Genetic Analysis of Complex Diseases

Neil Risch, in a series of seminal papers in 1990 (*1*, *2*), demonstrated the utility of sib-pair linkage analysis in identifying genes for complex genetic traits. In doing so, he defined what is the current paradigm for the genetic dissection of common complex genetic diseases. The recent Perspective by Risch and Kathleen Merikangas (*3*) again shapes the future of human disease gene mapping by defining what will undoubtedly become the statistical "state of the art." Risch and Merikangas advocate conducting genomic screens based on association studies of candidate genes using the transmission disequilibrium test (TDT). It is important, however, not to infer from their arguments that current linkage analysis methods cannot detect most genes underlying complex disease.

Risch and Merikangas extend the current paradigm to include anticipated technological advances. However, the magnitude of γ (defined as the relative risk in the heterozygote) in complex traits and the application of this approach using current molecular technology must be considered.

In their formulation, Risch and Merikangas show that a TDT approach is more powerful than a sib pair approach, particularly for disease alleles with small genetic effects. This conclusion is based on the sample size required to detect a gene with a $\gamma \leq 4$. They show that, while sib pair analysis requires a practical (for example, 100 to 400) number of sib pairs to detect a gene with $\gamma = 4$ and a disease allele frequency p, between 0.1 and 0.5, the number of sib pairs required becomes impractical (for example, more than 1000) when $\gamma \leq 2$.

Previously, Risch (*1*, *2*) established the use of a sibling recurrence risk ratio (λ_s) to estimate the power of a sib pair design to detect linkage. Estimable from epidemiologic data, λ_s is calculated as the ratio of the recurrence risk in siblings of an affected individual and the population prevalence of the disorder. This λ_s represents the overall recurrence risk ratio, which may result from the actions of a single gene or multiple genes acting additively or epistatically. If a number of genes are hypothesized, the gene-specific λ_s (referred to here as λ_{gs}) may be estimated by assuming a model (additive or epistatic), the number of genes, and partitioning λ_s accordingly.

Many researchers are accustomed to evaluating the magnitude of genetic effects by using λ_s rather than γ. We calculated λ_{gs} corresponding to $\gamma \leq 2$ and p = 0.01, 0.10, 0.5, and 0.8, respectively. The results indicated that for $\gamma \leq 2$, $\lambda_{gs} < 1.3$. It has previously been shown that genes with $\lambda_{gs} < 1.3$ would be difficult to detect using sib pair methods (*2*, *4*). The curve comparing λ_{gs} with γ shows that even genes with moderate effect (for example, $\lambda_{gs} < 2$) may produce γ that can be detected by linkage analysis in reasonably sized samples of sib pairs.

These results indicate that linkage analysis of complex disease based on genomic screens using current microsatellite markers can be a fruitful enterprise in complex genetic diseases. An excellent example is the discovery of the late onset Alzheimer's disease susceptibility gene APOE. Using Risch and Merikangas's formulas, we calculated the number of sib pairs that would have been necessary to detect the effect of APOE on the risk of AD. The γ in individuals heterozygous for APOE-4 is 4.5 (*5*) and the frequency of APOE-4 in the general population is about 15% (*6*); the resulting probability of allele sharing is $Y = 0.625$, and the minimum number of affected sib pairs required to detect linkage is 164. Alzheimer's disease has an overall λ_s of 5, with a λ_{gs} of only 2 for APOE (*7*). Other complex diseases, such as multiple sclerosis ($\lambda_s = 30$) (*8*) and autism ($\lambda_s = 150$) (*9*) have substantial genetic components. Even if there are 10 epistatic genes of equal multiplicative effect underlying multiple sclerosis ($\lambda_s = 30$; $\lambda_{gs} = 1.4$), linkage analysis should be able to detect them. Because it is difficult to determine a priori which disease alleles have minor or moderate genetic effects, linkage analysis should not be arbitrarily abandoned.

Risch and Merikangas point out that genomic screening of candidate genes is several years from becoming reality. When the molecular resources become available, the advantages of genomic screening using TDT, such as increasing power to detect minor genetic effects, allowing the use of singleton cases, and testing effects of functional polymorphisms in genes, will make this the method of choice. Until then, well-designed linkage studies of complex traits will still be able to detect genes of major or moderate effect.

William K. Scott
Margaret A. Pericak-Vance
Section of Medical Genetics, Department of Medicine, Duke University Medical Center, Durham, NC 27710, USA
Jonathan L. Haines
Molecular Neurogenetics Unit, Massachusetts General Hospital, Charlestown, MA 02129, USA

REFERENCES AND NOTES

1. N. Risch, *Am. J. Hum. Genet.* **46**, 222 (1990); *ibid.*, p. 229.
2. ———, *ibid.*, p. 242.
3. ——— and K. Merikangas, *Science* **273**, 1516 (1996).
4. E. R. Hauser *et al.*, *Genet. Epidemiol.* **13**, 117 (1996).
5. J. L. Haines *et al.*, *Genomics* **33**, 53 (1996).
6. E. H. Corder *et al.*, *Neurology* **45**, 1323 (1995).
7. A. D. Roses *et al.*, *Am. J. Hum. Genet.* **57**, A202 (1995).
8. Multiple Sclerosis Genetics Group, *Nature Genet.* **13**, 469 (1996).
9. A. Bailey *et al.*, *Psychol. Med.* **25**, 63 (1995).
10. Supported by grants NS31153 from the National Institutes of Health and a grant from the Muscular Dystrophy Association. We thank M. Jordan for helpful discussion.

22 October 1996; accepted 8 January 1997

Risch and Merikangas (*1*) point out the efficiency of association studies for statistical power in identifying genetic markers of disease. But they limit themselves to studies of family-based association, affected sib-pairs, and parental transmission of alleles and do not mention population-based association studies (either cohort or case-control). While family-based association studies do have certain strengths, population-based studies can be far more efficient in terms of time, money, and logistics. It can take much longer to identify and collect samples from a single affected family than to collect samples from 10 or 100 patients with disease. Some studies, such as those of parental transmission, may not be practical in adult onset diseases where parents are deceased. Such practical issues, as well as our ability to generalize the results to the larger population, favor the use of population-based studies.

Perhaps as important, population-based studies commonly measure environmental exposures and can assess gene-environment interaction, data for which are nearly always lacking in family-based studies. An association between a susceptibility gene and a disease may not be apparent if there is a second factor required to initiate the disease process, such as an environmental exposure. Similarly, one may detect the effect of exposure only among genetically susceptible subpopulations. There are a number of neurologic and other diseases in which this model functions, but the cases of genes that modulate carcinogen-induced cancers (such as the polymorphic glutathione S-transferases and N-acetyltransferases) are perhaps the best examples (*2*). Simple tests of gene-disease association are likely to be misleading without due attention to environmental factors.

Douglas A. Bell
Laboratory of Computational Biology and Risk Analysis, National Institute of

Environmental Health Sciences, Research Triangle Park, NC 27709, USA
Jack A. Taylor
Epidemiology Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC

REFERENCES

1. N. Risch and K. Merikangas, *Science* **273**, 1516 (1996).
2. H. Chen *et al.*, *Lancet* **347**, 295 (1996); D. Vineis *et al.*, *Nature* **369**, 154 (1994); D. A. Bell *et al.*, *J. Natl. Cancer Inst.* **85**, 1159 (1993).

27 September 1996; accepted 8 January 1997

Risch and Merikangas make the intriguing point that, given a relatively small number of families, the transmission-disequilibrium test (*1*) has enough statistical power to determine if any of a large (in some sense complete, genome-wide) set of diallelic markers is associated with a disease. Another approach, based on disequilibrium between marker alleles and disease in a randomly ascertained population sample, can be considered. Like Risch and Merikangas, we can show that, when the disease is relatively common, the disease-allele frequency is intermediate and its effect small, statistical power comparable to that of standard family-based linkage studies is achieved with a smaller number of randomly sampled individuals. The sample sizes required for the disequilibrium method are generally larger than those for transmission-disequilibrium, but the random-ascertainment scheme has practical advantages.

If one assumes that π is the probability that an individual with genotype *aa* has the disease, with *Aa* and *AA* individuals being, respectively, γ and γ^2 more likely to develop the disease than *aa* individuals, one can show the expected frequencies of four categories defined by allelic state and disease status in a random population sample (*2*). The association between marker allele *A* and disease can be tested with the "chi-square" statistic. The same statistic can be used to calculate sample sizes needed to detect such an association, if indeed it exists, with a given significance level and power, for fixed values of π, γ, and p, respectively (*3*). When the prevalence of the disease is greater than about 5% and the disease allele is not rare, the random sample approach requires no more than 10 times the number of genotyped individuals in an affected offspring study (*4*). Once genotyped, the same sample can be used to study a number of diseases for the additional, small cost of ascertaining the presence or absence of a disease in each individual. The random sample approach could be especially useful for efficiently diagnosed late onset diseases, where it may not be possible to type parents for affected offspring studies. Non-insulin-dependent diabetes and hypertension, with prevalences of 6% (*5*) and 23% (*6*), respectively, could be effectively studied using this approach.

A realistic program for mapping disease markers using the random-ascertainment scheme may require a prospective design in which a cohort is fully genotyped and monitored for disease. The successes of the Framingham Study (*7*) and others like it show that large-scale prospective studies are not beyond reach. The effort required to genotype a large sample at many marker loci seems formidable, but the automated methods envisioned by Risch and Merikangas greatly reduce the labor for the random-ascertainment scheme as well. The utility of both approaches depends on the existence of marker alleles strongly associated with disease-causing polymorphisms, but as yet the nature and extent of such associations in the human genome are not well understood (*8*). Population structure (the result of admixture or other factors) introduces complications for simple disequilibrium methods that are minimized in family-based transmission-disequilibrium studies (*9*). However, if study populations are defined carefully and data are examined for the effects of population structure, these difficulties may be balanced by gains in efficiency that accrue when a single large sample is used to study several diseases.

Anthony D. Long
Mark N. Grote
Charles H. Langley
Center for Population Biology, University of California, Davis, CA 95616, USA
E-mail: tdlong@ucdavis.edu

REFERENCES AND NOTES

1. R. Spielman, R. E. McGinnis, W. J. Ewens, *Am. J. Hum. Genet.* **52**, 506 (1993).
2. The disease marker *A* has a population frequency of p, and *a* has a frequency of $q = 1\text{-}p$. Using the Hardy-Weinberg (H-W) law, we calculate the frequencies of the three affected genotypes (*AA, Aa,* and *aa*) as $\pi\gamma^2p^2$, $\pi\gamma 2pq$, and πq^2, respectively. Similarly, the frequencies of the three nonaffected genotypes are $(1 - \pi\gamma^2)p^2$, $(1 - \pi\gamma)2pq$, and $(1 - \pi)q^2$. Allele frequencies are obtained by adding, within each disease category, the frequencies of homozygotes for the allele plus half the frequency of heterozygotes. The H-W law may not accurately give genotype frequencies in some cases, such as severe early-onset diseases, where a different parametrization would be required. We assume the H-W law mainly to facilitate sample size calculations, as Risch and Merikangas did in calculating parental heterozygosities.
3. If one assumes the null hypothesis that allele and disease status are independent (equivalent to the hypothesis $\gamma = 1$), expected counts are formed by multiplying marginal frequencies together with the sample size n, and the statistic

$$X^2 = \sum_{i,j} \frac{(\text{observed}_{ij} - \text{expected}_{ij})^2}{\text{expected}_{ij}}$$

can be compared to quantiles of the χ^2_1 distribution (the χ^2 distribution on one degree of freedom). When allele *A* is associated with disease, $\gamma > 1$, and the statistic X^2 follows the noncentral $\chi^2_{1,\lambda}$ distribution, with noncentrality parameter

$$\lambda = n\,\pi\,[\gamma^2p + q - (\gamma p + q)^2]/[1 - \pi(\gamma p + q)^2]$$

{See A. Agresti's book [*Categorical Data Analysis* (Wiley, New York, 1990)] for a description of the noncentrality parameter}. For sample size calculations, γ, π, and p are taken as fixed values, and n is a variable in λ. As in Risch and Merikangas, the significance level for a given marker locus is set at $\alpha = 5 \times 10^{-8}$, to give a genome-wide significance level of 5%. β, the power of the test for a single marker, is then the probability that a $\chi^2_{1,\lambda}$ variable exceeds $Q = 29.72$, the quantile of the χ^2_1 distribution corresponding to the per locus significance level. As the $\chi^2_{1,\lambda}$ variable is equal in distribution to Y^2, where Y is a normal $(\sqrt{\lambda}, 1)$ variable, it follows that $\beta = 1 - pr\,(-\sqrt{Q} < Y < \sqrt{Q})$. Converting to standard deviates and taking the area added in the lower tail as negligible, β is approximately $1 - \Phi(\sqrt{Q} - \sqrt{\lambda})$, where Φ is the cumulative distribution function for a standard normal variable. Setting β at 0.8 determines the value that $\sqrt{Q} - \sqrt{\lambda}$ must equal, and then one can solve for n. The χ^2 distributions have been used mainly to facilitate sample size calculations; given marker-disease data, one would probably use Fisher's exact test to detect associations.
4. For a table comparing sample sizes, see http://dmitri.ucdavis.edu/association_study
5. *Online Mendelian Inheritance in Man* (Johns Hopkins University, Baltimore, MD, 1995; MIM No. 125853, http://www3.ncbi.nlm.nih.gov/omim/).
6. *Health, United States, 1995* (National Center for Health Statistics, Hyattsville, MD, 1995).
7. *The Framingham Study; An Epidemiological Investigation of Cardiovascular Disease* (Government Printing Office, Washington, DC, 1995).
8. L. B. Jorde, *Am. J. Hum. Genet.* **56**, 11 (1995).
9. Reviewed in E. S. Lander and N. J. Schork, *Science* **265**, 2037 (1994); D. E. Weeks and M. Lathrop, *Trends Genet.* **11**, 513 (1995).

9 October 1996; accepted 8 January 1997

Risch and Merikangas (*1*) show the great power of genetic association studies such as the TDT in the detection of genes with modest effects. As they mention, all TDT computations were based on the optimal assumption that the analyzed allele was the disease allele itself. A more common situation is, and could well remain, the analysis of polymorphisms which have a low prior probability to be the disease allele even if they are within the actual disease gene. The power of the TDT is highly dependent not only on the linkage disequilibrium between the disease allele and the analyzed allele but also on the relative frequencies of both these alleles.

With the same genetic model as that used by Risch and Merikangas—a disease locus with two alleles, *A* and *a*, with population frequencies of p and $1\text{-}p$, respectively, and a multiplicative model with genotypic relative risks of γ and γ^2 for *Aa* and *AA* subjects, respectively—one can assume a closely linked diallelic marker (recombination fraction = 0) with alleles *B* and *b* of

respective frequencies m and 1-m. The coefficient of linkage disequilibrium, δ, is defined as freq(AB)-pm, and the maximum value of δ, δ_{max}, is reached with freq(AB) is the lowest of the two frequencies m and p. The probability that a heterozygous Bb subject carries A in coupling when B is $\alpha_1 = p + \delta/m$, and the probability that the same subject carries A in coupling with b is $\alpha_2 = p - \delta/(1 - m)$ (*2*). In a sample of single affected individuals with their parents, the probability for a Bb parent to transmit B to his affected child is $P(\text{tr} - B) = [1 + (\gamma - 1)\alpha_1]/[2 + (\gamma - 1)(\alpha_1 + \alpha_2)]$ (*3*). The situation described by Risch and Merikangas corresponds to complete linkage disequilibrium, that is, $\delta = \delta_{max}$ with $m = p$, with $P(\text{tr} - B)$ reducing to $\gamma/(1 + \gamma)$. In other cases, the number of necessary families increases dramatically as p differs from m even when $\delta = \delta_{max}$, and also as δ decreases. Thus, the power of association studies such as the TDT can be quite strong when there is a high probability that the allele studied is the causal allele as shown by Risch and Merikangas. In other cases, researchers should be aware that the power of such association studies can be greatly diminished as soon as the ratio m/p departs from unity and the linkage disequilibrium becomes weaker.

Bertram Müller-Myhsok
Department of Molecular Genetics, Bernhard Nocht Institute for Tropical Medicine, D-20359 Hamburg, Germany
Laurent Abel
Institut National de la Santé et de la Re cherche Médicale (INSERM) U.436, Mathematical and Statistical Modeling in Biology and Medicine, Hôpital Pitié-Salpêtrière, F-75013 Paris, France

REFERENCES AND NOTES

1. N. Risch and K. Merikangas, *Science* **273**, 1516 (1996).
2. We have $\alpha_1 = p(A/B) = p(AB)/P(B)$. Given that $p(AB)$ is $\delta + pm$ and $p(B) = m$, we find $\alpha_1 = \delta/m + p$. The value for α_2 is obtained in an analogous fashion.
3. Let $p(aff/B)$ be the probability for a child of a Bb parent to be affected given allele B is transmitted and $p(aff/b)$ be the corresponding probability given allele b is transmitted. By Bayes theorem $P(\text{tr} - B) = p(\text{aff}/B)/[p(\text{aff}/B) + p(\text{aff}/b)]$ since the prior probabilities of transmitting B and b are equal to 0.5 $p(\text{aff}/B)$ is $[\gamma\alpha_1 + (1 - \alpha_1)]D$, and $p(\text{aff}/b)$ is $[\gamma\alpha_2 + (1 - \alpha_2)]D$, where D is the probability that a subject is affected given he carries allele a. Thus, after some algebra, $P(\text{tr} - B) = [1 + (\gamma - 1)\alpha_1]/[2 + (\gamma - 1)(\alpha_1 + \alpha_2)]$.

11 November 1996; accepted 8 January 1997

Response: We agree with Scott *et al.* that linkage analysis will be able to identify genes of major, but not genes of modest, effect. As such, we also agree that linkage analysis should not be arbitrarily abandoned, because undoubtedly it will lead to the discovery of some important disease susceptibility genes. However, we do not agree that linkage analysis can detect most genes underlying complex diseases, and we anticipate that few genes for complex disorders will be identified in this fashion.

As indicated by Scott *et al.*, one measure of the total genetic effect for a complex disease is λ_s, the sibling risk ratio (*1*). However, it is generally impossible to determine the number of loci contributing to that total; if the number is large, even for a large value of λ_s, then none of the loci may be easily detected by linkage analysis.

We showed in our Perspective (*2*) that loci which confer a genotypic relative risk γ less than 4 would be difficult or impossible to identify with current linkage strategies. The numbers of sib pairs required to detect linkage that were given in the table in our Perspective (*2*, p. 1516) were actually underestimates, for two reasons: (i) There was an error in the computer program producing the required number of sib pairs for linkage; the actual numbers are approximately 50% larger than given (*3*); and (ii) the numbers given correspond to the ideal case of completely informative markers and no recombination. Allowing for more realistic circumstances of reduced marker informativity and moderate recombination, the corrected numbers probably would be about two to three times larger than given in the table. Thus, while it is still possible to detect a locus with γ of 4 or greater in a large family collection (say 500 or more), loci with smaller values of γ are unlikely to be detected.

How many loci are likely to exist for complex diseases with $\gamma > 4$? While it is difficult to know beforehand, animal models might offer a clue. As an example, the non-obese diabetic (NOD) mouse provides a useful model for human insulin dependent diabetes mellitus in being genetically complex, having an autoimmune etiology, and in the importance of the major histocompatibility loci. However, backcross experiments have shown that at least 10 other loci are probably involved in susceptibility, and only one of these loci had a value of γ greater than 4, with the rest in the range of 2 or less (*4*). We also note that animal backcross experiments are more analogous to human association studies than linkage studies, and this is why they have been more successful in identifying susceptibility loci than human linkage studies.

As indicated by Scott *et al.*, multiple sclerosis is a complex disease with a presumed substantial genetic component (*5*). However, three recently published genome screens (*6*) of moderate size did not produce clear and replicable evidence of linkage in any chromosomal region. This lack of susceptibility loci of large effect in this disease suggests that a very large number of families may be required to detect linkage.

The discovery of apoE as a major risk factor for late onset Alzheimer's disease is surely one of the major success stories of modern human genetics. Thus, it is important to evaluate the means by which this discovery was made. As indicated by Scott *et al.*, it has been estimated that apoE confers a λ_s value of around 2, with some modification for age of onset (*7*). Thus, in theory, this locus would be identifiable by linkage analysis with a sufficient number of sib pairs (several hundred minimum). In fact, the initial linkage observation on chromosome 19 (*8*), which produced a lod score of 4, was based on an analysis with markers that were likely to be in linkage disequilibrium with apoE. Performing linkage analysis with a marker associated with disease leads to an increase in the lod score (*9*). Similar linkage analysis with a nearby marker with little or no linkage disequilibrium (for example, the apo CII microsatellite) in the same material does not produce significant evidence for linkage (*8, 10*). Thus, in reality, the "linkage" discovery on chromosome 19 was actually based on an association between marker loci and the disease.

We agree with Scott *et al.* that genome-wide association studies will be based on future rather than current technology (as indicated in our title), and for the present we are still limited to the technology that exists. Although we agree that linkage studies should continue to be pursued, we also believe that this approach will produce only a modest number of loci for complex diseases.

We agree with Bell and Taylor that candidate genes are best tested in the framework of a biological hypothesis, often involving an interaction with a predisposing environmental agent, and the examples they provide are illuminating [for others, see (*11*)]. Also, as they point out, classic epidemiologic study designs, such as case-control or cohort, are excellent for testing such gene-environment interaction effects. The primary drawback from such designs for detecting genetic effects, however, is the potential for confounding, leading to an incorrect inference of causality for an observed association (*12*). Specifically, consider a population that has ethnic stratification and a tendency toward endogamy within strata. Further suppose these strata differ both in disease prevalence and allele frequencies at an unrelated locus. When performing a case-control study from such an admixed population, if the cases and controls are unbalanced for these strata, an allele frequency difference between cases and controls may emerge which is artifac-

tual and not causal. The solution, of course, is to precisely match the cases and controls according to these strata, or to perform a stratified analysis; such would be possible with the major ethnic groups such as exist in the United States. However, further strata are likely to exist within the major ethnic groupings (for example, European subgroups of Caucasians) for which matching and stratification might generally be quite difficult. Of course, this problem disappears in a completely randomly mating population.

This problem can also be solved by resorting to family-based association tests, such as the TDT we used in our analysis. This test has been shown to be immune to confounding due to population stratification (*13*). Also, in the absence of population stratification, this test has similar power to the usual case-control design (*14*). Furthermore, cases or families (or both) can also be classified according to a relevant environmental exposure and allelic transmission compared across these classes to search for gene-environment interactions. We also showed that unless the disease predisposing allele frequency is high, families with more than one affected child can be substantially more powerful than singletons, although they are also likely to be more difficult to find.

Because of the potential problem of genetic stratification, the optimal design for searching for genes of modest effect, especially in the absence of a clear biological model, is the family-based design, such as singleton or multiple affected sibs with parents. For early onset diseases, such samples should not be difficult to obtain, and are likely worth the potential additional cost. We would add that precise ethnic matching in a case-control paradigm can also lead to increased expense, if achievable at all. In the situation of late onset diseases, where parents are usually unavailable, an alternative design is discordant sib pairs, where effectively an unaffected sib serves as a control for the affected sib. This design also protects against genetic stratification artifact, but may lead to somewhat reduced power because of the genetic correlation between sibs (*14*).

Long *et al.* suggest a prospective study design where a random population sample is subsequently followed for development of disease. Presumably, at initiation, everyone in the study is genotyped for a large number of loci. They show that if the disease is sufficiently common, reasonable power is obtained by contrasting the allele frequencies in those who develop the disease with those that do not. The primary benefit from this approach is that multiple diseases can be studied using the same population of subjects, again provided the diseases are sufficiently common. It would appear that a minimum frequency of 10% is required to obtain plausible sample sizes for sufficient power.

There are also several drawbacks to this approach. First, as for the typical epidemiologic paradigms, such as case-control studies, there is the problem of population substructure as we have described (in our response to Bell and Taylor) and also mentioned by Long *et al.* Second, with this approach, sample pooling is not possible, because it is unknown a priori which individuals will become affected. Thus, this approach requires construction of individual genotypes, which can greatly magnify the technical effort. By contrast, for a typical case-control design, two pools can be formed—one for affected individuals, another for those unaffected, and overall allele frequencies within the two groups determined. Thus, for a study of n cases and n controls and t loci, genotypes for only $2t$ samples need to be determined as opposed to $2nt$ samples (*15*). The same efficiency may obtain for a family-based design, such as affected individuals and their parents, where those affected are pooled and contrasted to the pooled group of parents. While this approach cannot give the precise data needed for a TDT analysis, it still provides a robust, powerful, and efficient means for initial screening; any positive loci can subsequently be subjected to individual genotyping (*14*).

The approach of Long *et al.* would not be practical for rare diseases, for example, those with a population frequency less than 5%. However, a compromise is possible. Numerous studies already exist that sample affected individuals, with parents or unaffected sibs, for a variety of diseases. The subjects from these studies can be followed for a variety of other diseases and then subjected to analysis as they develop these other, more frequent diseases. Pooling across studies could then provide sufficient material.

As indicated by Müller-Myhsok and Abel, our analysis was based on association studies where the actual disease predisposing polymorphism is in hand. This is why we incorporated such a large number of tested alleles (1,000,000). We also indicated that the number of loci to be tested might be reduceable substantially if one allows for linkage disequilibrium. However, as pointed out by Müller-Myhsok and Abel, depending on linkage disequilibrium is not without risk. The power of the association test can decline dramatically as linkage disequilibrium diminishes or if the tested allele has a substantially different frequency than the disease allele. To a large extent, the expectation with regard to linkage disequilibrium across the genome is uncharted territory, and thus it is difficult to predict the power of using a less dense map at this point in time. However, we can present two cases that provide some degree of optimism. The first pertains to apoE and late onset Alzheimer's disease. Several polymorphisms in the apoE region show strong linkage disequilibrium and comparable allele frequencies, allowing association to be readily detected with other neighboring polymorphisms (*16*). A second example is the insulin VNTR region of chromosome 11p. Several polymorphisms in this region have been identified showing strong disequilibrium and similar allele frequencies, leading to comparable degrees of association with disease (*17*).

As genome-wide linkage studies are supplanted by genome-wide association studies, and the distribution of linkage disequilibrium across chromosomes and populations is further explored, the degree to which linkage disequilibrium as opposed to direct causality can be utilized to locate disease susceptibility loci in the genome will become more apparent.

Neil Risch
Department of Genetics,
Stanford University School of Medicine,
Stanford, CA 94305–5120, USA
E-mail: risch@lahmed.stanford.edu
Kathleen Merikangas
Department of Epidemiology,
Yale University School of Medicine,
New Haven, CT 06510, USA
E-mail: kath@zeus.psych.yale.edu

REFERENCES AND NOTES

1. N. Risch, *Am. J. Hum. Genet.* **46**, 222 (1990).
2. _____ and K. Merikangas, *Science* **273**, 1516 (1996).
3. In the computer program based on the formula given in reference 6 of (*2*), the value of σ was inadvertently fixed at 0. Because in most calculations the value of σ was close to 1, the correct values are approximately 1.5 times the values given in the table.
4. N. Risch, S. Ghosh, J. Todd, *Am. J. Hum. Genet.* **53**, 702 (1993).
5. G. Ebers, D. Sadovnick, N. Risch, *Nature* **377**, 150 (1995); A. D. Sadovnick *et al.*, *Lancet* **347**, 1728 (1996).
6. S. Sawcer *et al.*, *Nature Genet.* **13**, 464 (1996); Multiple Sclerosis Genetics Group, *ibid.* p. 469; G. Ebers *et al.*, *ibid.* p. 472.
7. A. D. Roses *et al.*, *Am. J. Hum. Genet.* **57**, A202 (1995).
8. M. A. Pericak-Vance *et al.*, *ibid.* **48**, 1034 (1991).
9. F. Clerget-Darpoux *et al.*, *Biometrics* **42**, 393 (1986).
10. M. A. Pericak-Vance, personal communication.
11. R. Ottman, *Genet. Epidemiol.* **7**, 177 (1990).
12. R. Spielman, R. E. McGinnis, W. J. Ewens, *Am. J. Hum. Genet.* **52**, 506 (1993).
13. W. J. Ewens *et al.*, *ibid.* **57**, 455 (1995).
14. N. Risch and J. Teng, in preparation.
15. N. Arnheim, C. Strange, H. Ehrlich, *Proc. Natl. Acad. Sci.* **82**, 6970 (1985); V. C. Sheffield, D. Y. Nishimura, E. M. Store, *Curr. Opin. Genetic Devel.* **5**, 335 (1995).
16. M-C. Chartier-Harlin *et al.*, *Hum. Mol. Genet.* **3**, 569 (1994).
17. C. Julier *et al.*, *Nature* **354**, 155 (1991); S. T. Bennett *et al.*, *Nature Genet.* **9**, 284 (1995).

20 January 1997; accepted 27 January 1997