The Future of Genetic Studies of Complex Human Diseases

Neil Risch and Kathleen Merikangas

Geneticists have made substantial progress in identifying the genetic basis of many human diseases, at least those with conspicuous determinants. These successes include Huntington's disease, Alzheimer's disease, and some forms of breast cancer. However, the detection of genetic factors for complex diseases—such as schizophrenia, bipolar disorder, and diabetes—has been far more complicated. There have been numerous reports of genes or loci that might underlie these disorders, but few of these findings have been replicated. The modest nature of the gene effects for these disorders likely explains the contradictory and inconclusive claims about their identification. Despite the small effects of such genes, the magnitude of their attributable risk (the proportion of people affected due to them) may be large because they are quite frequent in the population, making them of public health significance.

Has the genetic study of complex disorders reached its limits? The persistent lack of replicability of these reports of linkage between various loci and complex diseases might imply that it has. We argue below that the method that has been used successfully (linkage analysis) to find major genes has limited power to detect genes of modest effect, but that a different approach (association studies) that utilizes candidate genes has far greater power, even if one needs to test every gene in the genome. Thus, the future of the genetics of complex diseases is likely to require large-scale testing by association analysis.

How large does a gene effect need to be in order to be detectable by linkage analysis? We consider the following model: Suppose a disease susceptibility locus has two alleles A and a, with population frequencies p and $q = 1 - p$, respectively. There are three genotypes: AA, Aa, and aa. We define genotypic relative risks (GRR, the increased chance that an individual with a particular genotype has the disease) as follows: Let the risk for individuals of genotype Aa be γ times greater than the risk for individuals with genotype aa, a GRR of γ. We assume a multiplicative relation for two A alleles, so that the GRR for genotype AA is γ^2. The method of linkage analysis we have chosen for this argument is a popular current paradigm in which pairs of siblings, both with the disease, are examined for sharing of alleles at multiple sites in the genome defined by genetic markers. The more often the affected siblings share the same allele at a particular site, the more likely the site is close to the disease gene. Using the formulas in (*1*), we calculate the expected proportion Y of alleles shared by a pair of affected siblings for the best possible case—that is, a closely linked marker locus (recombination fraction $\theta = 0$) that is fully informative (heterozygosity = 1) (*2*)—as

$$Y = \frac{1 + w}{2 + w} \text{ where } w = \frac{pq\,(\gamma - 1)^2}{(p\gamma + q)^2}$$

If there is no linkage of a marker at a particular site to the disease, the siblings would be expected to share alleles 50% of the time; that is, Y would equal 0.5. Values of Y for various values of p and γ are given in the third column of the table. For an allele of moderate frequency (p is 0.1 to 0.5) that confers a GRR (γ) of fourfold or greater, there is a detectable deviation of Y from the null value of 0.5. On the other hand, for an allele conferring a GRR of 2 or less, the expected marker-sharing only marginally exceeds 50%, for any allele frequency (p). Thus, it is clear that the use of linkage analysis for loci conferring GRR of about 2 or less will never allow identification because the number of families required (more than ~2500) is not practically achievable.

Although tests of linkage for genes of modest effect are of low power, as shown by the above example, direct tests of association with a disease locus itself can still be quite strong. To illustrate this point, we use the transmission/disequilibrium test of Spielman *et al.* (*3*). In this test, transmission of a particular allele at a locus from heterozygous parents to their affected offspring is examined. Under Mendelian inheritance, all alleles should have a 50% chance of being transmitted to the next generation. In contrast, if one of the alleles is associated with disease risk, it will be transmitted more often than 50% of the time.

For this approach, we do not need families with multiple affected siblings, but can focus just on single affected individuals and their parents. For the same model given above, we can calculate the proportion of heterozygous parents as $pq(\gamma + 1)/(p\gamma + q)$(*4*). Similarly, the probability for a heterozygote parent to transmit the high risk A allele is just $\gamma/(1 + \gamma)$. Association tests can also be performed for pairs of affected siblings. When the locus is associated with disease, the transmission excess over 50% is the same as for single offspring, but the probability of parental heterozygosity is increased at low values of p; for higher values of p, the probability of parental heterozygosity is decreased. The formula for parental heterozygosity for an affected pair of siblings for the same genetic model as used in the first example is

$$h = \frac{pq\,(\gamma + 1)^2}{2\,(p\gamma + q)^2 + pq(\gamma - 1)^2}$$

		Linkage			Association			
					Singletons		Sib pairs	
Genotypic risk ratio (γ)	Frequency of disease allele A (p)	Probability of allele sharing (Y)	No. of families required (N)	Probability of transmitting disease allele A P(tr-A)	Proportion of heterozygous parents (Het)	(N)	(Het)	(N)
4.0	0.01	0.520	4260	0.800	0.048	1098	0.112	235
	0.10	0.597	185	0.800	0.346	150	0.537	48
	0.50	0.576	297	0.800	0.500	103	0.424	61
	0.80	0.529	2013	0.800	0.235	222	0.163	161
2.0	0.01	0.502	296,710	0.667	0.029	5823	0.043	1970
	0.10	0.518	5382	0.667	0.245	695	0.323	264
	0.50	0.526	2498	0.667	0.500	340	0.474	180
	0.80	0.512	11,917	0.667	0.267	640	0.217	394
1.5	0.01	0.501	4,620,807	0.600	0.025	19,320	0.031	7776
	0.10	0.505	67,816	0.600	0.197	2218	0.253	941
	0.50	0.510	17,997	0.600	0.500	949	0.490	484
	0.80	0.505	67,816	0.600	0.286	1663	0.253	941

Comparison of linkage and association studies. Number of families needed for identification of a disease gene.

N. Risch is in the Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305–5120, USA. E-mail: risch@lahmed.stanford.edu. K. Merikangas is in the Departments of Epidemiology and Psychiatry, Unit, Yale University School of Medicine, New Haven, CT 06510, USA. E-mail: kath@zeus.psych.yale.edu

On the right side of the table, we present the proportion of heterozygous parents (Het) and the probability of transmission of the A allele from a heterozygous parent to an affected child [P(tr-A)] for the same values of GRR as considered above for the example of linkage analysis. The deviation from the null hypothesis of 50% transmission from heterozygous parents is substantially greater than the excess allele sharing that is found by linkage analysis in sibling pairs. This disparity between the methods is particularly true for lower values of γ (that is, with lower relative risk). For example, for $\gamma = 1.5$, allele sharing is at most 51%, while the A allele is transmitted 60% of the time from heterozygous parents.

In this respect then, association studies seem to be of greater power than linkage studies. But of course, the limitation of association studies is that the actual gene or genes involved in the disease must be tentatively identified before the test can be performed. In fact, the actual polymorphism within the gene (or at least a polymorphism in strong disequilibrium) must be available. However, we show that this requirement is only daunting because of limitations imposed by current technological capabilities, not because sufficient families with the disease are not available or the statistical power is inadequate (5). For example, imagine the time when all human genes (say 100,000 in total) have been found and that simple, diallelic polymorphisms in these genes have been identified. Assume that five such diallelic polymorphisms have been identified within each gene, so that a total of $10 \times 10^5 = 10^6$ alleles need to be tested. The statistical problem is that the large number of tests that need to be made leads to an inflation of the type 1 error probability. For a linkage test with pairs of affected siblings, we use a lod score (logarithm of the odds ratio for linkage) criterion of 3.0, which asymptotically corresponds to a type 1 error probability α of about 10^{-4}. In a linkage genome screen with 500 markers, this significance level gives a probability greater than 95% of no false positives. The equivalent false positive rate for 1,000,000 independent association tests can be obtained with a significance level $\alpha = 5 \times 10^{-8}$.

We illustrate the power of linkage versus association tests at different significance levels by determining the sample size N (number of families) necessary to obtain 80% power (the probability of rejecting the null hypothesis when it is false) (6) (see table). With a linkage approach and a disease gene with a GRR of 4 or greater, the number of affected sibling pairs necessary to detect linkage is realistic (185 or 297), provided the allele frequency p is between 5 and 75%. For a gene with a GRR of 2 or less, however, the sample sizes are generally beyond reach (well over 2000), precluding their identification by this approach. In contrast, the required sample size for the association test, even allowing for the smaller significance level, is vastly less than for linkage, especially for affected sibling pair families when the value of p is small. Even for a GRR of 1.5, the sample sizes are generally less than 1000, well within reason.

Thus, the primary limitation of genome-wide association tests is not a statistical one but a technological one. A large number of genes (up to 100,000) and polymorphisms (preferentially ones that create alterations in derived proteins or their expression) must first be identified, and an extremely large number of such polymorphisms will need to be tested. Although testing such a large number of polymorphisms on several hundred, or even a thousand families, might currently seem implausible in scope, more efficient methods of screening a large number of polymorphisms (for example, sample pooling) may be possible. Furthermore, the number of tests we have used as the basis for our calculations (1,000,000) is likely to be far larger than necessary if one allows for linkage disequilibrium, which could substantially reduce the required number of markers and families needed for initial screening.

Some of the important loci for complex diseases will undoubtedly be found by linkage analysis. However, the limitations to detecting many of the remaining genes by linkage studies can be overcome; numerous genetic effects too weak to identify by linkage can be detected by genomic association studies. Fortunately, the samples currently collected for linkage studies (for example, affected pairs of siblings and their parents) can also be used for such association studies. Thus, investigators should preserve their samples for future large-scale testing.

The human genome project can have more than one reward. In addition to sequencing the entire human genome, it can lead to identification of polymorphisms for all the genes in the human genome and the diseases to which they contribute. It is a charge to the molecular technologists to develop the tools to meet this challenge and provide the information necessary to identify the genetic basis of complex human diseases.

References and Notes

1. N. Risch, *Am. J. Hum. Genet.* **40**, 1 (1987); *ibid.* **46**, 229 (1990).
2. From the formulas in (*1*), we have $\lambda_O = 1 + 0.5V_A/K^2$ and $\lambda_S = 1 + (0.5V_A + 0.25V_D)/K^2$, where $K = p^2\gamma^2 + 2pq\gamma + q^2 = (p\gamma + q)^2$, $V_A = 2pq(\gamma - 1)^2 (p\gamma + q)^2$, and $V_D = p^2q^2(\gamma - 1)^2$. Hence, $\lambda_O = 1 + w$ and $\lambda_S = (1 + 0.5w)^2$, where $w = pq(\gamma - 1)^2$. The proportion of alleles shared is given by $Y = 1 - 0.5z_1 - z_o$, where z_1 and z_o are the probabilities of the sib pair sharing 1 and 0 disease alleles ibd, respectively. From (*1*), $z_o = 0.25/\lambda_S$ and $z_1 = 0.5\lambda_O/\lambda_S$. Thus, after some algebra, $Y = 1 - 0.25(\lambda_O + 1)/\lambda_S = (1 = w)/(2 + w)$.
3. R. Spielman, R. E. McGinnis, W. J. Ewens, *Am. J. Hum. Genet.* **52**, 506 (1993).
4. By Bayes theorem, the probability of a parent of an affected child being heterozygous is given by P(Het|Aff child) = P(Het)P(Aff Child|Het)/P(Aff Child) $= 2pq[0.5p(\gamma^2 + \gamma)] + 0.5q(\gamma + 1)/(p\gamma + q)^2 = pq(\gamma + 1)/(p\gamma + q)$.
5. E. S. Lander and N. J. Schork, *Science* **265**, 2037 (1994).
6. Consider a set of M independent, identically distributed random variables B_i of discrete value. Under the null hypothesis H_0, assume $E(B_i) = 0$ and $\mathrm{Var}(B_i) = 1$. Under the alternative hypothesis H_1, let $E(B_i) = \mu$ and $\mathrm{Var}(B_i) = \sigma^2$. For a sample of size M, let $T = \Sigma B_i/\sqrt{M}$. Then under H_0, T also has mean 0 and variance 1, while under H_1, it has mean $\sqrt{M}\,\mu$ and variance σ^2. We assume that T is approximately normally distributed both under H_0 and H_1. Then the sample size M required to obtain a power of $1 - \beta$ for a significance level α is given by

$$M = (Z_\alpha - \sigma Z_{1-\beta})^2/\mu^2 \quad (1)$$

For each affected sib pair, we score the number of alleles shared ibd from each of $2N$ parents. Define $B_i = 1$ if an allele is shared from the ith parent and $B_i = -1$ if unshared. Under the null hypothesis of no linkage, $P(B_i = 1) = P(B_i = -1) = 0.5$, so $E(B_i) = 0$ and $\mathrm{Var}(B_i) = 1$. For the genetic model described above with genotypic relative risks of γ^2, γ, and 1, allele sharing by affected sibs is independent for the two parents; thus, we can consider sharing of alleles one parent at a time. Thus, for affected sib pairs assuming $\theta = 0$ and no linkage disequilibrium, the formula is

$$N = \frac{(Z_\alpha - \sigma Z_{1-\beta})^2}{2\mu^2}$$

where

$$\mu = 2Y - 1$$

$$\sigma^2 = 4Y(1 - Y)$$

$$Y = \frac{1 + w}{2 + w}$$

$$w = \frac{pq\,(\gamma - 1)^2}{(p\gamma + q)^2}$$

$Z_\alpha = 3.72$ (corresponding to $\alpha = 10^{-4}$), and $Z_{1-\beta} = -0.84$ (corresponding to $1 - \beta = 0.80$). For an association test using the transmission/disequilibrium test, with the disease locus or a nearby locus in complete disequilibrium, the number (N) of families with affected singletons required for 80% power is also calculated from formula 1. For this case, we score the number of transmissions of allele A from heterozygous parents. Let h be the probability a parent is heterozygous under the alternative hypothesis, namely, $h = pq(\gamma + 1)/(p\gamma + q)$. Then define $B_i = h^{-0.5}$ if the parent is heterozygous and allele A is transmitted; $B_i = 0$ if the parent is homozygous; and $B_i = -h^{-0.5}$ if the parent is heterozygous and transmits allele a. Under the null hypothesis, $E(B_i) = 0$ and $\mathrm{Var}(B_i) = 1$. Under the alternative hypothesis, $\mu = E(B_i) = \sqrt{h}(\gamma - 1)/(\gamma + 1)$ and $\sigma^2 = \mathrm{Var}(B_i) = 1 - h(\gamma - 1)^2/(y + 1)^2$. In this case, there are two parents per family and they act independently, so the required number (N) of families is given by half of formula 1 where μ and σ^2 are given above. Here, $Z_\alpha = 5.33$ (corresponding to $\alpha = 5 \times 10^{-8}$). For the same test but with affected sib pairs instead of singletons, the number of families required is given by half of formula 1 (transmissions from two parents to two children) with the same formulas for μ and σ^2 as for singleton families but now using the heterozygote frequency for parents of affected sib pairs. Using the above formulas, we can calculate sample sizes for the three study designs.

27 October 1995; accepted 6 June 1996.