Am. J. Hum. Genet. 47:266–274, 1990

Effective Testing of Gene-Disease Associations

Michael Swift,*,† Lawrence L. Kupper,‡ and Charles L. Chase*

*Biological Sciences Research Center, †Genetics Division, Department of Medicine, School of Medicine, and ‡Department of Biostatistics, School of Public Health, University of North Carolina, Chapel Hill, NC

Summary

We propose a method for testing any hypothesized association between a candidate allele, for which there is a specific laboratory test, and a common chronic disease. Families in which this allele is segregating are identified through index individuals who are homozygous or heterozygous for the allele. The sample consists of the subset of identified families who also have at least one member with the common disease of interest. For each independent family in this subset, select one person with the disease and determine if he or she is heterozygous for the allele. The observed proportion of heterozygotes in this sample is compared to the proportion expected on the basis of each diseased relative's null probability of being heterozygous for the allele; this null probability depends only on the relative's relationship to the index individual and the population allele frequency. We provide these null probabilities, develop appropriate inference procedures, discuss sample size requirements, and compare this method to a standard case-control design. Results using this method are unlikely to be influenced by confounders, systematic bias, or genetic heterogeneity.

Introduction

Currently available molecular genetic techniques have greatly enhanced the power of genetic and epidemiologic strategies used to identify specific genes that predispose to common chronic diseases. For example, the increasingly detailed genomic map of DNA polymorphisms may make it possible to adapt linkage methods, highly successful for recognized Mendelian syndromes, to map some genes for non-Mendelian chronic disorders (Lander and Botstein 1986). However, because of clinical and genetic heterogeneity (Goldin and Gershon 1988), the practical usefulness of linkage analysis in this setting remains to be determined.

Another well-known strategy for testing the hypothesized association of a "candidate allele" with a disease is to compare the allele frequency among diseased persons with the frequency in a well or control population (Cooper and Clayton 1988). Establishing an association through such population studies is easy when the association is very strong and the population frequency of the candidate allele is sufficiently high, as for ankylosing spondylitis and HLA B27 (Ebringer 1980). In many instances, however, different population studies have yielded conflicting results for potentially important associations, such as those of constitutive h-*ras* variants with specific cancers (Krontiris et al. 1985; Gerhard et al. 1987). Establishing gene-disease associations promptly and reliably is important for identifying persons at high risk for the disease and for elucidating the biochemical mechanisms, including gene-environment interactions, underlying the disease.

We developed an alternative method, described below, for assessing gene-disease associations for non-Mendelian disorders when we sought the most efficient strategy for using molecular methods to identify mutant alleles in a specific setting: testing the predisposition of ataxia-telangiectasia (A-T; a cancer-prone autosomal recessive syndrome) heterozygotes to specific cancers (Swift et al. 1987, 1990). This new method is general in that it applies to *any* hypothesized gene-disease association; it has specific advantages over alternative strategies.

General Method

Consider an autosomal locus at which there are two alleles, A and a. Further suppose that the rarer allele,

Received November 17, 1989; final revision received April 20, 1990.

Address for correspondence and reprints: Michael Swift, M.D., 325 Biological Sciences Research Center #7250, University of North Carolina, Chapel Hill, NC 27599.

0002-9297/90/4702-0011$02.00

a, has been hypothesized to be associated, in the heterozygous state Aa, with disease D. Disease D typically would be a non-Mendelian common disease such as a specific cancer, mental illness, or diabetes mellitus. Such disorders rarely show clear Mendelian patterns in families; for a common disease in which a typical Mendelian pattern *is* evident, linkage analysis is more suitable than our proposed method. Select, from the population in which you wish to test the association of a with D, a sample of N' unrelated index individuals who are heterozygous or homozygous for allele a. The index individuals may be homozygous or heterozygous for the allele, since they serve only to identify families in which the allele is segregating; index individuals, who may be found through a population survey, need not, and typically will not, have disease D.

The next step is to compile pedigrees with reliable clinical information for each of the N' families identified through the index individuals heterozygous or homozygous for a. Using the clinical information, select N blood relatives who have disease D from the N' families, with the following rules. The selected diseased blood relative cannot be an index person. Select only one diseased blood relative from the family of each Aa index person. If the index person is aa and his parents are not related, one individual with disease D may be selected from the maternal and one from the paternal lineage; this is because allele a is segregating in each lineage and the allele status of each maternal relative is independent of that of each paternal relative. Then, for each of the N selected blood relatives with disease D, determine his or her genotype at the locus.

The observed proportion of these N diseased relatives with genotype Aa is then compared to the proportion expected from the familial relationships and from the (assumed to be known) population gene frequency q ($0 < q \leqslant .5$) of a. A statistically significant elevation of the observed over the expected proportion provides evidence in support of an association between disease D and Aa status. For most chronic diseases, the observed odds ratio estimates reasonably well the relative risk, which is the ratio of the risk of disease for Aa carriers to the risk for AA noncarriers.

Null Probabilities of Being Heterozygous for Allele a for Blood Relatives of a Known Aa Heterozygote or aa Homozygote

We will consider three scenarios: (1) the index subject is an Aa heterozygote; (2A) the index subject is an aa homozygote who is clinically normal; (2B) the index subject is an aa homozygote and aa homozygotes are all clinically abnormal (or, aa is "genetically lethal"). Scenario 2B applies when the index individuals all have a specific autosomal recessive syndrome, such as A–T. The individuals with the syndrome need not have disease D, although it is likely that they themselves do have a high risk of developing it. For the first two scenarios, standard methods (Li and Sacks 1954; Campbell and Elston 1971) provide exact expressions (functions of q; see table 1, parts A and B) for the conditional probabilities that blood relatives of the index subject are Aa assuming random mating and that no mating pair has genes identical by descent. These probabilities are the proportions of the diseased blood relatives that are expected to be heterozygous for allele a under the null hypothesis of no association between the heterozygous state Aa and disease D. Standard conditional probability arguments are used to derive the null probabilities under scenario 2B (table 1, part C), in which the rare aa relative will be readily recognized and excluded from the analysis because he or she will have the distinctive phenotype of the autosomal recessive syndrome of the index individuals.

Tests of significance and related assessments of statistical power depend on these null probabilities. Table 1 shows that inaccuracy in the measurement of q is unlikely to affect statistical inferences in most practical circumstances. First, even varying q twofold to fivefold has little effect on the null probabilities for close relatives of the index individual (e.g., parents, children, and siblings in scenario 1). Second, for almost all degrees of relationship, there is little variation in the null probabilities for all values of $q \leqslant .03$. The population frequencies of hypothesized disease-predisposing alleles typically fall in this range. The accuracy of the measurement of q is of much greater importance for allele frequencies higher than 0.03 and when relatives as distant as first cousins are included in the sample. For alleles whose frequency is 0.10 or higher *and* somewhat uncertain, alternatives to our method should be used to test hypothesized associations. If the source population is composed of two or more subpopulations with widely discrepant values of q, it is appropriate to analyze such subpopulations separately.

If allele a predisposes heterozygotes to a common disease, it is likely that aa homozygotes are also predisposed to the same common disease. However, in most practical applications, q will be small, and homozygous blood relatives will be encountered rarely. When it is of interest to examine *both* the Aa and aa genotypes as potentially important genetic risk factors for

Table 1

Conditional Probability of Heterozygosity

A. Conditional Probability θ That a Relative of a Known Heterozygote (Aa) Is Heterozygous (Aa) for Selected Values of q in the Population under Study (Scenario 1)[a]

q	Sibling[b]	Second-Degree Relative[c]	First Cousin[d]
.001	.5005	.2510	.1265
.005	.5025	.2550	.1325
.010	.5049	.2599	.1399
.030	.5145	.2791	.1687
.050	.5237	.2975	.1963
.100	.5450	.3400	.2600
.200	.5800	.4100	.3650

B. Conditional Probability θ That a Relative of a Clinically Normal Homozygote (aa) Is Heterozygous (Aa) for Selected Values of q in the Population under Study (Scenario 2A)

q	Child or Parent[e]	Sibling[f]	Second-Degree Relative[g]	First Cousin[h]
.001	.9990	.5000	.5005	.2512
.005	.9950	.5000	.5025	.2562
.010	.9900	.4999	.5049	.2623
.030	.9700	.4996	.5141	.2862
.050	.9500	.4988	.5225	.3088
.100	.9000	.4950	.5400	.3600
.200	.8000	.4800	.5600	.4400

C. Conditional Probability θ That a Relative of a Clinically Abnormal Homozygote (aa) Is Heterozygous (Aa), Conditional on Such Relatives and Their Mates Not Being aa, for Selected Values of q in the Population under Study (Scenario 2B)[i]

q	Grandparent[j]	Aunt or Uncle[k]	Niece or Nephew[l]	First Cousin[m]
.001	.5005	.5001	.3338	.2506
.005	.5025	.5006	.3355	.2531
.010	.5050	.5013	.3377	.2562
.030	.5150	.5038	.3464	.2683
.050	.5250	.5063	.3548	.2800
.100	.5500	.5128	.3750	.3079
.200	.6000	.5263	.4118	.3578

[a] The conditional probability that a parent or child of a known heterozygote (Aa) is heterozygous (Aa) is .5, independently of q.
[b] $(1 + q - q^2)/2$.
[c] $(1 + 4q - 4q^2)/4$ (grandparent, aunt, uncle, niece or nephew).
[d] $(1 + 12q - 12q^2)/8$.
[e] $(1 - q)$.
[f] $(1 - q^2)/2$.
[g] $(1 + q - 2q^2)/2$.
[h] $(1 + 5q - 6q^2)/4$.
[i] Under Scenario 2B, the conditional probabilities that a parent and a sibling of the index person are heterozygous (Aa) are 1.0 and 2/3, respectively, independently of q.
[j] $(1 + q)/2$.
[k] $2/(4 - q)$.
[l] $(1 + 2q)/(3 + 2q)$.
[m] $(1 + 3q - q^2)/(4 + 2q - q^2)$.

the disease under study, appropriate modifications can be made to the binomial-based analysis discussed below for the Aa genotype.

Statistical Inference and Sample-Size Considerations

As stated earlier, to avoid certain dependency-related complications, our proposed study design dictates choosing exactly one blood relative (of any type) with disease D for each identified Aa heterozygote (or one diseased maternal relative and one diseased paternal relative of each aa homozygote).

Probability Model

For the ith diseased blood relative ($i = 1,2, \ldots, N$), define the dichotomous random variable R_i to be 1 if the ith diseased blood relative is heterozygous (Aa), and 0 if not. Note that $R_i = 0$ if the ith diseased blood relative is either AA or aa under scenarios 1 and 2A. The aa genotype in relatives is excluded under scenario 2B, since these relatives will have the distinctive phenotype of the autosomal recessive syndrome in the index individuals.

Now, let $\theta_i = \text{pr}(R_i = 1 \mid D)$. And, under H_0 (no association between being a heterozygote and disease risk), let θ_{0i} denote the null value of θ_i. Given the familial relationship of the ith selected diseased blood relative and the known value of q, the specific numerical value of θ_{0i} (the null probability that this ith blood relative is heterozygous Aa) is determined using the methods of the previous section. Under H_A (positive association between being a heterozygote and disease risk), θ_i is greater than θ_{0i} for $i = 1,2, \ldots, N$. The methods to be developed apply also to testing for a protective, as opposed to a detrimental, effect of heterozygosity. When an allele protects against disease D, θ_i is less than θ_{0i} for $i = 1,2, \ldots, N$.

The effect measure of interest will be the odds ratio

$$\phi = \frac{\theta_i/(1 - \theta_i)}{\theta_{0i}/(1 - \theta_{0i})}, \text{ where } 0 < \phi < +\infty .$$

This odds ratio will not depend on i unless other risk factors for disease D are nonrandomly distributed with respect to allele a. Note that

$$\theta_i = \frac{\phi\theta_{0i}}{1 + \theta_{0i}(\phi - 1)},$$

so that the mean

$$\bar{\theta} = \frac{\phi}{N} \sum_{i=1}^{N} \frac{\theta_{0i}}{1 + \theta_{0i}(\phi - 1)} .$$

In terms of ϕ, the null hypothesis of interest is $H_0:\phi=1$ and the alternative hypothesis is either $H_A:\phi>1$ for a hypothesized detrimental effect of heterozygosity or $H_A:\phi_A<1$ for a protective effect.

Maximum-Likelihood Methods

Since the R_i's are mutually independent 0–1 random variables, the likelihood function L for these data is

$$L = \prod_{i=1}^{N} \left\{ \left[\frac{\phi\theta_{0i}}{1 + \theta_{0i}(\phi - 1)}\right]^{R_i} \times \left[\frac{(1 - \theta_{0i})}{1 + \theta_{0i}(\phi - 1)}\right]^{1 - R_i} \right\} .$$

Equating $d\ln L/d\phi$ with zero gives $\hat{\phi}$, the maximum-likelihood estimator of ϕ, as the solution to the likelihood equation

$$\sum_{i=1}^{N} R_i = \sum_{i=1}^{N} \left[\frac{\hat{\phi}\theta_{0i}}{1 + \theta_{0i}(\hat{\phi} - 1)}\right] . \qquad (1)$$

In general, equation (1) must be solved iteratively. A good starting value for this iteration process is

$$\frac{\bar{R}/(1 - \bar{R})}{\bar{\theta}_0/(1 - \bar{\theta}_0)} ,$$

where $\bar{R}$ is the observed proportion of diseased blood relatives who are heterozygous and $\bar{\theta}_0$ is the expected value of $\bar{R}$ under H_0, or

$$\bar{R} = N^{-1} \sum_{i=1}^{N} R_i, \text{ and } \bar{\theta}_0 = N^{-1} \sum_{i=1}^{N} \theta_{0i} .$$

Since $[\ln\hat{\phi} - \ln\phi]/[\hat{\text{Var}}(\ln\hat{\phi})]^{1/2}$ is approximately $N(0,1)$ for moderate to large N (Woolf 1955), where $\hat{\text{Var}}(\ln\hat{\phi})$ is the estimated variance of $\ln\hat{\phi}$, it can be shown that an approximate 100 $(1-\alpha)$% confidence interval for ϕ is

$$\hat{\phi} \exp\{\pm Z_{1-\alpha/2} [\hat{\text{Var}}(\ln\hat{\phi})]^{1/2}\} , \qquad (2)$$

where $Z_{1-\alpha/2}$ is the $100(1-\alpha/2)$ percentile point of the standard normal distribution and where

$$\hat{\text{Var}}(\ln\hat{\phi}) = \left\{\hat{\phi} \sum_{i=1}^{N} \frac{\theta_{0i}(1-\theta_{0i})}{[1 + \theta_{0i}(\hat{\phi}-1)]^2}\right\}^{-1} . \qquad (3)$$

Power and Sample-Size Considerations

To test H_0 versus H_A, we recommend the use of the score test statistic

$$(\bar{R} - \bar{\theta}_0)/[\text{Var}(\bar{R}|H_0)]^{1/2} ,$$

where

$$\text{Var}(\bar{R}|H_0) = N^{-2} \sum_{i=1}^{N} \theta_{0i}(1 - \theta_{0i})$$

is the variance of $\bar{R}$ under H_0. Under H_0, this test statistic will have an approximate standard normal distribution for moderate to large N. Hence, the approximate power of a size α test of $H_0:\phi = 1$ versus $H_A:\phi>1$ equals

$$\text{pr}\left\{\frac{(\bar{R} - \bar{\theta}_0)}{[N^{-2} \sum_{i=1}^{N} \theta_{0i}(1 - \theta_{oi})]^{1/2}} > Z_{1-\alpha} | \phi > 1\right\} ,$$

where pr $(Z > Z_{1-\alpha}) = \alpha$ when $Z \sim N(0,1)$. Thus, to achieve a power of at least $(1 - \beta)$, standard statistical arguments dictate that N should be the smallest positive integer satisfying the inequality

$$N \geq \frac{Z_{1-\alpha}[\sum_{i=1}^{N} \theta_{0i}(1 - \theta_{0i})]^{1/2} + Z_{1-\beta}[\sum_{i=1}^{N} \theta_i(1 - \theta_i)]^{1/2}}{\bar{\theta} - \bar{\theta}_0} .$$

Since

$$N\bar{\theta}_0(1 - \bar{\theta}_0) \geq \sum_{i=1}^{N} \theta_{0i}(1 - \theta_{0i})$$

and

$$N\bar{\theta}(1 - \bar{\theta}) \geq \sum_{i=1}^{N} \theta_i(1 - \theta_i) ,$$

an upper bound for N is that positive integer, say N^*, satisfying the inequality

$$N^* \geq \frac{\{Z_{1-\alpha}[\bar{\theta}_0(1 - \bar{\theta}_0)]^{1/2} + Z_{1-\beta}[\bar{\theta}(1 - \bar{\theta})]^{1/2}\}^2}{(\bar{\theta} - \bar{\theta}_0)^2} .$$

Let us assume that $\bar{\theta} \doteq [\phi\bar{\theta}_0] / [1 + \bar{\theta}_0(\phi - 1)]$. This assumption is exactly true when the diseased blood relatives are all of the same type. In most practical situations, the absolute difference between the exact and approximate values of $\bar{\theta}$ should be less than .05.

Combining the above approximation with the previous inequality involving N^* gives the inequality

$$N^* \geq \frac{\{Z_{1-\alpha}[1 + \bar{\theta}_0(\phi - 1)] + Z_{1-\beta}\phi^{1/2}\}^2}{\bar{\theta}_0(1 - \bar{\theta}_0)(\phi - 1)^2}. \quad (4)$$

The motivation behind the development leading to expression (4) is now apparent. To use expression (4) to help design a study of the kind we are advocating (i.e., to obtain an approximate idea of the needed number N^* of diseased blood relatives), it is only necessary to specify values for the following quantities: α, the size of the rejection region for a one-sided test of $H_0:\phi=1$ versus either $H_A:\phi>1$ or $H_A:\phi<1$; $(1 - \beta)$, the desired power of the test; ϕ, the anticipated value of the population odds ratio (a value greater than 1 for a detrimental effect of being a heterozygote, and a value less than 1 for a protective effect); and $\bar{\theta}_0$, the mean of the N θ_{0i} values based on the types of diseased blood relatives likely to be chosen and on the value of q. Table 2 contains values of N^* based on (4) for some combinations of values of α, $(1 - \beta)$, ϕ, and $\bar{\theta}_0$. When the anticipated true value of ϕ is greater than 2.0, note that N^* does not vary much with changes in $\bar{\theta}_0$. Table 2 also demonstrates that a relatively small increase in sample size results in a significant increase in power.

Also, from expression (4) and from the entries in table 2, it is clear that, for a given value of ϕ, the required sample size increases as $\bar{\theta}_0$ gets closer to zero. Hence, our proposed "enrichment process" of making the null probabilities of heterozygosity larger than $2q(1 - q)$ via the use of relatives of index persons known to carry allele a leads to a statistically powerful design.

Finally, for fixed values of α, β, and ϕ, the choice of $\bar{\theta}_0$, say $\bar{\theta}_0^*$, which minimizes the right-hand side of inequality (4) is

$$\bar{\theta}_0^* = \left[1 + \frac{(Z_{1-\alpha}\phi + Z_{1-\beta}\phi^{1/2})}{(Z_{1-\alpha} + Z_{1-\beta}\phi^{1/2})}\right]^{-1}.$$

Since this optimal $\bar{\theta}_0^*$ varies inversely with ϕ, the power of our method to detect the effects of genes that protect against a disease is much greater when the persons with disease D are close, rather than more distant, relatives of the index individuals (table 2, part B).

Illustrative Example

The following hypothetical example shows how the proposed methodology will be used to assess whether females heterozygous for the A-T gene are at elevated risk for developing breast cancer (Swift et al. 1987) when there is a reliable test for the A-T heterozygote. The A-T

Table 2

Power-based Sample-Size Requirements

A. Sample Sizes (N^*) from Equation (4) Required to Test $H_0:\phi=1$ versus $H_A:\phi>1$ with Power at Least $(1 - \beta)$ for $\alpha = .01$, for Selected Values of the Odds Ratio (ϕ), and of the Expected Proportion of Heterozygotes under the Null Hypothesis ($\bar{\theta}_0$)

	$1 - \beta = .75$			$1 - \beta = .80$			$1 - \beta = .85$		
ϕ	$\bar{\theta}_0 = .15$	$\bar{\theta}_0 = .30$	$\bar{\theta}_0 = .55$	$\bar{\theta}_0 = .15$	$\bar{\theta}_0 = .30$	$\bar{\theta}_0 = .55$	$\bar{\theta}_0 = .15$	$\bar{\theta}_0 = .30$	$\bar{\theta}_0 = .55$
1.5	348	234	233	392	262	259	445	296	290
2.0	104	76	84	118	85	93	135	96	104
2.5	54	42	51	61	47	56	70	54	63
3.0	35	29	37	40	32	41	46	37	45
3.5	25	22	30	29	25	33	34	28	36
4.0	20	18	26	23	20	28	26	23	31
4.5	16	15	23	19	17	25	22	19	27
5.0	14	14	21	16	15	23	18	17	25

B. Sample Sizes (N^*) from Equation (4) Required to Test $H_0:\phi=1$ versus $H_A:\phi<1$ with Power at Least $(1 - \beta)$ for $\alpha = .01$, for Selected Values of the Odds Ratio (ϕ), and of the Expected Proportion of Heterozygotes under the Null Hypothesis ($\bar{\theta}_0$)

	$1 - \beta = .75$			$1 - \beta = .80$			$1 - \beta = .85$		
ϕ	$\bar{\theta}_0 = .15$	$\bar{\theta}_0 = .30$	$\bar{\theta}_0 = .55$	$\bar{\theta}_0 = .15$	$\bar{\theta}_0 = .30$	$\bar{\theta}_0 = .55$	$\bar{\theta}_0 = .15$	$\bar{\theta}_0 = .30$	$\bar{\theta}_0 = .55$
.25	81	39	21	87	42	23	93	46	26
.50	217	115	76	237	126	85	261	140	95

gene has been mapped to chromosome 11q22-23 (Gatti et al. 1988), so that A-T heterozygotes can be identified in A-T families once closely linked markers or allele-specific probes become available.

Suppose that $N' = 60$ families of A-T homozygotes are surveyed and that 10 grandmothers and 10 aunts with breast cancer are found. Assuming that $q = .01$, then the null probabilities of heterozygosity for the grandmothers and aunts are, respectively, $(1 + q)/2 = 0.5050$ and $2/(4 - q) = 0.5013$ (see table 1, part C). Hence $\bar{\theta}_0 = .5031$ and $\text{Var}(\bar{R}|H_0) = .0125$.

If $\sum_{i=1}^{20} R_i = 16$ of these $N = 20$ female breast cancer cases are found to be heterozygous for the A-T gene, then $\bar{R} = 16/20 = .80$. The score statistic value is $(.80 - .5031)/(.0125)^{1/2} = 2.66$ ($P \doteq .004$), indicating a highly significant association for these data.

From expression (1), the maximum-likelihood estimate $\hat{\phi}$ of the population odds ratio ϕ is $\hat{\phi} = 3.95$. And, from expressions (2) and (3), an approximate 95% confidence interval for ϕ is (1.32, 11.82).

Comparison with an Alternative Strategy

We have compared sample-size requirements based on expression (4) to those for a study design that is the most realistic competitor to the design we have proposed. The obvious alternative design is the 1-to-M matched case-control design, where each case is a family member with the disease of interest, each set of M controls consists of M randomly chosen nondiseased blood relatives for each case, the dichotomous exposure outcome variable pertains to being heterozygous or not, and the basic matching variable pertains to family membership.

The differences between sample-size requirements based on expression (4) and those based on the standard conditional analysis of 1-to-M matched data (Breslow and Day 1987, table 7.9) are so large that adjustments for possible intrafamilial correlation effects (Liang 1985) would not alter the obvious conclusion. In particular, for $\alpha = .05$, $(1 - \beta) = .80$, and $\bar{\theta}_0 = .30$, table 3 illustrates that the N^* values are much less than the corresponding N^+ values of Breslow and Day, where the required number of 1-to-M matched case-control sets equals $N^+/(M + 1)$.

One obvious reason for the large discrepancies between the N^* and N^+ values is the statistical fact that the value of q (or, equivalently, $\bar{\theta}_0$) is assumed to be known when determining N^*, while N^+ reflects the necessity to estimate background rates.

Table 3

Sample-Size Requirements for $\alpha = .05$, $(1 - \beta) = .80$, and $\bar{\theta}_0 = .30$ Using our Proposed Method (N^*) Compared to 1-to-M Matched Data Method (N^+)

		N^+ Values		
Odds Ratio	N^* Values	$M = 1$	$M = 2$	$M = 4$
1.50	163	710	792	1,095
2.00	53	244	270	370
2.50	30	142	156	215
3.00	20	100	111	150
3.50	16	78	87	115
4.00	13	64	72	95
4.50	11	56	60	85
5.00	10	50	54	75

Note.—For a discussion of the 1-to-M matched data method, see Breslow and Day (1987, chap. 7).

Other Characteristics of this Method

Effects of Sample Stratification and of Other Risk Factors

Undetected stratification is likely to affect population-based tests of associations but not our proposed method, even though there is no explicit control group matched, for potential confounders, to the diseased relatives. In our method, the individuals with the disease D are selected from the extended families of the index individuals. Within either the maternal or paternal family of each index individual, heterozygotes and nonheterozygotes for allele a are first-, second-, and third-degree relatives who share both a high proportion of their other genes and the familial environment, including ethnic origins and socioeconomic status (SES). Alleles at other loci and environmental risk factors (unless causally related to the allele) are expected to be distributed randomly among allele a heterozygotes and nonheterozygotes in each family. In general, selection bias is unlikely because relatives with disease D are chosen without any knowledge of their status at the locus of interest.

Using the example of an association between allele a and breast cancer, consider how risk factors such as SES or parity might influence the assessment of this association. Any risk factor is either distributed randomly with respect to allele a in each family, or it is not. If the risk factor is randomly distributed with respect to allele a, the association cannot be influenced by the factor. Suppose, on the other hand, that allele a predisposes to low parity, which in turn predisposes to breast cancer. Then an association between allele

a and breast cancer will be found. This is a true association, mediated through the effect of allele a on parity. Establishing a gene-disease association does not explicate its biological mechanism, which must be determined from further studies or from information already available about how the gene acts.

A true association between allele a and breast cancer might be missed only if the allele had two opposing effects. The allele might predispose to the cancer on the basis of some cellular mechanism related to tumor growth, and also predispose to high parity, which reduces the chance that a breast cancer will develop. This unlikely situation would be detected only by comparing the parity of allele a heterozygotes to nonheterozygotes in the families.

Our proposed strategy can be adapted to test hypotheses of the form "allele a is associated with significant excess mortality (Schiliro et al. 1989) by (or enhanced survival to) age A" by selecting at random from the family of each index individual one living person slightly older than age A. If allele a leads to excess mortality, the observed proportion of allele a heterozygotes in this group will be less than expected. Hypotheses about the effect of specific alleles on mortality are, of course, of substantial biological interest. If such an effect is suspected, it is important to detect and evaluate it before evaluating hypotheses about the allele and a specific disease, since tests of the latter association could be misleading if the allele differentially affects mortality.

When there is incomplete cooperation from the selected set of relatives with disease D, the observed odds ratio will be unbiased unless the allele a influences the chance that a relative will be available to, or cooperate in, the study. Thus, if allele a is associated with a severe personality disorder, the sample of N individuals with disease D might not accurately represent the proportion of allele a heterozygotes in the population of interest. This potential source of bias must always be taken into account in testing hypotheses about alleles that predispose to mental disorders.

While the selection of index individuals can be arbitrary, it is important, since this set of individuals defines the source population. For example, if the index individuals are selected from a population in which there are two subpopulations, in one of which the allele a is not present, then the assessment of heterozygote risk will apply only to the subpopulation in which it is prevalent. As always, any association found in one population should be generalized with great care to other populations.

Another way in which the test of an association depends on the source population is illustrated by the test of the hypothesized h-*ras* breast-cancer association in a set of families in which three or more first- or second-degree relatives have had breast cancer (Hall et al. 1990). In these families the breast cancers may result from a gene or environmental factor whose effect is so pronounced that effects of alleles at the h-*ras* locus are not detectable. On the other hand, if the N' families are selected from a general population and the observed proportion of heterozygotes is compared to that expected in this sample, then a positive association of the h-*ras* allele with breast cancer might be detected.

Specificity of the Hypothesis

The particular group of relatives with disease D will specify the disease phenotype whose association with allele a is being tested. For example, if disease D is breast cancer and all relatives with breast cancer in the particular sample have had premenopausal onset, then the observed proportion who are heterozygous for allele a pertains only to the association of a with premenopausal breast cancer.

Detecting Genes That Protect Against Disease

The power of this proposed method is particularly great when compared to that of population tests of associations for disease-*protecting* effects of specific alleles, because, for population tests, the allele frequency must be determined in a very large sample of diseased persons if a significant deviation *downward* from the general population frequency is to be found (Fleiss 1981).

Indifference to Genetic Heterogeneity

It is likely that alleles at several or many different loci predispose to a specific common disease such as breast cancer. Since the statistical power of the method we propose for testing the association of a single allele with the common disease depends only on the frequency of that particular allele and the strength of the association (as measured by ϕ), this power is not affected by the presence of disease-predisposing alleles at other loci. The power of linkage analysis to detect a single locus involved in the etiology of a common disease is, on the other hand, greatly influenced by the degree of underlying genetic heterogeneity (Lander and Botstein 1986; Goldin and Gershon 1988) unless a single large pedigree is studied.

Availability of Appropriate Families

For some studies of alleles hypothesized to predis-

pose heterozygotes to common diseases, sets of families in which an allele is segregating have already been identified or can be easily identified. For example, clinical information about relatives has already been collected for about 200 families of A-T patients (Swift et al. 1990). Any hypothesis about a specific cancer that may be associated with A-T heterozygosity can be tested rapidly by collecting DNA from the blood relatives with that cancer and determining the proportion of A-T heterozygotes. Because of an excess of cancer in families of retinoblastoma patients (Strong et al. 1984), the retinoblastoma allele has been hypothesized to predispose to certain other specific cancers. This hypothesis could be tested with convincing rigor in families with retinoblastoma by the method we are proposing. Conflicting analyses of disease-predisposition associated with alpha$_1$-antitrypsin deficiency heterozygosity (Hutchinson 1988) could be definitively resolved in families of known homozygotes. On the other hand, for alleles at some loci, surveys of defined populations may be necessary to assemble the set of N' families in which the allele is segregating.

A sample of families appropriate to our method can also be constructed conveniently by beginning with a set of DNA samples, from individuals with disease D, from which allele status could be determined. An example would be a set of stored specimens from patients who have had a breast cancer. For each specimen, find a living relative and determine first if that relative is heterozygous for the allele of interest. For each of the N relatives found to be heterozygous, then, and only then, test the allele status of the corresponding breast-cancer relative, and compare the observed to the expected proportion as described above.

Availability of DNA Samples

When assessing the heterozygosity status of the relatives with disease D, blood is the most convenient source of DNA. For diseases with a high mortality rate, it will be possible to use formalin-fixed tissue from surgical or autopsy specimens. Although DNA in such specimens is often too degraded for Southern blots (Dubeau et al. 1986), the presence of a specific allele can usually be determined with the polymerase chain reaction (PCR) (Burmer et al. 1989). This determination should be done on noncancerous tissue whenever possible, since there may be allele loss in cancers.

Conclusion

When testing hypotheses about the association of specific alleles with common chronic diseases, results using this statistically powerful method are unlikely to be influenced by confounders, systematic bias, or genetic heterogeneity.

Acknowledgments

This work was supported by National Institutes of Health grants CA14235 and HD03110. We thank Daniel Weeks, Robert Elston, Jean MacCluer, Barry Margolin, and Daphne Morrell for reading an earlier version of this manuscript and for their many helpful suggestions.

References

Breslow NE, Day NE (1987) Statistical methods in cancer research, vol 2: The design and analysis of cohort studies. International Agency for Research on Cancer, no 82. Oxford University Press, Oxford

Burmer GC, Rabinovitch RS, Loeb LA (1989) Analysis of c-Ki-*ras* mutations in human colon carcinoma by cell sorting, polymerase chain reaction, and DNA sequencing. Cancer Res 49:2141–2146

Campbell MA, Elston RC (1971) Relatives of probands: models for preliminary genetic analysis. Ann Hum Genet 35:225–236

Cooper DN, Clayton JF (1988) DNA polymorphism and the study of disease associations. Hum Genet 78:299–312

Dubeau L, Chandler LA, Gralow JR, Nichols PW, Jones PA (1986) Southern blot analysis from DNA extracted from formalin-fixed pathology specimens. Cancer Res 46:2964–2969

Ebringer RW (1980) HLA-B27 and the link with rheumatic diseases: recent developments. Clin Sci 59:405–410

Fleiss JL (1981) Statistical methods for rates and proportions, 2d ed. Wiley, New York

Gatti RA, Berkel I, Boder E, Braedt G, Charmley P, Concannon P, Ersoy F, Foroud T, Jaspers NGJ, Lange K, Lathrop GM, Leppert M, Nakamura Y, O'Connell P, Paterson M, Salser W, Sanal O, Silver J, Sparkes RS, Susi E, Weeks DE, Wei S, White R, Yoder F (1988) Localization of an ataxia-telangiectasia gene to chromosome 11q22-23. Nature 336:577–580

Gerhard DS, Dracopoli NC, Bales SJ, Houghton AN, Watkins P, Payne CE, Greene MH, Housman DE (1987) Evidence against Ha-ras-1 involvement in sporadic and familial melanoma. Nature 325:73–75

Goldin LR, Gershon ES (1988) Power of the affected-sib-pair method for heterogeneous disorders. Genet Epidemiol 5:35–42

Hall J, Huey B, Morrow J, Newman B, Carter C, Buehring T, King MC (1990) Rare HRAS alleles and susceptibility to human breast cancer. Genomics 6:188–191

Hutchison DCS (1988) Natural history of alpha-1-protease inhibitor deficiency. Am J Med 84:3–12

Krontiris TG, DiMartino NA, Colb M, Parkinson DR (1985) Unique allelic restriction fragments of the human Ha-*ras* locus in leukocyte and tumor DNAs of cancer patients. Nature 313:369–374

Lander ES, Botstein D (1986) Strategies for studying heterogeneous genetic traits in humans by using a linkage map of restriction fragment length polymorphisms. Proc Natl Acad Sci USA 83:7353–7357

Li CC, Sacks L (1954) The derivation of joint distribution and correlation between relatives by the use of stochastic matrices. Biometrics 10:347–360

Liang KY (1985) Odds ratio inference with dependent data. Biometrika 72:678–682

Schiliro G, Li Volti S, Marino S, Dibenedetto SP, Samperi P, Testa R, Mollica F (1989) Increase with age in the prevalence of β-thalassemia trait among Sicilians. N Engl J Med 321:762

Strong LC, Herson J, Haas C, Elder K, Chakraborty R, Weiss KM, Majumder P (1984) Cancer mortality in relatives of retinoblastoma patients. J Natl Cancer Inst 73:303–311

Swift M, Chase CL, Morrell D (1990) Cancer predisposition of ataxia-telangiectasia heterozygotes. Cancer Genet Cytogenet 46:21–27

Swift M, Reitnauer PJ, Morrell D, Chase CL (1987) Breast and other cancers in families with ataxia-telangiectasia. N Engl J Med 316:1289–1294

Woolf B (1955) On estimating the relationship between blood group and disease. Ann Hum Genet 19:251–253