times \log_{10}\left(4 \times \frac{1}{25}\right) + 12 \times \log_{10}\left(2 \times \frac{12}{25}\right) + 12 \times \log_{10}\left(4 \times \frac{12}{25}\right) = 2{\cdot}39.$$

By only using sib pair IBD sharing, the ASP lod score excludes information about the marker contribution from affected parents. Model-based methods compute the likelihood of the entire pedigree, conditional on the genetic model of inheritance. Consider, for example, family 7. The allele shared IBD (allele 1) is inherited from the affected parent, which provides additional support in favour of linkage. If the allele shared IBD had been inherited from the unaffected

parent, support in favour of linkage would have been reduced. The ASP lod score makes no distinction between these two cases. The model-based dominant lod score and the ASP lod score gave about the same *p*-value. If the disease had been solely due to the disease locus (no sporadics), the model-based lod score would have exceeded the ASP lod score by a larger amount. In our data set, however, five families had offspring whose disease status was not due to the disease allele, but to random effects uncorrelated with genetic status. In these five families, the model-based lod score contribution was negative and cancelled out some of the positive evidence for linkage from the other families. For the traditional ASP lod score, such evidence against linkage is included only by failing to increase the lod score, not by subtracting from it. Had the sporadic rate been higher, the ASP lod score may have been more powerful than the model-based lod score.

Transmission/disequilibrium tests (TDTs) were also performed. TDTs detect linkage only in the presence of association (or association only in the presence of linkage). Because we treated all affected offspring as independent, our results are valid tests of linkage, but not association, as affected sibs in the same family are independent only under the null hypothesis of no linkage. We report the results of two tests, a McNemar χ^2_1 test of allele 1 versus all others, and a marginal homogeneity χ^2_{19} test of all alleles. Both tests are highly significant for linkage, reflecting the fact that TDT tests include information about both linkage and association, while the linkage tests above ignored the strong among-family association between marker allele 1 and disease status (that is, disequilibrium). For these data, TDTs, which require the presence of both linkage and association, proved most powerful but, depending on the history and structure of the population under study, power may fall to zero if a different marker had been chosen, even if the new marker were also tightly linked.

The *p*-values we report here are referred to as *pointwise p*-values because they concern only the marker locus being tested. In larger studies, researchers may perform similar tests on hundreds of marker loci, leading to multiple testing concerns that continue to be debated.[22] It has been suggested that, in the context of scanning the human genome for evidence of linkage to any of 200–300 markers, a pointwise *p*-value be less than 2×10^{-5} to be considered statistically significant at a *genomewise* level of $\alpha = 0{\cdot}05$.[21]

We simulated these data using a model of a highly heritable disease. Most tests designed to detect linkage or association would have detected linkage in these data. In fact, detection of linkage to highly heritable diseases is now routinely performed with great success. Many real diseases, however, are considerably more complex and therefore more challenging. The usefulness of model-free as opposed to model-based methods, and of linkage methods as opposed to disequilibrium methods, for detecting genes underlying complex diseases remains the subject of vigorous discussion. As our discussion of the results of our analyses indicate, even data in which the presence of linkage is obvious give different answers for a variety of reasons. In practice, it is difficult to predict in any given data set which of the many methods available will be best able to detect a disease gene.

7. DISCUSSION

Recently there has been an explosion of genetic mapping studies. Human diseases and traits of all types and levels of complexity are the subject of gene searches. However, there is much that is not known about the statistical properties of these methods, including the effects of sampling design and proper procedure for genome-wide inference. Further, new methods and new ideas for

statistical approaches to complex disease mapping regularly appear in the literature. As a result, there are many opportunities for statistical theorists to contribute in meaningful ways. The analyses of the example data set were intended to demonstrate both the diversity of methods and the practical challenges the genetic statistician faces when searching for disease susceptibility genes. Unfortunately, because genetic terminology and the specialized nature of the methodology can appear at the same time overwhelming and too narrowly focused, few general medical statisticians become familiar with this area. We hope that this overview of genetic mapping methods will encourage more statisticians to pursue the subject further. The recently published *Encyclopedia of Biostatistics*[78] contains many entries concerning genetic analysis and serves as an excellent starting point.

ACKNOWLEDGEMENTS

This work was supported in part by U.S. Public Health Service grants HG01577 from the National Center for Human Genome Research, RR03655 from the National Center for Research Resources, CA73270 from the National Cancer Institute and GM28345 from the National Institute of General Medical Sciences.

REFERENCES

1. Elston, R. C. 'Some recent developments in the theoretical aspects of segregation analysis', *in* Majumder, P. P. (ed.), *Human Population Genetics*, Plenum Publishing, New York, 1993, pp. 117–137.
2. Elston, R. C. 'Modern methods of segregation analysis, *in* Moolgavkar, S. H. and Prentice, R. L. (eds), *Modern Statistical Methods in Chronic Disease Epidemiology*, Wiley, New York, 1986, pp. 213–224.
3. Thompson, E. A. 'Genetic epidemiology: a review of the statistical basis', *Statistics in Medicine*, **5**, 291–302 (1986).
4. Elston, R. C. and Stewart, J. 'A general model for the analysis of pedigree data', *Human Heredity*, **21,** 523–542 (1971).
5. Amos, C. I., Dawson, D. V. and Elston, R. C. 'The probabilistic determination of identity-by-descent sharing for pairs of relatives from pedigrees', *American Journal of Human Genetics*, **47**, 842–853 (1990).
6. Haseman, J. K. and Elston, R. C. 'The investigation of linkage between a quantitative trait and a marker locus', *Behavior Genetics*, **2**, 3–19 (1972).
7. Idury, R. M. and Elston, R. C. 'A faster and more general hidden Markov model algorithm for multipoint likelihood calculations', *Human Heredity*, **47**, 197–202 (1997).
8. Kruglyak, L. and Lander, E. S. 'Complete multipoint sib-pair analysis of qualitative and quantitative trait data', *American Journal of Human Genetics*, **57**, 439–454 (1995).
9. Kruglyak, L., Daly, M. J. and Lander, E. S. 'Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping', *American Journal of Human Genetics*, **56**, 519–527 (1995).
10. Lander, E. S. and Green, P. 'Construction of multilocus genetic maps in humans', *Proceedings of the National Academy of Sciences USA*, **84**, 2363–2367 (1987).
11. Sobel, E. and Lange, K. 'Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics', *American Journal of Human Genetics*, **58**, 1323–1337 (1996).
12. Kruglyak, L. and Lander, E. S. 'Faster multipoint linkage analysis using Fourier transforms', *Journal of Computational Biology*, **1**, 1–7 (1998).'
13. Jennings, H. S. 'The numerical results of diverse systems of breeding with respect to two pairs of characters, linked or independent, with special relation to the effects of linkage', *Genetics*, **12**, 97–154 (1917).
14. Robbins, R. B. 'Some applications of mathematics to breeding problems. III', *Genetics*, **3**, 375–389 (1918).
15. Geiringer, H. 'Further remarks on linkage in Mendelian heredity', *Annals of Mathematical Statistics*, **16**, 390–393 (1945).

16. Chakravarti, A. 'Linkage disequilibrium', *in* Armitage, P. and Colton, T. (eds), *Encyclopedia of Biostatistics*, Wiley, Chichester, 1998.
17. Barnard, G. A. 'Statistical inference', *Journal of the Royal Statistical Society*, **11**, 116–139 (1949).
18. Ott, J. *Analysis of Human Genetic Linkage*, Johns Hopkins University Press, Baltimore, MD, 1991.
19. Morton, N. E. 'Sequential tests for the detection of linkage', *American Journal of Human Genetics*, **7**, 277–318 (1955).
20. Morton, N. E. 'Significance levels in complex inheritance', *American Journal of Human Genetics*, **62**, 690–697 (1998).
21. Lander, E. S. and Kruglyak, L. 'Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results', *Nature Genetics*, **11**, 241–247 (1995).
22. Witte, J. S., Elston, R. C. and Schork, N. S. 'Genetic dissection of complex traits', *Nature Genetics*, **12**, 355–358 (1996).
23. Elston, R. C. 'William Allan Award Address. Algorithms and inferences: the challenge of multifactorial diseases', *American Journal of Human Genetics*, **60**, 255–262 (1997).
24. Williamson, J. A. and Amos, C. I. 'On the asymptotic behavior of the estimate of the recombination fraction under the null hypothesis of no linkage when the model is misspecified', *Genetic Epdemiology*, **7**, 309–318 (1990).
25. Williamson, J. A. and Amos, C. I. 'Guess lod approach: Sufficient conditions for robustness', *Genetic Epidemiology*, **12**, 163–176 (1995).
26. Elston, R. C. 'Man bites dog? The validity of maximizing lod scores to determine mode of inheritance', *American Journal of Medical Genetics*, **34**, 487–488 (1989).
27. Hodge, S. E. and Elston, R. C. 'Lods, wrods and mods: The interpretation of lod scores under different models', *Genetic Epidemiology*, **11**, 329–342 (1994).
28. Morton, N. E. 'The detection and estimation of linkage between the genes for elliptocytosis and the Rh blood types', *American Journal of Human Genetics*, **8**, 80–96 (1956).
29. Smith, C. A. B. 'Testing for heterogeneity of recombination fraction values in human genetics', *Annals of Human Genetics*, **27**, 175–182 (1963).
30. Faraway, J. J. 'Distribution of the admixture test for the detection of linkage under heterogeneity', *Genetic Epidemiology*, **10**, 75–83 (1993).
31. Chernoff, H. and Lander, E. 'Asymptotic distribution of the likelihood ratio test that a mixture of two binomials is a single binomial', *Journal of Statistical Planning and Inference*, **43**, 19–40 (1995).
32. Green, J. R. and Woodrow, J. C. 'Sibling method for detecting HLA-linked genes in a disease', *Tissue Antigens*, **9**, 31–35 (1977).
33. Blackwelder, W. C. and Elston, R.C. 'A comparison of sib-pair linkage tests for disease susceptibility loci', *Genetic Epidemiology*, **2**, 85–97 (1985).
34. Schaid, D. J. and Nick, T. G. 'Sib-pair linkage tests for disease susceptibility loci: Common tests vs. the asymptotically most powerful test', *Genetic Epidemiology*, **7**, 359–370 (1990).
35. Knapp, M., Seuchter, S. A. and Baur, M. P. 'Linkage analysis in nuclear families. 1: Optimality criteria for affected sib-pair tests', *Human Heredity*, **44**, 37–43 (1994).
36. Risch, N. 'Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs', *American Journal of Human Genetics*, **46**, 242–253 (1990).
37. Holmans, P. 'Asymptotic properties of affected-sib-pair linkage analysis', *American Journal of Human Genetics*, **52**, 362–374 (1993).
38. Risch, N. 'Linkage strategies for genetically complex traits. I. Multilocus models', *American Journal of Human Genetics*, **46**, 222–228 (1990).
39. Cordell, H. J., Todd, J. A., Bennett, S. T., Kawaguchi, Y. and Farrall, M. 'Two-locus maximum lod score analysis of a multifactorial trait: joint consideration of IDDM2 and IDDM4 in type 1 diabetes', *American Journal of Human Genetics*, **57**, 920–934 (1995).
40. Olson, J. M. 'Likelihood-based models for linkage analysis using affected sib pairs', *Human Heredity*, **47**, 110–120 (1996).
41. Kruglyak, L., Daly, M. J., Reeve-Daly, M. P. and Lander, E. S. 'Parametric and nonparametric linkage analysis: a unified multipoint approach', *American Journal of Human Genetics*, **58**, 1347–1363 (1996).
42. Whittemore, A. S. and Halpern, J. 'A class of tests for linkage using affected pedigree members', *Biometrics*, **50**, 118–127 (1994).

43. Amos, C. I. and Elston, R. C. 'Robust methods for the detection of genetic linkage for quantitative data from pedigrees', *Genetic Epidemiology*, **6**, 349–360 (1969).
44. Olson, J. M. and Wijsman, E. M. 'Linkage between quantitative trait and marker loci: methods using all relative pairs', *Genetic Epidemiology*, **10**, 87–102 (1993).
45. Amos, C. I. 'Robust variance-components approach for assessing genetic linkage in pedigrees', *American Journal of Human Genetics*, **54**, 535–543 (1994).
46. Almasy, L. and Blangero, J. 'Multipoint quantitative-trait linkage analysis in general pedigrees', *American Journal of Human Genetics*, **62**, 1198–1211 (1998).
47. Cardon, L. R. and Fulker, D. W. 'The power of interval mapping of quantitative trait loci using selected sib pairs', *American Journal of Human Genetics*, **55**, 825–833 (1994).
48. Carey, G. and Wiliamson, J. A. 'Linkage analysis of quantitative traits: increased power by using selected samples', *American Journal of Human Genetics*, **49**, 786–796 (1991).
49. Risch, N. and Zhang, H. 'Extreme discordant sib pairs for mapping quantitative trait loci in humans', *Science*, **268**, 1584–1589 (1995).
50. Risch, N. and Zhang, H. 'Mapping quantitative trait loci with extreme discordant sib pairs: samplings considerations', *American Journal of Human Genetics*, **58**, 836–843 (1996).
51. Jorde, L. B. 'Linkage disequilibrium as a gene mapping tool (editorial)', *American Journal of Human Genetics*, **56**, 11–14 (1995).
52. Hastbacka, J., de la Chapelle, A., Kaitila, I. Sistonen, Pl, Weaver, A. and Lander, E. 'Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland', *Nature Genetics*, **2**, 204–211 (1992).
53. Bodmer, W. F. 'Human Genetics: the molecular challenge', *Cold Spring Harbor Symposium in Quantitative Biology*, **51**, 1–13 (1986).
54. Xiong, M. and Guo, S. W. 'Fine-scale genetic mapping based on linkage disequilibrium: theory and applications', *American Journal of Human Genetics*, **60**, 1513–1531 (1997).
55. Lazzaroni, L. C. 'Linkage disequilbrium and gene mapping: an empirical least-squares approach', *American Journal of Human Genetics*, **62**, 159–170 (1998).
56. Spielman, R. S., McGinnis, R. E. and Ewens, W. J. 'Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM)', *American Journal of Human Genetics*, **52**, 506–516 (1993).
57. Spielman, R. S. and Ewens, W. J. 'The TDT and other family-based tests for linkage disequilibrium and association', *American Journal of Human Genetics*, **59**, 983–989 (1996).
58. Schaid, D. J. 'Relative-risk regression models using cases and their parents', *Genetic Epidemiology*, **12**, 813–818 (1995).
59. Cleves, M. A., Olson, J. M. and Jacobs, K. B. 'Exact transmission-disequilibrium tests with multiallelic markers', *Genetic Epidemiology*, **14**, 337–347 (1997).
60. Hill, W. G. and Weir, B. S. 'Maximum-likelihood estimation of gene location by linkage disequilibrium', *American Journal of Human Genetics*, **54**, 705–714 (1994).
61. Lynch, M. and Walsh, B. *Genetics and Analysis of Quantitative Traits*, Sinauer, Sunderland MA, 1997, pp. 413–418.
62. Hill, W. G. and Robertson, A. 'Linkage disequilbrium in finite populations', *Theoretical and Applied Genetics*, **38**, 226–231 (1968).
63. Jorde, L. B., Watkins, W. S., Carlson, M., Groden, J., Albertsen, H., Thliveris, A. and Leppert, M. 'Linkage disequilibrium predicts physical distance in the adenomatous polyposis coli region', *American Journal of Human Genetics*, **54**, 884–898 (1994).
64. Devlin, B. and Risch, N. 'A comparison of linkage disequilibrium measures for fine-scale mapping', *Genomics*, **29**, 311–322 (1995).
65. Guo, S.-W. 'Linkage disequilibrium measures for fine-scale mapping: a comparison', *Human Heredity*, **47**, 310–314 (1997).
66. Luria, S. E. and Delbrück, M. 'Mutations from bacteria from virus sensitivity to virus resistance', *Genetics*, **28**, 491–511 (1943).
67. Kaplan, N. L. and Weir, B. S. 'Are moment bounds on the recombination fraction between a marker and a disease locus too good to be true? Allelic association mapping revisited for simple genetic diseases in the Finnish population', *American Journal of Human Genetics*, **57**, 1486–1498 (1995).

68. Kaplan, N. L., Hill, W. G. and Weir, B. S. 'Likelihood methods for locating disease genes in nonequilibrium populations', *American Journal of Human Genetics*, **56**, 18–32 (1995).
69. Rannala, B. and Slatkin, M. 'Likelihood analysis of disequilibrium mapping, and related problems', *American Journal of Human Genetics*, **62**, 459–473 (1990).
70. De la Chapelle, A. and Wright, F.A. 'Linkage disequilibrium mapping in isolated populations: the example of Finland revisited', *Proceedings of the National Academy of Sciences*, **95**, 12416–12423 (1998).
71. Hill, W. G. 'Estimation of linkage disequilbrium in randomly mating populations', *Heredity*, **33**, 229–239 (1974).
72. Slatkin, M. and Excoffier, L. 'Testing for linkage disequilibrium in genotypic data using the EM algorithm', *Heredity*, **76**, 377–383 (1996).
73. Brown, A. H. D. 'Sample sizes required to detect linkage disequilibrium between two or three loci', *Theoretical Population Biology*, **8**, 184–201 (1975).
74. Weir, B. S. and Cockerham, C. C. 'Testing hypotheses about linkage disequilibrium with multiple alleles', *Genetics*, **88**, 633–642 (1978).
75. Terwilliger, J. D. 'A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci', *American Journal of Human Genetics*, **56**, 777–787 (1995).
76. Delvin, B., Risch, N. and Roeder, K. 'Disequilibrium mapping: composite likelihood for pairwise disequilibrium', *Genomics*, **36**, 1–16 (1996).
77. Weir, B. S. *Genetic Data Analysis*, Sinauer, Sunderland MA, 1990.
78. Armitage, P. and Colton, T. *Encyclopedia of Biostatistics*, Wiley, Chichester, 1998.

