

Genetic Association Analysis Using Data from Triads and Unrelated Subjects

Michael P. Epstein,¹ Colin D. Veal,³ Richard C. Trembath,³ Jonathan N. W. N. Barker,⁴ Chun Li,⁵ and Glen A. Satten²

¹Department of Human Genetics, Emory University, and ²Centers for Disease Control and Prevention, Atlanta; ³Division of Medical Genetics, University of Leicester, Leicester, United Kingdom; ⁴St. John's Institute of Dermatology, King's College London, London; and ⁵Center for Human Genetics Research, Vanderbilt University, Nashville

The selection of an appropriate control sample for use in association mapping requires serious deliberation. Unrelated controls are generally easy to collect, but the resulting analyses are susceptible to spurious association arising from population stratification. Parental controls are popular, since triads comprising a case and two parents can be used in analyses that are robust to this stratification. However, parental controls are often expensive and difficult to collect. In some situations, studies may have both parental and unrelated controls available for analysis. For example, a candidate-gene study may analyze triads but may have an additional sample of unrelated controls for examination of background linkage disequilibrium in genomic regions. Also, studies may collect a sample of triads to confirm results initially found using a traditional case-control study. Initial association studies also may collect each type of control, to provide insurance against the weaknesses of the other type. In these situations, resulting samples will consist of some triads, some unrelated controls, and, possibly, some unrelated cases. Rather than analyze the triads and unrelated subjects separately, we present a likelihood-based approach for combining their information in a single combined association analysis. Our approach allows for joint analysis of data from both triad and case-control study designs. Simulations indicate that our proposed approach is more powerful than association tests that are based on each separate sample. Our approach also allows for flexible modeling and estimation of allele effects, as well as for missing parental data. We illustrate the usefulness of our approach using SNP data from a candidate-gene study of psoriasis.

Introduction

With the recent availability of high-density maps of SNPs, association studies are increasingly popular choices for identifying genetic variants that influence disease. Such studies rely on the concept of linkage disequilibrium (LD)—that is, the statistical association of alleles at two tightly linked loci (here, a SNP and a disease-influencing gene) in a population. The tight linkage required typically exists over short distances, although such distances can be variable (Abecasis et al. 2001). Nevertheless, identification of SNPs in LD with disease susceptibility variants should narrow the location of a disease-influencing gene and should facilitate positional cloning efforts.

Statistical methods for testing association between SNPs and disease (reviewed by Thomas [2004]) gen-

erally consist of a comparison of SNP alleles or genotypes from a sample of affected cases with those from an appropriate sample of unaffected controls. One common choice of controls is a collection of unrelated subjects. In this situation, one can test for association by comparing the SNP allele or genotype frequencies of unrelated cases with those of controls by use of standard goodness-of-fit statistics. An attractive feature of the use of unrelated controls is that they generally are easy to collect, which facilitates collection of a sufficient number of participants to detect genes of small effect that contribute to the risk of complex diseases.

A major issue regarding the use of unrelated cases and controls in association models is that one cannot distinguish valid association due to linkage from spurious association due to confounding effects. One of the discussed confounders in genetic association studies is population stratification, which occurs if the population from which the cases and controls were sampled consists of latent subpopulations, each with different SNP allele frequencies and risks of disease. A spurious association due to this confounding effect will occur for any SNP allele that is at an elevated frequency in the subpopulation with the greatest disease prevalence. Examples of this type of confounding have been discussed

Received October 26, 2004; accepted for publication January 27, 2005; electronically published February 14, 2005.

Address for correspondence and reprints: Dr. Michael P. Epstein, Department of Human Genetics, Emory University School of Medicine, 615 Michael Street, Suite 301, Atlanta, GA 30322. E-mail: mepstein@genetics.emory.edu

© 2005 by The American Society of Human Genetics. All rights reserved. 0002-9297/2005/7604-0007\$15.00

in case-control studies of diabetes (Knowler et al. 1988) and alcoholism (Gelernter et al. 1993). If hidden population substructure exists, one can correct the bias in statistical tests by using genotype information from unlinked markers (Devlin and Roeder 1999; Pritchard et al. 2000; Satten et al. 2001). However, the number of unlinked SNPs that is needed to properly correct for population stratification is uncertain.

To avoid spurious association from population stratification, Falk and Rubinstein (1987) recommended collecting the parents of each case subject and using the nontransmitted parental alleles as a control sample. In doing this, the cases and controls are matched in genetic ancestry and are therefore robust to population stratification. Using this idea, Spielman et al. (1993) constructed a joint test of linkage and association, called the “transmission/disequilibrium test” (TDT), that attempts to identify preferential transmission of alleles from parent to affected child within different triads (comprising an affected child plus two parents) by use of a McNemar statistic. As a generalization of the TDT, Schaid and Sommer (1993, 1994) developed a likelihood procedure for triads, called the “conditional on parental genotypes” (CPG) approach, that models the probability of an affected offspring’s genotype conditional on parental genotypes as a function of the genotype relative risks (RRs) of the offspring. This CPG likelihood approach allows flexible modeling of the genotype RRs, which can be estimated using standard maximum-likelihood procedures.

By being robust to population stratification, parental controls are ideal choices for association modeling. However, the sampling of parental controls is often more difficult and more expensive than that of unrelated controls, since studies must identify and sample the two parents of an affected subject. Many parents may not be available for analysis, because of death, refusal to participate, or false paternity. Statistical methods exist for handling triads with missing parental data (Sun et al. 1999; Weinberg 1999; Rabinowitz 2002; Allen et al. 2003), but, even if parental controls are properly handled, the difficulty in sampling them may result in samples of insufficient size for detection of genes that have small effects on disease risk.

The choice to use either unrelated or parental controls for association analysis requires serious deliberation. Situations may arise in which a study collects both parental and unrelated controls for association analysis, rather than choose one specific type of control. There are different reasons for using such a sampling design. For example, in a candidate-gene study of disease, one might use the triads to test for LD between certain SNPs and disease while using the unrelated controls to investigate patterns of LD between the SNPs, in an effort to reduce potential confounding. Veal et al. (2002) used

such a strategy in a study that investigated the influence of candidate genes within the major histocompatibility complex (MHC) on susceptibility to psoriasis. Studies might also collect a sample of triads to confirm previous association results that were found using a sample of unrelated cases and controls (Martin and Kaplan 2000), since significant replication by use of the triad sample ensures that significant association results from the unrelated sample are due to LD and not stratification. Studies might also sample both parental and unrelated controls for association analysis, since each set of controls provides insurance against the weaknesses of the other set. The parental controls would provide insurance if the use of unrelated controls yields an association-test bias due to stratification, whereas the unrelated controls would provide insurance if the number of sampled parental controls among the triads falls below anticipated numbers, which could lead to underpowered TDTs and CPG tests.

In such situations, the overall sample for association analysis will consist of triads, unrelated controls, and, perhaps, unrelated cases. Recently, Nagelkerke et al. (2004) proposed a joint analysis of such data by use of a likelihood-based approach and showed that the resulting analyses yield an increase in power, compared with methods that analyze triads and unrelated subjects separately. Nagelkerke et al. (2004) also provided ad hoc procedures to determine whether triad data and unrelated data can be safely combined. This is important because triad, control, and case data could differ by confounding factors (such as population stratification) that could potentially lead to biased inference if the data are naively combined.

Nagelkerke et al. (2004) advocated an approximate analysis that, although having the advantage of being easily conducted by use of standard logistic-regression software, makes the strong assumptions of Hardy-Weinberg equilibrium (HWE), random mating, and a multiplicative model of allele effect on disease. Here, we advocate a likelihood-based approach that modifies the approach of Nagelkerke et al., to allow for more flexible modeling of allele effects and less-restrictive assumptions about the distribution of parental mating types and genotypes. We show here that the power gains achieved by Nagelkerke et al. (2004) under the assumption of HWE and random mating are preserved when a more general parental mating-type distribution is used, whereas violation of these assumptions in the approximate procedure of Nagelkerke et al. can lead to biased inference. In addition, we also provide formal tests to determine whether data types (triads, unrelated controls, and unrelated cases) can be combined; these tests do not require genotyping at additional loci. Finally, we also consider triad samples with missing parental data. Here, a new subtlety arises: whereas

likelihood-based analyses of incomplete triad data (Weinberg 1999) are robust to population stratification because a saturated model is fit to parental genotypes, joint analysis of incomplete triads with unrelated controls can yield biased estimates of genotype RRs if population stratification exists between the two samples. We discuss how to protect against this bias as part of our procedures for testing whether data types should be combined.

In subsequent sections, we develop the likelihood approach of Nagelkerke et al. (2004) and describe our modifications to their estimation procedures and statistical tests for detection of association. We also describe our formal procedures for testing whether data types may be combined and further describe our approach for accommodating incomplete triad data. We evaluate the performance of our approach, using both simulated data and real data from a study of psoriasis.

Material and Methods

Assumptions and Notation

We assume a sample of triads, unrelated controls, and unrelated cases that are all genotyped at a SNP of interest. We denote the two alleles of the SNP as *A* and *a*. For a given triad, we define $G_p = (G_{p,1}, G_{p,2})$ as the unordered genotypes of the two parents and define G_o as the genotype of the offspring. We define G_u as the genotype of an unrelated subject. We code each genotype to equal the number of copies of *A* carried by the subject of interest. Finally, we define D_o and D_u as disease outcome variables (1 = affected and 0 = unaffected) for a triad offspring and unrelated subject, respectively. For triad offspring, we assume $D_o = 1$ on the basis of sampling design. For unrelated subjects, D_u equals 1 for affected cases and 0 for unaffected controls.

Likelihood Derivation

The combined association procedures that we consider are based on the likelihood proposed by Nagelkerke et al. (2004) for combining data on parental genotypes G_p , offspring genotypes G_o , and unrelated genotypes G_u conditional on the disease outcomes of the offspring D_o and of the unrelated subjects D_u . Assuming a sample of *I* triads, *J* unrelated controls, and *K* unrelated cases, we can write this likelihood as

$$L = \prod_{i=1}^I P(G_{pi}, G_{oi} | D_{oi} = 1) \times \prod_{j=1}^J P(G_{uj} | D_{uj} = 0) \times \prod_{k=1}^K P(G_{uk} | D_{uk} = 1), \tag{1}$$

where *i* indexes the triad probabilities, *j* indexes the control probabilities, and *k* indexes the case probabilities.

The construction of *L* requires specification of the three probabilities in equation (1). We first specify the probability of the triad genotypes in $P(G_p, G_o | D_o = 1)$. We rewrite this probability as $P(G_p, G_o | D_o = 1) = P(G_o | G_p, D_o = 1) \times P(G_p | D_o = 1)$. Here, $P(G_o | G_p, D_o = 1)$ denotes the probability of an affected offspring’s genotype conditional on parental genotypes and corresponds to the probability in the CPG approach described by Schaid and Sommer (1993). Define the RR as

$$\psi_g = \frac{P(D_o = 1 | G_o = g)}{P(D_o = 1 | G_o = 0)}, \quad g = 1, 2.$$

Schaid and Sommer (1993) showed that, if offspring disease risk is independent of parental genotype given offspring genotype, then

$$P(G_o = g | G_p = g_p, D_o = 1) = \frac{\psi_g P(G_o = g | G_p = g_p)}{\sum_{g^*} \psi_{g^*} P(G_o = g^* | G_p = g_p)},$$

where $P(G_o = g | G_p = g_p)$ is the Mendelian proportion of offspring with *g* copies of allele *A*, given parental mating type g_p . In table 1, we show $P(G_o | G_p, D_o = 1)$ values for all possible triad genotype combinations (Schaid and Sommer 1993).

To model the RR parameters, we could assume a general model that allows for arbitrary values of ψ_1 and ψ_2 . Alternatively, we could select nongeneral models for the RR, including multiplicative ($\psi_1 = \psi$ and $\psi_2 = \psi^2$), additive ($\psi_1 = \psi$ and $\psi_2 = 2\psi - 1$), dominant ($\psi_1 = \psi_2 = \psi$), and recessive ($\psi_1 = 1$ and $\psi_2 = \psi$) models. Here, ψ denotes a scalar RR parameter to be estimated.

To complete the construction of $P(G_p, G_o | D_o = 1)$ in equation (1), we next specify $P(G_p | D_o = 1)$, which is the probability of the parental genotypes in the triads. We note that the frequency of G_p among triads differs from that in the general population because of selective sampling through affected offspring. However, we can evaluate this probability as

$$P(G_p = g_p | D_o = 1) = \frac{\sum_g \psi_g P(G_o = g | G_p = g_p) P(G_p = g_p)}{\sum_{g_p^*} \sum_{g^*} \psi_{g^*} P(G_o = g^* | G_p = g_p^*) P(G_p = g_p^*)}, \tag{2}$$

where ψ_g and $P(G_o | G_p)$ are as defined above. $P(G_p)$ in equation (2) denotes the frequency of the parental genotypes in the general population. We calculate $P(G_p)$ using the parental genotype distribution described by

Table 1
Evaluation of $P(G_o|G_p, D_o = 1)$ for a SNP

| G_p and G_o | $P(G_o G_p, D_o = 1)$ |
|-----------------|--|
| $G_p = (2,2):$ | |
| $G_o = 2$ | 1 |
| $G_o = 1$ | 0 |
| $G_o = 0$ | 0 |
| $G_p = (2,1):$ | |
| $G_o = 2$ | $\frac{\psi_2}{\psi_1 + \psi_2}$ |
| $G_o = 1$ | $\frac{\psi_1}{\psi_1 + \psi_2}$ |
| $G_o = 0$ | 0 |
| $G_p = (2,0):$ | |
| $G_o = 2$ | 0 |
| $G_o = 1$ | 1 |
| $G_o = 0$ | 0 |
| $G_p = (1,1):$ | |
| $G_o = 2$ | $\frac{\psi_2}{1 + 2\psi_1 + \psi_2}$ |
| $G_o = 1$ | $\frac{2\psi_1}{1 + 2\psi_1 + \psi_2}$ |
| $G_o = 0$ | $\frac{1}{1 + 2\psi_1 + \psi_2}$ |
| $G_p = (1,0):$ | |
| $G_o = 2$ | 0 |
| $G_o = 1$ | $\frac{\psi_1}{1 + \psi_1}$ |
| $G_o = 0$ | $\frac{1}{1 + \psi_1}$ |
| $G_p = (0,0):$ | |
| $G_o = 2$ | 0 |
| $G_o = 1$ | 0 |
| $G_o = 0$ | 1 |

NOTE.— $G_p = (G_{p,1}, G_{p,2})$ denotes unordered parental genotypes, and G_o denotes offspring genotype. Each subject's genotype equals the number of copies of high-risk allele A that the individual possesses.

Weinberg et al. (1998). For a SNP, G_p takes one of six possible mating types in the set $\{(2,2), (2,1), (2,0), (1,1), (1,0), (0,0)\}$. Define μ_l as the probability of the l th mating type ($l = 1, \dots, 6$) in the population. As shown in table 2, $P(G_p|D_o = 1)$ for each mating type is then specified in terms of the RR parameters ψ_1 and ψ_2 and the mating-type parameters $\mu = (\mu_1, \mu_2, \dots, \mu_6)$. Because the values

of μ may be any positive values whose sum is 1, we avoid assumptions of HWE or random mating among the parents.

To incorporate unrelated controls in the combined analysis, we make a rare-disease assumption such that $P(G_u|D_u = 0) \approx P(G_u)$; hence, we can write $P(G_u|D_u = 0)$ as the marginal probability of a single parent's mating on the basis of the probabilities given in table 2. In particular, we have

$$\begin{aligned}
 P(G_u = 2|D_u = 0) &\approx P(G_{p,1} = 2) = \mu_1 + \frac{\mu_2}{2} + \frac{\mu_3}{2} \\
 P(G_u = 1|D_u = 0) &\approx P(G_{p,1} = 1) = \frac{\mu_2}{2} + \mu_4 + \frac{\mu_5}{2} \\
 P(G_u = 0|D_u = 0) &\approx P(G_{p,1} = 0) = \frac{\mu_3}{2} + \frac{\mu_5}{2} + \mu_6. \quad (3)
 \end{aligned}$$

Because the model for $P(G_p)$ is saturated, fitting $P(G_p, G_o|D_o = 1)$ to triad data alone yields no additional information on ψ_1 and ψ_2 , compared with fitting $P(G_o|G_p, D_o = 1)$. However, equation (3) shows that the inclusion of unrelated control data in equation (1) yields additional information on the mating-type parameters μ . As a result, the addition of unrelated controls to the triad data in equation (1) increases the efficiency of inference on ψ_1 and ψ_2 . Further efficiency gains can be achieved by incorporating genotype data from unrelated cases into the joint analysis. To do this, we again make a rare-disease approximation such that we can model the genotype probability of a case as

$$\begin{aligned}
 P(G_u = g|D_u = 1) &= \frac{\psi_g P(G_{p,1} = g)}{\sum_{g^*} \psi_{g^*} P(G_{p,1} = g^*)} \\
 &\approx \frac{\psi_g P(G_u = g|D_u = 0)}{\sum_{g^*} \psi_{g^*} P(G_u = g^*|D_u = 0)} \quad (4)
 \end{aligned}$$

Examination of this probability demonstrates that case genotypes will provide additional information on ψ_1 , ψ_2 , and μ .

Testing for Linkage and Association Between a SNP and Disease

We can use L in equation (1) to estimate ψ_1 , ψ_2 , and μ by using standard maximum-likelihood procedures. We can also use L to construct likelihood-ratio (LR) statistics for testing the null hypothesis of no linkage or association between a SNP and disease. Testing this null hypothesis, H_0 , corresponds to testing $\psi_1 = \psi_2 = 1$ for a general RR model or to testing $\psi = 1$ for a nongeneral RR model. Under H_0 , the LR statistic asymptotically

Table 2
Evaluation of $P(G_p|D_o = 1)$ for a SNP

| G_p | $P(G_p)$ | $P(G_p D_o = 1)$ |
|-------|----------|--|
| (2,2) | μ_1 | $\frac{\psi_2\mu_1}{R}$ |
| (2,1) | μ_2 | $\frac{(\psi_1 + \psi_2)\mu_2}{2R}$ |
| (2,0) | μ_3 | $\frac{\psi_1\mu_3}{R}$ |
| (1,1) | μ_4 | $\frac{(1 + 2\psi_1 + \psi_2)\mu_4}{4R}$ |
| (1,0) | μ_5 | $\frac{(1 + \psi_1)\mu_5}{2R}$ |
| (0,0) | μ_6 | $\frac{\mu_6}{R}$ |

NOTE.— R = normalization factor.

follows either a χ^2_2 distribution (for a general RR model) or a χ^2_1 distribution (for a nongeneral RR model).

Testing Whether Data Sources Can Be Combined

Before making inferences based on the sample, one must first ensure that the data from triads, unrelated controls, and unrelated cases can be safely combined. Tests and estimators based on equation (1) are valid only if triads, unrelated controls, and unrelated cases are sampled from populations having the same allele frequencies and RR parameters. For example, comparison of $P(G_u|D_u = 0)$ in equation (3) with $P(G_p|D_o = 1)$ in equation (2) shows that additional information on ψ_1 and ψ_2 is gained by comparing the triad parents with unrelated controls. However, this information is only valid when the same mating-type parameters μ describe both samples. Suppose the true value of μ_1 in the unrelated controls is smaller than the true value of the corresponding parameter in the triad parents. The estimate of μ_1 in the parents will then be smaller than its true value, which, as shown by the values for $P(G_p|D_o = 1)$ in table 2, will result in an inappropriate inflation of the value of ψ_2 from the parental data. Similarly, comparison of $P(G_u|D_u = 0)$ in equation (3) with $P(G_u|D_u = 1)$ in equation (4) shows that additional information on ψ_1 and ψ_2 is gained by comparing the cases with controls (which corresponds to a retrospective analysis of a case-control study); again, this inference is valid only when the same set of mating-type parameters μ describe both samples.

The most direct test of whether data from unrelated subjects can be safely combined with triad data would be to test the equality of the mating-type distribution (the parameters μ) for each data source. Unfortunately, the mating-type distribution is not identifiable from unrelated cases or controls without an assumption like random mating. Furthermore, without an additional HWE assumption, this approach would involve tests

with >1 df, which is a potential threat to efficiency. Thus, in accordance with Nagelkerke et al. (2004), we test whether the data from unrelated subjects may be combined with the triad data by assessing whether the information on ψ_1 and ψ_2 obtained by combining the data from triad parents and unrelated subjects is compatible with the valid information on ψ_1 and ψ_2 obtained by comparing the transmitted alleles with untransmitted alleles in triads by use of the CPG approach. For this reason, we rewrite likelihood equation (1) as

$$L = \prod_{i=1}^I P(G_{oi}|G_{pi}, D_{oi} = 1; \psi_1, \psi_2, \mu) \times P(G_{pi}|D_{oi} = 1; \psi_1^{(p)}, \psi_2^{(p)}, \mu) \times \prod_{j=1}^J P(G_{uj}|D_{uj} = 0; \mu) \times \prod_{k=1}^K P(G_{uk}|D_{uk} = 1; \psi_1^{(c)}, \psi_2^{(c)}, \mu), \tag{5}$$

where $(\psi_1^{(p)}, \psi_2^{(p)})$ and $(\psi_1^{(c)}, \psi_2^{(c)})$ are each initially treated as a distinct set of RR parameters that corresponds to information on (ψ_1, ψ_2) obtained by comparing the triad parents with controls and by comparing the cases with controls, respectively. This approach offers some protection against the effect of population stratification that may be introduced if genotype frequencies in the unrelated cases or controls do not match those of the population from which triads were sampled. We can test whether data from triads and unrelated subjects may be combined by testing the hypotheses about the equality of (ψ_1, ψ_2) , $(\psi_1^{(p)}, \psi_2^{(p)})$, and $(\psi_1^{(c)}, \psi_2^{(c)})$. If supported by the data, $(\psi_1^{(p)}, \psi_2^{(p)})$ and $(\psi_1^{(c)}, \psi_2^{(c)})$ can each be constrained to equal (ψ_1, ψ_2) , resulting in efficiency gains for estimators of the RR parameters.

We can easily test the hypotheses about the equality of (ψ_1, ψ_2) , $(\psi_1^{(p)}, \psi_2^{(p)})$, and $(\psi_1^{(c)}, \psi_2^{(c)})$ by using LR statistics; score tests may also be constructed. We recommend first testing $H_0^{(p)}$, that $(\psi_1, \psi_2) = (\psi_1^{(p)}, \psi_2^{(p)})$. If this hypothesis is not rejected, then we set $(\psi_1, \psi_2) = (\psi_1^{(p)}, \psi_2^{(p)})$ and test $H_0^{(c|p)}$, that $(\psi_1, \psi_2) = (\psi_1^{(c)}, \psi_2^{(c)})$. We recommend this testing scheme because $H_0^{(p)}$ tests whether data from unrelated controls can be safely combined with triad data; if this is not possible, it is difficult to justify combining the data from unrelated cases and triads. In the examples we looked at, it is possible to estimate (ψ_1, ψ_2) by use of only triads and unrelated cases, but (ψ_1, ψ_2) and $(\psi_1^{(c)}, \psi_2^{(c)})$ cannot be separately estimated using such data. An alternative strategy is to directly test $H_0^{(p,c)}$, that $(\psi_1, \psi_2) = (\psi_1^{(p)}, \psi_2^{(p)}) = (\psi_1^{(c)}, \psi_2^{(c)})$. If a general (i.e., 2 df) model for the RR parameters is used, then test statistics for $H_0^{(p)}$ and $H_0^{(c|p)}$ each follow a χ^2 distribution with 2 df, whereas the test statistic $H_0^{(p,c)}$ has a χ^2 distribution

with 4 df. If a nongeneral (i.e., 1 df) model for the RR parameters is used (e.g., a multiplicative, dominant, or recessive model) then the degrees of freedom are half those in the general case.

These tests generalize the proposals of Nagelkerke et al. (2004) in two ways. First, we recommend initially testing whether the information on RR parameters from the comparison of parents with unrelated controls can be safely combined with the triad data and then subsequently testing whether the information on the RR parameters from the comparison of cases with controls can be safely combined with the triad data. This is important because we may find that, for example, the unrelated controls may be combined with the triad data but not with the unrelated cases. Second, we base our inference on likelihood equation (5), which contrasts the inferential procedure of Nagelkerke et al. (2004) that looked for overlap in CIs for the separate RR parameters, a procedure that is not recommended (Schenker and Gentleman 2001). In addition, by using the likelihood for inference, we can apply selection procedures, such as the Akaike information criteria (AIC), Bayesian information criteria (BIC), or backwards selection to formalize hypothesis testing and model selection. Backward selection can be used by starting with the richest possible model and removing parameters after hypothesis testing. For example, we could start with a general-risk model and separate estimates for (ψ_1, ψ_2) , $(\psi_1^{(p)}, \psi_2^{(p)})$, and $(\psi_1^{(c)}, \psi_2^{(c)})$. We could then conduct tests for equality of the different sets of RR parameters, as well as tests of specific genetic models (e.g., the multiplicative model). Under the assumption that at least one hypothesis is rejected, the appropriate parameters can be constrained (e.g., $[\psi_1, \psi_2] = [\psi_1^{(p)}, \psi_2^{(p)}]$), and the model can be refit. We then repeat this procedure until no additional hypotheses about parameters can be rejected.

Incorporation of Missing Parental Data

A practical problem in using triads for association analysis is that marker data for members of the triad may be unavailable due to factors such as refusal to participate, death, or false paternity. The unavailability of such data results in incomplete triads, missing either one or both parents (missing the triad offspring is also possible but will not be considered further). Using the terminology of Weinberg (1999), we refer to a triad missing one parent as a “dyad” and a triad missing both parents as a “monad.”

Various methods have been developed for association analysis of triads in the presence of dyads and monads. Approaches include the analysis of only the complete triad data or reconstruction of parental genotypes from informative offspring genotypes. Schaid (2004) provides an excellent overview of these approaches and

their limitations. Here, we follow the approach of Weinberg (1999) to account for missing parental genotype information. This approach corresponds to replacing $P(G_o|G_p, D_o = 1; \psi_1, \psi_2, \mu)P(G_p|D_o = 1; \psi_1^{(p)}, \psi_2^{(p)}, \mu)$ in likelihood equation (5) with

$$\sum_{G_p \in S} P(G_o|G_p, D_o = 1; \psi_1, \psi_2, \mu) \times P(G_p|D_o = 1; \psi_1^{(p)}, \psi_2^{(p)}, \mu) \quad (6)$$

for dyads or monads, where the set S corresponds to all parental genotypes consistent with the observed parental genotype data. When only complete triads, dyads, and monads are analyzed, this procedure is robust to population stratification, as long as a flexible (nonparametric) model for $P(G_p|D_o = 1; \psi_1, \psi_2, \mu)$ is chosen, because the μ parameters are fit to the distribution of parental genotypes. In combining equation (6) with data from unrelated controls, however, the parameters μ are estimated jointly with parental and control genotypes. Hence, inference based on equations (6) and (3) may be biased. For this reason, in tests for whether family-based data can be safely combined with data from unrelated controls, we recommend replacing (ψ_1, ψ_2) with $(\psi_1^{(p)}, \psi_2^{(p)})$ in both terms of equation (6). In the appendix, we give an expectation-maximization (EM) algorithm for evaluating the likelihood when equation (6) is used for dyads and monads.

We wish to point out that we consider monads arising from a sample of triads to be distinct from unrelated cases arising from a case-control sample in analysis. If we fail to make this distinction in the analysis, then we implicitly assume that monads and unrelated cases come from the same population, whereas, in fact, monads were sampled in the family-based arm of the study. Therefore, we treat monads—but not unrelated cases—as having missing parental data. For a study design in which all unrelated cases and triad data are collected simultaneously through the same recruitment method from the same population, monads could be treated as unrelated cases, and there would be no need for a separate set of $(\psi_1^{(c)}, \psi_2^{(c)})$ parameters in the analysis.

Application to Psoriasis Data Set and Simulations

We applied our combined association test to a subset of data from a genetic study of psoriasis described in Veal et al. (2002). This candidate-gene study genotyped 59 SNPs found throughout the psoriasis susceptibility 1 locus (*PSORS1*), which is contained within the MHC on chromosome 6p21. Using a collection of triads of European ancestry, Veal et al. (2002) performed TDTs on each of the 59 SNPs. To illustrate our likelihood approach, we focus attention on the results for SNP *CDSN1243*, which was found to be significantly asso-

ciated with disease in the initial analysis of the study ($P < .001$).

In addition to collecting the triads for association analysis, Veal et al. (2002) also collected an additional sample of unrelated controls of European ancestry for the analysis of LD conservation along the MHC. Here, we investigate whether incorporating the unrelated controls into the analysis increases the evidence for association between *CDSN1243* and a psoriasis-influencing variant. Our study sample consists of 149 triads (130 complete triads, 17 dyads, and 2 monads) and 269 unrelated controls. For this SNP, we constructed the likelihood equation (6) under additive, dominant, and recessive models for the RR. We therefore present results in terms of the previously defined scalar risk ψ , as well as the scalar risk $\psi^{(p)}$, which corresponds to information on ψ obtained by comparing the genotypes of triad parents with those of unrelated controls. For each likelihood, we implemented the EM algorithm given in the appendix.

Using an LR statistic, we first treated ψ and $\psi^{(p)}$ as separate parameters and tested hypothesis $H_0^{(p)}$, that $\psi = \psi^{(p)}$, to assess whether we could safely combine the triads and controls for analysis. If we failed to reject $H_0^{(p)}$, then we maximized ψ from the triads and controls together and tested the null hypothesis of no linkage or association. To determine the most likely mechanism of genetic action, we calculated the AIC under each RR model and chose as the best model the one with the lowest AIC value (Akaike 1985).

Using the psoriasis study sample as a starting point, we conducted additional simulations to investigate the type I error and power of our combined association test under the assumption of a sample of triads and unrelated controls. Assuming the same number of triads and controls as in the psoriasis data set, we simulated parental and unrelated control information by using estimates of μ from the best model (based on the AIC) in the psoriasis analyses. We varied the true value of ψ in a range from 1 (null) to 2.5 and varied the true RR model among additive, dominant, and recessive mechanisms. To investigate the effect of missing parental data on results, we varied the percentage of missing triad parents from 0% to 40%. To study power changes that occur when controls are added to the analysis, we analyzed each data set twice: once by using our combined association approach and once by using the CPG likelihood (which ignores information from the unrelated controls). Each result is based on 10,000 replicates of the data.

We performed two additional sets of simulations to assess the power of the combined association test for testing hypothesis $H_0^{(p)}$ (that $\psi = \psi^{(p)}$) when it is inappropriate to combine triads and unrelated controls for analysis. The first set of simulations corresponds to the situation in which triads and controls come from completely different populations with different sets of mat-

ing-type frequencies. For the triads, we simulated genotypes by using the estimates of μ from the best model (on the basis of the AIC) in the psoriasis analyses. However, for the controls, we simulated genotypes under HWE that varied the frequency of the *A* and *a* alleles among values that would make the sample unsuitable to combine with the triads. For each data set, we then tested $H_0^{(p)}$ to assess whether it was appropriate to combine the two samples. Each result is based on 10,000 replicates of the data.

The second set of simulations for testing $H_0^{(p)}$ corresponds to the situation in which population stratification exists in the sample. For these simulations, we assumed that we sampled triads and unrelated controls from two discrete strata. We assumed the first stratum was of European origin (like the psoriasis sample) and had the *CDSN1243* allele frequencies from the best model (on the basis of the AIC) in the psoriasis analyses. We assumed the second stratum was of Thai origin and had *CDSN1243* allele frequencies corresponding to those found in a psoriasis study conducted by Romphruk et al. (2003). For simplicity, we assumed that the *CDSN1243* alleles were in HWE in both strata. We induced stratification by sampling triads and controls in different proportions from the two strata. We sampled controls in equal proportions from the two strata but sampled triads in unequal proportions. For each data set, we then tested $H_0^{(p)}$ to assess whether it was appropriate to combine the triads and controls. Each result is based on 10,000 replicates of the data.

We also investigated the type I error and power of our combined association test if one were to jointly analyze SNP data from both triad and case-control studies. We again assumed a sample of 149 triads but now assumed that half of the 269 unrelated subjects were cases and the other half were controls. We varied the true value of ψ in a range from 1 (null) to 2.5 and varied the true RR model among additive, dominant, and recessive mechanisms. To determine the power differences between our combined test and tests based on the separate samples, we analyzed each data set three times: once by using our combined association approach, once by using the CPG likelihood (which ignores information from the unrelated cases and controls), and once by using a 1-df likelihood-based association test for unrelated subjects (which ignores information from the triads). Each result is based on 5,000 replicates of the data.

Results

Analysis of Psoriasis Data Set

Table 3 presents the results of the analyses of the association of psoriasis and *CDSN1243*. We first tested whether the triads and unrelated controls could be com-

Table 3
Results from CDSN1243 Analysis of Psoriasis Data Set

| ASSUMPTION AND VALUE | MODEL | | |
|--------------------------|----------|----------|-----------|
| | Additive | Dominant | Recessive |
| $\psi \neq \psi^{(p)}$: | | | |
| Parameter estimate: | | | |
| ψ | 3.39 | 3.40 | 1.99 |
| $\psi^{(p)}$ | 3.07 | 5.04 | 3.82 |
| μ_1 | .047 | .059 | .034 |
| μ_2 | .281 | .294 | .257 |
| μ_3 | .097 | .077 | .151 |
| μ_4 | .180 | .185 | .176 |
| μ_5 | .309 | .272 | .337 |
| μ_6 | .086 | .113 | .045 |
| AIC | 1,149.15 | 1,157.36 | 1,161.07 |
| P value ^a | .84 | .57 | .15 |
| $\psi = \psi^{(p)}$: | | | |
| Parameter estimate: | | | |
| ψ | 3.26 | 3.90 | 2.41 |
| μ_1 | .046 | .060 | .042 |
| μ_2 | .279 | .298 | .290 |
| μ_3 | .097 | .079 | .132 |
| μ_4 | .180 | .185 | .180 |
| μ_5 | .310 | .273 | .309 |
| μ_6 | .088 | .105 | .047 |
| AIC | 1,147.18 | 1,155.6 | 1,161.13 |
| Combined LR statistic | 32.68 | 24.19 | 18.72 |
| CPG LR statistic | 21.92 | 14.94 | 10.18 |

^a P values for hypothesis $H_0^{(p)}$, that $\psi = \psi^{(p)}$.

combined safely in the combined association analysis. Modeling ψ and $\psi^{(p)}$ as separate parameters, we constructed an LR statistic under each RR model to test hypothesis $H_0^{(p)}$, that $\psi = \psi^{(p)}$. P values for testing $H_0^{(p)}$ for additive, dominant, and recessive models were .84, .57, and .15, respectively. These results suggested that it was appropriate to let $\psi = \psi^{(p)}$ and thereby safely combine the data from the triads and unrelated controls within the combined association analysis.

The lower portion of table 3 shows results obtained by use of both our combined association analysis that used the triads and unrelated controls and the CPG approach that used triads only. Under all three models considered, results found using both methods suggest that CDSN1243 is significantly associated with a psoriasis-influencing variant that has a large effect on the disease (range of ψ estimates, 2.41–3.90). Under an additive model, the P value of the combined test was 1.08×10^{-8} , whereas the P value for the CPG approach was 2.84×10^{-6} . Under a dominant model, the combined P value was 8.72×10^{-7} , whereas the CPG P value was 1.10×10^{-4} . Under a recessive model, the combined P value was 1.51×10^{-5} , whereas the CPG P value was 1.41×10^{-3} . These results suggest that incorporating the unrelated controls into the analysis increases our

ability to detect association. On the basis of the AIC values, we chose the additive model with the assumption $\psi = \psi^{(p)}$ as the best model for the analysis of CDSN1243. Note that the selected model is the same as that obtained by picking the model with lowest AIC value among all genetic models in which $\psi = \psi^{(p)}$ and all genetic models in which ψ and $\psi^{(p)}$ are estimated separately.

Power Comparisons: Triads and Unrelated Controls

Figure 1 presents power curves at $\alpha = 0.05$ for the combined association test and the CPG approach under the assumption of a sample of 149 complete triads and 269 unrelated controls. Under the null model of $\psi = 1.0$, the type I error rates for both methods appeared appropriate at $\alpha = 0.05$. For $\psi > 1.0$, the results show that our combined association test has improved power to detect association, relative to the CPG approach, across all three model types considered for analysis. In general, we observed the largest increase in power for additive models (e.g., from 0.53 to 0.73 for $\psi = 1.5$), followed by dominant (from 0.32 to 0.43) and recessive models (from 0.42 to 0.51).

The power figures for the combined association test in figure 1 assume no missing parental data in the triads. Such missing parental data could potentially affect the power of our combined association analysis. Figure 2 plots the estimated power of our combined association test for different models with $\psi = 1.5$ at $\alpha = 0.05$, with the percentage of missing parental data in the range 0%–40%. These simulation results suggest that using likelihood equation (6) for monads and dyads can recover much of the power that would be lost if we were to exclude data from these incomplete triads from the analysis. For an additive model, the power only decreases from 0.73, for complete data, to 0.72, for data with 40% of parental data missing. We find a similar power difference for dominant models. For recessive models, the reduction in power for missing data was greater than that for the other models, but still only decreased from 0.51, for complete data, to 0.47, for data with 40% of parental data missing. These results suggest that missing parental data does not seriously impact the power of our combined association analysis.

We next investigated the power of our combined association test as a function of the number of controls used in the analysis. We simulated data under the assumption of 149 complete triads but varied the number of controls from 0 to 400. Figure 3 shows the power results at $\alpha = 0.05$ for an additive model with $\psi = 1.5$. The figure shows that the power increases from 0.53, when there are no controls, to 0.76, when there are 300 controls. However, the figure also shows that the power appears to plateau for samples with >300

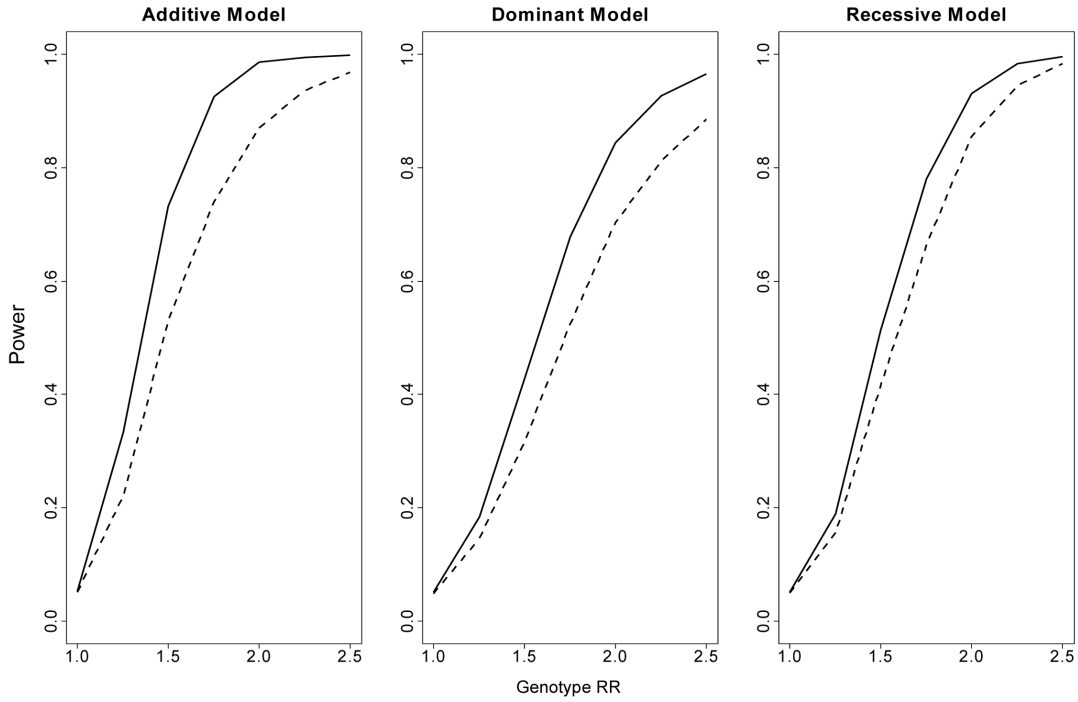


Figure 1 Power at $\alpha = 0.05$ of the combined association test (*solid line*) and the CPG approach (*dashed line*) under the assumption of 149 triads and 269 unrelated controls. Power results are based on 10,000 replicates of data simulated using the estimated mating types from the psoriasis study given in table 3.

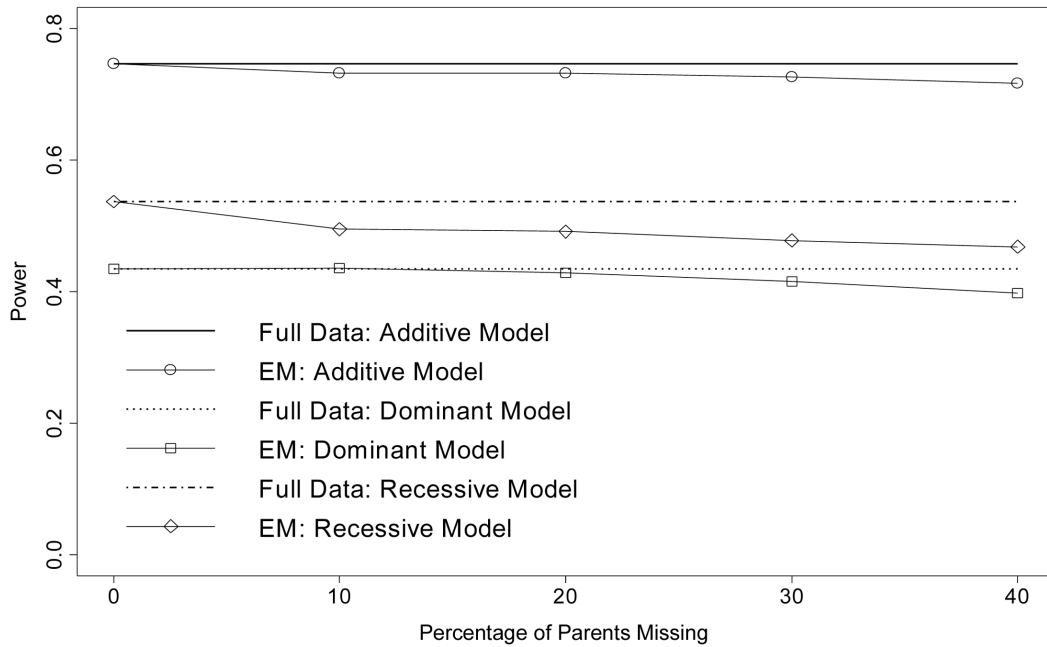


Figure 2 Power at $\alpha = 0.05$ of the combined association test with missing parental data. Power results are based on 10,000 replicates of data simulated using the estimated mating types from the psoriasis study given in table 3, with the assumption $\psi = 1.5$.

controls. For 400 controls, the power was 0.77. When we simulated and analyzed data sets with 800 controls, the power increased to only 0.80. Since controls provide information about μ only and not RRs, they fail to provide additional power to detect association once the mating-type frequencies are well estimated. Therefore, we expect the power to plateau once we incorporate enough controls in the combined association analysis to accurately estimate μ in the triads.

We also investigated the power of our combined association test as a function of the number of triads used in the analysis. We simulated data under the assumption of 269 unrelated controls but varied the number of triads from 50 to 350. Figure 4 shows the power results for an additive model with $\psi = 1.5$ when $\alpha = 0.05$. The power of our combined association test increased from 0.39, for 50 triads, to 0.95, for 350 triads. Our test also appears to be more powerful than the CPG approach, for all triad analyses that we considered. We also failed to notice any plateau in the power to detect association for either method, which we expected because triads provide information about both RRs and mating-type frequencies. Therefore, we expect the power of both the combined association test and the CPG approach to increase with an increasing number of triads.

We next assessed the power of the combined association test for testing hypothesis $H_0^{(p)}$ (that $\psi = \psi^{(p)}$) in the situations for which it is inappropriate to combine data from triads and unrelated controls. We first considered the situation in which the triads and controls come from different populations with different mating-type frequencies. We simulated data for 149 triads on the basis of the mating-type frequencies from the psoriasis analysis. We then simulated data for 269 unrelated controls under HWE. If the controls come from the same population as the triad parents, the estimated frequency of A would be 0.477. We therefore performed simulations using this assumed frequency of A to investigate the type I error rate for testing $H_0^{(p)}$. To investigate the power to reject $H_0^{(p)}$ under the alternative, we considered additional simulations that varied the frequency of the A allele in a range of 0.30–0.70.

Figure 5A shows the power to reject $H_0^{(p)}$ at $\alpha = 0.05$, under the assumption of an additive model with $\psi = 1.0$. When the frequency of A is 0.477, our combined association test has appropriate type I error. Under alternative models, the power to reject $H_0^{(p)}$ increases steadily with an increase in the magnitude between the assumed frequency and the null frequency of 0.477. We have >80% power to reject $H_0^{(p)}$ when the frequency of the A allele is either <0.37 or >0.58.

Figure 5B shows the impact on RR estimates that occurs when allele-frequency differences exist between the populations from which the triad parents and unrelated controls are sampled. When ψ and $\psi^{(p)}$ are esti-

mated as separate parameters under an additive model with $\psi = 1.0$, one can see that the former parameter is unbiased across all allele frequencies considered, which is expected, since this parameter estimate comes only from the CPG-based component of our likelihood. On the other hand, estimates of $\psi^{(p)}$ can be quite biased. We found an upward bias as the frequency of A decreases from the null frequency (0.477) and found a downward bias when the frequency increases from the null frequency. The situation of upward bias appears more severe, although we note that, under an additive model, ψ has a lower bound of 0.50. Figure 5B also shows mean RR estimates if one naively combines the triad and control data (letting $\psi = \psi^{(p)}$) when allele-frequency differences occur between the two samples. When substantial allele-frequency differences exist between the two samples, mean estimates of ψ in this situation can substantially differ from the true value, which can lead to biased inference. However, figure 5B also shows that the mean RR estimates based on the best model (ψ from the CPG approach if $H_0^{(p)}$ is rejected; otherwise, ψ from the combined test) showed little or no bias in these situations. These results show that testing $H_0^{(p)}$ prior to combining the data is necessary for valid association analysis.

We next assessed the power of the combined association test for testing $H_0^{(p)}$ when population stratification exists between the two samples. We sampled *CDSN1243* genotypes for 149 triads and 269 controls from two distinct strata. We assumed that the first stratum is of European origin and that the frequency of the A allele in the stratum is 0.477 (on the basis of the results of the psoriasis analysis). We assumed that the second stratum is of Thai origin and that the frequency of the A allele in the stratum is 0.90 (on the basis of results reported by Romphruk et al. [2003]). For simplicity, we assumed the *CDSN1243* alleles were in HWE in each stratum. We then induced stratification by sampling triads and controls in different proportions from the two strata. We assumed the controls were sampled in equal proportions but assumed triads were sampled from the first stratum with probability q . If $q = 0.50$, then no stratification exists between the triads and unrelated controls. We therefore performed simulations using $q = 0.50$ to investigate the type I error rate for testing $H_0^{(p)}$. To investigate the power to reject $H_0^{(p)}$ under the alternative when stratification exists, we considered additional simulations that varied the value of q from 0.10 to 0.90.

Figure 6A shows the power to reject $H_0^{(p)}$ at $\alpha = 0.05$ under an additive model with the assumption $\psi = 1.0$. When $q = 0.50$, our combined association analysis has appropriate type I error. Under alternative models, the power to reject $H_0^{(p)}$ increases steadily with an increase in the magnitude of the difference between the assumed value of q and the null value of $q = 0.50$. We have >80% power to reject $H_0^{(p)}$ when $q < 0.23$ or

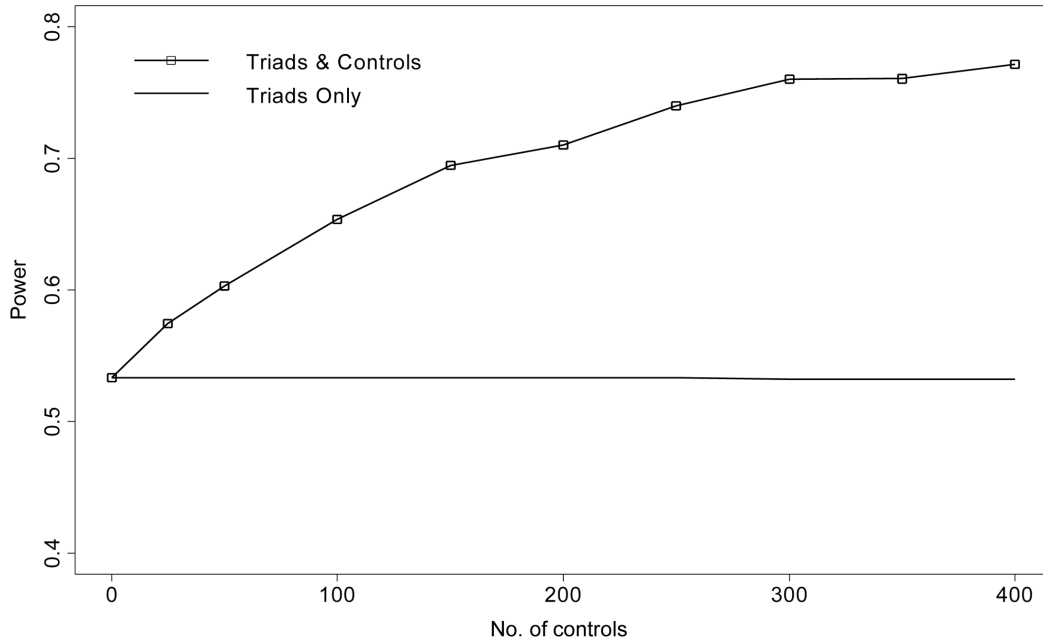


Figure 3 Power at $\alpha = 0.05$ of the combined association test under the assumption of 149 triads and a variable number of unrelated controls. Results are based on 10,000 replicates of data simulated using the estimated mating types from the psoriasis study given in table 3, under an additive model with the assumption $\psi = 1.5$.

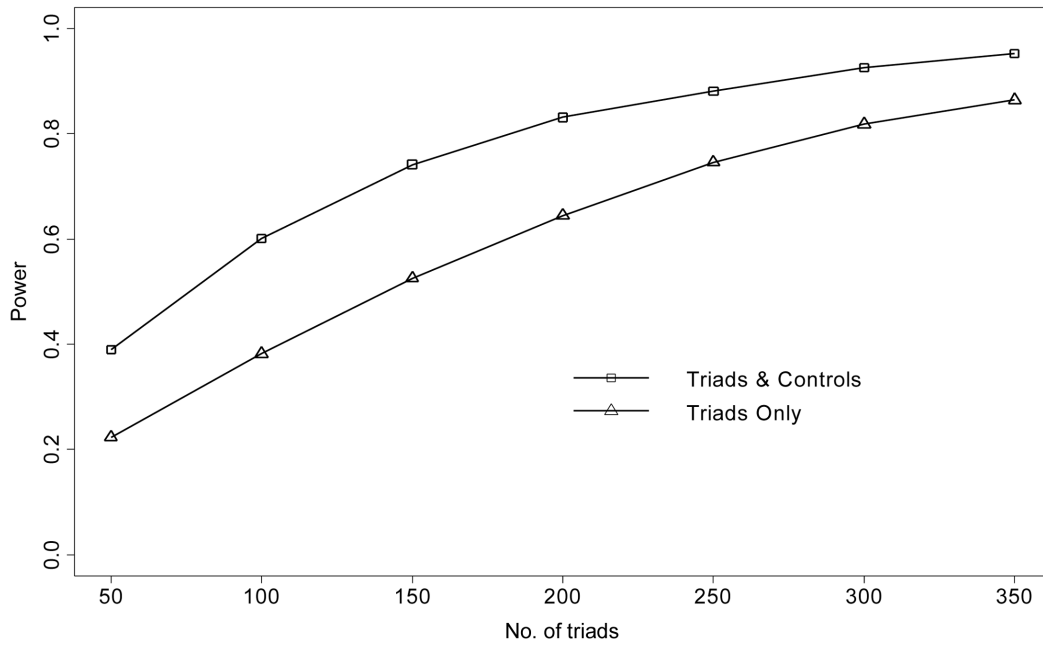


Figure 4 Power at $\alpha = 0.05$ of the combined association test under the assumption of 269 unrelated controls and a variable number of triads. Results are based on 10,000 replicates of data simulated using the estimated mating types from the psoriasis study given in table 3, under an additive model with the assumption $\psi = 1.5$.

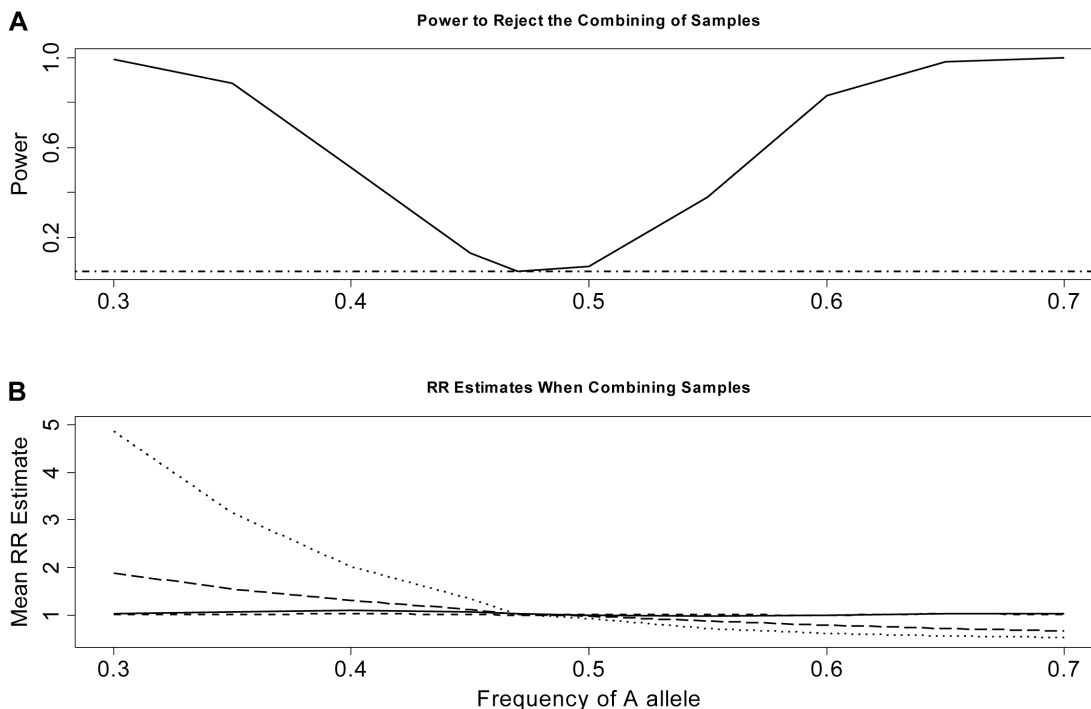


Figure 5 Suitability of combining SNP data from 149 triads and 269 unrelated controls. *A*, Power of the combined association test to reject hypothesis $H_0^{(p)}$: $\psi = \psi^{(p)}$ at $\alpha = 0.05$, under an additive model with the assumption $\psi = 1.0$. The dashed line denotes power equal to 0.05. *B*, Mean estimates of $\psi^{(p)}$ (short-dashed line) and ψ (medium-dashed line) when the two parameters are estimated separately and estimates of ψ (long-dashed line) when the data are combined. Also shown is ψ (solid line) for the best model (ψ from the CPG approach if $H_0^{(p)}$ is rejected; otherwise, ψ from the combined test). Results are based on 10,000 replicates of data. The frequency of the A allele is 0.477 when controls come from the same population as the triads.

$q > 0.77$. Figure 6B shows the impact on RR estimates that occurs when population stratification exists between the triads and unrelated controls. When ψ and $\psi^{(p)}$ are estimated as separate parameters under an additive model with the assumption $\psi = 1.0$, one can see that the former parameter is unbiased across all values of q being considered, which is expected, since this parameter estimate comes only from the CPG-based component of our likelihood (which is robust to stratification). However, estimates of $\psi^{(p)}$ can be quite biased when the sampling proportions differ between the triad and control samples. We found a downward bias for $q < 0.50$ and an upward bias for $q > 0.50$. Figure 6B also shows mean estimates of ψ if one naively combines the triad and control data (letting $\psi = \psi^{(p)}$) when $q \neq 0.50$. When stratification exists, mean estimates of ψ in this situation can substantially differ from the true value, which can lead to biased inference. However, figure 6B shows that the mean RR estimates from the best model (ψ from the CPG approach if $H_0^{(p)}$ is rejected; otherwise, ψ from the combined test) show little or no bias across different values of q . This again demonstrates that a valid association analysis of a combined data set requires prior testing of $H_0^{(p)}$.

Power Comparisons: Triads, Unrelated Controls, and Unrelated Cases

Figure 7 presents power curves at $\alpha = 0.05$ for our combined association test, the CPG approach, and the likelihood-based case-control association test under the assumption of a sample of 149 complete triads, 135 controls, and 134 cases. Under the null model of $\psi = 1.0$, the type I error rates for the three methods appeared appropriate at $\alpha = 0.05$. For $\psi > 1$, the performance of the combined association analysis is clearly superior to the other two tests. These results show that combining both triad and case-control studies in association analyses can substantially increase the power to identify disease-influencing variants. We observed the largest power increases for additive and dominant models, followed by recessive models. Relative to the next most powerful test, we found that the power at $\psi = 1.5$, when using our combined association test, increased from 0.53 to 0.85 under an additive model, from 0.32 to 0.58 under a dominant model, and from 0.42 to 0.66 under a recessive model. In general, we found that the CPG and case-control tests had similar power under all three models, with the CPG test being slightly less powerful than

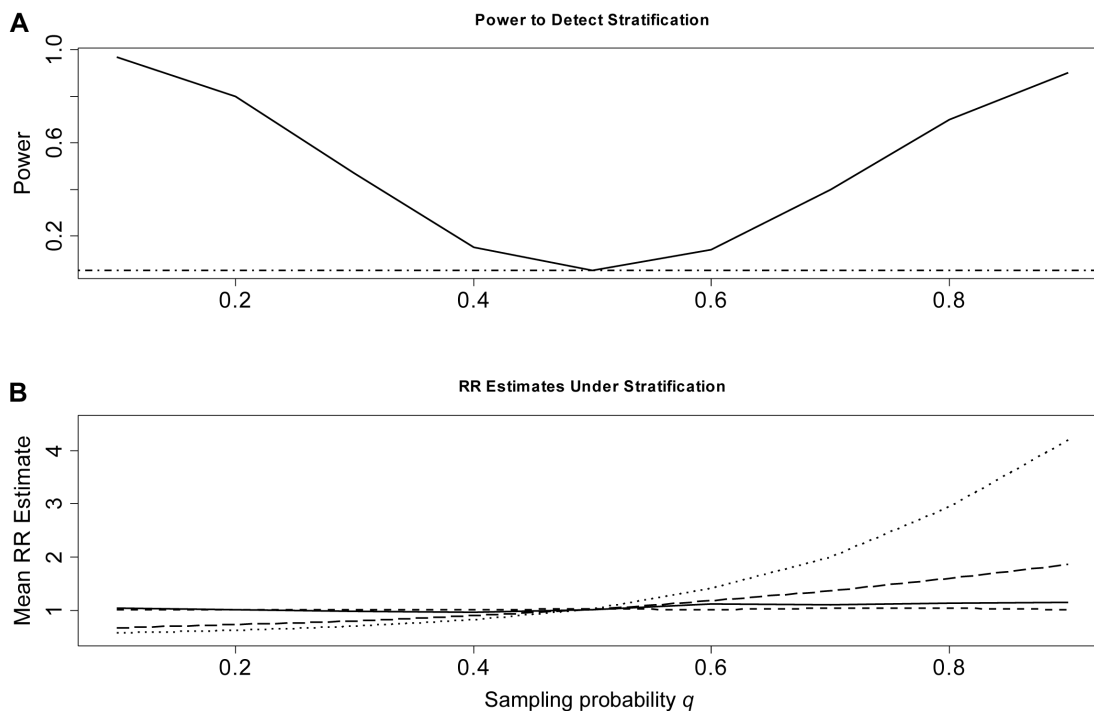


Figure 6 Suitability of combining SNP data from 149 triads and 269 unrelated controls under population stratification. Controls are sampled from two strata in equal proportions. Triads are sampled from stratum 1 with probability q . *A*, Power of the combined association test to reject hypothesis $H_0^{(p)}: \psi = \psi^{(p)}$ at $\alpha = 0.05$, under an additive model with the assumption $\psi = 1.0$. The dashed line denotes power equal to 0.05. *B*, Mean estimates of $\psi^{(p)}$ (short-dashed line) and ψ (medium-dashed line) when the two parameters are estimated separately and estimates of ψ (long-dashed line) when the data are combined. Also shown is ψ (solid line) for the best model (ψ from the CPG approach if $H_0^{(p)}$ is rejected; otherwise, ψ from the combined test). Results are based on 10,000 replicates of data. $q = 0.50$ corresponds to no stratification existing between triads and unrelated controls.

the case-control test under an additive model and moderately more powerful under dominant and recessive models.

Discussion

The choice of an appropriate control sample for association analysis often requires serious debate. Situations may arise in which a study might have both triads and unrelated subjects available for analysis. Rather than analyze these two samples separately, Nagelkerke et al. (2004) developed a joint analysis approach that allows for the combination of SNP data from triads, unrelated controls, and unrelated cases. Their approach emphasized assumptions and approximate analyses restricted to the multiplicative model that allows the use of standard logistic-regression software packages for analyses. Here, we take the viewpoint that the difficult step in gene discovery is the gathering of genetic data; in data analysis, the best possible method should be used even if specialized software is required. Furthermore, we believe it may be of interest to consider model types other than multiplicative, such as additive, dominant, or re-

cessive. For this reason, we have developed combined tests of association that are based on the likelihood of Nagelkerke et al. (2004) but that do not require assumptions like HWE, random mating, and a multiplicative model of allele effect. Analyses based on both real and simulated data indicate that our combined association approach has improved power over statistical methods that analyze triads and unrelated subjects separately, even when we assume a fairly general model for parental mating types in the target population. Our combined approach allows for flexible modeling of allele effects on disease and missing parental data. We are currently implementing our combined association analysis procedures in a Windows-based software package for public use, which can be downloaded free of charge from our Web site (see Epstein software Web site).

In addition, we also developed formal statistical tests to determine when it is appropriate to combine triads, unrelated controls, and unrelated cases together in our combined association analysis. Using simulated data based on the psoriasis data set, we found that substantial bias in RR estimates can arise when allele frequencies differ between the unrelated controls and the pop-

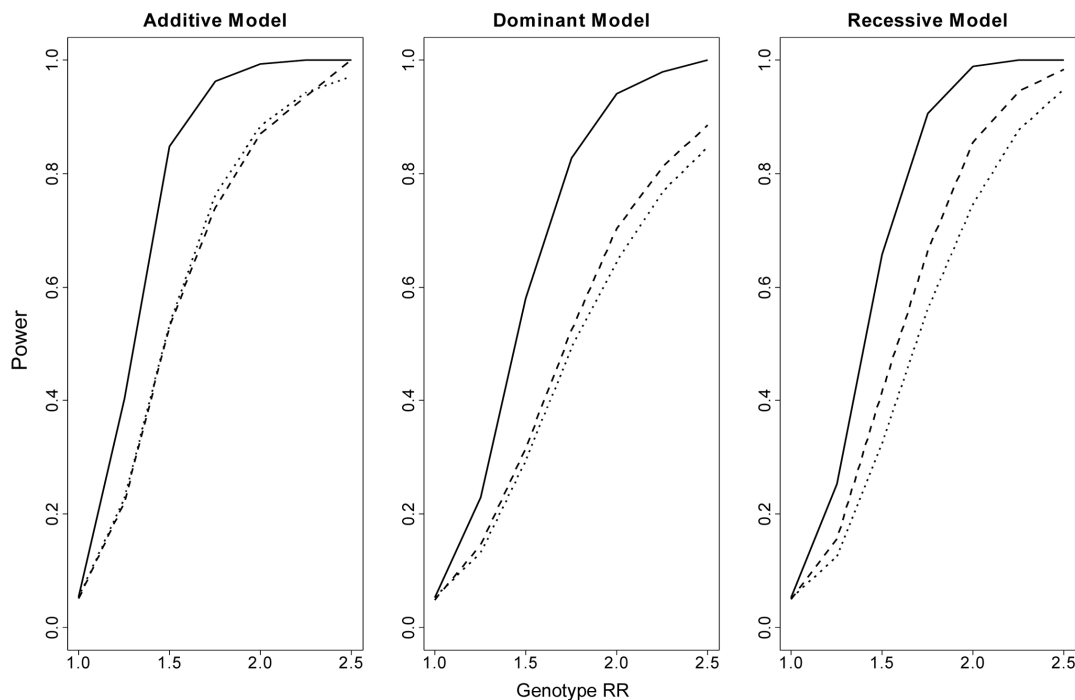


Figure 7 Power at $\alpha = 0.05$ of combined association test (solid line), CPG approach (long-dashed line), and case-control association test (short-dashed line), under the assumption of 149 triads, 135 controls, and 134 cases. Results are based on 10,000 replicates of data simulated using the estimated mating types from the psoriasis study given in table 3.

ulation from which the triads were sampled. We see this bias because information on (ψ_1, ψ_2) from triad parents comes from a direct comparison of their genotype distribution with that of unrelated controls (adjusted for RR with eq. [3]). If the distributions differ between the two samples, then estimates of (ψ_1, ψ_2) obtained using information from the parents can be grossly inflated or deflated. Note that only allele-frequency differences between the two samples are required to induce bias. This is opposed to bias in case-control studies, which originates from differences in both allele frequencies and disease risk. These simulation results suggest that special care must be taken to ensure that triads and controls are comparable and highlight the importance of testing whether the information from these two sources can be safely combined.

Because Nagelkerke et al. (2004) assumed HWE and random mating among parents, we determined the effect of a violation of these assumptions in joint analysis of data from triads and unrelated subjects. We find that when these assumptions are violated, biased RR estimates and erroneous inference may occur. To illustrate this, we note that the mating-type frequencies of the *CDSN1243* data in table 3 suggest nonrandom mating. Therefore, we used these estimated mating types from the psoriasis study to simulate 149 triads and 269 controls under additive, dominant, and recessive models

with the assumption $\psi = 1.0$. We analyzed each data set twice: once by using our combined association analysis (which does not require the assumptions of random mating and HWE) and once by using a modification of our approach that explicitly assumes random mating and HWE. Results are based on 10,000 replicates.

Table 4 provides empirical type I error rates and mean parameter estimates of ψ under these two methods. The table shows that association methods that assume random mating and HWE will have either deflated or inflated type I error relative to the nominal level, depending on the assumed genetic mechanism. The results also indicate that estimates of ψ may be biased for dominant and recessive models. In addition, the empirical type I error rates for testing whether samples should be combined may also be either deflated or inflated with respect to the nominal levels (results not shown), depending on the genetic model. In contrast, our combined association method has appropriate type I error for testing and yields unbiased estimates of ψ . Our test also yields a valid test for testing whether samples can be safely combined (results not shown) for all genetic models considered.

Without the explicit assumption of random mating and HWE, our combined method yields robust tests of linkage and association when these two conditions are violated. At the same time, when these two conditions

Table 4

Type I Error and Bias of Combined Association Tests when Random Mating Does Not Hold: The CDSN1243 Example with the Assumption $\psi = 1.0$

| TEST, ERROR, AND BIAS | MODEL | | |
|--|----------|----------|-----------|
| | Additive | Dominant | Recessive |
| Association test assuming random mating: | | | |
| Type I error, $\alpha = 0.05$ | .041 | .084 | .092 |
| Type I error, $\alpha = 0.01$ | .008 | .019 | .027 |
| Mean value of ψ | 1.012 | .909 | 1.145 |
| Association test not assuming random mating: | | | |
| Type I error, $\alpha = 0.05$ | .050 | .051 | .050 |
| Type I error, $\alpha = 0.01$ | .010 | .010 | .010 |
| Mean value of ψ | 1.017 | 1.022 | 1.015 |

NOTE.—Results are based on 10,000 replicates. Simulated data are based on mating-type frequencies from CDSN1243 analyses in table 3.

actually hold, we find our method has similar power to a method that explicitly assumes these conditions. To demonstrate this, we simulated 149 triads and 269 controls under a model that assumed random mating and HWE. We let the frequency of *A* be 0.477 (similar to the frequency of the susceptibility allele in the psoriasis sample) and simulated 5,000 replicate data sets under additive, dominant, and recessive models with the assumption $\psi = 1.5$. Under an additive model, the power of our combined test (0.73 at $\alpha = 0.05$) was equivalent to the power of an analogous test that assumed random mating and HWE. Our test was only slightly less powerful under both dominant (power of our combined test, 0.44; power with assumption of random mating and HWE, 0.48) and recessive (power of our combined test, 0.46; power with assumption of random mating and HWE, 0.49) models. We therefore conclude that the robustness of our method does not come at the expense of a substantial power loss under the ideal situation of HWE and random mating.

A useful extension of our combined association test would be allowance for genotype data from additional affected and unaffected siblings within a family. The accommodation of additional affected siblings is challenging, since a likelihood formulation is difficult if the locus under study is a marker locus rather than a susceptibility locus (Tu et al. 2000; Whittemore and Tu 2000). Given a family with *N* affected offspring, we instead suggest entering the likelihood of each of the *N* possible triads into the analysis separately (as if the data were independent). Because the correlation between affected siblings is not accounted for, the resulting analysis corresponds to a composite likelihood (Lindsay 1988), which can be used for inference as long as robust (sandwich) tests and estimators of variance are applied. For this reason, note that the LR statistic and AIC are not available, and hence all inference must be based on either robust Wald tests or generalized score tests (Boos 1992). A similar approach using generalized estimating

equations has been proposed by H. Putter, J. J. Houwing-Duistermaat, and N. J. D. Nagelkerke (unpublished data) using the same approach advocated by Nagelkerke et al. (2004). To allow for unaffected siblings such as those seen in discordant sib pair study designs, we make a rare-disease assumption such that we can multiply the likelihood in equation (1) by a factor corresponding to the likelihood for conditional logistic regression of matched-pair data. Data from multiplex sibships can be included in a similar way, although a variance adjustment is required to account for the correlation between sibs (Siegmund et al. 2000). Finally, we note that, if no triads are available, then combining unrelated controls with data from sibships in the manner described above does not result in any gains in efficiency, whereas combining data from sibships with case and control data is more like a meta-analysis than the combined approach described here.

Although we have focused on combined association analysis of single SNPs, we intend to extend our approach to allow for the analysis of haplotypes. Since haplotypes combine LD information from multiple markers simultaneously, we feel that such an approach could be more powerful than our current approach. Direct extension of the likelihood-based analysis described here, to accommodate haplotypes, is not trivial, because of the increase in the number of parameters needed to model the haplotypes. In particular, fitting a saturated model of parental haplotype pairs is not possible, since this distribution is not identifiable from unphased genotype data. Alternatively, we could reduce the number of mating-type parameters by assuming HWE and random mating among parents. However, as we have already showed for the psoriasis data set, such assumptions may not be valid in practice and may lead to erroneous inference. Therefore, we intend to develop a semiparametric procedure for estimating the haplotype RRs, as described by A. S. Allen, G. A. Satten, and A. A. Tsiatis (unpublished data).

Acknowledgments

We thank the members of the psoriasis study for allowing us to present results from the analysis of the psoriasis data. This research was supported by the University Research Committee of Emory University (to M.P.E.).

Appendix

EM Algorithm for Maximizing Likelihood Equation (6) in the Presence of Missing Parental Data

Assuming that the genotype data are complete, we can use L in equation (5) to estimate the complete set of unknown parameters, which we denote by $\theta = \{\psi_1, \psi_2, \psi_1^{(p)}, \psi_2^{(p)}, \psi_1^{(c)}, \psi_2^{(c)}, \mu_1, \mu_2, \dots, \mu_6\}$. However, if any parental genotype data are missing, then we choose to use an EM algorithm (Dempster et al. 1977) similar to the one described in Weinberg (1999) for estimation. Proper implementation of the EM algorithm begins by construction of the log likelihood of L in equation (6) under the assumption of complete data. We let l_c denote this complete log likelihood.

The EM algorithm proceeds iteratively, with each iteration comprising an E step and an M step. The E step requires the calculation of the expected value of l_c conditional on current estimates of θ , as well as the observed triad genotype and phenotype data. The subsequent M step then maximizes this expected value of l_c to update θ . The EM algorithm then cycles between the E and M steps until the convergence of θ . Convergence is declared when the sum of the squares of the parameter estimates at successive iterations is less than a small positive number, such as 1×10^{-12} .

To construct l_c , we let I_{rst} denote the number of triads with unordered parental mating type (r,s) ($r,s = 0,1,2$; $r \geq s$) and offspring genotype t ($t = 0,1,2$) and let p_{rst} denote the frequency of this genotype combination in the triads. We also let J_w denote the number of unrelated controls with genotype w ($w = 0,1,2$) and let γ_w denote the frequency of this genotype in the unrelated control sample. We let K_w denote the number of unrelated cases with genotype w and let η_w denote the frequency of this genotype in the unrelated case sample. It is straightforward to show that p_{rst} , γ_w , and η_w are functions of different elements of θ . Using this notation, we construct the complete log likelihood:

$$l_c = \sum_{r=0}^2 \sum_{s=0}^r \sum_{t=0}^2 I_{rst} \log \{p_{rst}\} + \sum_{w=0}^2 J_w \log \{\gamma_w\} + \sum_{w=0}^2 K_w \log \{\eta_w\}. \tag{A1}$$

At the start of iteration $k + 1$, we have estimates of $p_{rst}^{(k)}$, $\gamma_w^{(k)}$, and $\eta_w^{(k)}$ from the previous k th iteration. We take the expectation of l_c by using $p_{rst}^{(k)}$, $\gamma_w^{(k)}$, $\eta_w^{(k)}$, and the observed triad data. The observed data consist of (1) complete triads for which both the parental mating type and offspring genotype are observed, (2) dyads for which the parental mating type is partially missing and the offspring genotype is observed, and (3) monads for which the parental mating type is completely missing and the offspring genotype is observed. Let C_{rst} denote the total observed number of complete triads with observed parental mating type (r,s) and offspring genotype t . Let D_{rt} denote the total observed number of dyads with observed parental genotype r and offspring genotype t . Finally, let M_t denote the total observed number of monads with offspring genotype t .

Define $Y_{rst}^{(k+1)}$ as the expected number of triads with parental mating type (r,s) and offspring genotype t at iteration $k + 1$. Using the observed data, we can evaluate $Y_{rst}^{(k+1)}$ for $r \neq s$:

$$Y_{rst}^{(k+1)} = C_{rst} + D_{rt} \frac{p_{rst}^{(k)}}{\sum_{s^*=0}^{r-1} p_{rs^*t}^{(k)} + 2p_{rrt}^{(k)} + \sum_{s^*=r+1}^2 p_{s^*rt}^{(k)}} + D_{st} \frac{p_{rst}^{(k)}}{\sum_{r^*=0}^{s-1} p_{s^*rt}^{(k)} + 2p_{sst}^{(k)} + \sum_{r^*=s+1}^2 p_{r^*st}^{(k)}} + M_t \frac{p_{rst}^{(k)}}{\sum_{r^*=0}^2 \sum_{s^*=0}^{r^*} p_{r^*s^*t}^{(k)}}$$

and for $r = s$:

$$Y_{rrt}^{(k+1)} = C_{rrt} + D_{rt} \frac{2p_{rrt}^{(k)}}{\sum_{s^*=0}^{r-1} p_{rs^*t}^{(k)} + 2p_{rrt}^{(k)} + \sum_{s^*=r+1}^2 p_{s^*rt}^{(k)}} + M_t \frac{p_{rst}^{(k)}}{\sum_{r^*=0}^2 \sum_{s^*=0}^{r^*} p_{r^*s^*t}^{(k)}}.$$

The E step of the EM algorithm involves substituting the value of $Y_{rst}^{(k+1)}$ for I_{rst} in equation (A1) for all r, s , and t . The M step then maximizes the expected value of l_c with respect to θ by use of a quasi-Newton algorithm to obtain $\theta^{(k+1)}$, which can then be used to calculate $p_{rst}^{(k+1)}$, $\gamma_w^{(k+1)}$, and $\eta_w^{(k+1)}$. The EM algorithm then proceeds to the start of iteration $k + 2$, and the process is repeated until convergence.

Electronic-Database Information

The URL for data presented herein is as follows:

Epstein software, <http://server2k.genetics.emory.edu/mepstein/software.php>

References

- Abecasis GR, Noguchi E, Heinzmann A, Traherne JA, Bhat-tacharyya S, Leaves NI, Anderson GG, Zhang Y, Lench NJ, Carey A, Cardon LR, Moffatt MF, Cookson WOC (2001) Extent and distribution of linkage disequilibrium in three genomic regions. *Am J Hum Genet* 68:191–197
- Akaike H (1985) Prediction and entropy. In: Atkinson AC, Fienberg SE (eds) *A celebration of statistics*. Springer, New York, pp 1–24
- Allen AS, Rathouz PJ, Satten GA (2003) Informative missingness in genetic association studies: case-parent designs. *Am J Hum Genet* 72:671–680
- Boos DD (1992) On generalized score tests. *Am Stat* 46:327–333
- Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B* 39:1–22
- Devlin B, Roeder K (1999) Genomic control for association studies. *Biometrics* 55:997–1004
- Falk CT, Rubinstein P (1987) Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Ann Hum Genet* 57:455–464
- Gelernter J, Goldman D, Risch N (1993) The A1 allele at the D2 dopamine receptor gene and alcoholism. *JAMA* 269:1673–1677
- Knowler WC, Williams RC, Pettitt DJ, Steinberg AG (1988) GM 3,5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *Am J Hum Genet* 43:520–526
- Lindsay BG (1988) Composite likelihood methods. *Comtemp Math* 80:221–239
- Martin ER, Kaplan NL (2000) A Monte Carlo procedure for two-stage tests with correlated data. *Genet Epidemiol* 18:48–62
- Nagelkerke NJ, Hoebee B, Teunis P, Kimman TG (2004) Combining the transmission disequilibrium test and case-control methodology using generalized logistic regression. *Eur J Hum Genet* 12:964–970
- Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. *Am J Hum Genet* 67:170–181
- Rabinowitz D (2002) Adjusting for population heterogeneity and misspecified haplotype frequencies when testing non-parametric null hypotheses in statistical genetics. *J Am Stat Assoc* 97:742–758
- Romphruk AV, Oka A, Romphruk A, Tomizawa M, Choonhakarn C, Naruse TK, Puapairoj C, Tamiya G, Leelayuwat C, Inoko H (2003) Corneodesmosin gene: no evidence for PSORS 1 gene in North-eastern Thai psoriasis patients. *Tissue Antigens* 62:217–224
- Satten GA, Flanders WD, Yang Q (2001) Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am J Hum Genet* 68:466–477
- Schaid DJ (2004) Transmission disequilibrium methods for family-based studies. Mayo Clinic technical report number 72
- Schaid DJ, Sommer SS (1993) Genotype relative risks: methods for design and analysis of candidate-gene association studies. *Am J Hum Genet* 53:1114–1126
- (1994) Comparison of statistics for candidate-gene association studies using case and parents. *Am J Hum Genet* 55:402–409
- Schenker N, Gentleman J (2001) On judging the significance of differences by examining the overlap between confidence intervals. *Am Statistician* 3:182–186
- Siegmund KD, Langholtz B, Kraft P, Thomas DC (2000) Testing linkage disequilibrium in sibships. *Am J Hum Genet* 67:244–248
- Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 52:506–516
- Sun F, Flanders WD, Yang Q, Khoury MJ (1999) Transmission disequilibrium test (TDT) when only one parent is available: the 1-TDT. *Am J Epidemiol* 150:97–104
- Thomas DC (2004) *Statistical methods in genetic epidemiology*. Oxford University Press, New York
- Tu I-P, Balise RR, Whittemore AS (2000) Detection of disease genes by use of family data. II. Application to nuclear families. *Am J Hum Genet* 66:1341–1350
- Veal CD, Capon F, Allen MH, Heath EK, Evans JC, Jones A, Patel S, Burden D, Tillman D, Barker JNWN, Trembath RC (2002) Family-based analysis using a dense single-nucleotide polymorphism-based map defines genetic variation at *PSORS1*, the major psoriasis-susceptibility locus. *Am J Hum Genet* 71:554–564
- Weinberg CR (1999) Allowing for missing parents in genetic studies of case-parent triads. *Am J Hum Genet* 64:1186–1193
- Weinberg CR, Wilcox AJ, Lie RT (1998) A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. *Am J Hum Genet* 62:969–978
- Whittemore AS, Tu I-P (2000) Detection of disease genes by use of family data. I. Likelihood-based theory. *Am J Hum Genet* 66:1328–1340