Genotype Relative Risks: Methods for Design and Analysis of Candidate-Gene Association Studies

D. J. Schaid* and S. S. Sommer†

*Section of Biostatistics, Department of Health Sciences Research, and †Department of Biochemistry and Molecular Biology, Mayo Clinic/Foundation, Rochester, MN

Summary

Design and analysis methods are presented for studying the association of a candidate gene with a disease by using parental data in place of nonrelated controls. This alternative design eliminates spurious differences in allele frequencies between cases and nonrelated controls resulting from different ethnic origins and population stratification for these two groups. We present analysis methods which are based on two genetic relative risks: (1) the relative risk of disease for homozygotes with two copies of the candidate gene versus homozygotes without the candidate gene and (2) the relative risk for heterozygotes with one copy of the candidate gene versus homozygotes without the candidate gene. In addition to estimating the magnitude of these relative risks, likelihood methods allow specific hypotheses to be tested, namely, a test for overall association of the candidate gene with disease, as well as specific genetic hypotheses, such as dominant or recessive inheritance. Two likelihood methods are presented: (1) a likelihood method appropriate when Hardy-Weinberg equilibrium holds and (2) a likelihood method in which we condition on parental genotype data when Hardy-Weinberg equilibrium does not hold. The results for the relative efficiency of these two methods suggest that the conditional approach may at times be preferable, even when equilibrium holds. Sample-size and power calculations are presented for a multitiered design. The purpose of tier 1 is to detect the presence of an abnormal sequence for a postulated candidate gene among a small group of cases. The purpose of tier 2 is to test for association of the abnormal variant with disease, such as by the likelihood methods presented. The purpose of tier 3 is to confirm positive results from tier 2. Results indicate that required sample sizes are smaller when expression of disease is recessive, rather than dominant, and that, for recessive disease and large relative risks, necessary sample sizes may be feasible, even if only a small percentage of the disease can be attributed to the candidate gene.

Introduction

Linkage analysis has been demonstrated to be a powerful method for mapping genes responsible for some human diseases (i.e., Huntington disease [Gusella et al. 1983], familial adenomatous polyposis [Bodmer et al. 1987; Leppert et al. 1988], and cystic fibrosis [Beaudet et al. 1986]). In particular, these types of diseases follow Mendelian patterns of inheritance typified by segregation of a single major gene. In contrast, linkage analysis for multifactorial diseases with more complex patterns of inheritance, such as schizophrenia (Risch 1990), have limitations because (a) the assumption of a single major gene may be incorrect, (b) genetic heterogeneity and sporadic cases may be present, (c) the penetrance of susceptible genotypes may be reduced, and (d) the genetic etiology may not be well defined. Furthermore, large pedigrees with multiple affected members may be difficult to obtain.

Association studies offer an alternative approach to linkage analyses for studying the association between a
particular genetic marker and disease. For this design, the frequency of a particular marker is compared between cases (unrelated affected individuals) versus controls (unrelated unaffected individuals). If the marker is an anonymous genetic marker which is not likely to be directly causative of the disease, such as polymorphisms identified by their restriction fragment length or VNTRs, then the marker is associated with disease only if one or a few disease mutations predominate and if the disease mutation and marker alleles are in linkage disequilibrium. If linkage exists between the anonymous marker and disease loci, yet the alleles at these two loci are in linkage equilibrium, then the alleles at these loci are not associated, which emphasizes the limitation of association studies using anonymous markers.

Alternatively, mutations in a candidate gene can be found and tested for association with disease in a case-control study. To this end, rapid and efficient methods are becoming available to examine specific genes for any abnormal nucleic acid sequence. Sequence changes that alter the amino acid sequence of the protein or alter a sequence implicated in splicing or gene expression are often, but not always, of functional significance. Sobell et al. (1992, 1993) have operationally defined these changes as variations affecting protein structure or expression (VAPSEs) and describe a multitiered research strategy in which candidate genes first are examined at the nucleic-acids level to identify VAPSEs. Once VAPSEs are identified, the association of the aberrant allele with a disease is assessed by comparing the prevalence of the allele in a large group of patients and ethnically similar controls.

A crucial consideration for the design of an association study is the selection of appropriate controls. If the frequency of the sequence variant differs across various ethnic groups and if the cases and controls differ in ethnic background, then a false-positive association could result. As an alternative choice of “controls,” Rubinstein et al. (1981) and Falk and Rubinstein (1987) proposed that the two parental genes not transmitted to their diseased offspring could be used as the control sample. This would assure that both samples of the case’s genes and its matched control genes (i.e., nontransmitted parental genes) would come from the same genetic population. As a measure of association between the marker and disease, Falk and Rubinstein (1987) defined the term haplotype relative risk (HRR) as an odds ratio—the odds of the presence of a particular marker among either of the two alleles transmitted to the affected child divided by the odds that the marker is among either of the two nontransmitted alleles.

Ott (1989) further defined the statistical properties of the HRR for diseases with a recessive mode of inheritance. One of his main findings was that, when random chance is ignored, the HRR differs from unity, which implies an association, only when disequilibrium exists between the disease and marker loci and the two loci are linked. However, the HRR groups individuals who are homozygous for the marker with heterozygous subjects and computes the odds ratio contrasted to the baseline group, which does not have the marker. Hence, the HRR is implicitly a weighted average of two odds ratios: (1) an odds ratio for those homozygous for the marker versus those homozygous without the marker and (2) an odds ratio for those heterozygous for the marker versus those homozygous without the marker. Our basic premise is that, when assessing a candidate gene for association with disease, it may at times be more informative to examine each of these two odds ratios, or their corresponding relative risks, which we will refer to as genotype relative risks (GRRs).

We present two new methods for assessing the association of a candidate gene with disease. These methods depend on obtaining data for the candidate gene on both cases and their parents. We then use likelihoods appropriate for the GRRs. The first method is appropriate when Hardy-Weinberg equilibrium (HWE) holds (i.e., mating is random with respect to the candidate gene, there is no heterozygote advantage, etc.). The second approach is a conditional likelihood method appropriate when the frequency of parental mating types deviates from that predicted when HWE holds, such as when assortative mating exists. We then apply the likelihood methods for analysis of such an association study to a multitiered strategic design that allows screening for an abnormal variant in tier 1, testing for association of the abnormal variant with disease in tier 2, and a confirmatory association study in tier 3.

**Statistical Methods**

**Likelihood When HWE Holds**

For simplicity we will define only two alleles at the candidate-gene locus. Let M be the mutant disease allele, and let N be the normal allele. When considering VAPSEs, M may be the set of all abnormal variants
affecting protein structure or expression, whereas N is the predominant normal sequence or any normal neutral variant. If several types (i.e., alleles) of abnormal sequences are found, our method could be directly extended to account for additional GRRs, although sample size may limit the number of relative risks that can be reasonably estimated. We define \( p \) to be the population frequency of M and \( q \) the frequency of N. To develop the likelihood of the data, we define the following notation. Let \( D \) be the event that an individual has the disease, and define the probabilities of disease, conditional on the genotypes at the candidate-gene locus, as \( f_2 = P(D|MM), f_1 = P(D|MN), \) and \( f_0 = P(D|NN) \). Note that the subscript \( "j" \) of \( f_j \) indicates the number of copies of \( M \). We will use this subscript notation repeatedly for convenience. These conditional probabilities would be the penetrances of the candidate gene if it were the one and only disease-causing gene. We also define the relative-risk parameters \( \psi_1 = f_1/f_0 \) and \( \psi_2 = f_2/f_0 \), which measure the relative increase in disease probabilities for heterozygous MN individuals and homozygous MM individuals, respectively, when compared to that probability for homozygous NN individuals. Note that these relative risks, \( \psi_i (i = 1,2) \), measure the relative increase in the probability of disease when only the candidate gene is considered. For complex disorders caused by multiple loci, the \( \psi \) values represent the marginal values averaged over all multilocus genotypes not measured that are associated with the disorder. Under the null hypothesis of no association of the candidate gene with disease, all conditional probabilities are equal \( (f_2 = f_1 = f_0) \), which implies \( \psi_1 = \psi_2 = 1 \). The likelihood methods we develop will be used to test \( (1) \) no association of marker with disease, \( H_0: \psi_1 = \psi_2 = 1, (2) \) dominant disease expression, \( H_D: \psi_1 = \psi_2 \), and \( (3) \) recessive disease expression, \( H_R: \psi_1 = 1 \).

For this single-locus, two-allele model, there are three possible genotypes for each parent, resulting in nine possible mating types, of which six are unique when the order of the parents is ignored. These six mating types are given in table 1. For simplicity of presentation, we will initially assume that mating is at random with respect to the candidate gene and define the genotypes as \( g_0 = NN, g_1 = MN, \) and \( g_2 = MM \). This allows us to write the probability of each mating type, conditional on the child (i.e., case) having disease, as a simple function of the population gene frequency of the mutant gene and the two relative risks. This may be calculated by conditional probabilities as

\[
P(\text{mating type } i|D) = P(\text{mating type } i) \frac{P(D|\text{mating type } i)}{P(D)}
\]

(1)

\[
= P(\text{mating type } i) \frac{\sum_{j=0}^{2} f_j P(g_j|\text{mating type } i)}{f_2 p^2 + f_1 2p q + f_0 q^2}.
\]

In equation (1) we calculate \( P(g_j|\text{mating type } i) \) according to Mendelian proportions, and we calculate \( P(\text{mating type } i) \) assuming HWE. We may then write equation (1) in terms of \( \psi_1 \) and \( \psi_2 \) by dividing numerator and denominator of equation (1) by \( f_0 \), these probabilities are presented in column two of table 1. If the disease is not associated with the candidate gene, then \( \psi_1 = \psi_2 = 1 \), and these conditional probabilities in column 2 of table 1 reduce to the unconditional probabilities predicted by HWE alone (Li 1978, p. 4).

For each mating type, a diseased child may be classified into a unique category on the basis of the child’s genotype at the candidate-gene locus. This genotype category is denoted as \( x_{ij} \) in column 3 of table 1, where \( x_{ij} \) denotes the offspring category from the \( i \)th mating type that has \( j \) marker genes. Since Falk and Rubinstein (1987) and Ott (1989) compare the frequency of M on transmitted genes versus the frequency of M on the parental nontransmitted genes, we present in column 5 of table 1 the corresponding labels for nontransmitted genes.

The conditional probability of each case’s genotype, conditional on the parental mating type and the case being affected, is \( P(g_j|D, \text{mating type } i) \). By conditional probability calculations, we find

\[
P(\text{g}_j|D, \text{mating type } i) = \frac{\sum_{j=0}^{2} f_j P(g_j|\text{mating type } i)}{\sum_{j=0}^{2} f_j P(g_j|\text{mating type } i)}.
\]

(2)

Assuming \( P(g_j|\text{mating type } i) \) is determined solely by Mendelian proportions and dividing numerator and denominator by \( f_0 \), we can write the conditional probability of each \( x_{ij} \) in terms of \( \psi_1 \) and \( \psi_2 \); these probabilities are given in column 6 of table 1. The probability of each \( x_{ij} \) can be found by multiplying column 2 of table 1 times column 6. The likelihood equation is simply a multinomial probability density of dimension 10 for the 10 \( x_{ij} \) categories in table 1 (see Appendix, section A).
Table 1

Classification of Diseased Cases according to Their Genotype and Their Parent's Genotypes at the Candidate-Gene Locus

<table>
<thead>
<tr>
<th>Parental Mating Type</th>
<th>Probability of Mating Type, Given Diseased Case</th>
<th>Notation</th>
<th>Case Genotype</th>
<th>Nontransmitted Parental Genes</th>
<th>Probability of Case Genotype, Given Mating Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. MM × MM .......</td>
<td>( p^4 \frac{\psi_1}{R} )</td>
<td>( x_{12} )</td>
<td>MM</td>
<td>MM</td>
<td>1</td>
</tr>
<tr>
<td>2. MM × MN .......</td>
<td>( 4p^3q \frac{(\psi_1 + \psi_2)}{2R} )</td>
<td>( x_{22} )</td>
<td>MM</td>
<td>MN</td>
<td>( \frac{\psi_1}{\psi_1 + \psi_2} )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( x_{31} )</td>
<td>MN</td>
<td>MM</td>
<td>1</td>
</tr>
<tr>
<td>3. MM × NN .......</td>
<td>( 2p^2q^2 \frac{\psi_1}{R} )</td>
<td>( x_{42} )</td>
<td>MM</td>
<td>NN</td>
<td>( \frac{\psi_1}{\psi_1 + \psi_2} )</td>
</tr>
<tr>
<td>4. MN × MN .......</td>
<td>( 4p^3q \frac{(\psi_1 + 2\psi_2 + 1)}{4R} )</td>
<td>( x_{41} )</td>
<td>MN</td>
<td>NN</td>
<td>( \frac{2\psi_1}{\psi_2 + 2\psi_1 + 1} )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( x_{40} )</td>
<td>MN</td>
<td>MM</td>
<td>1</td>
</tr>
<tr>
<td>5. MN × NN .......</td>
<td>( 4pq \frac{(\psi_2 + 1)}{2R} )</td>
<td>( x_{51} )</td>
<td>MN</td>
<td>NN</td>
<td>( \frac{\psi_1}{\psi_1 + 1} )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( x_{50} )</td>
<td>NN</td>
<td>NN</td>
<td>1</td>
</tr>
<tr>
<td>6. NN × NN .......</td>
<td>( q^4 \frac{1}{R} )</td>
<td>( x_{60} )</td>
<td>NN</td>
<td>NN</td>
<td>1</td>
</tr>
</tbody>
</table>

Note that the likelihood depends on three parameters, \( \psi_1, \psi_2, \) and \( p \). Standard likelihood procedures can then be used both to find maximum-likelihood estimates of \( \psi_1, \psi_2, \) and \( p \) and to develop likelihood-ratio statistics for testing various hypotheses.

For large samples, the likelihood-ratio statistic for testing the hypothesis of no association between the marker and disease \( (H_0: \psi_1 = \psi_2 = 1) \) is \( \chi^2 = 2 \log[L(\hat{\psi}_1, \hat{\psi}_2, \hat{p})/L(1, 1, p)] \), which has an approximate \( \chi^2 \) distribution with 2 df. Furthermore, one can test (1) whether the two relative risks are equal, which we shall call a dominant model \( (H_D: \psi_1 = \psi_2 = \psi_1) \), with the 1-df likelihood-ratio statistic \( \chi^2_D = 2 \log[L(\hat{\psi}_1, \hat{\psi}_2, \hat{p})/L(\psi_1, \psi_1, \hat{p})] \) or (2) whether the heterozygotes are not at an elevated risk, which we shall call a recessive model \( (H_R: \psi_1 = 1) \), with the 1-df likelihood-ratio statistic \( \chi^2_R = 2 \log[L(\hat{\psi}_1, \hat{\psi}_2, \hat{p})/L(1, \psi_2, p)] \). All of these likelihood-ratio tests require maximizing the likelihood both when the parameters are restricted to their hypothesized values and when they are not restricted. This is easily performed using iterative computer methods.

The approximate power to test the various hypotheses may be determined by using the noncentral \( \chi^2 \) distribution. When there exists an association between the marker and disease, i.e., the alternative hypothesis, \( H_A: \psi_1 \neq 1 \) and \( \psi_2 \neq 1 \), the likelihood-ratio statistic has an approximate noncentral \( \chi^2 \) distribution with noncentrality parameter \( \lambda_{\text{effHW}} = n \cdot \text{effHW} \) (given in the Appendix, section A) where \( \text{effHW} \), the efficiency of the test statistic, measures how much the relative risks deviate from their hypothesized values and also considers the variance and covariance of the two relative risks, and \( n \) is the number of cases. Power increases as \( \lambda_{\text{effHW}} \) increases, which occurs when either \( n \) or \( \text{effHW} \), or both, increase. To determine sample size for a given probability of a type I error of \( \alpha \), one needs to determine the value of
\( \lambda_{HWE} \) for a noncentral \( \chi^2 \) distribution with 2 df so that power is equal to the specified value, \( \gamma \). Then determine \( ef_{HWE} \) from the equations given in the Appendix, so that

\[ n = \lambda_{HWE} / ef_{HWE}. \]

We refer to the above likelihood as the Hardy-Weinberg likelihood, or HW likelihood, since we have assumed HWE. We will next condition on parental marker information and define the resulting likelihood as conditional on parental genotypes (CPG) likelihood.

**CPG Likelihood**

If the candidate gene is associated with disease and there is assortative mating or other deviations from HWE, then our parameter estimates based on the above unconditional likelihood analysis would be biased because the mating-type probabilities would not be correct. An analysis conditional on parental mating types will result in unbiased estimates of \( \psi_1 \) and \( \psi_2 \), even if assortative mating exists. For this reason, we develop CPG likelihood methods.

When both parents are homozygous (mating types 1, 3, and 6), their offspring can have only one genotype, so the conditional genotype probabilities in column 6 of Table 1 are unity regardless of the true association between disease and candidate gene. These categories are noninformative for association when we condition on parental mating types and are excluded from analyses. The offspring from mating types 2 and 5 can be of two genotypes (i.e., binomial probability densities), the offspring from mating type 4 can be of three genotypes (i.e., trinomial probability density), and the product of these three probability densities gives the likelihood equation

\[
L(\psi_1, \psi_2) = C \frac{\psi_2}{\psi_1 + \psi_2}^{x_{22}} \frac{\psi_1}{\psi_1 + \psi_2}^{x_{23}} \frac{\psi_2}{\psi_2 + 2 \psi_1 + 1}^{x_{42}} \frac{2 \psi_1}{\psi_2 + 2 \psi_1 + 1}^{x_{41}} \frac{1}{\psi_2 + 2 \psi_1 + 1}^{x_{40}} \frac{\psi_1}{\psi_1 + 1}^{x_{51}} \frac{1}{\psi_1 + 1}^{x_{30}},
\]

where

\[
C = \begin{pmatrix} n_2 & n_4 \\ x_{22} & x_{42} & x_{40} & x_{51} \end{pmatrix},
\]

\( n_i \) = number of cases from mating type \( i \), and

\( x_{ij} \) = number of cases from mating type \( i \) with \( j \) mutant alleles.

Standard likelihood methods may then be used to estimate \( \psi_1 \) and \( \psi_2 \), their standard errors, and construct test statistics. As an alternative to iterative methods, Rao’s (1973, p. 417) efficient scores test may be used to construct a statistic to test for no association. This approach requires the first partial derivatives of the log likelihood with respect to \( \psi_1 \) and \( \psi_2 \), but instead of iteratively finding the maximum-likelihood estimates which equate these partial derivatives to zero, the unknown parameter values are simply set equal to their hypothesized values; these equations are called “efficient score equations” and are

\[
S_1 = \left( x_{21} - \frac{n_2}{2} \right) + \left( x_{41} - \frac{n_4}{2} \right) + \left( x_{51} - \frac{n_5}{2} \right),
\]

\[
S_2 = \left( x_{22} - \frac{n_2}{2} \right) + \left( x_{42} - \frac{n_4}{4} \right),
\]

which are totals of deviations between observed counts of heterozygous cases from its expected value based on Mendelian segregation \( (S_1) \), and that for homozygous MM cases \( (S_2) \). Hence, the conditional likelihood procedure measures perturbations in the frequencies of offspring genotypes among mating types 2, 4, and 5 from their Mendelian expectations. When there is no association between the mutant allele and disease, the Mendelian ratios are 1 MM:1 MN for mating type 2, 1 MM:2 MN:1 NN for mating type 4, and 1 MN:1 NN for mating type 5. When \( \psi_1 = \psi_2 = 1 \) for our dominant model, the distortion is for mating types 4 and 5. For our recessive model, the distortion is for mating types 2 and 4. The resulting score statistic is

\[
\text{score statistic} = \frac{4[S_1(n_2 + 3n_4/4) + 2S_2(n_2 + n_4/2) + S_2^2(n_2 + n_4 + n_5)]}{(n_2 + n_4 + n_3)(n_2 + 3n_4/4) - (n_2 + n_4/2)^2},
\]

which has a large sample \( \chi^2 \) distribution with 2 df. Note that now we are conditioning on \( n_2, n_4, \) and \( n_5 \) so that
power computations are conditional. Also note that this conditional score statistic is only defined when \( n_2 + n_4 \neq 0 \). That is, the genotypic frequencies for offspring from mating types 2 and 4 contribute information for both \( \psi_1 \) and \( \psi_2 \), whereas mating type 5 only contributes information for \( \psi_1 \). If \( n_2 + n_4 = 0 \), then the appropriate conditional test of association would be the simple binomial statistic for mating type 5, \( 4(x_{51} - n_5/2)^2/n_5 \), which has an approximate \( \chi^2 \) distribution with 1 df. Hence our procedure would be to use the 2-df score statistic given by equation (5) when \( n_2 + n_4 > 0 \), or use the 1-df statistic when \( n_2 + n_4 = 0 \), as long as \( n_2 > 0 \). For large samples, under \( H_2 \); \( \psi_1 \neq 1 \), \( \psi_2 \neq 1 \), the score statistic in equation (5) has a noncentral \( \chi^2 \) distribution with noncentrality parameter \( \lambda_{\text{CPG}} \) (see Appendix, section B), and this parameter depends on \( n_2, n_4, \) and \( n_5 \).

When planning a study with intent to use the CPG likelihood for analysis, the values of \( n_i \) for each of the six mating types, in particular the informative mating types, are not known. In this case, the average power of our procedure for a given total sample size is relevant. This is the sum of the conditional powers for all possible realizations of \( n_2, n_4, \) and \( n_5 \), weighted by the probabilities of each of these realizations. On the basis of the multinomial probabilities in table 1, it is possible to compute the exact average power, without depending on large sample approximations. We have written programs in the C language to compute exact average power, but we find it to be too time consuming for sample sizes larger than 50. Alternatively, we may approximate power by replacing \( n_2, n_4, \) and \( n_5 \) by their expectations (total number of cases times mating-type probabilities in table 1) in the calculation of the noncentrality parameter \( \lambda_{\text{CPG}} \), so that \( \lambda_{\text{CPG}} = n \text{eff}_{\text{CPG}} \), where \( \text{eff}_{\text{CPG}} \) depends only on the probabilities for mating types 2, 4, and 5 (see Appendix, section B). We have compared the exact average power with power approximated by a noncentral \( \chi^2 \) distribution using \( \lambda_{\text{CPG}} \) and have found that power computations were close when \( p \) was not too small (i.e., \( p > .01 \)) and when \( \psi_1 \) and \( \psi_2 \) were not too large. Hence, we use \( \text{eff}_{\text{CPG}} \) for relative efficiency comparisons.

**Missing Parental Genotype Data**

Undoubtedly it will be difficult to obtain complete data for both parents for all cases. If relatives of the missing parent(s) (i.e., other offspring, sibs, and/or parents) are available for study of the candidate gene, then it may be possible to infer the missing parent’s genotype so that the case can be uniquely classified according to table 1. The likelihood procedures are the same as above when the missing parent’s genotype can be uniquely inferred. If the missing parent’s genotype cannot be uniquely inferred, the case should not be excluded, since there may still be information available.

Here we present the general form of the appropriate likelihood when one or both parental genotypes for the candidate gene may be missing but when additional relatives of the missing parent(s) are available for evaluation of the candidate gene. Let \( I_m \) be the marker information available on the candidate gene for relatives of the mother of the \( i \)th case, and let \( I_f \) be that for the father of the \( i \)th case. The likelihood is then

\[
L = \prod_{i=1}^{n} \left( \sum_{j=1}^{6} P(\text{mating type } j | I_m, I_f, D_i) \right) \times P(g_i | \text{mating type } j D_i),
\]

where \( P(\text{mating type } j | I_m, I_f, D_i) \) can be calculated using the Elston-Stewart (1971) approach for calculating likelihoods for pedigrees (for details, see Amos et al. 1990) and where \( P(g_i | \text{mating type } j, D_i) \) is given by equation (2) and in column 6 of table 1. Maximum-likelihood procedures can then be used for estimating parameters and testing hypotheses by using the HW likelihood.

The following example illustrates how partial missing data can be included. Suppose that a case has genotype MM, one of its parents has genotype MN, and the genotype of the other parent is missing. From table 1 we observe that the case could have come from mating types 2 or 4. Without any additional family members, the contribution of this case to the likelihood equation (6) is the sum of two probabilities, \( [4p^2q^2\psi_2 / 2R] + [4p^2q^2\psi_2 / 4R] \). Furthermore, if a sib of the case has genotype NN, then mating type 2 could be excluded, and the contribution of the case to likelihood equation (6) would then be \( 4p^2q^2\psi_2 / 4R \). This example illustrates that genotype information that is partially missing at random can be accounted for and that additional family members may improve the precision of the estimates of \( \psi_1, \psi_2, \) and \( p \).

**Single-Locus Model for Attributable Risk**

When planning a study that uses either the HW- or CPG-likelihood methods incorporating parental geno-
type information, it is necessary to specify \( p, \psi_1, \) and \( \psi_2 \) for power calculations. One possibility is to specify a range of likely values for each of these parameters and to choose the sample size so that the minimum power over these ranges is the desired value of, say, 80%-90%. Reasonable values of \( \psi_1 \) and \( \psi_2 \) may be easy to specify, but values of allele frequency \( p \) may be more difficult to specify. Alternatively, it may be easier to specify a genetic model, relative risks, and the percentage of disease attributable to the candidate gene, the attributable risk (AR). AR for the candidate-gene locus is defined as \( [P_D - P(D|NN)]/P_D \), where \( P_D \) is the population lifetime prevalence of disease. After substituting \( p^2/2 + 2pqf_i + q^2f_0 \) for \( P_D \) and canceling \( f_0 \) from numerator and denominator, AR may be written as

\[
AR = \frac{p^2(\psi_2 - 1) + 2pq(\psi_1 - 1)}{p^2\psi_2 + 2pq\psi_1 + q^2}. \tag{7}
\]

If a dominant model is hypothesized such that \( \psi_1 = \psi_2 = \psi \), then we can find, by rearranging terms in equation (7), that

\[
p = 1 - \sqrt{(1-AR)(\psi - 1) - AR} \tag{8}
\]

For recessive disease, \( \psi_1 = 1 \), and

\[
p = \sqrt{\frac{AR}{(1-AR)(\psi_2 - 1)}}. \tag{9}
\]

Therefore, if we specify a model of disease expression, with \( \psi_1 \) and \( \psi_2 \) and \( AR \), then we can determine \( p \) to compute relative efficiency and sample size.

**Results**

**Relative Efficiency Comparisons**

The large-sample relative efficiency for the CPG-likelihood method relative to the HW-likelihood method represents the inverse of the ratio of sample sizes required to achieve the same power for given magnitudes of relative risks. The ratio we present is the sample size for the HW-likelihood method divided by the sample size for the CPG-likelihood method, which is computed as the ratio of the inverse of efficiencies, \( \text{eff}_{\text{CPG}}/\text{eff}_{\text{HW}} \). Hence, 100% relative efficiency implies that both methods require the same sample size to achieve the same power. Greater than 100% relative efficiency implies that the conditional likelihood method requires a smaller sample size than does the HW-likelihood method. Conversely, less than 100% relative efficiency implies that the conditional likelihood method requires a larger sample size than does the HW-likelihood method. Relative efficiency calculations are given in figure 1 for dominant and recessive models, for relative risks of 2, 4, and 8, and AR ranging from 1%-49%. For both dominant and recessive models, the relative efficiency was less than 100% for relative risks of 2, and this loss in efficiency increased with increasing AR. The relative efficiency was greater than 100% for relative risks of 8. Although the actual values of relative efficiency differed between dominant and recessive mod-
els, the general patterns were similar, as depicted in figure 1. However, these efficiency calculations are merely approximations to give insight into the relative merits of the conditional likelihood method. It should be understood that these approximations are appropriate for large sample sizes and small relative risks. These may not always represent the true relative efficiencies, especially since the noncentrality parameter for the CPG likelihood is approximated by using the expected number of informative matings. Simulations to achieve desired power may give a better description of the relative merits of these two likelihood methods, and these are presented in the following section.

**Sample-Size Calculations**

Before presenting sample sizes for association studies using HW- and CPG-likelihood methods, we present sample-size estimates for tier 1 of our multitiered strategy. The purpose of tier 1 is to screen diseased individuals to find at least one abnormal variant. This is relevant when VAPSEs are sought, since it is labor intensive to find a VAPSE, but, once found, much faster methods are available for finding known VAPSEs among the larger group of cases and their parents (Sobell et al. 1992).

On the basis of the probabilities given in table 1, the probability that a case will carry an abnormal variant (i.e., MM or MN), given that the individual has disease, can be written as \( \Theta = AR + (1 - q^2)(1 - AR) \). After substituting \( q = 1 - p \) from equations (8) and (9), the probability that a case will carry an abnormal variant, \( \Theta \), may be written as \( \Theta = AR\,\psi/(\psi - 1) \), for dominant expression of disease, and \( \Theta = AR(\psi - 2)/(\psi - 1) + 2[AR(1 - AR)/(\psi - 1)]^{\psi} \), for recessive expression of disease. The sample size of \( n \) cases required to detect at least one VAPSE with power \( \gamma \), using a binomial probability equal to \( \Theta \), is given by \( n = \log(1 - \gamma)/\log(1 - \Theta) \). When the candidate gene is not associated with disease, \( AR = 0 \) and \( \Theta = 1 - q^2 \). It can then be shown that the probability of finding at least one VAPSE among \( n \) cases is \( \alpha_1 = 1 - (1 - p)^n \). Hence, for a given \( p \), we can assess the probability of a false-positive result, \( \alpha_1 \), at tier 1.

Sample sizes rounded up to the nearest integer are plotted in figure 2, for power of 90%, for relative risks of 2, 4, 10, and 100, and for \( AR \) of 1%-25%; several example sample sizes are also given. From figure 2 it may be seen that tier 1 sample size is less than 50 when \( AR \) is at least 5% for both dominant and recessive expression of disease. When \( AR \) is 1%-3%, the sample size required for dominant expression may be much larger than that required for recessive expression. The sample sizes required for recessive disease are often feasible.

We now present results for sample-size calculations for tiers 2 and 3, the association studies. Since exact power calculations for the conditional statistic were too time consuming for even moderate sample sizes, we simulated power by sampling from a multinomial distribution with probabilities given by the product of columns 2 and 6 of table 1. As a result, this procedure gives the expected power for the conditional likelihood approach assuming HWE. A sample unit is defined as a
Table 2

Sample Sizes to Achieve 90% Power to Detect Association of Disease with Candidate-Gene Locus, with Type I Error of 5%

<table>
<thead>
<tr>
<th>LIKELIHOOD AND AR*</th>
<th>Dominant Model, ψι =</th>
<th>Recessive Model, ψί =</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2  4  10  30  100</td>
<td>2  4  10  30  100</td>
</tr>
<tr>
<td>HW:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>&gt;3,500 1,840 1,280 870 800</td>
<td>1,825 965 565 390 325</td>
</tr>
<tr>
<td>5%</td>
<td>675 385 250 195 165</td>
<td>470 215 130 90 75</td>
</tr>
<tr>
<td>10%</td>
<td>400 190 125 100 85</td>
<td>285 110 70 45 40</td>
</tr>
<tr>
<td>25%</td>
<td>195 80 50 35 25</td>
<td>175 60 30 20 15</td>
</tr>
<tr>
<td>CPG✿</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>3,005 1,405 970 730 670</td>
<td>2,255 1,210 690 485 400</td>
</tr>
<tr>
<td>5%</td>
<td>720 325 225 160 145</td>
<td>545 250 150 110 90</td>
</tr>
<tr>
<td>10%</td>
<td>395 180 115 85 70</td>
<td>320 135 85 55 45</td>
</tr>
<tr>
<td>25%</td>
<td>215 85 50 35 25</td>
<td>200 65 35 25 15</td>
</tr>
</tbody>
</table>

*Relationship between AR, ψι, ψυ, and p is given in eqn. (8) and (9).

**ψι = ψυ = ψ**

✿ ψι = 1; ψυ = ψ

dSample sizes were calculated assuming HWE to predict the frequency of informative mating types.

Multitier Design Considerations

In the previous sections we have considered power and sample-size calculations separately for each tier. A global design strategy considering all three tiers may be conducted in several ways, but we present a method to minimize the total cost of the study. Because we may likely evaluate multiple candidate genes, we wish to control the probability that at least one of them is a false-positive, the experiment-wise error rate. One simple approach is to use the Bonferroni correction; if αγ is the desired experiment-wise error rate and m is the number of candidate genes evaluated, then a given marker association is not considered statistically significant unless the p-value < α = αγ/m.

For each marker evaluated, we use the rule that a marker is claimed to be significantly associated with disease only if (1) at tier 1 the marker is detected in at least one of the ni cases, in which case we go on to tier 2, and (2) at tier 2 a “promising” association is found among n2 cases and their parents, in which case we go on to tier 3, and (3) at tier 3 a significant association is found among n3 cases and their parents. For this rule we wish to maintain the probability of the aggregate type I error at α and determine the values of ni (i = 1, 2, 3) so that the aggregate power, γ, is large, say, 80%. Let αi and γi be the probabilities of a false-positive and true-
positive finding at the $i^{\text{th}}$ tier. If cases are not used for more than a single tier, then the aggregate $\alpha$ and $\gamma$ are the product of terms across tiers, $\alpha = \Pi \alpha_i$ and $\gamma = \Pi \gamma_i$. Hence, we have flexibility in choosing $\alpha_i$ and $\gamma_i$ as long as these constraints are met. There is a trade-off between tiers in terms of sample sizes and cost. We wish to minimize cost while maintaining the desired values of $\alpha$ and $\gamma$. If $c_i$ is the cost per unit sample (i.e., time, dollars), then the total cost for a single marker is $c_T = \sum c_i n_i$. If we assume that the cost of attaining a case and both of its parents is the same for tiers 2 and 3, $c_2 = c_3$, and the cost for a tier 1 case is $R$ times the cost for a case and both of its parents for tiers 2 and 3, then $c_T$ is proportional to $R n_1 + n_2 + n_3$. We can numerically minimize this cost function by allowing $\alpha_i$ and $\gamma_i$ ($i = 1, 2, 3$) to vary, while constraining the products to be $\alpha = \Pi \alpha_i$ and $\gamma = \Pi \gamma_i$. However, when reporting results from tier 3, the scientific community has generally accepted $p$-values $<.05$ as significant, so we fix $\alpha_3 = 0.05$. Since $\alpha_i = 1 - (1 - p)^{2n_i}$ is determined when we specify allele frequency $p$ and $n_i$, $\alpha_3$ is also fixed at $\alpha_3 = \alpha / (\alpha, \alpha_3)$. Hence, only the $\gamma_i$'s are free to vary. We use the asymptotic approximations to determine sample sizes, $n_2$ and $n_3$, for the CPG-likelihood method. To illustrate this design approach, consider an example. The required parameters—and example values—are $m = 10$ candidate genes, aggregate $\alpha = 0.005$, aggregate $\gamma = 0.80$, a dominant model with $\psi_1 = \psi_2 = 10$, $AR = 0.05$, allele frequency of $p = .01$ when the null hypothesis is true, and a cost ratio of $R = 10$. Then the design with the minimum cost is: (1) for tier 1 evaluate 41 cases, and if at least one VAPSE is found, continue to tier 2 ($\alpha_1 = 0.56, \gamma_1 = 0.91$), (2) for tier 2, accrue 302 cases and their parents, and if, by the CPG-likelihood method, $p$-value $<\alpha_2 = 0.18$, continue to tier 3 ($\gamma_2 = 0.97$), and (3) for tier 3, accrue 322 cases and their parents, and if $p$-value $<\alpha_3 = 0.05$, claim a significant association ($\gamma_3 = 0.91$). If the cost ratio were $R = 100$, the values of ($n_i, \alpha_i, \gamma_i$) would be: tier 1, $n_1 = 29, \alpha_1 = 0.44, \gamma_1 = 0.80$; tier 2, $n_2 = 406, \alpha_2 = 0.23, \gamma_2 = 0.994$; and tier 3, $n_3 = 732, \alpha_3 = 0.05, \gamma_3 = 0.999$. The impact of larger values of $R$ is to reduce $n_i$, which means smaller values of $\gamma_i$, and consequently larger values of $n_2$ and $n_3$ to be sure that the aggregate power is as desired.

**Discussion**

We have presented likelihood methods for analysis of association studies using parental data for the candidate gene. When it is difficult to insure that nonrelated controls will be appropriately matched according to ethnicity with diseased cases, parental genotype data offer an unbiased method to estimate relative risks and test for association because cases are matched with their parents. The relative benefits of using parental genotype data in place of unrelated controls depend on several key issues: (1) the accuracy of ethnicity data and the effort in obtaining it when using unrelated controls; (2) the cost of obtaining appropriate controls, with unrelated controls often much easier and cheaper to obtain than parents; (3) obtaining parental genotype data may be impossible for some individual cases, with greater difficulties for older cases; and (4) the magnitude of variation of candidate-gene frequencies across ethnic groups. Unrelated controls are likely to result in either false-positive or false-negative associations when gene frequencies have large variability and when relative risks are small.

The advantages that the HW-likelihood method offers are that population gene frequencies can be estimated when the assumption of HWE holds; the HW-likelihood method tends to be more powerful than the CPG-likelihood method when relative risks are small; and genotype data for relatives of missing parents can be included in a straightforward manner to reduce loss of information. Collection of data on parents, and possibly other relatives, is desirable also for subsequent family studies, to further examine cosegregation of a candidate gene with disease, once a positive association is detected. For the HW likelihood, we used the parameter $p$ to determine mating-type probabilities. If the grandparents of cases have heterogeneous ethnic backgrounds, then the allele frequency $p$ will represent an average frequency, which may or may not be adequate for the mating-type probabilities; if mating-type probabilities are not correct, then $\psi_i$ and $\psi_j$ will be biased. For the CPG likelihood, Mendelian proportions for offspring genotypes are distorted when relative risks are not unity, so relative risks remain unbiased even if the parents have mixed ancestry, as long as the magnitudes of the relative risks do not depend on the ethnic origin of the mutation. Hence, the CPG-likelihood method is the more robust approach, when dealing with cases from widely different ethnic backgrounds. In addition, the power of the CPG-likelihood method tends to be greater than that of the HW-likelihood method when relative risks are large.

It should be understood that if the candidate gene is
not the true underlying gene responsible for disease, yet is linked to the true gene, then the relative risks depend on the recombination fraction and linkage disequilibrium between the candidate gene and true disease gene. Our likelihood methods allow estimation of relative risks and differ from the \( \chi^2 \) association methods recently proposed by Terwilliger and Ott (1992) and Spielman et al. (1993). The magnitudes of \( \psi_1 \) and \( \psi_2 \), as averages over unmeasured genotypes associated with disease, will depend on the relative contributions of the other unmeasured genes. Although, in general, the magnitudes of relative risks are difficult to specify a priori, genotypic odds ratios using nontransmitted parental alleles as controls have been reported to be as high as 60 for the association of HLA-DR antigens with insulin-dependent diabetes mellitus (Falk and Rubin-stein 1987), and Lathrop (1983) reported a genotypic odds ratio of 19, using nonrelated controls, for the association of HLA-DR antigens with rheumatoid arthritis. Hence, it may very likely be possible to observe relative risks much larger than those in the traditional case-control study.

When multiple cases are related as sibs, their inclusion in our likelihood equations is valid if the disease phenotype is dependent solely on the genotype of the candidate gene, so that sibs are conditionally independent, given parental genotypes. In this case, affected sibs can be treated as if they are nonrelated cases in the likelihood equations, when care is taken to represent the parental mating type only once for each sibship in the HW likelihood. However, if there are additional genes responsible for the disease phenotype, then the affected sibs are not independent, even when conditioning on the parental mating type for the candidate gene. This becomes problematic and requires further research on how best to include related cases.

Although it is difficult to formally test for deviations from HWE by our methods, a pragmatic approach is to compare the relative-risk estimates from both HW-likelihood and CPG-likelihood methods and to assume that gross deviations are due to violation of the assumption of HWE. An ad hoc statistical test could be constructed as follows: The likelihood of the six mating types, \( L_{6}(p, \psi_1, \psi_2) \), is a multinomial probability density with probabilities given in column two of Table 1. Let \( \psi_1^{\text{CPG}} \) and \( \psi_2^{\text{CPG}} \) be the maximum-likelihood estimates of relative-risk parameters from the CPG likelihood. The likelihood-ratio statistic we propose to test for equality of HW and CPG relative-risk estimates is 

\[
2 \log L_{6}(p, \psi_1^{\text{CPG}}, \psi_2^{\text{CPG}}) - \log L_{6}(p, \psi_1, \psi_2) \]

where \( p \), \( \psi_1 \), and \( \psi_2 \) are values that maximize the appropriate likelihoods. For large samples, the distribution of this likelihood-ratio statistic should be approximately \( \chi^2 \) with 2 df. The problem with this method is that the variability of \( \psi_1^{\text{CPG}} \) and \( \psi_2^{\text{CPG}} \) is not accounted for. Hence, one should be conservative when using this approach. Other evidence, such as assortative mating, may suggest that deviation occurs for the disease studied.

We presented in Table 2 sample-size estimates for the CPG likelihood when the frequency of informative matings is predicted by HWE. Any other method that one specifies for proportions which are not in HWE can be incorporated for sample-size calculations. The main point is that the number of informative mating types 2, 4, and 5 will determine power. If HWE is assumed for initial sample-size calculations and if large deviations are subsequently found, one can adjust sample-size estimates on the basis of observed frequencies of the six mating types.

The advantage that our likelihood method offers is that not only can statistical tests of association be conducted but formal statistical tests of Mendelian hypotheses (i.e., dominant and recessive) can also be performed. However, we view these tests as a method to assist interpretation of the relative risks for complex disorders and not as proof of simple Mendelian inheritance. Furthermore, although a study may be designed to have sufficient power to detect association, there is no guarantee that power will be sufficient to test whether \( \psi_1 = \psi_2 \) or whether \( \psi_1 = 1 \). Hence, one should be cautious when giving genetic interpretations to \( \psi_1 \) and \( \psi_2 \).

The multiterraced strategy that we propose allows an investigator to design a study with minimum cost, so that there is sufficient aggregate power to detect an abnormal candidate gene at the first tier, to determine whether the association with disease is promising in the second tier, and then to more rigorously test associations in the third tier. The significance level and power used for each tier should be considered in light of the amount of effort and sample size required for each tier. One crucial point that may need further study is the choice of aggregate type I error, \( \alpha \). In the traditional likelihood linkage analysis for two loci, a lod score 3 has been accepted as evidence for linkage. This corresponds to a \( p \)-value <.0001. This conservative approach prevents false claims of significant linkage from
occurring too often when the prior probability of linkage is approximately .05 (Ott 1991). If any one of many genes is responsible for the disease, then the prior probability that the candidate gene is truly responsible for the disease may be much less than .05, which implies that a conservative value of α be used for each candidate gene. We have taken the more traditional statistical approach and used a conservative value of α to adjust for evaluating m candidate genes. Simulation experiments may be useful in order to determine a level of α that maintains a low probability of false-positive results when there is a small prior probability that a candidate gene is truly associated with disease and when many candidate genes are to be evaluated, yet that allows feasible sample sizes.

Acknowledgments
We thank Robert Elston, Janet Sobell, and anonymous reviewers for their critical comments which improved this paper. This work was supported in part by U.S. Public Health Service training grant HL07567 from the National Heart Lung and Blood Institute, a Mayo Foundation Scholarship, and MH44276.

Appendix
Derivations of Noncentrality Parameters λ_{HW} and λ_{CPG}
We define sample size in terms of the number of cases, n. Furthermore, define I as Fisher's expected information matrix with elements $n_{11} = nE[-df \log L/ \partial \psi_1^2]$, $n_{12} = nE[-df \log L/ \partial \psi_1 \partial \psi_2]$, and $n_{22} = nE[-df \log L/ \partial \psi_2^2]$ and where L is the likelihood for either the unconditional or conditional methods. For large samples, the likelihood-ratio statistic for testing associations has an approximate $\chi^2$ distribution with 2 df and noncentrality parameter $\lambda = nB^2$, where $B^2 = (\psi_1 - 1, \psi_2 - 1)$ (Stuart and Ord 1991, pp. 870–872); we define $eff = B^2$ so that $\lambda = n eff$. We have assumed Pitman alternatives of the form $\psi_1 = 1 + c_1/\sqrt{n}$ and $\psi_2 = 1 + c_2/\sqrt{n}$, for arbitrary constants $c_1$ and $c_2$, so that information I can be approximated under $H_0$: $\psi_1 = \psi_2 = 1$. The following sections give results necessary for calculating the appropriate noncentrality parameters.

A. Hardy-Weinberg Likelihood Using Parental Genotype Data
The likelihood equation is a multinomial with probabilities found from multiplying columns 2 and 6 of table 1. It follows that $\log L = constant + a \log(p) + b \log(1 - p) + c \log(\psi_2) + d \log(\psi_3) - n \log(\psi_2p^2 + \psi_3(1 - p^2) + (1 - p^2)^2)$, where

\[ a = 4x_{12} + 3x_{22} + 3x_{32} + 2x_{31} + 2x_{42} + 2x_{41} + 2x_{40} + x_{51} + x_{50}, \]
\[ b = x_{22} + x_{31} + 2x_{42} + 2x_{41} + 2x_{40} + 3x_{51} + 3x_{50} + 4x_{60}, \]
\[ c = x_{12} + x_{22} + x_{42}, \]
\[ d = x_{21} + x_{31} + x_{41} + x_{51}, \]

and the elements of the information matrix are

\[ i_{11} = 2pq(1/\psi_1 - 2pq/r)/r, \]
\[ i_{12} = -2pq^2q/r^2, \]
\[ i_{22} = p^2(1/\psi_2 - p^2/r)/r, \]

where $r = p^2\psi_2 + 2pq\psi_1 + q^2$, and $q = 1 - p$.

B. Likelihood Conditional on Parental Mating Types
The likelihood equation is given by equation (3), and it follows that

\[ i_{11} = n_2\psi_2/[\psi_1 + \psi_2]\psi_1], \]
\[ + n_2[\psi_2 + 1]/[\psi_2 + 2\psi_1 + 1]\psi_1, \]
\[ + n_{12}/[\psi_1 + 1]^2]\psi_1, \]
\[ i_{12} = -n_2/(\psi_1 + \psi_2^2) - n_2/[\psi_1 + 2\psi_1 + 1]^2, \]
\[ i_{22} = n_2\psi_1/[\psi_1 + \psi_2]^2\psi_2], \]
\[ + n_2/[\psi_1 + 1]^2\psi_2]. \]

If we replace $n_2$, $n_4$, and $n_6$ by their expected values, $nP(matting type i)$, then $n$ can be factored out of the above equations for the conditional likelihood method, similar to that for the unconditional likelihood method.

References
Li CC (1978) First course in population genetics. Pacific Grove, CA
Risch N (1990) Genetic linkage and complex diseases, with special reference to psychiatric disorders. Genet Epidemiol 7:3–16