Am. J. Hum. Genet. 68:457–465, 2001

Problems in the Definition, Interpretation, and Evaluation of Genetic Heterogeneity

Alice S. Whittemore and Jerry Halpern

Stanford University School of Medicine, Department of Health Research and Policy, Stanford

Suppose that we wish to classify families with multiple cases of disease into one of three categories: those that segregate mutations of a gene of interest, those which segregate mutations of other genes, and those whose disease is due to nonhereditary factors or chance. Among families in the first two categories (the *hereditary* families), we wish to estimate the proportion, *p,* of families that segregate mutations of the gene of interest. Although this proportion is a commonly accepted concept, it is well defined only with an unambiguous definition of "family." Even then, extraneous factors such as family sizes and structures can cause *p* to vary across different populations and, within a population, to be estimated differently by different studies. Restrictive assumptions about the disease are needed, in order to avoid this undesirable variation. The assumptions require that mutations of all disease-causing genes (i) have no effect on family size, (ii) have very low frequencies, and (iii) have penetrances that satisfy certain constraints. Despite the unverifiability of these assumptions, linkage studies often invoke them to estimate *p,* using the admixture likelihood introduced by Smith and discussed by Ott. We argue against this common practice, because (1) it also requires the stronger assumption of equal penetrances for all etiologically relevant genes; (2) even if all assumptions are met, estimates of *p* are sensitive to misspecification of the unknown phenocopy rate; (3) even if all the necessary assumptions are met and the phenocopy rate is correctly specified, estimates of *p* that are obtained by linkage programs such as HOMOG and GENEHUNTER are based on the wrong likelihood and therefore are biased in the presence of phenocopies. We show how to correct these estimates; but, nevertheless, we do not recommend the use of parametric heterogeneity models in linkage analysis, even merely as a tool for increasing the statistical power to detect linkage. This is because the assumptions required by these models cannot be verified, and their violation could actually decrease power. Instead, we suggest that estimation of *p* be postponed until the relevant genes have been identified. Then their frequencies and penetrances can be estimated on the basis of population-based samples and can be used to obtain more-robust estimates of *p* for specific populations.

Introduction

For a given hereditary disease, we need to know whether some families segregate a disease-causing mutation of one gene whereas other families segregate mutations of other genes—or whether all hereditary cases of the disease are due to mutations of a single gene. In the case of multiple genes, we also need to know the contribution of each gene to the total hereditary-disease burden.

However, there are difficulties in the definition and interpretation of the proportion, *p,* of hereditary families that segregate mutations of a particular gene. (We call a family with multiple cases of disease *hereditary* if its disease is due to heritable genetic mutations.) Specifically, for a population of interest, "family" must be defined in a way that permits enumeration of the families in that population; for example, the definition that allows both nuclear families (parents and at least one child) and two-generation families (parents, children and their spouses, and grandchildren) would have to avoid double counting of the nuclear families within the two-generation families. Even then, there are problems in the interpretation of *p*. One problem is that the probability that a family segregates a mutation of the gene of interest depends on the family's pedigree structure and phenotype. It depends on pedigree structure through the number of founders: families with many founders are more likely to segregate mutations than are families with just two founders. It also could depend on pedigree structure through the number of nonfounders, if the mutation affects fertility. The probability of segregation of a mutation depends on family phenotype: when the disease-causing genes have different penetrances, hereditary families with few affected members are more likely to segregate mutations of the less penetrant genes than are families with many affected members. Similarly, for quantitative traits, families with high

Received September 7, 2000; accepted for publication November 29, 2000; electronically published January 19, 2001.

Address for correspondence and reprints: Dr. Alice S. Whittemore, Stanford University School of Medicine, Department of Health Research and Policy, Redwood Building, Room T204, Stanford, CA 94305-5405. E-mail: alicesw@leland.stanford.edu

0002-9297/2001/6802-0016$02.00

mean values of the trait may segregate mutations of genes that are different than those in families with lower mean values segregate mutations. Thus, any measure of heterogeneity, such as the parameter α introduced by Smith (1963), varies across types of families, and consequently it varies across populations with different types of families. A further problem is that, for common mutations, a family may segregate mutations of more than one gene. In this situation it is not clear which gene(s) causes disease occurrence in the family.

Here we show that interpretation of p, independently of the structures and phenotypes of the families in a population, is possible only for diseases caused by mutations of very low population frequency that do not affect family size and whose penetrances satisfy certain conditions. For diseases that do not satisfy these assumptions, p varies across populations with different types of families. Nevertheless, for any given population, samples of families recruited according to a well-defined ascertainment scheme could be used to apportion responsibility among the relevant genes after they have been identified and after the frequencies and penetrances of their mutations have been estimated.

When the relevant genes are not known, their contributions are often estimated on the basis of linkage data, by means of Smith's heterogeneity parameter α . This parameter is commonly interpreted as the proportion of families with linkage to (i.e., that segregate mutations of) the gene of interest. Indeed, during the 54-mo period between January 1, 1996, and June 30, 2000, Smith's heterogeneity analysis was used to analyze linkage data reported by 64 papers (an average of more than one paper per month) published in the *Journal*. We show that heterogeneity assessed on the basis of linkage data before the genes are known requires not only the assumptions described above but also the stronger assumption that all mutations of all relevant genes be equally penetrant. It also requires correct specification of disease probabilities among those who carry no mutation (the phenocopy rate). We use simulations to demonstrate that heterogeneity estimates are sensitive to misspecification of this rate, the values of which cannot be deduced from linkage data. We also show that, even when the assumptions are valid and the penetrances and phenocopy rate are correctly specified, commonly used methods for the estimation of p are based on the wrong likelihood function and thus are biased in the presence of phenocopies. We show how to correct these estimates; but, nevertheless, we conclude that, for the complex diseases facing geneticists today, assessment of the relative contributions of multiple genes should be postponed until the genes have been identified and their frequencies and penetrances have been estimated on the basis of population-based data.

Assumptions Needed for Interpretation of Heterogeneity

We wish to quantify the probability that a family with multiple cases of disease segregates a disease-predisposing mutation of a gene of interest, hereafter called "gene 1." We want to accommodate the possibility that, instead, the family may either (*a*) segregate a mutation at one of other unlinked disease genes, collectively called "gene 2," or (*b*) contain multiple cases of disease because of other, nonhereditary reasons. In particular, we wish to determine the probability that family i segregates a mutation of gene 1, given the family's pedigree structure S_i and phenotype Φ_i and given that the family is hereditary—that is, the family segregated a disease-predisposing mutation of some gene. (By a family's *pedigree structure* we mean its genealogy, in the sense used by Thompson (1986, section 2.2). By a family's *phenotype* we mean the disease data for that family's members, where the data might be a binary indicator for disease occurrence, a censored time to disease variable, or a quantitative-trait measurement). Let G_k denote the event that the family segregates a mutation of gene k, and let

$$\pi_{ki} = P(G_k|\Phi_i,S_i) \tag{1}$$

denote the probability of this event, given the family's structure and phenotype, $k = 1,2$. In particular, π_{1i} is the probability that the family segregates a mutation of gene 1, given its structure and phenotype. In a linkage study, π_{1i} is the probability that the family's alleles at markers linked to gene 1 segregate with the disease. Let

$$\gamma_i = P(G_1 \cup G_2|\Phi_i,S_i) \tag{2}$$

denote the probability that the disease in the family is hereditary; values $\gamma_i < 1$ occur if, in some families, the disease is due to nonhereditary factors or chance. We wish to know the value of

$$\frac{\pi_{1i}}{\gamma_i} = P(G_1|G_1 \cup G_2,\Phi_i,S_i) \ , \tag{3}$$

which is the probability that the family segregates a mutation of gene 1, given that the disease in the family is hereditary. In general, all of these probabilities depend on the family's pedigree structure S_i and phenotype Φ_i in complex and largely unknown ways. The dependence is simpler under the following assumptions.

ASSUMPTION A.1: *The probabilities* $P(G_1|S_i)$ *and* $P(G_2|S_i$ *depend on family structure* S_i *only through the number* n_i *of founders.*

This assumption would be violated if, for example, mutations of one of the genes altered fertility (for further discussion, see Janssen et al. 1997).

ASSUMPTION A.2: *The frequencies of predisposing mutations of all genes are low.*

Letting q_k denote the frequency of mutations of gene k, $k = 1,2$, we can express this assumption as $q_k \ll 1$ for $k = 1,2$. Assumption A.2 implies that terms of order q_1^2, q_2^2, and $q_1 q_2$ are negligible. Accordingly, we shall ignore such terms whenever we use assumption A.2. Assumption A.2 also implies that there is negligible probability that a family segregates a mutation of more than one gene. Thus we write the probability (2) as $\gamma_i = \pi_{1i} + \pi_{2i}$. Substitution of this relation into equation (3) shows that the probability that a family with hereditary disease segregates a mutation of gene 1 is

$$\frac{\pi_{1i}}{\gamma_i} = \frac{\pi_{1i}}{\pi_{1i} + \pi_{2i}} \ . \tag{4}$$

Assumption A.2 also implies negligible probability that more than one family founder carries a mutation—or that a founder is homozygous for a mutation. Thus, the probability that a founder carries a mutation of gene k is $2q_k$. Using Assumption A.1, we can write the probability that a family with structure S_i segregates a mutation of gene k as

$$P(G_k|S_i) = 1 - (1 - 2q_k)^{n_i} = 2n_i q_k, \ k = 1,2 \ , \tag{5}$$

where the equal signs ignore terms of quadratic order in q_1 and q_2. Probability (5) gives

$$\frac{P(G_1|S_i)}{P(G_2|S_i)} = \frac{q_1}{q_2} \ . \tag{6}$$

Assumption A.2, although plausible for rare Mendelian diseases, is unlikely to hold for diseases caused by common low-penetrance polymorphisms.

To evaluate the relative contributions of genes 1 and 2 to the disease, we also need to consider their penetrances. To do so, we let $\rho_i = [P(\Phi_i|G_1,S_i)]/[P(\Phi_i|G_2,S_i)]$ denote the probability of the ith family's phenotype, given that it segregates a mutation of gene 1, divided by the corresponding probability, given that the family segregates a mutation of gene 2. Then, from equation (1), the relative probability that this family segregates a mutation of gene 1 and not of gene 2 is

$$\frac{\pi_{1i}}{\pi_{2i}} = \frac{P(G_1|\Phi_i,S_i)}{P(G_2|\Phi_i,S_i)} = \frac{\rho_i P(G_1|S_i)}{P(G_2|S_i)} \ . \tag{7}$$

Substitution of the right side of relation (6) into equation (7) gives

$$\frac{\pi_{1i}}{\pi_{2i}} = \frac{\rho_i q_1}{q_2} \ . \tag{8}$$

From equations (4) and (8) we find that the probability that the family segregates a mutation of gene 1, given that it is hereditary, is

$$\frac{\pi_{1i}}{\gamma_i} = \frac{\rho_i q_1}{\rho_i q_1 + q_2} \ . \tag{9}$$

In some applications it is reasonable to make the following assumption:

ASSUMPTION A.3: *The phenotype probability ratios* $\rho_i \equiv \rho$ *are constant, independent of pedigree structure and phenotype.*

In this case, we can use equation (9) to define the fraction of all hereditary families that segregate mutations of gene 1, as

$$p = \frac{\pi_{1i}}{\gamma_i} = \frac{\rho q_1}{\rho q_1 + q_2} \ . \tag{10}$$

This equation shows that, when assumptions A.1–A.3 hold, the proportion p does not vary with the family index i through the family's structure or phenotype—and, thus, that values for p are comparable across populations involving different types of families. Specifically, p represents the fraction of all heritable familial aggregation due to the gene of interest. The value $p = 1$ corresponds to no heterogeneity (i.e., all families whose disease is hereditary can be explained by gene 1), and the value $p = 0$ corresponds to no etiological role for gene 1. Moreover, within a population, estimates of p are comparable across studies involving different types of families.

Multiplication of both sides of equation (10) by γ_i gives

$$\pi_{1i} = p\gamma_i \ . \tag{11}$$

When $\gamma_i = 1$ for all family phenotypes and structures—that is, when the disease in all multiple-case families is hereditary—relation (11) becomes $\pi_{1i} = p$ for all multiple-case families, regardless of structure or phenotype.

When all the etiologically relevant genes and their disease-causing mutations have been identified, population-based studies can be used to estimate mutation frequencies for the genes, and epidemiological studies of disease risks in carriers can be used to estimate mutation penetrances. Then assumptions A.1–A.3 would not be needed, because these quantities would, for hereditary families, allow estimation of the proportion of families, of any given structure and phenotype, that segregate mutations of gene 1. They also would allow estimation of several other useful measures of the impact that the gene has on the disease burden of the population, such as (*a*) the fraction of all diseased cases that

carry its mutations and (*b*) the proportion of diseased cases that could have been avoided if no one in the population carried a mutation of the gene.

Problems in Estimating *p* from Linkage Data

The method that Smith (1963) used to estimate α, the proportion of linked families, in a particular linkage study was developed for diseases for which assumptions A.1 and A.2 are reasonable and for which the disease in all multiple-case families is hereditary (i.e., $\gamma_i = 1$ for all *i*). The method also makes, implicitly, an assumption stronger than A.3, an assumption that we designate as A.3′:

ASSUMPTION A.3′: *The penetrances of mutations of genes 1 and 2 are equal.*

This assumption means that carriers of mutations of genes 1 and 2 have either the same distribution of trait values, for quantitative traits, or the same age-specific incidence rates of disease, for qualitative traits. Clearly, this restrictive assumption is unverifiable until the genes have been identified.

Appendix A describes the fitting of parametric heterogeneity models to linkage data for qualitative diseases that satisfy Assumptions A.1–A.3′. However even when these assumptions are met, estimates of *p* are sensitive to misspecification of the unknown phenocopy rate. Specifically, when the phenocopy rate is overestimated, too many families without linkage are attributed to nonhereditary factors rather than to other genes, so *p* is overestimated; conversely, when the phenocopy rate is underestimated, then too many unlinked families are attributed to other genes rather than to nongenetic factors, and *p* is underestimated.

To illustrate this sensitivity, we simulated linkage data for families with the structures and phenotypes shown in figure 1. Details of the simulations can be found in Appendix C. The results are shown in table 1. As shown in the first and third rows of table 1, the estimates of *p* are too large when a true phenocopy rate of 0 is misspecified as 6%. As noted above, this overestimation occurs because some of the families that segregate mutations of gene 2 are incorrectly attributed to nonhereditary factors. Conversely, rows two and four of the table show that too small a specified phenocopy rate produces an underestimate of *p*. This occurs because some of the families whose disease is due to nonhereditary factors are incorrectly classified as families that segregate mutations of gene 2. Table 1 also shows that the magnitude of the bias increases with the increasing proportion of families having only two typed, affected members, because the disease in such families is more likely to be due to nonhereditary factors. Other simulations (results not shown) suggest that the estimates are less sensitive to misspecification of the mutation frequency q_1.

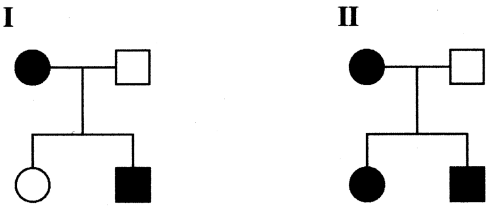

Figure 1 Family structures and phenotypes used in simulations

Table 1

Mean and SE of Estimates $\hat{p}$ of *p* with Incorrect Specification of the Phenocopy Rate f_0

True Parameter Value	Specified f_0[a]	Mean (SE) of $\hat{p}$, When Proportion of Type I Families =[b]		
		0%	20%	80%
$p = .50$:				
$f_0 = 0$	.06	.69 (.05)[c]	.71 (.05)	.85 (.01)
$f_0 = .06$	0	.37 (.02)	.34 (.02)	.24 (.03)
$p = .09$:				
$f_0 = 0$	.06	.12 (.05)	.12 (.06)	.13 (.01)
$f_0 = .06$	0	.07 (.03)	.07 (.03)	.07 (.03)

[a] Based on 400 replications of data simulated for 4,000 families with structures and phenotypes shown in figure 1, under the assumptions of dominant, completely penetrant mutations of genes 1 and 2, which have frequencies of $q_1 = .003$ and either $q_2 = .003$ ($p = .5$) or $q_2 = .03$ ($p = .09$), and phenocopy rate f_0. The diallelic marker is linked to gene 1 at $\theta = .01$.

[b] As described in figure 1.

[c] 100 × SE.

Finally, we show in Appendix B that, even if assumptions A.1, A.2, and A.3′ are valid and a positive phenocopy rate is correctly specified, the estimate $\hat{p}_H$ for *p*, an estimate obtained by the software programs HOMOG (Ott 1983, 1996) and GENEHUNTER (Kruglyak et al. 1996), is based on the wrong likelihood and therefore is biased. The bias occurs when there is a positive probability that (*a*) a family segregates mutations at genes other than gene 1 ($p < 1$) and (*b*) the disease in some families is due to nonhereditary factors ($\gamma_i < 1$ for some *i*). Moreover, studies using families whose phenotypes are more likely to have occurred by chance will have greater bias.

Appendix B describes a procedure for that uses either HOMOG or GENEHUNTER iteratively to obtain the correct maximum-likelihood estimate $\hat{p}$. We used simulations to evaluate the magnitude of the bias in $\hat{p}_H$ and to check the performance of the corrected estimate. We generated the data by using assumptions A.1, A.2, and A.3′ and analyzed them by using the correct values of the penetrance parameters, the mutation frequency q_1,

and the recombination fraction θ. Table 2 shows the mean and 100 times the empirical standard error (SE) of (*a*) the HOMOG estimate $\hat{p}_H$, (*b*) the iteratively corrected estimate $\hat{p}_c$, and (*c*) the maximum-likelihood estimate $\hat{p}$. The first and fifth rows of table 2 show the results when the phenocopy rate is 0. In this case the common estimate $\hat{p}_H = \hat{p} = \hat{p}_c$ performs well. However, when the phenocopy rate is 6% the estimate $\hat{p}_H$ is too large. The upward bias is small when all of the families have three typed, affected members, but it increases as the number of families containing only two such members increases.

In contrast, both the maximum-likelihood estimate and the corrected estimate are unbiased. Thus, in principle, it is possible, in the presence of a positive phenocopy rate, to obtain unbiased estimates of p by iterative use of the HOMOG/GENEHUNTER software; however, the strong assumptions and parametric heterogeneity–model specifications needed to obtain this validity should discourage one from doing so.

Table 2

Mean and SE of Three Estimates of *p*, with Correct Model Specification of the Phenocopy Rate f_0

True Parameter Value[a]	Mean (SE) for Three Estimates of p When Proportion of Type I Families =[b]		
	0%	20%	80%
p = .50:			
f_0 = 0:			
$\hat{p}$	.50 (.02)	.50 (.02)	.50 (.02)
f_0 = .06:			
$\hat{p}_H$	.53 (.05)	.54 (.06)	.61 (.15)
$\hat{p}_c$	.50 (.05)	.50 (.06)	.50 (.22)
$\hat{p}$	.50 (.05)	.50 (.06)	.50 (.19)
p = .09:			
f_0 = 0:			
$\hat{p}$	.09 (.03)	.09 (.03)	.09 (.03)
f_0 = .06:			
$\hat{p}_H$	.10 (.05)	.11 (.07)	.15 (.15)
$\hat{p}_c$	.09 (.04)	.09 (.05)	.09 (.06)
$\hat{p}$	.90 (.04)	.09 (.05)	.09 (.06)

[a] f_0 is as defined in footnote "a" to table 1. $\hat{p}_H$ is the HOMOG/GENEHUNTER estimate (reported as α in the literature) ($\hat{p}_H = \hat{p}_c = \hat{p}$ when $f_0 = 0$); $\hat{p}_c$ is the corrected estimate; and $\hat{p}$ is the estimate obtained by maximization of likelihood (A10).

[b] SE values shown (in parentheses) are 100 × SE. Type I families are as described in figure 1.

Discussion

We have examined the problem of defining the proportion p of all hereditary families that segregate mutations of a particular gene of interest. We have shown that, under certain assumptions, p is independent of the structures and phenotypes of the families in a particular population. The assumptions require that mutations of all disease-causing genes have (1) no effect on family structure, (2) very low frequencies, and (3) penetrance ratios that are independent of family structure and phenotype. These assumptions, although plausible for rare Mendelian diseases, are unlikely to hold for diseases caused by common low-penetrance polymorphisms, and, for these latter diseases, comparison of p across populations with different types of families is problematic.

Evaluating heterogeneity from linkage data before the genes have been identified requires not only the assumptions described above but also the stronger assumption of equal penetrance for all the etiologically relevant genes. Even when all these assumptions are met, there are problems. One problem is that estimates of p that are based on linkage data are sensitive to misspecification of the phenocopy rate. Another problem is that, even when the assumptions are met and the phenocopy rate is correctly specified, the estimates of p that are produced by HOMOG or GENEHUNTER are biased when disease occurrence in some families is due to nonhereditary factors (including chance). The bias is small in linkage studies involving only families whose disease is highly likely to be hereditary, but it increases as the proportion of families whose disease is nonhereditary increases. We have shown how to obtain consistent estimates of p, either by maximizing directly an appropriate likelihood function or by correcting iteratively the estimates produced by the software. Nevertheless, we do not recommend attempting to evaluate genetic heterogeneity in a linkage analysis.

It might be argued that a heterogeneity analysis with p (or α) treated merely as a meaningless nuisance parameter provides increased power to detect linkage. However, we know of no evidence that a parametric heterogeneity analysis has more power than a simple nonparametric analysis when the data violate one or more of the many assumptions needed for the former. Thus, we suggest that linkage data be analyzed by nonparametric methods. These methods might include subgroup analyses when mutations of the gene of interest are considered more likely to segregate in families with certain attributes (e.g., early ages at onset, certain races or ethnicities, or absence of other known disease-causing mutations).

In conclusion, we have shown that estimation of the proportion of multiple-case families attributable to a given gene before the gene has been identified and characterized requires strong and generally unverifiable assumptions about the behavior of all predisposing genes in relation to the disease of interest. The restrictive na-

ture of these assumptions, together with the sensitivity of estimates to model misspecification, suggests that attempts to quantify heterogeneity should be postponed until the frequencies and penetrances of mutations in the relevant genes have been characterized by use of population-based data. Then, one can use this information to evaluate the fraction of families, of any given structure and phenotype, that segregate mutations of the gene of interest. Indeed, the information can be used to evaluate the gene's impact on populations of individuals, rather than on populations of families; for example, one can estimate the *population attributable risk* for the gene, which is the fraction of the total disease burden that would be prevented if no family member carried its deleterious mutations (Miettinen 1974). This measure is based on individuals rather than on families—and thus, apart from its conceptual advantages, it may have more public-health utility than does a family-based measure.

Acknowledgments

This research was supported by National Institutes of Health grant number R35 CA47448. The authors are grateful to Joseph B. Keller and to the reviewers of earlier versions of the manuscript, for helpful comments and suggestions.

Appendix A

Parametric Heterogeneity Models for Linkage Analysis

Suppose that we have typed a sample of N unrelated families for markers in a small region containing gene 1 and that we are willing to make assumptions A.1, A.2, and A.3′ for the data. We assume that each family chosen for study is representative of all families in the population that have the same structure and phenotype. (This assumption excludes model-based ascertainment schemes, such as expansion of the pedigree until no more affected members are found).

At position t in the region (a test locus for gene 1), the likelihood of disease in the family is the probability $P(m_i|\Phi_i,S_i)$ of its marker data m_i, given its pedigree structure S_i and phenotype Φ_i and given that t is the locus of gene 1. This probability depends on the probability of the family's identity-by-descent (IBD) configuration at t, given the family's structure and disease phenotype. Let $j = 1,...,J_i$ index the J_i possible configurations for family i, and let

$$z_{ij} = P(\text{IBD} = j|S_i,\Phi_i) \tag{A1}$$

denote the probability that the family has configuration j for alleles of gene 1, given its structure and phenotype, $j = 1,...,J_i$. We assume that the family's phenotype Φ_i is independent of its marker data m_i, given the family's IBD configuration of alleles of gene 1. Then,

$$P(m_i|\Phi_i,S_i) = \sum_{j=1}^{J_i} P[m_i|\text{IBD}(t) = j,S_i]z_{ij} . \tag{A2}$$

By Bayes's rule,

$$P[m_i|\text{IBD}(t) = j,S_i] = P(m_i|S_i)\frac{g_{ij}(t)}{r_{ij}} , \tag{A3}$$

where $g_{ij}(t) = P[\text{IBD}(t) = j|m_i,S_i]$ is the probability that the family has IBD configuration j at locus t, given its marker data, and r_{ij} is the marginal probability that the family has IBD configuration j. Substitution of relation (A3) into equation (A2) gives

$$P(m_i|\Phi_i,S_i) = P(m_i|S_i)\sum_{j=1}^{J_i}\frac{g_{ij}(t)}{r_{ij}}z_{ij} . \tag{A4}$$

Under the null hypothesis of no association between the disease and gene 1, we have $z_{ij} = r_{ij}$, and, from equation (A4), we have $P(m_i|\Phi_i,S_i) = P(m_i|S_i)$. Thus, the family's LOD score at position t is the logarithm of the likelihood ratio

$$L_i(t) = \frac{P(m_i|\Phi_i,S_i)}{P(m_i|S_i)} = \sum_{j=1}^{J_i} g_{ij}(t)\frac{z_{ij}}{r_{ij}} . \tag{A5}$$

We now write the unknown probabilities z_{ij} of equation (A1) as

$$z_{ij} = \pi_{1i}z^*_{ij} + (1 - \pi_{1i})r_{ij} , \tag{A6}$$

where $z^*_{ij} = P(\text{IBD} = j|G_1,\Phi_i,S_i)$ is the family's probability of configuration j for alleles of gene 1, given that it segregates a mutation of this gene. Use of relation (11) in equation (A6) gives

$$z_{ij} = p\gamma_i z^*_{ij} + (1 - p\gamma_i)r_{ij} . \tag{A7}$$

Substitution of the right side of likelihood (A7) for z_{ij} in likelihood ratio (A5) gives the family's LOD score as the logarithm of

$$L_i(p,t) = p\gamma_i L^*_i(t) + 1 - p\gamma_i , \tag{A8}$$

where

$$L^*_i(t) = \sum_{j=1}^{J_i} g_{ij}(t)\frac{z^*_{ij}}{r_{ij}} . \tag{A9}$$

We wish to obtain joint estimates of p and the position of gene 1 by maximizing the product of the terms (A8)

over the N families. This task is more difficult when $\gamma_i < 1$ for some or all of the sampled families, and, to accomplish it, we must specify each $\gamma_i = \gamma_i(p)$ in terms of specified values for the frequencies and penetrances of genotypes of gene 1 and the unknown parameter p and then must estimate p and t as those values that maximize the product of the likelihoods (A8):

$$L(p,t) = \prod_{i=1}^{N} [p\gamma_i(p)L_i^*(t) + 1 - p\gamma_i(p)] . \quad \text{(A10)}$$

Let $G = G_1 \cup G_2$ denote the event that the disease in a family is hereditary, and let $\tilde{G}$ denote the event that disease occurrence in a family is due to nonhereditary factors (including chance). From Bayes's rule, we have

$$\gamma_i = P(G|\Phi_i,S_i) = \frac{P(G|S_i)P(\Phi_i|S_i,G)}{P(G|S_i)P(\Phi_i|G,S_i) + P(\tilde{G}|S_i)P(\Phi_i|\tilde{G},S_i)} . \quad \text{(A11)}$$

Assumption A.2 and equation (11) imply that

$$P(G|S_i) = 1 - P(\tilde{G}|S_i) = 2n_i(q_1 + q_2) = 2n_i\frac{q_1}{p} . \quad \text{(A12)}$$

Here we have assumed that $q_1 > 0$ (and thus that $p > 0$). Assumption A.3′ implies that

$$P(\Phi_i|G,S_i) = P(\Phi_i|G_1,S_i) . \quad \text{(A13)}$$

Substitution of probabilities (A12) and (A13) into equation (A11) gives

$$\gamma_i(p) = \frac{2n_iP(\Phi_i|G_1,S_i)}{2n_iP(\Phi_i|G_1,S_i) + (pq_1^{-1} - 2n_i)P(\Phi_i|\tilde{G},S_i)} , \quad q_1 > 0 .$$

When $q_1 = 0$, application of l'Hôpital's rule gives

$$\gamma_i(0) = \frac{2n_iP(\Phi_i|G_1,S_i)}{2n_iP(\Phi_i|G_1,S_i) + (q_2^{-1} - 2n_i)P(\Phi_i|\tilde{G},S_i)} .$$

The quantities n_i and $q_1 > 0$ are specified as part of the input in parametric analysis, and the quantities $P(\Phi_i|G_1,S_i)$ and $P(\Phi_i|\tilde{G},S_i)$ can be approximated in terms of specified penetrances of genotypes of gene 1. Thus, in a parametric heterogeneity model, the $\gamma_i(p)$ values are specified functions of the unknown parameter p. Note that $\gamma_i(1)$ is the probability that the family segregates a mutation of gene 1 when there is no gene 2—that is, when $q_2 = 0$ and $p = 1$. With assumptions A.1, A.2, and A.3′ and correct model specification, one can obtain consistent and asymptotically efficient estimates $\hat{p}$ by maximizing likelihood function (A10).

Appendix B

Correction of the HOMOG/GENEHUNTER Estimate $\hat{p}_H$

For a given gene 1 locus t, the estimate $\hat{p}_H$ for p, produced by HOMOG and GENEHUNTER, maximizes the function

$$L_H(p,t) = \prod_{i=1}^{N} [pL_i^{(1)}(t) + 1 - p] , \quad \text{(B1)}$$

where

$$L_i^{(1)}(t) = \gamma_i(1)L_i^*(t) + 1 - \gamma_i(1) \quad \text{(B2)}$$

and the $L_i^*(t)$ values are given by likelihood (A9). The logarithms of $L_i^{(1)}(t)$ are the LOD scores produced by any of several standard software programs for parametric linkage analysis. Substitution of the likelihood (B2) into the right side of equation (B1) gives

$$L_H(p,t) = \prod_{i=1}^{N} [p\gamma_i(1)L_i^*(t) + 1 - p\gamma_i(1)] . \quad \text{(B3)}$$

The $\hat{p}_H$ value that maximizes this function is reported as "alpha" in all published studies that use either HOMOG or GENEHUNTER to calculate heterogeneity LOD scores.

Comparison of likelihood function (B3) versus the correct likelihood function (A10) shows that the two are equal if and only if $\gamma_i(p) = \gamma_i(1)$—that is, if and only if the ratio $\gamma_i(p)/\gamma_i(1)$ is 1, $i = 1,...,N$. On the basis of equation (A14) this ratio is

$$\lambda_i(p) = \frac{\gamma_i(p)}{\gamma_i(1)} = \frac{2n_iP(\Phi_i|G_1,S_i) + (q_1^{-1} - 2n_i)P(\Phi_i|\tilde{G},S_i)}{2n_iP(\Phi_i|G_1,S_i) + (pq_1^{-1} - 2n_i)P(\Phi_i|\tilde{G},S_i)} \geq 1 .$$

The inequality on the right holds because p in the denominator has a positive coefficient and because $0 < p \leq 1$. Equation (B4) shows that $\lambda_i = 1$ if and only if either $p = 1$ or $P(\Phi_i|\tilde{G},S_i) = 0$ (i.e., $\gamma_i = 1$). In conclusion, when linkage in some families is due to genes other than gene 1 and in other families is due to nonhereditary factors, then the estimate $\hat{p}_H$ is based on the wrong likelihood function and therefore is biased.

One can use either HOMOG or GENEHUNTER iteratively to obtain the correct maximum-likelihood estimate $\hat{p}$, provided that the correct model is used to

calculate the functions $\gamma_i(\cdot)$ of equation (A14) for the family structures and phenotypes in the data. To describe this iterative procedure, let $\hat{p}^{(1)} = \hat{p}_H$ denote the usual estimate obtained by use, in either HOMOG or GENEHUNTER, of the values $L_i^{(1)}(t)$ of likelihood (B2).

1. Given an estimate $\hat{p}^{(k)}$, compute

$$L_i^{(k+1)}(t) = 1 + \frac{\gamma_i(\hat{p}^{(k)})}{\gamma_i(1)}[L_i^{(1)}(t) - 1] \ . \qquad \text{(B5)}$$

2. Use the values $L_i^{(k+1)}(t)$ in place of $L_i^{(1)}(t)$ in either HOMOG or GENEHUNTER to obtain a new estimate $\hat{p}^{(k+1)}$for p.

Repeat steps 1 and 2 until successive estimates do not change appreciably, and denote by $\hat{p}_c$ the final estimate.

Notice that, if the phenocopy rate is specified as 0, then $\gamma_i(p) = \gamma_i(1) = 1$ for all i. In this case, equation (B5) shows that $L_i^{(k+1)}(t) = L_i^{(1)}(t)$, $k = 1,2,...$, and the corrected estimate $\hat{p}_c$ equals the original estimate $\hat{p}^{(1)} = \hat{p}_H$ produced by the software.

Appendix C

Simulations

We simulated linkage data satisfying assumptions A.1–A.3′, using various true values of p. Specifically, for each of 400 replications, we generated data at a single diallelic marker for $N = 4{,}000$ nuclear families, each with two offspring and each with one affected and one unaffected parent, as shown in figure 1. (We used an unrealistically large number of families to ensure a small SE in the estimates, so that their bias could be seen more clearly.) The marker was assumed to be linked to gene 1, at recombination fraction $\theta < \frac{1}{2}$, and to be unlinked to gene 2. We assumed a binary disease outcome and chose N_1 of the families to have phenotype $\Phi = I$ (one offspring affected), with the remaining N_2 families having phenotype $\Phi = II$ (both offspring affected), as shown in figure 1. We assigned marker genotype AB to the affected parent and marker genotype BB to the unaffected parent, in each family. For these parental genotypes, there are four possible pairs of offspring marker genotypes: AB,AB; AB,BB; BB,AB; and BB,BB. We generated one of these pairs for each family of type ℓ, $\ell = 1,2$, as a multinomial variate with probabilities

$$\begin{aligned} P(\text{AB,AB}|\text{parents' markers}, \Phi_\ell,\theta) &= P(\text{BB,BB}|\text{parents' markers}, \Phi_\ell,\theta) \\ &= \frac{1}{2}\varphi_\ell(\theta) \ , \end{aligned}$$

and

$$\begin{aligned} P(\text{AB,BB}|\text{parents' markers}, \Phi_\ell,\theta) &= P(\text{BB,AB}|\text{parents' markers}, \Phi_\ell,\theta) \\ &= \frac{1}{2}[1 - \varphi_\ell(\theta)] \ . \end{aligned}$$

Here the probabilities $\varphi_1(\theta)$ and $\varphi_2(\theta)$ are calculated under the assumption of a dominant model with $f_2 = f_1 = 1$ and with specified values for the phenocopy rate f_0 and for the mutation frequencies—q_1 and q_2—of the two genes. Specifically,

$$\varphi_1(\theta) = \frac{q_1 f_0[\theta^2 + (1-\theta)^2] + 2q_1\theta(1-\theta) + \frac{1}{2}q_2(1+f_0) + f_0^2(1-4q_1-4q_2)}{(q_1+q_2)(1+f_0) + 2f_0^2(1-4q_1-4q_2)} \ ,$$

and

$$\varphi_2(\theta) = \frac{q_1(1+f_0^2)[\theta^2 + (1-\theta)^2] + 4q_1 f_0\theta(1-\theta) + \frac{1}{2}q_2(1+f_0)^2 + f_0^3(1-4q_1-4q_2)}{(q_1+q_2)(1+f_0)^2 + 2f_0^3(1-4q_1-4q_2)} \ .$$

The estimates $\hat{p}$ were obtained by maximization of the likelihood (A10), with $\gamma_i(p) = \gamma_\ell(p)$, when family i was of type ℓ, $\ell = 1,2$, and $\gamma_\ell(p)$ was approximated by $\gamma_\ell(p) = 4/[4 + (pq_1^{-1} - 4)c_\ell]$. Here, $c_1 = 8f_0^2/[1+f_0]$ and $c_1 = 8f_0^3/[1+f_0]^2$. These expressions are approximate because they use the penetrance of the normal genotype at

gene 1 (which represents disease risk due not only to nonhereditary factors but also to the rare mutations of gene 2) to calculate the phenotype probabilities $P(\Phi_i|\tilde{G},S_i)$ of families whose disease is due only to nonhereditary factors.

References

Janssen B, Halley D, Sandkuijl L (1997) Linkage analysis under gene heterogeneity: behavior of the A-test in complex analyses. Hum Hered 47:223–233

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

Miettinen OS (1974) Proportion of disease caused or prevented by a given exposure, trait or intervention. Am J Epidemiol 99:325–332

Ott J (1983) Linkage analysis and family classification under heterogeneity. Ann Hum Genet 47:311–320

——— (1996) Analysis of human genetic linkage, rev ed. Johns Hopkins University Press, Baltimore

Smith CAB (1963) Testing for heterogeneity of recombination fraction values in human genetics. Ann Hum Genet 27: 175–182

Thompson EA (1986) Pedigree analysis in human genetics. Johns Hopkins University Press, Baltimore