Genetic Linkage: Interpreting Lod Scores

NEIL RISCH

THE ADVENT OF SOPHISTICATED MOLECULAR GENETIC tools and discoveries of the genes identified with some inherited disorders such as retinoblastoma, Duchenne muscular dystrophy, cystic fibrosis, and neurofibromatosis have resulted in increased interest in human genetics by both experimental and social scientists. When the biochemical or physiologic basis for a genetic disease is unknown, a sequential search, at random, of all human chromosomes is necessary if the implicated mutant gene is to be located.

This search is conducted through linkage analysis. Genetic linkage reflects the fact that two genes near one another on the same chromosome are not inherited independently in families. If two loci are close to each other on the same chromosome, alleles at these loci will tend to be inherited together. However, if the loci are some distance apart, crossing over or recombination between homologous chromosomes in meiosis creates new combinations of alleles. The frequency with which recombination occurs (called the recombination fraction, denoted by θ) increases with the distance between loci. If the loci are far apart, the probability of recombinant and parental (nonrecombinant) chromosomes is equal, just as when loci lie on different chromosomes. Recombination events along a chromosome that has multiple, closely spaced, genetic markers can be used to localize a disease gene to within a fairly narrow length of DNA, usually a few megabases at most.

The assertion of linkage is necessarily derived from statistical analysis in that it is based on an observed association between two traits in families. In testing for linkage, we must distinguish between two hypotheses: no linkage (or $\theta = 1/2$; the null hypothesis) and linkage at a recombination fraction $\theta < 1/2$ (the alternative hypothesis). The statistical criterion for concluding linkage between two traits is based on an odds ratio L provided by the data, or the ratio of the probability of observing the distributional pattern of the two traits in a given set of families under the hypothesis of linkage (at θ), to the same probability under the hypothesis of no linkage ($\theta = 1/2$). The decimal logarithm of L (called the lod score) is usually reported, at several values of θ, for convenience, and the maximum value of $\log L$ is obtained. Traditionally (*1*), a maximum lod score of 3, or an odds ratio of 1000 to 1, is required to assert linkage. The strictness of this criterion is due to the low prior probability that two traits are linked; that is, that the alternative hypothesis is true. For example, the prior probability that two randomly selected loci lie less than $\theta = 0.3$ apart from each other is only about 2 percent (*2*). Hence, the posterior odds for linkage, given a lod score of 3, is simply the prior odds of linkage (0.02) times the odds provided by the data (1000), or 20:1, giving a posterior probability of linkage of 20 out of 21, or 95 percent, and a posterior probability of no linkage, or a false positive probability, of 5 percent. Testing of multiple markers across the genome before a lod score of 3 is achieved actually increases the posterior probability of linkage, although the increase is very slight until many markers have been tested (*3*).

The primary assumption underlying lod score analysis is that the traits being analyzed are single-locus, Mendelian traits, with known mode of inheritance, which is correctly specified in the analysis. While this assumption is generally valid for standard genetic markers—such as blood groups, serum proteins, and restriction fragment length polymorphisms—and for some diseases, it may not be valid for the analysis of other diseases.

Mendelian (monogenic) inheritance of human disease is characterized by certain hallmarks. First, and perhaps most important, the recurrence risk in families is very high, of the order of 100 to 1000 times the frequency in the general population. Mutant alleles that cause disease tend to be rare because they are eliminated from the population through selection. This is particularly true for dominant diseases, where expression of disease stems from a single mutant disease allele, as in heterozygotes. By contrast, recessive alleles can persist in the population at higher frequency because most of the alleles occur in heterozygotes that are uniformly unaffected and in some special cases may actually have an advantage. Dominant diseases tend to show vertical transmission through a family, where multiple generations are affected. Offspring of affected individuals are each at 50 percent risk of developing the disease, regardless of gender. Recessive diseases tend to occur only in sibships with one or several sibs affected, but not in other family members, including parents. Another telltale sign of a recessive disease is an increased incidence of consanguinity among the parents of affected individuals, especially when the disease is rare. Expression of simple Mendelian diseases can vary, even within families. Not all individuals with the predisposing genotype may be affected (reduced penetrance), and the probability of becoming affected (penetrance) may increase with age, as in Huntington's disease. Another essential characteristic of Mendelian diseases is that the risk to relatives decreases by a factor of 1/2 with each degree of relationship, that is, from first- to second- to third-degree relatives. If the risk decreases more rapidly than that, a more complex genetic model is implicated (*4*).

Many common familial diseases do not conform to simple Mendelian expectations. Evidence for a genetic basis for such disorders can come, for example, from family, twin, and adoption studies. When the results of these different studies converge and lead to the same conclusion, a genetic role in disease etiology is strongly supported. However, such studies alone cannot specify the number of genes that may be involved in susceptibility, or the magnitude of their effects.

It is often difficult to determine the number of genes that have a role in disease susceptibility, even between the extreme cases of one and some large number of genes, each of small effect. A disease that is quite common, and potentially complex, can contain a subgroup defined on a clinical or biochemical basis that has the hallmarks of Mendelian inheritance. For example, the appearance of Alzheimer's disease is quite common in older individuals but is rare among 30- or 40-year-olds. Nevertheless, there have been pedigrees described with many very early onset cases (as many as 50) occurring in multiple generations in a pattern consistent with autosomal dominant inheritance (*5, 6*). The early onset disease in these pedigrees has been attributed to the effect of a single mutant allele segregating through the family. However, if there were individuals in the pedigree with late onset, it would be unclear whether they were carrying the same allele. A similar pattern appears in breast cancer, which is quite common, especially in older individuals. Several family studies suggest that a small subset of cases may be due to a dominant allele with high penetrance; these cases are distinguished by especially early onset (*7*).

Diabetes mellitus is another example. Early onset, prior to age 25,

The author is with the Departments of Epidemiology and Public Health and of Genetics, Yale University School of Medicine, New Haven, CT 06510.

diabetic disease is usually insulin-dependent (IDDM), is autoimmune in etiology, is HLA-associated, and is genetically complex. By contrast, noninsulin-dependent diabetes (NIDDM) usually has maturity onset, after age 50, is not HLA-associated, and is quite frequent. NIDDM or "maturity onset type diabetes" in youth (called MODY) is extremely uncommon and accounts for fewer than 5 percent of all early onset cases (before age 25). This early onset form of diabetes appears to have a Mendelian, autosomal dominant pattern of inheritance (*8*). However, because IDDM and maturity onset NIDDM are not infrequent, such cases may also appear in pedigrees with MODY, but are genetically unrelated (*8*).

When there is a well-defined Mendelian subgroup of a disease, analysis is like that for conventional Mendelian disease. This approach is effective if autosomal dominance is present because informative, extended pedigrees can be identified. A Mendelian recessive subgroup is more difficult to identify because recessive pedigrees generally consist of a small number of affected siblings only.

Even if no Mendelian subgroup can be defined a priori on a clinical (or age of onset) basis, it may still be possible to ascertain pedigrees with affected individuals in several generations. Such pedigrees, however, can be quite sparse, and the clinical definition of affected may be broadened to include less severe diagnoses to fill out pedigrees. If these diagnoses are common in the general population, the assumption of simple Mendelian (dominant) inheritance in these pedigrees is less secure.

Violation of the Mendelian assumption in lod score analysis may have serious consequences. If the genetic mechanism underlying a disease is complex, possibly involving several loci, successful detection of linkage may be more difficult than in the simple Mendelian case. The reason is that, unlike the Mendelian one-to-one correspondence between genotype and phenotype, the correspondence between phenotype (affected, unaffected) and genotype may be weak. Disease status alone does not allow clear discrimination among genotypes at a disease susceptibility locus. Only heterozygous parents are informative for linkage, and it may be difficult, or impossible, to determine who is heterozygous. For example, an affected individual with many affected children may have an increased probability of being homozygous (for the high risk allele), decreasing the usefulness of such a family for linkage analysis. Furthermore, for the non-Mendelian case, accurate specification of the mode of inheritance in lod score analysis is generally not possible. In the absence of linkage, the disease and marker are inherited independently, whatever genetic model is specified. Hence, the distribution of the lod score statistic under the null hypothesis (no linkage) remains the same. However, in the presence of linkage, the probability of rejecting the null hypothesis of no linkage (that is, the power) may be reduced by model misspecification, especially when dominance is misrepresented (*9*). Hence, the probability that a significant linkage result in the non-Mendelian case is actually false may be inflated. The false-positive rate can also be increased by testing multiple genetic models or disease classifications (or both), but not by testing multiple genetic markers (*3*).

Linkage evidence must be interpreted in the context of prior information regarding the genetics of the disorder. Prior evidence for a genetic role (from family, twin, and adoption studies) is important. Conventional lod score statistics are based on the Mendelian case and are well supported for that case. Such criteria should also serve well when applied to a Mendelian subgroup of a complex disease, where the subgroup has been defined clinically or when segregation analysis of systematically ascertained families has been made. By contrast, lod score statistics for non-Mendelian diseases in general require meticulous examination. A high lod score (>3) is unlikely in the absence of linkage, and therefore it may be tempting to conclude that linkage exists when a high lod score is obtained for a non-Mendelian trait. However, a high lod score is also unlikely in the presence of linkage when a locus has only a minor effect, and it is the ratio of these two likelihoods that determines the probability that the result may be a false positive.

In evaluating a linkage result, it is important to consider the plausibility of the genetic model used in the analysis. When a significant linkage finding depends on the use of a model known to be unlikely (for example, a rare autosomal dominant gene with high penetrance), the results should be viewed with caution. In the absence of any prior evidence for the major effect of a single locus (and particularly when there is evidence against such a major effect), a positive linkage should be viewed as a hypothesis-generating result (that is, the existence of a major locus) rather than a hypothesis-testing result. It is a hypothesis that can be confirmed or refuted by subsequent studies. Hence, for such cases, replication is indispensable. The inability to replicate a linkage result should not automatically be attributed to genetic heterogeneity without other evidence for such heterogeneity (*10*).

With these guidelines, how do we interpret recent linkage findings for various complex diseases? For early onset Alzheimer's, breast cancer, and NIDDM (MODY), prior evidence suggested that these are Mendelian subforms, and the reported findings are consistent with prior knowledge (*5, 11*). Replication of the Alzheimer's linkage on chromosome 21 now appears sound, although clearly in many families linkage with chromosome 21 has been excluded (*12*). Replication studies for breast cancer and MODY should be forthcoming soon. As yet, for many disorders that are not Mendelian and for which a clear Mendelian subgroup has not been characterized, the existence of major (mappable) loci remains unknown. Hence, initial linkage results for such disorders are more tenuous, and convincing replication is essential.

REFERENCES AND NOTES

1. N. E. Morton, *Am. J. Hum. Genet.* **7**, 277 (1955).
2. R. C. Elston and K. L. Lange, *Ann. Hum. Genet.* **38**, 341 (1975).
3. J. Ott, *Analysis of Human Genetic Linkage* (Johns Hopkins Univ. Press, Baltimore, 1985); N. Risch, *Am. J. Hum. Genet.* **48**, 1058 (1991).
4. ——— *Am. J. Hum. Genet.* **46**, 222 (1990).
5. P. H. St George-Hyslop *et al.*, *Science* **235**, 885 (1987).
6. G. D. Schellenberg *et al.*, *ibid.* **241**, 1507 (1988).
7. W. R. Williams and D. E. Anderson, *Genet. Epidemiol.* **1**, 7 (1984); D. T. Bishop, L. Cannon-Albright, T. McLellan, E. J. Gardner, M. H. Skolnick, *ibid.* **5**, 151 (1988); B. Newman, M. A. Austin, M. Lee, M-C. King, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 1 (1988); J. M. Hall *et al.*, *Am. J. Hum. Genet.* **44**, 577 (1989); E. B. Claus, N. Risch, W. D. Thompson, *ibid.* **48**, 232 (1991).
8. S. S. Fajans, *Diabetes Care* **13**, 49 (1990); R. B. Tattersall and S. S. Fajans, *Diabetes* **24**, 44 (1975).
9. F. Clerget-Darpoux, C. Bonaiti-Pellie, J. Hochez, *Biometrics* **42**, 393 (1986); D. A. Greenberg and S. E. Hodge, *Genet. Epidemiol.* **6**, 259 (1989); N. Risch and L. Giuffra, *Hum. Hered.*, in press.
10. N. Risch, *Genet. Epidemiol.* **7**, 3 (1990).
11. J. M. Hall *et al.*, *Science* **250**, 1684 (1990); G. I. Bell *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 1484 (1991).
12. A. M. Goate *et al.*, *Lancet* **335**, 352 (1989); P. H. St George-Hyslop *et al.*, *Nature* **347**, 194 (1990).

13. Supported by NIH grant HG00348.