STATISTICAL GENETICS '98
Methods of Linkage Analysis—and the Assumptions Underlying Them

Robert C. Elston

Department of Epidemiology and Biostatistics, Rammelkamp Center for Education and Research, Case Western Reserve University, Cleveland

In its simplest form, linkage analysis consists of counting recombinants and nonrecombinants, estimating the recombination fraction, and testing whether this fraction is significantly $<\frac{1}{2}$. By means of a general-likelihood formulation, this concept was extended, for the study of Mendelian diseases, to situations in which the number of recombinants cannot be counted directly, with allowance both for incompletely known genotypes—whether due to dominance, incomplete penetrance, or phenocopies—and for heterogeneity in the recombination fraction—whether due to the sex of the transmitting parent or to different loci controlling the appearance of disease. Logarithms to base 10 of the likelihood ratio (LODs), comparing the maximum likelihood to the likelihood when the recombination fraction is $\frac{1}{2}$, became the basis on which the evidence for linkage was judged. If the likelihood was maximized over only a single recombination fraction, a "LOD score" of 3 was taken to be significant. This corresponds to a P value (empirical type I error probability) of $<10^{-3}$—and, in large samples (as the number of *informative meioses* becomes large), to a P value of $\sim 10^{-4}$. Very approximately, using this "LOD score of 3" criterion would ensure, in most cases, that only 5% of declared linkages would be false. To account for maximization of the likelihood over more than one unknown parameter—for example, the two sex-specific recombination fractions or the recombination fraction and additional parameters such as penetrances—the critical value of the LOD was adjusted upward, to ensure that the large-sample P value remained the same, with the result that the latter was the actual currency being used—the exchange rate with LODs fluctuating according to circumstance. These methods of linkage analysis came to be known as "LOD score" methods and, more recently, have been called "parametric." I prefer to call these methods "*model based*," rather than by "LOD score" or "parametric," since the latter are terms that

are equally applicable to many of the other methods of linkage analysis, which are described briefly below. Of course, every statistical test must be based on a probability model, so I am using the term "model based" to indicate that details of the trait's mode of inheritance are being modeled. Typically, particular values of the allele frequencies and penetrance functions are specified, and we can assume, as is true for simple Mendelian traits, that in any one family there is segregation at only a single locus.

In the case of complex traits, it is impossible to model with any certainty all the causes of familial aggregation. Starting with Penrose (1935), many methods of linkage analysis have been developed that do not require the trait to be modeled in such detail, and recently these have been termed "allele sharing" methods (Lander and Schork 1994) and "parametric linkage analysis" (Kruglyak et al. 1996; Morton 1998). Most of these methods involve definite parameters that take on different values according to whether there is linkage, and so I prefer to call them "model free" rather than "nonparametric" (although, as we shall see, they also can involve a certain amount of genetic modeling). These methods are all based on little more than the premise that relatives who are similar with respect to the phenotype of interest will be similar at a marker locus, sharing identical marker alleles, only if a locus underlying the phenotype is linked to the marker. The more powerful of these methods are based on knowing or estimating the sharing of marker alleles that are identical by descent (IBD)—that is, that are direct copies of the same ancestral alleles. If, for some relatives, IBD sharing cannot be determined unequivocally on the basis of the data available, the methods may either ignore the data on those relatives, base tests directly on marker identity in state, or estimate IBD sharing probabilistically, using population marker-allele frequencies to do so. All these methods aim, with varying degrees of success, to provide valid tests of linkage, without the need to specify a detailed mode of inheritance for the phenotype of interest, and so it is not necessary to specify corresponding allele frequencies and penetrances. In large samples (in this case, as the number of *independent sets of relatives* becomes large), the type I error probability should be controlled properly.

The most common of the model-free methods use full sibs, who at any one locus can share 0, 1, or 2 alleles IBD. The three corresponding sharing probabilities, which sum to unity, are the parameters of interest. In a randomly mating population the proportions of sib pairs sharing 0, 1, or 2 marker alleles IBD at any locus are expected to be $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$, respectively, and the mean proportion of alleles that they share IBD (i.e., the proportion of sib pairs sharing two alleles IBD plus half the proportion sharing one allele IBD) is expected to be $\frac{1}{2}$. If the marker is linked to a locus underlying a trait of interest, then sibs similar in phenotype will tend to share $>\frac{1}{2}$ of their marker alleles IBD, whereas sibs who are dissimilar will tend to share $<\frac{1}{2}$ of their marker alleles IBD. If we are investigating a quantitative trait, we can test whether dissimilarity with respect to the quantitative trait—measured, for example, as the squared difference between the sibs' phenotypic values—is negatively correlated with the proportion of marker alleles shared IBD. A binary-disease outcome can be considered as a special case of a quantitative trait, by giving, without loss of generality, the value 1 to affected sibs and 0 to unaffected sibs. Then, testing for such a correlation is identical to testing whether the mean proportion of alleles shared is larger for similar (concordant) sibs than for dissimilar (discordant) sibs. Alternately, if similar and dissimilar sib pairs are not both available in the sample, we can test whether the mean proportion of alleles shared IBD is $>\frac{1}{2}$ for similar pairs, or $<\frac{1}{2}$ for dissimilar pairs. This "mean" test is the most powerful test of linkage if the gene effect on the phenotype is additive—that is, the heterozygote mean phenotype is halfway between the two corresponding homozygous mean phenotypes. In the case of a binary disease, we can consider the probability of being affected as being the phenotype. The gene effect is then additive either if (*a*) the penetrance of the heterozygote is halfway between the penetrances of the two homozygotes or (*b*) the homozygous genotype predisposing to disease is nonexistent. In either case, one parameter is estimated: either the difference between the mean proportions for similar and dissimilar sibs, if both types of sib pairs are being studied, or the mean proportion of marker alleles shared IBD, if only similar or dissimilar sib pairs are being studied. Because only one parameter is being estimated, we have the correspondence noted above—between a LOD score of 3 and a *P* value (in large samples) of $\sim 10^{-4}$. The same is true if the "proportion" test is used—that is, if we base our test on whether the proportion of sib pairs who share 2 alleles IBD is increased in sibs of similar phenotype and/or decreased in sibs of dissimilar phenotype, which can be a more powerful test of linkage when the alleles act nonadditively.

Affected sib pairs are often the only sibs studied, because they usually afford more power; but, if parental marker information is not available, then any test that does not compare the two types of sib pairs—similar and dissimilar—depends strongly, for its validity, on accurate knowledge of the marker-allele frequencies. A common test that is used to analyze affected-sibling data simultaneously tests whether all three IBD proportions deviate from the expected values—$\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$—in the expected direction if there is linkage and leads to a "maximum LOD score." In this case, because there is a (constrained) maximization over more than one parameter, it is no longer true that a LOD score of 3, even in very large samples, corresponds to a *P* value of 10^{-4}; and the threshold for a correspondingly "significant" LOD score is necessarily larger.

Another type of linkage analysis, which recently has been developed (Amos 1994) for quantitative phenotypes, is referred to as the "variance component" method. In this method we model the variance of the phenotype by decomposing it into (*a*) components due to linkage to individual marker locations and (*b*) residual polygenic and environmental components. Thus, instead of specifying the allele frequencies and penetrances for a trait locus, we model the familial covariances that it causes (regardless of how many alleles there are), in terms of a maximum of two parameters: an additive genetic-variance component and a dominant genetic-variance component. The latter is a zero component for all pairs of unilineal relatives—and, as a first approximation, may be assumed to be zero for other pairs of relatives as well. The variance component(s) for each trait locus, as well as the residual variance components, are estimated from the data at each chromosomal location, rather than being prespecified. The variance-component approach usually assumes a multivariate normal distribution for the data, which is a strong assumption: in addition to the usual assumption of a symmetric bell-shaped curve for the marginal distribution, it assumes that the joint distribution of the data for a family depends only on means, variances, and covariances. The variance-component approach can also be based on a "quasi-likelihood" method that does not make such a strong assumption, requiring only large samples for *P* values to be valid. The term "semiparametric" has been used in connection with this quasi-likelihood method, which allows for the presence of nonnormality—such as "skewness"—without the necessity to include further parameters in the likelihood in order to describe it.

Whatever particular method of linkage analysis is used—and regardless of the name given to it—the most important aspects to consider are validity, power, and robustness. A test is *valid* if the reported type I error rate, or *P* value, is correct. *Power* refers to the ability of the test to detect what is being sought—in our case, linkage. A test that correctly uses more of the infor-

mation that is in the data will be more powerful than one that ignores part of the data. On the other hand, a test that misuses part of the data will be less powerful—and, possibly, invalid. *Robustness* refers to a test having good properties, in terms of being valid and powerful, even though the assumptions underlying the test are not met. In large samples, simple model-based methods are extremely robust with respect to validity; under the null hypothesis of no linkage, the analysis makes no assumptions beyond those required for all linkage analyses—namely, (i) that all relationships among pedigree members are known without error, (ii) that marker-allele frequencies, which may or may not be required for analysis, are known without error, and, usually, (iii) that all marker-typing data are correct. A model-based analysis that uses a correct model always will be at least as powerful as any model-free analysis. However, the power of model-based analyses to detect linkage in samples of reasonable size can depend critically on how well the mode of inheritance specified for the trait approximates its true mode of inheritance (for discussion, see Jarvik 1998 [in this issue]). The fact that model-based analyses do not provide consistent estimates of the recombination fraction is of little importance when a global search of the whole genome is performed by efficient multipoint methods. The main disadvantages of model-based methods are as follows: (1) if multiple mode-of-inheritance models are investigated in the linkage analysis, in order not to miss a linkage because of model misspecification, then a more stringent criterion is necessary to detect linkage at a given significance level; and (2) it is computationally difficult to extend model-based methods to multilocus inheritance for the trait of interest. Model-free methods also tend to be robust with respect to validity, but these methods have the following advantages: (1) they require fewer tests, and (2) they can be extended more simply to allow for the simultaneous analysis of multiple trait loci. Each particular model-free method will have most power in a particular situation, but this does not imply that a set of underlying assumptions is necessary if the test is to be valid. The variance-component methods offer a parsimonious parametrization for multilocus models, and this tends to make them generally powerful. However, variance-component models that assume multivariate normality may not be validity robust, even for large samples.

For all methods of linkage analysis, accuracy of P values is of special concern. We can argue that P values are inappropriate as measures of evidence (Goodman 1993; Vieland and Hodge 1998). A common argument notes that, regardless of whether we find r recombinants when we look at n informative meioses or have looked at n informative meioses in order to find r recombinants, the evidence is the same—but that, if we use the strict definition of P values, then the two P values are different

(Berger and Berry 1988). However, the latter case is one of sequential sampling (i.e., we sample until we have found r recombinants), so that the sample size n is a random variable—a situation for which P values were never intended. For all their shortcomings, P values are probably the best measures that we have of the evidence for linkage in the analysis of complex diseases (Witte et al. 1996), especially when more than one parameter is estimated. Likelihood ratios—or, equivalently, LODs—are not comparable when different numbers of parameters are estimated. There is, of course, no magic P value below which linkage has been proved and above which it has yet to be proved. P values are, at best, only guides, because scientific inference is necessarily subjective (Malécot 1947). Nevertheless, they should be determined as accurately as possible, since the estimated power of a study is meaningless unless the type I error rate is correctly controlled. For many methods of linkage analysis, P values obtained on the basis of theoretical large-sample considerations are reasonably accurate if they are ∼.05 but are quite unreliable, for typical sample sizes, if they are much smaller. The only sure way to determine an accurate P value is by considering the sampling distribution of the statistic being used, for the sample size studied, when there is no linkage, either theoretically or by a Monte Carlo simulation procedure. In a study in which only affected persons have been typed for markers, this poses serious statistical difficulties. In any case, the smaller the P value, the larger the number of Monte Carlo replicates required for accurate P-value estimation.

Finally, there are differences of opinion about just which criterion should be used as a cutoff when we intend to report that "significant" linkage has been found (Lander and Kruglyak 1995; Witte et al. 1996; Morton 1998). If a whole-genome scan has been performed, then it makes sense to try to control the overall genomewide type I error rate, with allowance for the fact that multipoint linkage analysis can be used to test for linkage at all points between markers, as well as at the marker locations themselves. If we perform a test for linkage at each of k points along the genome, and if these k tests are independent, then statistical theory indicates that it is appropriate, to a close approximation, to allow for this "multiple testing" by use of a significance level k times as small. This would suggest that, to allow for an infinite number of points, an infinitesimally small significance level would be required. However, the tests along the length of the chromosome are not independent, and the theoretical arguments that have been made to find which single-location P value corresponds to a genomewide P value of .05 (Lander and Kruglyak 1995) are based on the assumption of a dependency structure that arises when there is no linkage interference; that is, they ignore the fact that crossing-over is

inhibited near points where crossovers have already occurred. This probably explains why the single-location *P* value that has been recommended on the basis of those considerations is more stringent than appears to be required in practice (Sawcer et al. 1997). It also should be noted that multipoint linkage analysis is typically performed on the assumption that there is no linkage interference—an assumption that may be problematic when markers are spaced far apart—and on the assumption that the intermarker distances are known without error. Despite these difficulties, it is probably best to report the most accurate single-location *P* values that we can, which would then be guides for future research.

Acknowledgments

References

Amos CI (1994) Robust variance-component approach for assessing genetic linkage in pedigrees. Am J Hum Genet 54: 535–543

Berger JO, Berry DA (1988) Statistical analysis and the illusion of objectivity. Am Sci 76:159–165

Goodman SN (1993) P values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. Am J Epidemiol 137:485–496

Jarvik GP (1998) Complex segregation analyses: uses and limitations. Am J Hum Genet 63:942–946 (in this issue)

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: a unified approach. Am J Hum Genet 58:1347–1363

Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11:241–247

Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265: 2037–2048

Malécot G (1947) Les critères statistiques et la subjectivité de la connaissance scientifique. Ann Univ Lyon 10:43–74

Morton NE (1998) Significance levels in complex inheritance. Am J Hum Genet 62:690–697

Penrose LS (1935) The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage. Ann Eugenics 6:133–138

Sawcer S, Jones HB, Judge D, Visser F, Compston A, Goodfellow PN, Clayton D (1997) Empirical genomewide significance levels established by whole genome simulations. Genet Epidemiol 14:223–229

Vieland VJ, Hodge SE (1998) *Statistical Evidence: a Likelihood Paradigm,* by Richard Royall. Am J Hum Genet 63:283–289

Witte JS, Elston RC, Schork NJ (1996) Genetic dissection of complex traits. Nat Genet 12:355–356