

Statistical Applications in Genetics and Molecular Biology

Volume 3, Issue 1

2004

Article 26

Increasing Power for Tests of Genetic Association in the Presence of Phenotype and/or Genotype Error by Use of Double-Sampling

Derek Gordon* Yaning Yang[†] Chad Haynes[‡]
Stephen J. Finch** Nancy R. Mendell^{††}
Abraham M. Brown^{‡‡} Vahram Haroutunian[§]

*Rockefeller University, gordon@linkage.rockefeller.edu

[†]Rockefeller University, yyang@linkage.rockefeller.edu

[‡]Rockefeller University, chad.haynes@mail.rockefeller.edu

**Stony Brook University, sfinch@gis.net

^{††}Stony Brook University, nmendell@notes.cc.sunysb.edu

^{‡‡}Burke Medical Research Institute, ambrown@med.cornell.edu

[§]Mount Sinai School of Medicine, vahram.haroutunian@mssm.edu

Copyright ©2004 by the authors. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher, bepress. *Statistical Applications in Genetics and Molecular Biology* is produced by The Berkeley Electronic Press (bepress). <http://www.bepress.com/sagmb>

Increasing Power for Tests of Genetic Association in the Presence of Phenotype and/or Genotype Error by Use of Double-Sampling *

Derek Gordon, Yaning Yang, Chad Haynes, Stephen J. Finch, Nancy R. Mendell, Abraham M. Brown, and Vahram Haroutunian

Abstract

Phenotype and/or genotype misclassification can: significantly increase type II error probabilities for genetic case/control association, causing decrease in statistical power; and produce inaccurate estimates of population frequency parameters. We present a method, the likelihood ratio test allowing for errors (LRTae) that incorporates double-sample information for phenotypes and/or genotypes on a sub-sample of cases/controls. Population frequency parameters and misclassification probabilities are determined using a double-sample procedure as implemented in the Expectation-Maximization (EM) method. We perform null simulations assuming a SNP marker or a 4-allele (multi-allele) marker locus. To compare our method with the standard method that makes no adjustment for errors (LRTstd), we perform power simulations using a 2k factorial design with high and low settings of: case/control samples, phenotype/genotype costs, double-sampled phenotypes/genotypes costs, phenotype/genotype error, and proportions of double-sampled individuals. All power simulations are performed fixing equal costs for the LRTstd and LRTae methods. We also consider case/control ApoE genotype data for an actual Alzheimer's study.

The LRTae method maintains correct type I error proportions for all null simulations and all significance level thresholds (10%, 5%, 1%). LRTae average estimates of population frequencies and misclassification probabilities are equal to the true values, with variances of 10e-7 to 10e-8. For power simulations, the median power difference LRTae-LRTstd at the 5% significance level is 0.06 for multi-allele data and 0.01 for SNP data. For the ApoE

*Dr. Derek Gordon and Dr. Yaning Yang contributed equally to this work. The authors gratefully acknowledge grants K01-HG00055 and R01-MH59492 from the National Institutes of Health. Collection of phenotype and genotype data for the ApoE Alzheimer's study is funded in part by NIH-AG14930 (A.M.B.) and the Winifred Masterson Burke Relief Foundation, P01-AG02219 (V.H.). The authors are indebted to Dr. Francisco De La Vega for informative discussions regarding genotyping technologies.

data example, the LRT_{ae} and LRT_{std} p-values are 5.8×10^{-5} and 1.6×10^{-3} , respectively. The increase in significance is due to adjustment in the LRT_{ae} for misclassification of the most commonly reported risk allele. We have developed freely available software that performs our LRT_{ae} statistic.

KEYWORDS: misclassification, case, control, likelihood ratio, study design, cost-benefits

Introduction

It has long been appreciated (Cochran 1968) that misclassification error can significantly affect the results of statistical tests of association. In the field of genetics, researchers may observe misclassification errors in phenotype and/or genotype. A major question in the field of statistical genetics is: how does one “deal” with such errors when performing tests of genetic association? Much work has been done to address the general question of errors in statistical tests and recently some work has been done to address the specific question of errors in genetic tests of association.

Breslow and Day (1980) attribute the first statistical work on errors in association tests applied to contingency tables to Bross (1954). In his work, Bross (1954) focused on the χ^2 test of independence applied to 2×2 contingency tables and what we describe as phenotype error, namely the effects of misclassifying a case subject as a control and vice versa. Bross’s findings were: assuming that the error procedure is independent of case/control status, there is no change in the level of significance (i.e., the type I error rate remains constant); power for the χ^2 test is reduced; and, there is a bias in point estimates of the population frequency parameters.

Mote and Anderson (1965) synthesized the work of Mitra (1958) and Bross (1954). They proved that the power of the χ^2 test with no error is always greater than or equal to the power of the test when errors are present and ignored.

Tenenbein (1970; 1972) presented a procedure that used perfect classification on a sub-sample of data (e.g., genotypes or affection status) to estimate misclassification rates for all categories and also provided asymptotically unbiased estimates of population parameters (e.g., genotype frequencies, proportion of cases and controls). He called this procedure a “double-sampling” procedure, because some observations are sampled twice – once with a perfect (or near perfect) classifier, and once with a fallible and usually less expensive classifier. Chen (1979) incorporated Tenenbein’s work into a log linear model. Hochberg (1977), also using Tenenbein’s work, developed both a least squares approach and a combined maximum likelihood and least squares approach for analyzing multi-dimensional cross-classified data subject to error. Espeland and Hui (1987) considered a unified log-linear approach to incorporating phenotype misclassification data through various methods (e.g., re-sampling or sampling in replicate populations to determine misclassification rates) in the analysis of epidemiology data.

Gustafson et al. (2001) used a Bayes analysis to adjust for uncertainty in misclassification probabilities as prior information and thereby improve estimates of the odds-ratio in case/control data. In related work for the 2×3 χ^2 test of genetic association with a di-allelic locus, Rice and Holmans (2003) incorporated

genotyping misclassification rates into the calculation of confidence interval estimates for parameters such as genotypic relative risk.

Gordon and Ott (2001) considered the analysis of genetic data in the presence of genotyping error. They confirmed that: (i) that there is no increase in type I error for certain tests of genetic association; (ii) that point estimates of the frequencies of SNP alleles are biased in the presence of errors [as was shown by Bross (1954) for phenotype error]; and (iii) that errors lead to a loss in power to detect association between a disease and a locus. Recently, Gordon et al. (2002; 2003) produced a quantification of the loss in power for case/control studies of genetic association due to genotyping errors. This quantification may be determined using the PAWE webtool (see Electronic Database Information).

A critical but unanswered question is how one can use information about misclassification to improve statistical power for genetic tests of association using case/control data. It is the purpose of this work, therefore, to develop a statistical method that actually increases power for association (as compared to the standard method that considers only fallible data) in the presence of both phenotype and genotype error. Our method assumes that double-sample data is available on a sub-sample of the case and control individuals. It has three major advantages over the standard method that only considers the fallible data: its power can be equal to or greater than the standard method when total costs are equal for both methods; it provides unbiased estimates of population parameters; and it provides maximum likelihood estimates of the phenotype and genotype misclassification probabilities.

Methods

Double sample data

We assume that we have double-sample data for phenotypes and/or genotypes (as defined by Tenenbein (1970; 1972)) for a subset of individuals. That is, we have two methods of measurement for either phenotype and/or genotype. The first method, labeled “fallible”, is cost-effective and has a substantial misclassification rate. The second method, labeled “infallible”, is typically more expensive and/or not feasible for an entire study and has a 0 or negligibly small misclassification rate. For the example of SNP genotypes, the fallible method is a standard genotyping technology and an infallible method could be sequencing (F. De La Vega; personal communication). Similarly with Alzheimer’s disease, the fallible method is phenotype diagnosis using a clinical dementia instrument and the infallible method is observation of plaques and tangles at autopsy.

We recognize that any classification technology may have errors or that there is no such thing as a perfect classifier (Hochberg 1977). However, it is certainly reasonable that one method of classification may have substantially lower misclassification rates than another method, for example a “gold-standard”

method. For the purposes of this work, we consider the method with the lower misclassification rate to be the “infallible” classifier. Here and elsewhere, we use the terms true and infallible (respectively, observed and fallible) interchangeably.

True and observed data

Throughout this work, we assume that we have observed phenotype and genotype data for a total of n individuals. The phenotype data is classified into two categories, case or control. The genotype data is classified into k categories where $k = a(a + 1) / 2$ and a is the number of alleles at a locus. We note that our method could easily be applied to phenotype data for multiple-discrete categories and for haplotype data as well (see Summary and Discussion).

Notation

We provide all notation for the mathematics presented in this work in table 1. For all terms, the indices i and i' are either 0 (case) or 1 (control) and the integer indices j and j' range from 1 through k inclusive, where k is the number of genotypes. In table 1, every term is listed under the sub-heading corresponding to the sub-heading in the main text in which the term first appears.

Throughout this work, we use prime superscripts to distinguish true categories from observed categories. For example, i' refers to the true phenotype classification for an individual. Also, we use the superscript t to denote “true” (as compared with observed) when referring to either an event or a parameter. For example, the notation Y_i^t (table 1) represents the event that an individual’s true phenotype classification is i' , whereas the notation Y_i (table 1) represents the event that an individual’s observed phenotype classification is i . Similarly, the notation $p_{i',j}^t$ represents the true probability of the genotype j' for individuals with phenotype classification i' , whereas the notation p_{ij} represents the observed probability of the genotype j for individuals with phenotype classification i . With this notation, we may distinguish between the events Y_0^t and Y_0 and the probabilities p_{01}^t and p_{01} .

Table 1. Notation for all formulas presented throughout the text

Log-likelihood of observed data and likelihood ratio test statistics

$n_{i'ij}^{(1)}$ = Number of individuals with true phenotype category i' , observed phenotype category i , true genotype category j' and observed genotype category j . (These individuals are double-sampled on both phenotype and genotype)

$n_{i'ij}^{(2)}$ = Number of individuals with true phenotype category i' , observed phenotype category i , and observed genotype category j . (These individuals are double-sampled on phenotype only)

$$n_{i'+}^{(2)} = \sum_i n_{i'ij}^{(2)} .$$

$n_{ij'j}^{(3)}$ = Number of individuals with observed phenotype category i , true genotype category j' and observed genotype category j . (These individuals are double-sampled on genotype only)

$$n_{ij'+}^{(3)} = \sum_j n_{ij'j}^{(3)} .$$

$n_{ij}^{(4)}$ = Number of individuals with observed phenotype category i and observed genotype category j .

$$n_{i+}^{(4)} = \sum_j n_{ij}^{(4)}$$

$$n_{+j}^{(4)} = \sum_i n_{ij}^{(4)}$$

Y_i = Event that an individual has observed phenotype i , ($i = 0,1$).

$Y_{i'}$ = Event that an individual has true phenotype i' , ($i' = 0,1$).

X_j = Event that an individual has observed genotype j , $1 \leq j \leq k$.

$X_{j'}$ = Event that an individual has true genotype j' , $1 \leq j' \leq k$.

X_{ij} = Event that an individual has observed phenotype i , ($i = 0,1$) and observed genotype j , $1 \leq j \leq k$.

$X_{i'j'}$ = Event that an individual has true phenotype i' , ($i' = 0,1$) and true genotype j' , $1 \leq j' \leq k$.

$q_i = \Pr(Y_i)$ = Observed sampling frequency of phenotype i .

$q_{i'} = \Pr(Y_{i'})$ = True sampling frequency of phenotype i' .

$p_{ij} = \Pr(X_j | Y_i)$ = Observed population frequency of genotype j for individuals with observed phenotype i .

$p_{i,j}^t = \Pr(X_{j'}^t | Y_{i'}^t) =$ True population frequency of genotype j' for individuals with phenotype i' .

$p_{*j'}^t = \Pr(X_{j'}^t) =$ True population frequency of genotype j' under the null hypothesis that $p_{0j'}^t = p_{1j'}^t = p_{*j'}^t$.

$p_{*j} = \Pr(X_j) =$ Observed population frequency of genotype j under the null hypothesis that $p_{0j'}^t = p_{1j'}^t = p_{*j'}^t$.

Note: For each i' , $\sum_{j'} p_{i',j'}^t = 1$; For each i , $\sum_j p_{ij} = 1$; Also, $q_0 + q_1 = q_0^t + q_1^t = 1$.

$\pi_{i'i} = \Pr(Y_i | Y_{i'})$

$\theta_{j'j} = \Pr(X_j | X_{j'})$

Note: When $i' \neq i, j' \neq j$, these parameters are referred to as misclassification parameters (Tenenbein 1972; Gordon et al. 2002). We make use of the double-sample data structure to determine estimates of phenotype and genotype misclassification values $\pi_{i'i}$ and $\theta_{j'j}$. The misclassification parameter estimates $\hat{\theta}_{j'j}$ are $\hat{\theta}_{j'j} = m_{j'j} / m_{j'+}$ (see below). Similarly, we estimate the phenotype misclassification terms $\pi_{i'i}$ by $\hat{\pi}_{i'i} = w_{i'i} / w_{i'+}$ (see below).

$m_{j'j}$ ($1 \leq j, j' \leq k$) = The number of individuals that have been classified by the fallible method as genotype j and by the infallible method as genotype j' .

$m_{j'+} = \sum_j m_{j'j}$.

$w_{i'i}$ ($0 \leq i, i' \leq 1$) = The number of individuals that have been classified by the fallible method as phenotype i and by the infallible method as phenotype i' .

$w_{i'+} = \sum_i w_{i'i}$.

$\ln(L_{1,ae}) =$ Log-likelihood of data as represented in equation (4), where genotype frequencies $p_{i',j'}^t$ are allowed to differ among different phenotype classes. (i.e., $p_{0j'}^t$ is not necessarily equal to $p_{1j'}^t$ for every j')

$\ln(L_{0,ae})$ = Log-likelihood of data as represented in equation (4), where genotype frequencies $p_{i,j}^t$ are constrained to be equal among different phenotype classes. (i.e., $p_{0,j}^t = p_{1,j}^t = p_{*,j}^t$, for every j')

$\ln(L_{1,std})$ = Log-likelihood of data when not correcting for misclassification, where genotype frequencies p_{ij} are allowed to differ among different phenotype classes. (i.e., $p_{0,j}$ is not necessarily equal to $p_{1,j}$ for every j) (also see equation 1b)

$\ln(L_{0,std})$ = Log-likelihood of data when not correcting for misclassification, where genotype frequencies p_{ij} are constrained to be equal among different phenotype classes. (i.e., $p_{0,j} = p_{1,j} = p_{*,j}$ for every j) (also see equation 1b)

Simulations (also see tables 2 and 3)

ε = Genotype error probability parameter for all simulations; for a given value of

ε , all genotype classification probabilities satisfy $\theta_{j'j} = \begin{cases} 1 - \varepsilon, j' = j \\ \varepsilon / (k - 1), j' \neq j \end{cases}$.

P = Allele frequency of SNP minor allele B for simulations.

α = Significance level of test statistic.

C_p = Cost of phenotyping an individual with fallible phenotyping measure.

C_G = Cost of genotyping an individual with fallible genotyping measure.

C_p^{DS} = Cost of phenotyping an individual with infallible phenotyping measure.

C_G^{DS} = Cost of genotyping an individual with infallible genotyping measure.

N_A = Number of case individuals collected for LRT_{std} method.

N_U = Number of control individuals collected for LRT_{std} method.

N_A^{DS} = Number of case individuals collected for LRT_{ae} method.

N_U^{DS} = Number of control individuals collected for LRT_{ae} method.

(Note: a proportion of the individuals collected for the LRT_{ae} method are double-sampled on phenotype, genotype, or both)

d_p^{DS} = Proportion of individuals (either case or control) collected for the LRT_{ae} method who are double-sampled on phenotype.

d_G^{DS} = Proportion of individuals (either case or control) collected for the LRT_{ae} method who are double-sampled on genotype.

EM-algorithm estimates of true parameters

$p_{i,j}^{t(r)}$ = r^{th} step estimate of parameter $p_{i,j}^t$.

$p_{*,j}^{t(r)}$ = r^{th} step estimate of parameter $p_{*,j}^t$.

$q_i^{t(r)}$ = r^{th} step estimate of parameter q_i^t .

We use the Expectation-Maximization (EM) method developed by Dempster et al. (Dempster et al. 1977) to estimate these parameters.

$E[]$ = Expectation operator

$I()$ = Indicator function

$X_{i,j}^{(2)}$ = Event that an individual in Group 2 (proof of appendix proposition 1) has true phenotype i' and observed genotype j .

$X_{ij}^{(3)}$ = Event that an individual in Group 3 (proof of appendix proposition 1) has observed phenotype i and true genotype j' .

$X_{ij}^{(4)}$ = Event that an individual in Group 4 (proof of appendix proposition 1) has observed phenotype i and observed genotype j .

Appendix

$X_{i,j}^{t,a}$ = Event that individual a has true phenotype i' and true genotype j' .

$I()$ = Indicator function

X_{ij}^a = Event that individual a has observed phenotype i and observed genotype j .

$X_{i,j}^{(1)a}$ = Event that individual a in Group 1 has true phenotype i' and true genotype j' .

$X_{i,j}^{(2)a}$ = Event that individual a in Group 2 has true phenotype i' and observed genotype j .

$X_{ij}^{(3)a}$ = Event that individual a in Group 3 has observed phenotype i and true genotype j' .

$X_{ij}^{(4)a}$ = Event that individual a in Group 4 has observed phenotype i and observed genotype j .

Legend for Table 1. In this table, we present all notation used throughout the text. Every notational term is listed under the sub-heading corresponding to the sub-heading in the main text in which the term first appears. Where necessary, dashed lines are used to separate notation.

Log-likelihood of observed data and likelihood ratio test statistics

We compute the log-likelihood of the observed data under the null and alternative hypotheses, allowing for error. The null hypothesis we test is $H_0 : p_{0j'} = p_{1j'} = p_{*j'}$ for all genotypes j' . The alternative hypothesis is $H_1 : p_{0j'} \neq p_{1j'}$ for at least one j' . Under either hypothesis, we have, by definition, the log-likelihood of the data given by:

$$\begin{aligned} \ln(L_{ae}) = & \sum_i \sum_{i'} \sum_j \sum_{j'} n_{i'ij'j}^{(1)} \ln[\Pr(Y_i, Y_{i'}, X_j, X_{j'})] + \sum_i \sum_{i'} \sum_j n_{i'ij}^{(2)} \ln[\Pr(Y_i, Y_{i'}, X_j)] \\ & + \sum_i \sum_j \sum_{j'} n_{ij'j}^{(3)} \ln[\Pr(Y_i, X_j, X_{j'})] + \sum_i \sum_j n_{ij}^{(4)} \ln[\Pr(Y_i, X_j)], \end{aligned} \quad (1a)$$

where the notation $\Pr(A, B, C, \dots)$ is the probability of observing event A and event B and event C and so forth and $n_{i'ij'j}^{(1)}, n_{i'ij}^{(2)}, n_{ij'j}^{(3)}, n_{ij}^{(4)}$ represent the counts for different categories of double-sample information (see table 1). For example, $n_{i'ij'j}^{(1)}$ is the number of individuals who have been double-sampled on both phenotype and genotype and who have true phenotype classification i' , observed phenotype classification i , true genotype classification j' , and observed genotype classification j . In equation (1a), the subscripts i, i' run over all phenotype classifications ($0 \leq i, i' \leq 1$) and the subscripts j, j' run over all genotype classifications ($1 \leq j, j' \leq k$).

When there are no double-sample data or when we assume that there is no error in the data, equation (1a) reduces to:

$$\begin{aligned} \ln(L_{std}) = & \sum_i \sum_j n_{ij}^{(4)} \ln[\Pr(Y_i, X_j)] \\ = & \sum_i \sum_j n_{ij}^{(4)} \ln(p_{ij} q_i) \\ = & \sum_i \sum_j n_{ij}^{(4)} [\ln(p_{ij}) + \ln(q_i)]. \end{aligned} \quad (1b)$$

A key assumption in our work is that, conditional on the underlying true data, the observed data are independent. That is, the measurement process for phenotype is

independent of the measurement process for genotype, so that $\Pr(X_j, Y_i | X_{j'}^t, Y_{i'}^t) = \Pr(X_j | X_{j'}^t) \Pr(Y_i | Y_{i'}^t)$. It follows that:

$$\begin{aligned} \Pr(Y_i, Y_{i'}^t, X_j, X_{j'}^t) &= \Pr(X_j, Y_i | X_{j'}^t, Y_{i'}^t) \Pr(X_{j'}^t, Y_{i'}^t) \\ &= \Pr(X_j | X_{j'}^t) \Pr(Y_i | Y_{i'}^t) \Pr(X_{j'}^t | Y_{i'}^t) \Pr(Y_{i'}^t) \\ &= \theta_{j',j} \pi_{i,i} p_{i',j}^t q_{i'}^t. \end{aligned} \quad (2)$$

Using equation (2) and the fact that

$$\begin{aligned} \Pr(Y_{i'}^t, Y_i, X_j) &= \sum_{j'} \Pr(Y_{i'}^t, Y_i, X_j, X_{j'}^t), \\ \Pr(Y_i, X_j, X_{j'}^t) &= \sum_{i'} \Pr(Y_{i'}^t, Y_i, X_j, X_{j'}^t), \\ \Pr(Y_i, X_j) &= \sum_{i'} \sum_{j'} \Pr(Y_{i'}^t, Y_i, X_j, X_{j'}^t), \end{aligned} \quad (3)$$

we may rewrite the log-likelihood (1a) as:

$$\begin{aligned} \ln(L_{ae}) &= \sum_i \sum_{i'} \sum_j \sum_{j'} n_{i'ij'j}^{(1)} \ln[\theta_{j',j} \pi_{i,i} p_{i',j}^t q_{i'}^t] + \sum_i \sum_{i'} \sum_j n_{i'ij}^{(2)} \ln[\sum_{v'} \theta_{v',j} \pi_{i,i} p_{i',v}^t q_{i'}^t] \\ &+ \sum_i \sum_{j'} \sum_j n_{ij'j}^{(3)} \ln[\sum_{u'} \theta_{j',j} \pi_{u,i} p_{u',j}^t q_{u'}^t] + \sum_i \sum_j n_{ij}^{(4)} \ln[\sum_{u'} \sum_{v'} \theta_{v',j} \pi_{u,i} p_{u',v}^t q_{u'}^t], \end{aligned} \quad (4)$$

where we have replaced the indices i' and j' in the sums (3) by the indices u' and v' respectively in equation (4) for purposes of clarity. From equation (4) we can determine the log-likelihood of the data under H_1 using the $r+1^{\text{st}}$ step EM algorithm estimates of $p_{i',j}^t$ and $q_{i'}^t$. Similarly, we can determine the log-likelihood of the data under H_0 using the $r+1^{\text{st}}$ step EM algorithm estimates of $p_{*,j}^t$ and $q_{i'}^t$. We comment that the $r+1^{\text{st}}$ step estimates of $q_{i'}^t$ may differ under the null and alternative hypotheses.

Formulas for $p_{i',j}^{t(r+1)}$, $p_{*,j}^{t(r+1)}$, and $q_{i'}^{t(r+1)}$ (the $r+1^{\text{st}}$ step estimates of $p_{i',j}^t$, $p_{*,j}^t$, and $q_{i'}^t$ respectively) under the alternative and null hypotheses are presented in the Results section (formulas (7)) and are derived in the appendix (equation (A.1)). It follows from the equation (4) that the log-likelihoods $\ln(L_{0,ae})$ and $\ln(L_{1,ae})$ (equation (1a)) are completely determined by misclassification parameters $\theta_{j',j}$ and $\pi_{i,i}$, the true parameters $p_{i',j}^t$, $p_{*,j}^t$, $q_{i'}^t$, and sample counts $(n_{i'ij'j}^{(1)}, n_{i'ij}^{(2)}, n_{ij'j}^{(3)}, n_{ij}^{(4)})$. In the previous sentence, $\ln(L_{0,ae})$ (respectively, $\ln(L_{1,ae})$) refers to the situation under the null (respectively alternative) hypothesis, where the terms $p_{i',j}^t$ in equation (4) are replaced (respectively, not replaced) by $p_{*,j}^t$. Our test of H_1 versus H_0 is a likelihood ratio

test (Kendall et al. 1994), which we call the *likelihood ratio test allowing for error*, or LRT_{ae} . It is given by

$$LRT_{ae} = 2[\ln(L_{1,ae}) - \ln(L_{0,ae})], \quad (5a)$$

where $\ln(L_{1,ae})$ and $\ln(L_{0,ae})$ are determined using equation (4) with the $r+1^{\text{st}}$ step estimates of the various parameters. Asymptotically, LRT_{ae} is distributed as χ^2_{k-1} , where the degrees of freedom (df) are $k - 1$ for a marker locus with k genotypes, or for a set of k observed haplotype pairs (see Summary and Discussion). For small samples or in situations where the asymptotic distribution may not hold, we can compute p-values via permutation (see Summary and Discussion).

To compare the performance of our test statistic that corrects for misclassification with the standard likelihood ratio test, denoted LRT_{std} , that does not make any correction, we compute log-likelihoods solely from the observed data. That is,

$$LRT_{std} = 2[\ln(L_{1,std}) - \ln(L_{0,std})], \quad (5b)$$

where the log-likelihoods under the null and alternative hypotheses are computed using the estimates $\hat{p}_{ij} = n_{ij}^{(4)} / n_{i+}^{(4)}$, $\hat{p}_{*j} = (n_{0j}^{(4)} + n_{1j}^{(4)}) / n$, $\hat{q}_i = n_{i+}^{(4)} / n$, ($n_{i+}^{(4)} = \sum_j n_{ij}^{(4)}$)

that are then substituted into equation (1b) (Rice and Holmans 2003). When there is no correction for misclassification, there is no need to compute \hat{q}_i under both the null and alternative hypothesis, as the terms with \hat{q}_i will cancel from the difference of the log-likelihoods (equation (5b)).

Simulations

Parameter settings, numbers of replicates, tolerance of EM estimates

We consider two types of simulations: (i) null and (ii) power comparison assuming a fixed budget (cost/benefits). In the null simulations, case and control genotype frequencies are equal. In the power simulations, there is at least one genotype for which case and control genotype frequencies differ. In table 2, we present settings considered for the sample size, phenotype misclassification probabilities, genotype misclassification probabilities (ε), the proportion of individuals who are double-sampled on phenotype (d_p^{DS}), and the proportion of individuals who are double-sampled on genotype (d_G^{DS}). Case and control genotype frequencies for the null and power comparison scenarios are provided below (see Methods – Empirical type I error rate of LRT_{ae} and table 3). In all of our simulations, we consider the following probabilities for genotype errors:

$$\theta_{j'j} = \begin{cases} 1 - \varepsilon, & j' = j \\ \varepsilon / (k - 1), & j' \neq j \end{cases}$$

While this error model is not the most general possible, even for SNPs (e.g., Kang et al. 2004), we choose it because it reduces the number of error model parameters to 1, even for multi-allele loci.

Table 2. Parameter settings for variables in simulations

Parameter	Description	Values considered in simulations
N_A, N_U	Case, Control sample sizes	500, 1000
π_{10}, π_{01}	Phenotype misclassification probabilities	0.25, 0.5
ε	Genotype misclassification parameter	0.01, 0.05
d_P^{DS}	Phenotype double-sample proportion	0.25, 0.5
d_G^{DS}	Genotype double-sample proportion	0.25, 0.5
Additional parameters for power (cost/benefits) analysis		
C_P	Cost of phenotyping an individual with fallible measure	1, 10
C_G	Cost of genotyping an individual with fallible measure	1
C_P^{DS} / C_P	(Cost to phenotype with infallible measure)/(Cost to phenotype with fallible measure)	5, 25
C_G^{DS} / C_G	(Cost to genotype with infallible measure)/(Cost to genotype with fallible measure)	5, 25

Legend for Table 2. In this table, we present settings for all parameters considered in the null and power (Cost/Benefits) simulations for SNP and multi-allele data. We consider a 2^k factorial design (Box et al. 1978), where $k = 6$ for null simulations, $k = 10$ for SNP power simulations, and $k = 9$ for multi-allele power simulations. This last value of k comes from the fact that there is only one set of genotype frequencies (table 2) for the multi-allele power simulations. Also, this k is not to be confused with the number of genotypes at a locus.

For any vector of simulation parameters, we simulate 10,000 replicates of data. Also, we stop our EM algorithm maximization when $|v^{(r+1)} - v^{(r)}| < 10^{-9}$ for all parameters v , where $v^{(r)}$ is the r^{th} -step estimate of the parameter.

Table 3. Case and control genotype frequencies considered for null and power (cost/benefits) simulations

Marker Locus	Marker Locus Genotype		
SNP	<i>AA</i>	<i>AB</i>	<i>BB</i>
Case genotype frequencies	$(1-P)^2 + 0.05$	$2P(1-P) - 0.10$	$P^2 + 0.05$
Control genotype frequencies	$(1-P)^2$	$2P(1-P)$	P^2
Multi-allele	<i>ii</i> (homozygote)	<i>ij</i> (heterozygote)	
Control genotype frequencies	P_i^2	$2P_iP_j$	
Case genotype frequencies	$P_i^2 + 0.06$	$2P_iP_j - 0.04$	

Legend for Table 3. This table reports the case and control genotype frequencies for the null and power (cost/benefits) simulations. Genotype frequencies for either group are given by the appropriately labeled rows. For SNP data, the settings of P considered are: $P = 0.2, 0.5$. For null simulations, both case and control genotype frequencies are provided in the row corresponding to control genotype frequencies. For power simulations, case and control genotype frequencies differ and are provided in the appropriately labeled rows.

Empirical type I error rate of LRT_{ae}

(a) SNPs

In the first set of simulations, we consider a SNP marker locus with three genotypes labeled *AA*, *AB*, and *BB* whose frequencies are in Hardy-Weinberg equilibrium (HWE). The true genotype frequencies p_{ij}^t , for both true cases and true controls are given by the values in table 3.

Empirical type I error rates for each simulation are defined as the proportion of replicates out of 10,000 that exceed the inverse of the one-tailed probability α of the central χ^2 distribution with 2 degrees of freedom; that is, the proportion that exceed the analytic cutoff for various significance levels. In our simulations we compute type I error rates at three different significant levels: $\alpha = 10\%$, 5% , 1% .

(b) Multi-allele data

We perform simulations assuming we have a locus with four alleles (labeled *A*, *B*, *C*, and *D*) and therefore ten genotypes. We present genotype frequencies for genotypes in the following order: *AA*, *AB*, *AC*, *AD*, *BB*, *BC*, *BD*, *CC*, *CD*, *DD*. The genotype frequency settings for each simulation are: 0.0625, 0.125, 0.125, 0.125, 0.0625, 0.125, 0.125, 0.0625, 0.125, and 0.0625. These are the frequencies under the assumption the each allele has frequency 0.25 and that the locus is in HWE. We also use these frequencies for the control population in our power simulations (see below). As with SNPs, empirical type I error rates for each

simulation are defined as the proportion of replicates out of 10,000 that exceed the inverse of the one-tailed probability α of the central χ^2 distribution with 9 degrees of freedom; that is, the proportion that exceed the analytic cutoff for various significance levels. In our simulations we compute type I error rates at three different significant levels: $\alpha = 10\%$, 5% , 1% .

Power comparison assuming a fixed budget (Cost/Benefits Analysis)

The LRT_{ae} statistic can be shown to be more powerful than LRT_{std} statistic when both statistics are applied to the same fallible data and double-sample data is available (results not shown). However, as was documented by Rice and Holmans (2003) for genotyping error, that power gain is due exclusively to the double-sample data. The question we seek to answer is thus whether this increase in power holds when budget is fixed for both statistics. That is, we consider power comparison for the two statistics when the total cost of phenotyping and genotyping some set of cases N_A and controls N_U with the fallible method and applying the LRT_{std} statistic is the same as the total cost of phenotyping and genotyping a portion of these cases and controls, N_A^{DS} and N_U^{DS} , obtaining phenotype and genotype double-samples on a subset, and applying the LRT_{ae} statistic. It is straightforward to show that the sample sizes for the LRT_{ae} method are given by the formulas:

$$\begin{aligned} N_A^{DS} &= \alpha_{DS} N_A, \\ N_U^{DS} &= \alpha_{DS} N_U, \end{aligned} \tag{6}$$

where $\alpha_{DS} = (C_P + C_G) / (C_P + C_G + d_P^{DS} C_P^{DS} + d_G^{DS} C_G^{DS})$ (notation defined in table 1). Here we make the additional assumption that the ratios N_A^{DS} / N_A and N_U^{DS} / N_U are equal. While this assumption is not necessary, it does appreciably reduce the number of simulations that we need perform. Note that α_{DS} is always less than 1, so that sample sizes analyzed with the fallible measures for the LRT_{ae} method will always be less than those for the LRT_{std} method.

We consider parameter settings for each of the variables $N_A, N_U, \pi_{01}, \pi_{10}, \varepsilon, C_P, C_G, C_P^{DS}, C_G^{DS}, d_P^{DS}, d_G^{DS}$ as given in table 2. Note that in this table, all costs are determined relative to the cost of genotyping with the fallible method C_G . Case and control genotype frequencies considered are given in table 3. As noted above, for each setting of the parameters in tables 2 and 3, we simulate 10,000 replicates of data and compute the proportion of replicates for which either statistic exceeds the analytic cutoff at the 10%, 5%, and 1% levels of significance. Each of these values is the *simulation power* at a given level of significance. We then compute the difference in simulation power between the two statistics for

each level of significance. A positive difference means that the simulation power for the LRT_{ae} method is greater than the simulation power for the LRT_{std} method for a given set of simulation parameters. A negative difference means the reverse.

We add a comment here about the values chosen for phenotype misclassification (π_{01} and π_{10}). We choose values of 0.25 and 0.5. While these values may be considered to be somewhat large, we note that they were chosen based on published results of phenotype misclassification in genetics studies (Press et al. 1994; Burd et al. 2001) (see Summary and Discussion for further comment on these parameter settings).

Estimation of misclassification and population frequency parameters

We document that the estimates of the parameter estimates such as $p_{i,j}^t$ and q_i^t are unbiased by reporting the averages and variances for each power simulation. Regarding the genotype error parameter $\theta_{j'j}$, we compute the average of the $k(k-1)$ terms $\theta_{j'j}$ ($j' \neq j$) for each replicate, and then compute the mean and variance of these averages, sampled over all 10,000 replicates. This procedure gives us the sampling distribution of the estimate of the parameter $\varepsilon/(k-1)$. Similarly, we find the sampling distribution of the estimate of the phenotype error parameters π_{01} and π_{10} .

Example - ApoE and Alzheimer's Disease

We apply our LRT_{ae} method to ApoE genotype data (Chromosome 19) in subjects ascertained for being affected or unaffected with late-onset Alzheimer's Disease (LOAD). These data come from previous Alzheimer's disease case/control studies (Sheu et al. 1999; Brown et al. 2004a; Brown et al. 2004b). In most populations there are three alleles at the ApoE locus. Conventionally, they are denoted $\varepsilon_2, \varepsilon_3$, and ε_4 and we label them 2, 3, and 4 respectively from this point forward. In a well known and often replicated association finding, every copy of the 4 allele in a person's genotype increases that person's risk of getting LOAD by a factor of 2.5-3 (Corder et al. 1993).

Genotypes of each patient sample were determined at two different time intervals. Initially genotyping was performed using polymerase chain reaction (PCR) followed by cleavage with the restriction enzyme HhaI (Hixson and Vernier 1990). The genotype data obtained the second time were determined using a more accurate double-digestion genotyping procedure (Zivelin et al. 1997). All conflicts between the first and second method were re-sampled to check for accuracy. In every instance, the second method was found to be correct. Thus, we consider the genotype data determined from the second time interval to be the true genotype data for our analyses.

In our application, we have 173 case individuals who are affected with LOAD and 118 control individuals who are not affected. All individuals are matched on age and ethnicity and all have ApoE genotypes determined from the first time interval. We randomly selected 119 individuals independent of disease status (approximately 40%) to be double-sampled for ApoE genotypes, in the sense that we have their genotypes from the second time interval as well. There is no double-sampling of phenotype in this example. We compute the LRT_{ae} and LRT_{std} values, genotype frequency estimates for each statistic, and estimates of the genotype error probabilities for these data.

Results

EM-algorithm estimates of true parameters

Our first results are formulas for the $r + 1^{\text{st}}$ step EM-algorithm estimates of the parameters $p_{i'j'}^t, p_{*j'}^t, q_{i'}^t$. From our computations (see Appendix), we derive the formulas:

$$\begin{aligned} p_{i'j'}^{t(r+1)} &= n_{i'j'}^{(r)} / n_{i'+}^{(r)}, \\ q_{i'}^{t(r+1)} &= n_{i'+}^{(r)} / n, \\ p_{*j'}^{t(r+1)} &= n_{+j'}^{(r)} / n, \end{aligned} \tag{7}$$

where $n_{i'j'}^{(r)} = n_{1i'j'}^{(r)} + n_{2i'j'}^{(r)} + n_{3i'j'}^{(r)} + n_{4i'j'}^{(r)}$, $n_{i'+}^{(r)} = \sum_j n_{i'j'}^{(r)}$, and

$$n_{1i'j'}^{(r)} = \sum_i \sum_j n_{i'ij'j}^{(1)},$$

$$n_{2i'j'}^{(r)} = \sum_j n_{i'+j}^{(2)} E[I(X_{i'j'}^t) | X_{i'j}^{(2)}] = \sum_j n_{i'+j}^{(2)} \left(\frac{\theta_{j'j} p_{i'j'}^{t(r)}}{\sum_{v'} \theta_{v'j} p_{i'v'}^{t(r)}} \right),$$

$$n_{3i'j'}^{(r)} = \sum_i n_{ij'+}^{(3)} E[I(X_{i'j'}^t) | X_{ij'+}^{(3)}] = \sum_i n_{ij'+}^{(3)} \left(\frac{\pi_{i'i} p_{i'j'}^{t(r)} q_{i'}^{t(r)}}{\sum_{u'} \pi_{u'i} p_{u'j'}^{t(r)} q_{u'}^{t(r)}} \right),$$

$$n_{4i'j'}^{(r)} = \sum_i \sum_j n_{ij}^{(4)} E[I(X_{i'j'}^t) | X_{ij}^{(4)}] = \sum_i \sum_j n_{ij}^{(4)} \left(\frac{\theta_{j'j} \pi_{i'i} p_{i'j'}^{t(r)} q_{i'}^{t(r)}}{\sum_{u'} \sum_{v'} \theta_{v'j} \pi_{u'i} p_{u'v'}^{t(r)} q_{u'}^{t(r)}} \right).$$

In these equations, $n_{i'+j}^{(2)} = \sum_i n_{i'ij}^{(2)}$, $n_{ij'+}^{(3)} = \sum_j n_{ij'j}^{(3)}$, $E[]$ is the expectation operator, $I()$

is the indicator function, $X_{i'j}^{(2)}$ is the event that an individual has true phenotype i' and observed genotype j (for individuals in ‘‘Group 2’’ – see proof of proposition 1

in appendix), $X_{ij'}^{(3)}$ is the event that an individual has observed phenotype i and true genotype j' (for individuals in “Group 3”), and $X_{ij}^{(4)}$ is the event that an individual has observed phenotype i and observed genotype j (individuals in “Group 4”). Note that, if every individual is double-sampled for both phenotype and genotype, then we have $n_{i'j'}^{(r)} = n_{1i'j'}$, i.e., perfect classification for every observation, and the formulas (7) reduce to the formulas

$$\hat{p}_{i'j'} = n_{i'ij'j}^{(1)} / n_{i'+j'+}, \hat{p}_{*j'} = (n_{0+j'+}^{(1)} + n_{1+j'+}^{(1)}) / n, \hat{q}_{i'} = n_{i'+j'+} / n, \quad (n_{i'+j'+}^{(1)} = \sum_i \sum_{j'} \sum_j n_{i'ij'j}^{(1)},$$

$$n_{i'+j'+}^{(1)} = \sum_i \sum_j n_{i'ij'j}^{(1)} = n_{1i'j'})$$

that we use with LRT_{std} statistic (Methods - Log-likelihood of observed data and likelihood ratio test statistics). For our simulations, most parameter estimates converged within 10 steps (results not shown).

Simulations

Empirical type I error rates

Standard asymptotic theory (Kendall et al. 1994) states that the asymptotic null distribution of LRT_{ae} is central χ^2 with $k-1$ df. Based on the simulation settings provided in Methods, we confirmed this result (results not shown). In particular, every median observed type I error rate contains the correct significance level in its 95% confidence interval, as computed using the method of confidence intervals implemented in the BINOM software (Electronic Database Information).

Results – Power comparison for a fixed budget (Cost/Benefits Analysis)

We present summary results of our power comparison in table 4. In that table, we report each of the three quartile values (1st, median, 3rd) (Tukey 1977) for difference in simulation power at the different levels of significance and different settings of the genotype error parameter ε (tables 2 and 4), as well as the minimum, 10th percentile, 90th percentile, and maximum difference values. For each setting of ε and each significance level (10%, 5%, 1%), these differences are computed over all remaining simulation parameters considered in tables 2 and 3. There were a total of $2^{10} = 2048$ data points for the SNP simulations and $2^9 = 1024$ data points for the Multi-allele simulations.

Studying this table, we see that differences in simulation power can be substantial. For SNPs, the minimum difference is -0.586 . This value occurs for the parameter settings: Significance level = 5%, $N_A = N_U = 1000$, $P = 0.2$, $C_P = C_G = 1$, $C_P^{DS} / C_P = C_G^{DS} / C_G = 25$, $d_P^{DS} = 0.25$, $d_G^{DS} = 0.5$, $\pi_{01} = \pi_{10} = 0.25$, and $\varepsilon = 0.01$ (full results not shown).

Table 4. Percentiles of simulation power differences $LRT_{ae} - LRT_{std}$ at various significance level thresholds for two different genotype error rates

Genotype	Percentile	$\varepsilon = 0.01$			$\varepsilon = 0.05$		
		Significance Level			Significance Level		
		10%	5%	1%	10%	5%	1%
SNP	Min	-0.566	-0.590	-0.510	-0.505	-0.500	-0.390
	10%	-0.263	-0.220	-0.160	-0.235	-0.190	-0.110
	1 st Q	-0.072	-0.060	-0.040	-0.052	-0.040	-0.030
	Median	0.007	0.005	0.001	0.016	0.011	0.004
	3 rd Q	0.144	0.113	0.047	0.136	0.099	0.040
	90%	0.251	0.213	0.105	0.235	0.185	0.091
	Max	0.645	0.579	0.384	0.571	0.509	0.307
Multi-allele	Min	-0.580	-0.657	-0.816	-0.570	-0.668	-0.790
	10%	-0.430	-0.506	-0.517	-0.440	-0.492	-0.465
	1 st Q	-0.080	-0.100	-0.094	-0.070	-0.085	-0.080
	Median	0.062	0.055	0.029	0.070	0.057	0.029
	3 rd Q	0.328	0.316	0.210	0.324	0.307	0.196
	90%	0.456	0.432	0.353	0.442	0.436	0.346
	Max	0.863	0.878	0.790	0.846	0.852	0.745

Legend for Table 4. Here we present simulation power differences corresponding to the minimum (Min), 10th percentile (10%), first quartile (1st Q), median, 3rd quartile (3rd Q), 90th percentile (90%) and maximum (Max) for the two settings of the genotype error parameter ε (0.01, 0.05) (table 2), three significance levels (10%, 5%, and 1%) and two types of genotype data (SNPs, Multi-allele).

The maximum difference of 0.645 occurs for parameter settings: Significance level = 10%, $N_A = N_U = 1000$, $P = 0.2$, $C_P = C_G = 1$, $C_P^{DS} / C_P = C_G^{DS} / C_G = 5$, $d_P^{DS} = 0.5$, $d_G^{DS} = 0.25$, $\pi_{01} = \pi_{10} = 0.5$, and $\varepsilon = 0.01$ (full results not shown). In general, the percentiles for each significance level do not seem to vary much with different settings of the genotype error parameter setting.

For our multi-allele simulations, we observe even greater differences in simulation power. The minimum difference of -0.816 occurs for the parameter settings: Significance level = 1%, $N_A = N_U = 1000$, $C_P = C_G = 1$, $C_P^{DS} / C_P = C_G^{DS} / C_G = 25$, $d_P^{DS} = 0.25$, $d_G^{DS} = 0.5$, $\pi_{01} = \pi_{10} = 0.25$, and $\varepsilon = 0.01$ (full results not shown). The maximum difference of 0.878 occurs for parameter settings: Significance level = 10%, $N_A = N_U = 1000$, $C_P = C_G = 1$, $C_P^{DS} / C_P = C_G^{DS} / C_G = 5$, $d_P^{DS} = 0.5$, $d_G^{DS} = 0.25$, $\pi_{01} = \pi_{10} = 0.5$, and $\varepsilon = 0.01$ (full results not shown).

We present histogram plots of the simulation power differences for the two types of loci (SNP or multi-allele) and three different significance levels (10%, 5%, 1%) in figures 1-3. In these figures, we merge results for the $\varepsilon = 0.01$ and $\varepsilon = 0.05$ settings because the results of table 4 suggest that differences in simulation power do not depend heavily upon settings of ε . All plots are produced using S-Plus Version 6.1 (Academic Site Edition) software (see Electronic Database Information). Perhaps the most informative of these plots is the one for multi-allele data at the 10% significance level (figure 1). We see clearly a mixture of at least three distributions of simulation power differences. When analyzing the raw data that generated that figure, we can assign the left most distribution to those simulations for which $\pi_{01} = \pi_{10} = 0.25$; that is, every simulation power difference that is less than -0.21 has $\pi_{01} = \pi_{10} = 0.25$. Every simulation power difference that is greater than 0.51 has $\pi_{01} = \pi_{10} = 0.5$. In general, every positive simulation power difference at the 10% significance level has at least one of π_{01} or π_{10} equal to 0.5 . A regression analysis (results not shown) indicates that the most significant factors in determining simulation power difference for the multi-allele data are (in order of importance): π_{01} , π_{10} , and C_p^{DS} / C_p , with the misclassification probabilities being of equal importance. Intuitively, it is clear that the higher the misclassification probability, the more benefit there is from performing double-sampling. Furthermore, because the phenotype misclassification probabilities are considerably larger than the genotype misclassification probabilities (table 2), they affect overall power more. Finally, the ratio C_p^{DS} / C_p is vitally important because it largely determines the number of samples available for the LRT_{ae} method, which directly affects power.

Estimation of misclassification and population frequency parameters

The results of our misclassification and population frequency parameter estimation are that the LRT_{ae} sample mean parameter estimates are highly accurate (maximum difference of 0.001 between mean parameter estimate and true parameter value) with sample variances for all estimates are on the order of 10^{-8} to 10^{-7} , indicating that the 99% sampling margin of error is less than 0.001 (full results not shown).

It should be noted that, as documented in previous papers (Bross 1954; Gordon et al. 2002), both the sample mean of the observed genotype frequencies p_{ij} and the sample frequency of the observed cases and controls q_i for the LRT_{std} statistic were biased away from the true values (results not shown).

Figure 1. Histograms of the simulation power differences $LRT_{ae} - LRT_{std}$ at 10% significance level for SNP and Multi-allele (Multi) genotypes

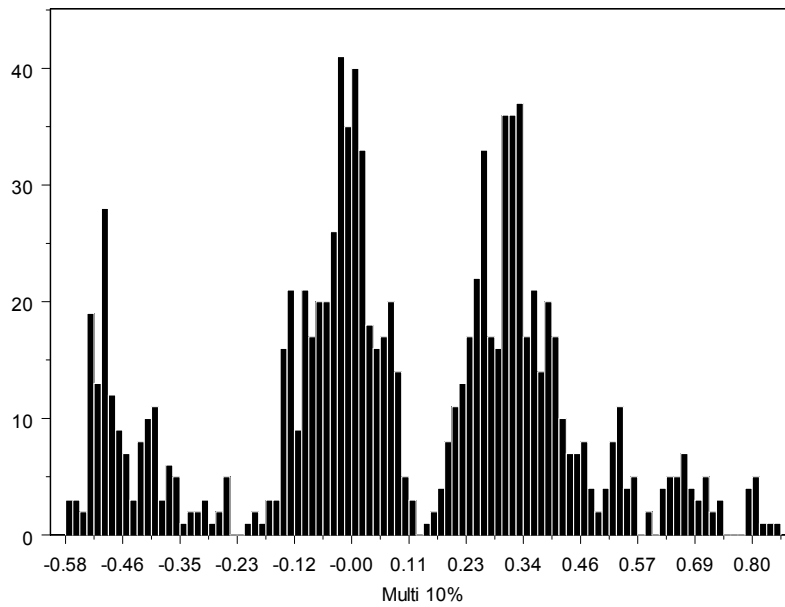
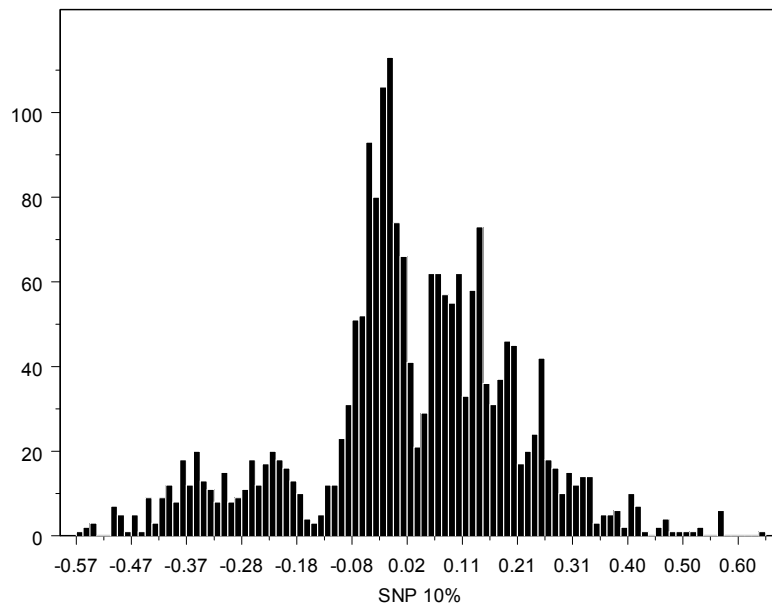


Figure 2. Histograms of the simulation power differences $LRT_{ae} - LRT_{std}$ at 5% significance level for SNP and Multi-allele (Multi) genotypes

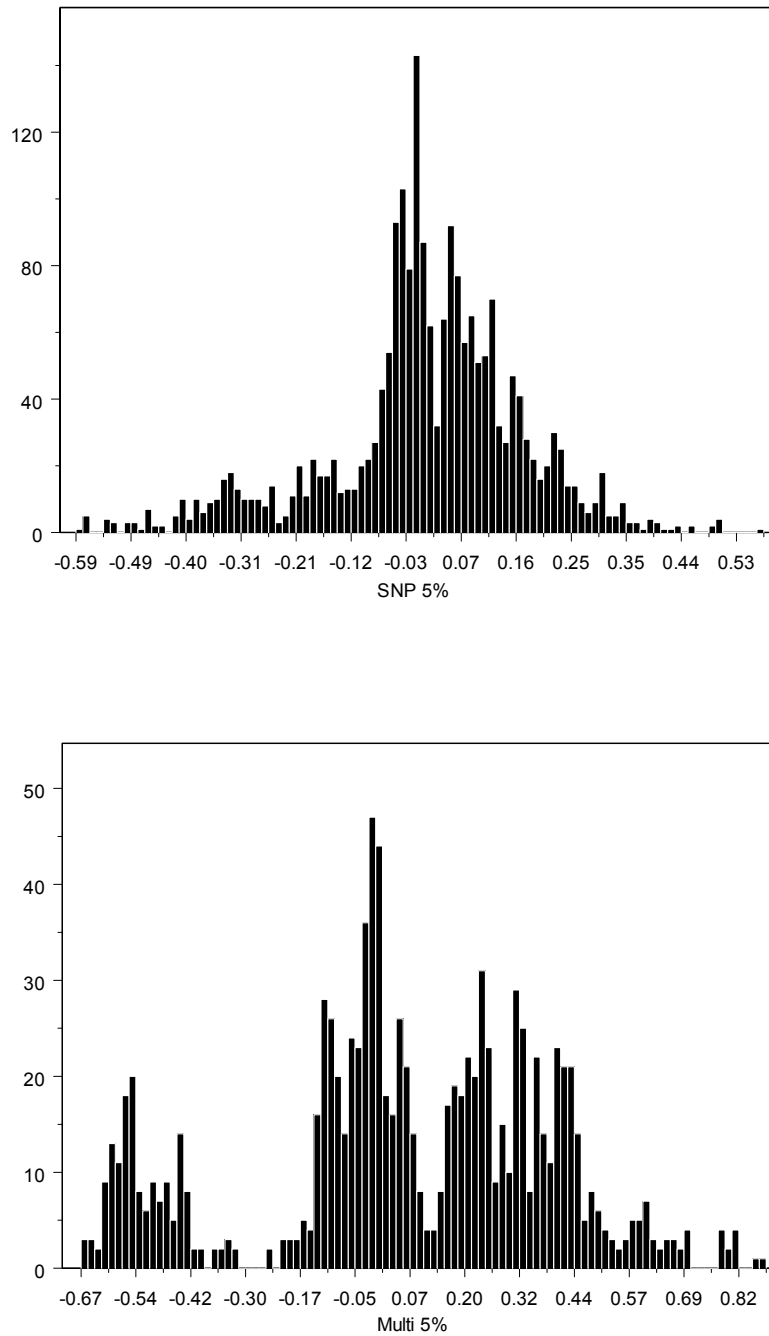
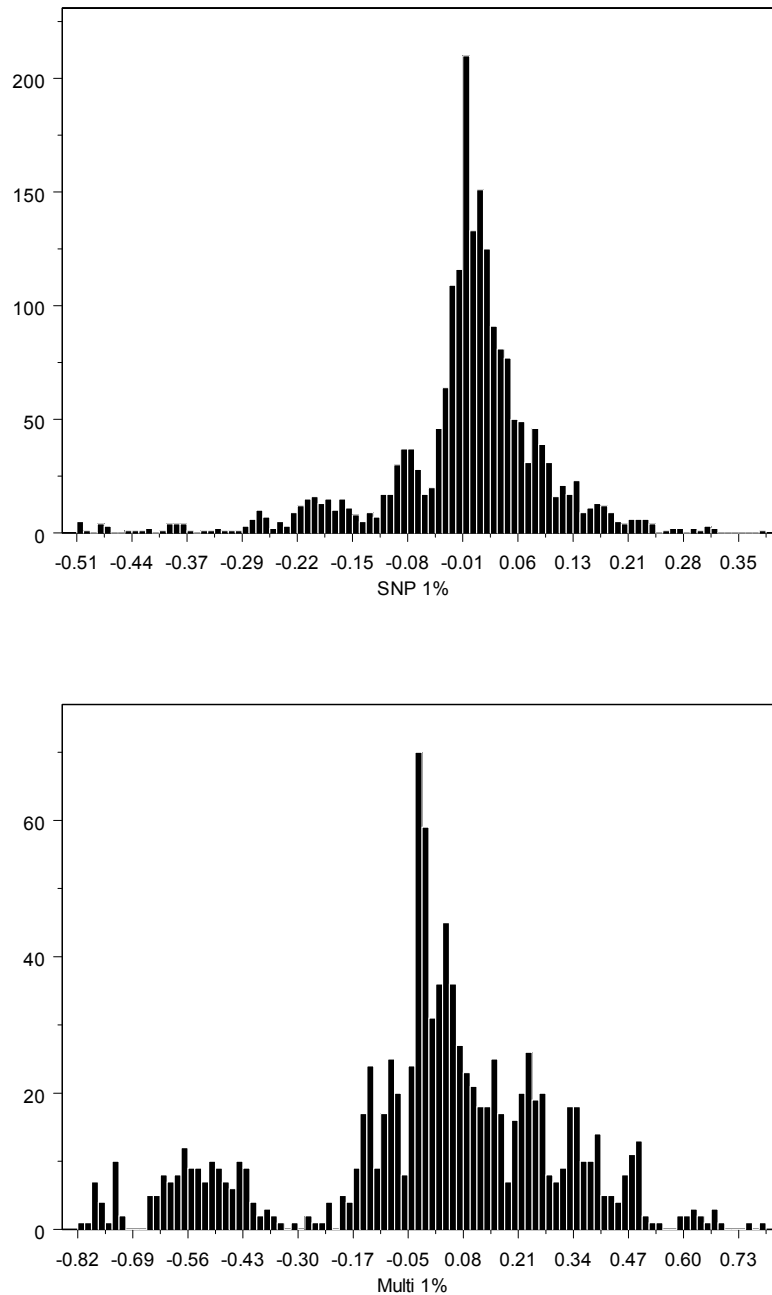


Figure 3. Histograms of the simulation power differences $LRT_{ae} - LRT_{std}$ at 1% significance level for SNP and Multi-allele (Multi) genotypes



Legend for Figures 1-3. In these figures, all histograms in the top half of the figure refer to SNP data and all histograms in bottom half refer to multi-allele data. The horizontal axes (SNP x %, Multi x %) are the simulation power differences at the x level of significance ($x = 10, 5, \text{ or } 1$) and the vertical axes are the counts of simulations whose simulation power differences lie within a given bin (100 bins for each histogram). Figures 1, 2, and 3 are histograms for 10%, 5%, and 1% significance levels, respectively.

Example - ApoE and Alzheimer's Disease

For the LOAD data, the LRT_{ae} statistic is 26.9787 and the LRT_{std} statistic is 19.4040. Each method is asymptotically distributed as χ^2 with 5 df under the null. Therefore, the corresponding asymptotic p-values are 5.8×10^{-5} and 1.6×10^{-3} , respectively.

By studying table 5, we may deduce the reason for the increased significance with the LRT_{ae} method. In that table, we see that the genotype carrying the greatest risk for causing LOAD, 44, is misclassified 67% of the time ($n = 3$) as the 33 genotype, based on the double-sample data. Obviously, this misclassification rate is based on too small a sample to have any degree of accuracy. Still, the effect of this misclassification is that, for the LRT_{std} method, the difference between the case and control genotype frequencies for 44 is $0.052 - 0.009 = 0.043$, as opposed to a difference of $0.096 - 0 = 0.096$ for the LRT_{ae} method. Similarly, the 34 genotype is misclassified 5.26% of the time as the 33 genotype, and the differences in the 34 genotype frequency between cases and controls for the LRT_{std} and LRT_{ae} methods are $0.318 - 0.144 = 0.174$ and $0.344 - 0.159 = 0.185$, respectively. Thus, for two risk genotypes, the effect of genotype errors is to decrease the genotype frequency difference between cases and controls with the concomitant loss of statistical power.

Table 5. EM-algorithm estimates of true genotype and phenotype misclassification parameters for ApoE LOAD data

	$\theta_{j'j}$	Observed Genotype						n
		22	23	24	33	34	44	
True Genotype	22	0.500	0.500	0.000	0.000	0.000	0.000	2
	23	0.091	0.727	0.000	0.091	0.000	0.091	11
	24	0.000	0.000	1.000	0.000	0.000	0.000	4
	33	0.000	0.016	0.000	0.984	0.000	0.000	61
	34	0.000	0.000	0.026	0.053	0.921	0.000	38
	44	0.000	0.000	0.000	0.667	0.000	0.333	3

$p_{i',j'}^t(LRT_{ae})$ and $p_{ij}(LRT_{std})$ estimates

Method	Aff	Genotype					
		22	23	24	33	34	44
LRT_{ae}	Control	0.000	0.118	0.024	0.699	0.159	0.000
	Case	0.019	0.057	0.019	0.465	0.344	0.096
LRT_{std}	Control	0.009	0.102	0.025	0.712	0.144	0.009
	Case	0.012	0.058	0.029	0.532	0.318	0.052

Legend for Table 5. In the upper portion of this table, we report estimated genotype misclassification probabilities $\hat{\theta}_{j',j}$ for all genotypes at the ApoE locus in 173 cases and 118 control individuals from a genetic case/control study of LOAD (see *Methods - Example - ApoE and Alzheimer's Disease* for a fuller description of the data set). In the lower portion of this table, we report maximum likelihood estimates of the true and observed genotype frequencies $p_{i',j'}^t$, and p_{ij} using the LRT_{ae} and LRT_{std} methods, respectively. The column labeled “Aff” refers to the affection status of the individual (either case or control).

Because of the small or 0 genotype frequency estimate for some genotypes, we confirmed the p-value estimate for the LRT_{ae} estimate using permutation. Observed case and control status were randomly permuted. Our permutation p-value, based on 1,000,000 replicates, is 4.6×10^{-5} . This estimate is consistent with the p-value based on asymptotic theory. Note that we can perform permutation testing here by randomly reassigning case and control status and keeping marginal totals fixed, since there is no double-sample information for phenotypes (see Summary and Discussion).

Summary and Discussion

Misclassification error can be a vexing problem in case/control association studies because of the loss in statistical power, the incorrect estimation of population frequency parameters, and the fact that errors cannot be detected short of some additional repeated sampling (Cochran 1968; Miller et al. 2002; Rice and Holmans 2003) or use of genetic information with tightly linked markers (Ehm et al. 1996; Abecasis et al. 2002; Sobel et al. 2002; Zou et al. 2003). Our method, the LRT_{ae} , corrects for these problems through the use of a sub-sample of individuals who are double-sampled on either phenotype and/or genotype. The results of our simulations are that the LRT_{ae} method appears to be more powerful than the LRT_{std} method for at least half of the simulation parameter settings even for equal fixed costs. Furthermore, the LRT_{ae} method has the added advantages that it provides asymptotically unbiased estimates of genotype frequencies and

phenotype and genotype misclassification error rates. Lastly, the results of our simulation studies suggest that we can achieve large power gains, given a fixed cost, using the LRT_{ae} method, even for larger phenotype misclassification rates.

We note that our phenotype misclassification probabilities used in the cost/benefits analysis (table 2) may appear large. As noted in the Methods section, however, we based these misclassification probabilities on published findings. In the spirit of a factorial design (Box et al. 1978), this work is a preliminary investigation as to whether there are situations for which the LRT_{ae} method is more powerful than the LRT_{std} method. Now that we have documented such situations, we plan to perform a more thorough analysis of the parameter settings for which either of these methods is more powerful, given fixed cost.

Beyond the question of power comparisons, we think there is a natural way to use the LRT_{ae} statistic. When “better” (i.e., lower misclassification rate) measurements for phenotype and/or genotype become available for some proportion of the individuals in one’s study, this information can *always* be used in the LRT_{ae} method. Sometimes researchers will simply replace the fallible classification with the infallible classification and remove the fallible classification from their database. However, applying the LRT_{std} method to such updated data will still produce biased population frequency estimates. Furthermore, data storage is usually inexpensive. We therefore recommend that researchers keep both fallible and infallible classifications on all individuals for whom such data are available and then apply the LRT_{ae} method.

One question that arises is how we apply permutation testing when we have double-sample data. In the situation where we have double sample data only for genotypes, we may determine p-values via permutation as is typically done; that is, we may randomly permute case and control status, keeping the marginal totals fixed, and report the proportion of permuted samples that exceed the LRT_{ae} value for the observed data. We may do this since we assume that the measurement process for phenotype is independent of the measurement process for genotype (equation (2)). When we apply this procedure to our LOAD data, we observe a permutation p-value that is the same order of magnitude as the asymptotic p-value. We hypothesize that if we have double sample data only for phenotypes, then we may randomly permute genotypes, keeping marginal totals fixed. We are currently investigating this hypothesis as well as permutation procedures when double-samples are available for both phenotype and genotype.

In this work, we assume that the sub-sample for which genotypes (or phenotypes) are double-sampled represents a random selection of the total. In practice, however, there may be a bias in selection of individuals for double-sampling. For example, with Alzheimer’s disease, autopsies may only be performed for individuals who have been diagnosed with the disease. For genotypes, double-sample information may only be available on individuals for

whom the quality of the genotype obtained by the fallible method is questionable. It is therefore important to be aware of the conditions under which double-sampling is being performed, so as to protect against biased estimates of the various parameters.

While we have not emphasized it here, our work may also be easily applied to haplotype analysis with cases and controls. For example, some researchers use multi-locus genotypes to infer haplotypes (e.g., Clark 1990; Excoffier and Slatkin 1995; Fallin and Schork 2000; Stephens et al. 2001; Zou and Zhao 2003). Those inferences may have misclassification error. With our method, one can use results from molecular haplotyping methods (e.g., Douglas et al. 2001) as the true measurement, thereby correcting haplotype misclassifications and potentially increasing power to detect association. We comment that, for autosomal data, any double-sample procedure would necessarily be applied to pairs of haplotypes for individuals rather than to the count of specific haplotypes in a data set.

Finally, we comment that we have software that performs our LRT_{ae} method. Researchers may obtain this software by using the URL: <ftp://linkage.rockefeller.edu/software/lrtae/>

Appendix

Here we present mathematical derivations of the formulas (7) for $p_{i'j'}^{t(r+1)}, q_{i'}^{t(r+1)}, p_{*j'}^{t(r+1)}$.

PROPOSITION 1. Let $n_{i'j'}^{(r)} = n_{1i'j'}^{(r)} + n_{2i'j'}^{(r)} + n_{3i'j'}^{(r)} + n_{4i'j'}^{(r)}$, where

$$n_{1i'j'}^{(r)} = \sum_i \sum_j n_{i'ij'}^{(1)}$$

$$n_{2i'j'}^{(r)} = \sum_j n_{i'+j}^{(2)} E[I(X_{i'j'}^t) | X_{i'j}^{(2)}] = \sum_j n_{i'+j}^{(2)} \left(\frac{\theta_{j'j} p_{i'j'}^{t(r)}}{\sum_{v'} \theta_{v'j} p_{i'v'}^{t(r)}} \right),$$

$$n_{3i'j'}^{(r)} = \sum_i n_{ij'+}^{(3)} E[I(X_{i'j'}^t) | X_{ij'+}^{(3)}] = \sum_i n_{ij'+}^{(3)} \left(\frac{\pi_{ii} p_{i'j'}^{t(r)} q_{i'}^{t(r)}}{\sum_{u'} \pi_{u'i} p_{u'j'}^{t(r)} q_{u'}^{t(r)}} \right),$$

$$n_{4i'j'}^{(r)} = \sum_i \sum_j n_{ij}^{(4)} E[I(X_{i'j'}^t) | X_{ij}^{(4)}] = \sum_i \sum_j n_{ij}^{(4)} \left(\frac{\theta_{j'j} \pi_{ii} p_{i'j'}^{t(r)} q_{i'}^{t(r)}}{\sum_{u'} \sum_{v'} \theta_{v'j} \pi_{u'i} p_{u'v'}^{t(r)} q_{u'}^{t(r)}} \right),$$

and $n_{i'+j}^{(2)} = \sum_i n_{i'ij}^{(2)}, n_{ij'+}^{(3)} = \sum_j n_{ij'j}^{(3)}$, $E[\]$ is the expectation operator, $I(\)$ is the indicator function, $X_{i'j}^{(2)}$ is the event that an individual has true phenotype i' and

observed genotype j (for individuals in “Group 2” – see proof below), $X_{ij}^{(3)}$ is the event that an individual has observed phenotype i and true genotype j' (for individuals in “Group 3”), and $X_{ij}^{(4)}$ is the event that an individual has observed phenotype i and observed genotype j (individuals in “Group 4”). Also, let $n_{i+}^{(r)} = \sum_{j'} n_{i'j'}^{(r)}$. Then the $r+1$ st iteration-estimates of $p_{i'j'}^t, q_{i'}^t$ and $p_{+j'}^t$ are:

$$\begin{aligned} p_{i'j'}^{t(r+1)} &= n_{i'j'}^{(r)} / n_{i+}^{(r)}, \\ q_{i'}^{t(r+1)} &= n_{i+}^{(r)} / n, \\ p_{+j'}^{t(r+1)} &= n_{+j'}^{(r)} / n. \end{aligned} \tag{A.1}$$

Note that estimates of $q_{i'}^{t(r+1)}$ may be different when maximizing under the null hypothesis versus the alternative hypothesis.

PROOF: We use the method of Expectation-Maximization. We can divide the sample of n individuals into four distinct groups; the first n_1 individuals have double sample data on both phenotypes and genotypes (Group 1); the next n_2 individuals have double sample data on phenotypes only (Group 2); the next n_3 individuals have double sample data on genotypes only (Group 3); and the last n_4 individuals have only fallible measures of both genotype and phenotype (Group 4). We may write the log-likelihood of the true data as:

$$\begin{aligned} \ln(L) &= \sum_{i'} \sum_{j'} [\ln(p_{i'j'}^t q_{i'}^t)] \\ &\times \left(\sum_{1 \leq a \leq n_1} I(X_{i'j'}^{t,a}) + \sum_{n_1+1 \leq a \leq n_1+n_2} I(X_{i'j'}^{t,a}) + \sum_{n_1+n_2+1 \leq a \leq n_1+n_2+n_3} I(X_{i'j'}^{t,a}) + \sum_{n_1+n_2+n_3+1 \leq a \leq n_1+n_2+n_3+n_4} I(X_{i'j'}^{t,a}) \right) \tag{A.2} \\ &= \sum_{i'} \sum_{j'} [\ln(p_{i'j'}^t) + \ln(q_{i'}^t)] \\ &\times \left(\sum_{1 \leq a \leq n_1} I(X_{i'j'}^{t,a}) + \sum_{n_1+1 \leq a \leq n_1+n_2} I(X_{i'j'}^{t,a}) + \sum_{n_1+n_2+1 \leq a \leq n_1+n_2+n_3} I(X_{i'j'}^{t,a}) + \sum_{n_1+n_2+n_3+1 \leq a \leq n_1+n_2+n_3+n_4} I(X_{i'j'}^{t,a}) \right). \end{aligned}$$

where $X_{i'j'}^{t,a}$ is the event that individual a has true phenotype i' and true genotype j' , and $I()$ is the indicator function.

Expectation Step

Examining equation (A.2) more closely, we see that computing $E[\ln(L) | OD]$, the expectation of the log-likelihood conditional on the observed data, reduces to computing $E[I(X_{i'j'}^{t,a}) | X_{ij}^a]$ for each individual a , where X_{ij}^a is the event that individual a 's observed phenotype and genotype are i and j , respectively. If either

the phenotype or genotype (or both) for an individual has (have) been double sampled, then we know that (those) measurement(s) with certainty. In such instances, we add a “prime” superscript, indicating that the category (phenotype or genotype) is known with certainty. For example, for individuals in Group 1, we replace X_{ij}^a with $X_{i'j'}^{(1)a}$, since both phenotype and genotype are known with certainty. Similarly, for individuals in Group 2, we replace X_{ij}^a with $X_{i'j}^{(2)a}$, since phenotypes are known with certainty, and for individuals in Group 3, we replace $X_{ij}^{(3)a}$ with X_{ij}^a . We show in the next lemma (Lemma 2), that the desired expectations using the r^{th} -iteration estimates of $p_{i'j'}$ and $q_{i'}$ are:

(Group 2)

$$E[I(X_{i'j'}^{t,a}) | X_{i'j}^{(2)a}] = \left(\frac{\theta_{j'j} p_{i'j'}^{t(r)}}{\sum_{v'} \theta_{v'j} p_{i'v'}^{t(r)}} \right);$$

(Group 3)

$$E[I(X_{i'j'}^{t,a}) | X_{ij'}^{(3)a}] = \left(\frac{\pi_{i'i} p_{i'j'}^{t(r)} q_{i'}^{t(r)}}{\sum_{u'} \pi_{u'i} p_{u'j'}^{t(r)} q_{u'}^{t(r)}} \right); \quad (\text{A.3})$$

(Group 4)

$$E[I(X_{i'j'}^{t,a}) | X_{ij}^{(4)a}] = \left(\frac{\theta_{j'j} \pi_{i'i} p_{i'j'}^{t(r)} q_{i'}^{t(r)}}{\sum_{u'} \sum_{v'} \theta_{v'j} \pi_{u'i} p_{u'v'}^{t(r)} q_{u'}^{t(r)}} \right).$$

Note that, for individuals in Group 1, because phenotypes and genotypes are known with certainty, $E[X_{i'j'}^{t,a} | X_{i'j'}^{(1)a}] = 1$. Using equation (A.2), the linear property of the expectation operator, and the fact that, for each group, the conditional expectations (A.3) are independent of the particular individual a , we have

$$\begin{aligned}
& E[\ln(L) | OD] \\
&= \sum_{i'} \sum_{j'} [\ln(p_{i'j'}^t) + \ln(q_{i'}^t)] \\
&\times \left(\sum_{1 \leq a \leq n_1} E[I(X_{i'j'}^{t,a}) | X_{i'j'}^{(1)a}] + \sum_{n_1+1 \leq a \leq n_1+n_2} E[I(X_{i'j'}^{t,a}) | X_{i'j'}^{(2)a}] \right. \\
&+ \left. \sum_{n_1+n_2+1 \leq a \leq n_1+n_2+n_3} E[I(X_{i'j'}^{t,a}) | X_{i'j'}^{(3)a}] + \sum_{n_1+n_2+n_3+1 \leq a \leq n_1+n_2+n_3+n_4} E[I(X_{i'j'}^{t,a}) | X_{i'j'}^{(4)a}] \right) \\
&= \sum_{i'} \sum_{j'} [\ln(p_{i'j'}^t) + \ln(q_{i'}^t)] \\
&\times \left(n_{i'j'} + \sum_j n_{i'+j}^{(2)} E[I(X_{i'j'}^t) | X_{i'j}^{(2)}] + \sum_i n_{ij'+}^{(3)} E[I(X_{i'j'}^t) | X_{ij'+}^{(3)}] + \sum_i \sum_j n_{ij}^{(4)} E[I(X_{i'j'}^t) | X_{ij}^{(4)}] \right) \\
&= \sum_{i'} \sum_{j'} n_{i'j'}^{(r)} [\ln(p_{i'j'}^t) + \ln(q_{i'}^t)],
\end{aligned}$$

using the definition of $n_{i'j'}^{(r)}$ given in the statement of Proposition 1.

Maximization Step

$$\begin{aligned}
\text{Let } f_{i'} &= E[\ln(L) | OD] - \lambda_{i'} (\sum_{j'} p_{i'j'}^t - 1) \\
&= \sum_{i'} \sum_{j'} n_{i'j'}^{(r)} [\ln(p_{i'j'}^t) + \ln(q_{i'}^t)] - \lambda_{i'} (\sum_{j'} p_{i'j'}^t - 1).
\end{aligned}$$

Then, using the standard Lagrange multiplier technique, we have

$$\frac{\partial f_{i'}}{\partial p_{i'j'}^t} = \frac{n_{i'j'}^{(r)}}{p_{i'j'}^t} - \lambda_{i'}. \text{ Setting this equation equal to 0 and solving for } p_{i'j'}^t, \text{ yields}$$

$p_{i'j'}^t = n_{i'j'}^{(r)} / \lambda_{i'}$. To determine $\lambda_{i'}$, note that $1 = \sum_{j'} p_{i'j'}^t = \frac{\sum_{j'} n_{i'j'}^{(r)}}{\lambda_{i'}} = \frac{n_{i'+}^{(r)}}{\lambda_{i'}}$. It follows that $\lambda_{i'} = n_{i'+}^{(r)}$ and thus $p_{i'j'}^{t(r+1)} = n_{i'j'}^{(r)} / n_{i'+}^{(r)}$. Applying the same technique to $g = E[\ln(L) | OD] - \lambda (\sum_{i'} q_{i'}^t - 1)$, taking the partial derivative with respect to $q_{i'}^t$, and noting that $\sum_{i'} \sum_{j'} n_{i'j'} = n$, we get $q_{i'}^{t(r+1)} = n_{i'+}^{(r)} / n$. This completes the proof.

LEMMA 2. Let $X_{i'j'}^{(2)a}$, $X_{ij'+}^{(3)a}$, $X_{ij}^{(4)a}$ be the observed events for an individual a in Groups 2, 3, and 4, respectively, as defined above (proof of proposition 1; also see

table 1). Using the r^{th} -iteration parameter estimates $p_{i'j'}^{t(r)}$ and $q_{i'}^{t(r)}$, the conditional expectations $E[I(X_{i'j'}^{t,a}) | X_{i'j}^{(2)a}]$, $E[I(X_{i'j'}^{t,a}) | X_{ij'}^{(3)a}]$, and $E[I(X_{i'j'}^{t,a}) | X_{ij}^{(4)a}]$ for individuals in Groups 2, 3, and 4 respectively, are given by the formulas in equation (A.3).

PROOF: For the first equation, note that

$$\begin{aligned}
 E[I(X_{i'j'}^{t,a}) | X_{i'j}^{(2)a}] &= \Pr(X_{i'j'}^{t,a} | X_{i'j}^{(2)a}) \\
 &= \Pr(X_{i'j}^{(2)a} | X_{i'j'}^{t,a}) \Pr(X_{i'j'}^{t,a}) / \sum_{v'} \Pr(X_{i'j}^{(2)a} | X_{i'v'}^{t,a}) \Pr(X_{i'v'}^{t,a}) \\
 &= \theta_{j'j} p_{i'j'}^{t(r)} q_{i'}^{t(r)} / \sum_{v'} \theta_{v'j} p_{i'v'}^{t(r)} q_{i'}^{t(r)} \\
 &= \theta_{j'j} p_{i'j'}^{t(r)} / \sum_{v'} \theta_{v'j} p_{i'v'}^{t(r)}.
 \end{aligned} \tag{A.4}$$

The first equality in equation (A.4) follows from the fact that the expectation of the indicator function is the probability of the event. The second equality follows from Bayes Rule, and the third follows from the definition of the genotype misclassification matrix θ and the definition of conditional probability.

For the equation in Group (3), we have

$$\begin{aligned}
 E[I(X_{i'j'}^{t,a}) | X_{ij'}^{(3)a}] &= \Pr(X_{i'j'}^{t,a} | X_{ij'}^{(3)a}) \\
 &= \Pr(X_{ij'}^{(3)a} | X_{i'j'}^{t,a}) \Pr(X_{i'j'}^{t,a}) / \sum_{u'} \Pr(X_{ij'}^{(3)a} | X_{u'j'}^{t,a}) \Pr(X_{u'j'}^{t,a}) \\
 &= \pi_{i'i} p_{i'j'}^{t(r)} q_{i'}^{t(r)} / \sum_{u'} \pi_{u'i} p_{u'j'}^{t(r)} q_{u'}^{t(r)}.
 \end{aligned} \tag{A.5}$$

As above, the third equality in formula (A.5) follows from the definition of the genotype misclassification matrix π and the definition of conditional probability. The equation for Group (4) follows as in the derivation (A.5). The only change necessary is the observation that $\Pr(X_{ij}^{(4)a} | X_{i'j'}^{t,a}) = \theta_{j'j} \pi_{i'i}$.

Electronic Database Information

BINOM - <ftp://linkage.rockefeller.edu/software/utilities/>

PAWE – <http://linkage.rockefeller.edu/pawe/>

S-PLUS – <http://www.insightful.com>

References

- Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 30:97-101
- Box GEP, Hunter WG, Hunter JS (1978) *Statistic for Experimenters*. John Wiley and Sons, New York

- Breslow NE, Day NE (1980) *Statistical Methods in Cancer Research*. Vol 1. International Agency for Research on Cancer, Lyon
- Bross I (1954) Misclassification in 2 x 2 tables. *Biometrics* 10:478-486
- Brown AM, Gordon D, Lee H, Caudy M, Haroutunian V, Blass JP (2004a) Substantial linkage disequilibrium across the dihydrolipoyl succinyltransferase gene region without Alzheimer's disease association. *Neurochem Res* 29:629-635
- Brown AM, Gordon D, Lee H, Xu Y, Caudy M, Hardy J, Haroutunian V, Blass JP (2004b) Association of the dihydrolipoamide dehydrogenase gene with Alzheimer's disease in an Ashkenazi Jewish population. *Am J Med Genet* (in press)
- Burd L, Kerbeshian J, Klug MG (2001) Neuropsychiatric genetics: misclassification in linkage studies of phenotype-genotype research. *J Child Neurol* 16:499-504
- Chen TT (1979) Log-linear models for categorical data with misclassification and double sampling. *J Am Stat Assoc* 74:481-488
- Clark AG (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol Biol Evol* 7:111-122
- Cochran WG (1968) Errors of measurement in statistics. *Technometrics* 10:637-666
- Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, Roses AD, Haines JL, Pericak-Vance MA (1993) Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* 261:921-923
- Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *J Roy Statist Soc B* 39:1-38
- Douglas JA, Boehnke M, Gillanders E, Trent JM, Gruber SB (2001) Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies. *Nat Genet* 28:361-364
- Ehm MG, Kimmel M, Cottingham Jr. RW (1996) Error detection for genetic data, using likelihood methods. *Am J Hum Genet* 58:225-234
- Espeland MA, Hui SL (1987) A general approach to analyzing epidemiologic data that contain misclassification errors. *Biometrics* 43:1001-1012
- Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* 12:921-927
- Fallin D, Schork NJ (2000) Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. *Am J Hum Genet* 67:947-959
- Gordon D, Finch SJ, Nothnagel M, Ott J (2002) Power and sample size calculations for case-control genetic association tests when errors are

- present: application to single nucleotide polymorphisms. *Hum Hered* 54:22-33
- Gordon D, Levenstien MA, Finch SJ, Ott J (2003) Errors and linkage disequilibrium interact multiplicatively when computing sample sizes for genetic case-control association studies. *Pac Symp Biocomput*: 490-501
- Gordon D, Ott J (2001) Assessment and management of single nucleotide polymorphism genotype errors in genetic association analysis. *Pac Symp Biocomput*:18-29
- Gustafson P, Le ND, Saskin R (2001) Case-control analysis with partial knowledge of exposure misclassification probabilities. *Biometrics* 57:598-609
- Hixson JE, Vernier DT (1990) Restriction isotyping of human apolipoprotein E by gene amplification and cleavage with HhaI. *J Lipid Res* 31:545-548
- Hochberg Y (1977) Use of double sampling schemes in analyzing categorical data with misclassification errors. *J Am Stat Assoc* 72:914-921
- Kang SJ, Gordon D, Finch SJ (2004) What SNP genotyping errors are most costly for genetic association studies? *Genet Epidemiol* 26:132-141
- Kendall M, Stuart A, Ord JK (1994) *The Advanced Theory of Statistics. Vol I.* Oxford University Press, New York
- Miller CR, Joyce P, Waits LP (2002) Assessing allelic dropout and genotype reliability using maximum likelihood. *Genetics* 160:357-366
- Mitra SK (1958) On the limiting power function of the frequency chi-square test. *Ann Math Stat* 29:1221-1233
- Mote VL, Anderson RL (1965) An investigation of the effect of misclassification on the properties of chisquare-tests in the analysis of categorical data. *Biometrika* 52:95-109
- Press MF, Hung G, Godolphin W, Slamon DJ (1994) Sensitivity of HER-2/neu antibodies in archival tissue samples: potential sources of error in immunohistochemical studies of oncogene expression. *Cancer Res* 54:2771-2777
- Rice KM, Holmans P (2003) Allowing for genotyping error in analysis of unmatched cases and controls. *Ann Hum Genet* 67:165-174
- Sheu KF, Brown AM, Haroutunian V, Kristal BS, Thaler H, Lesser M, Kalaria RN, Relkin NR, Mohs RC, Lilius L, Lannfelt L, Blass JP (1999) Modulation by DLST of the genetic risk of Alzheimer's disease in a very elderly population. *Ann Neurol* 45:48-53
- Sobel E, Papp JC, Lange K (2002) Detection and integration of genotyping errors in statistical genetics. *Am J Hum Genet* 70:496-508
- Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 68:978-989

- Tenenbein A (1970) A double sampling scheme for estimating from binomial data with misclassifications. *J Am Stat Assoc* 65:1350-1361
- Tenenbein A (1972) A double sampling scheme for estimating from misclassified multinomial data with applications to sampling inspection. *Technometrics* 14:187-202
- Tukey JW (1977) *Exploratory Data Analysis*. Addison-Wesley, Reading, MA
- Zivelin A, Rosenberg N, Peretz H, Amit Y, Kornbrot N, Seligsohn U (1997) Improved method for genotyping apolipoprotein E polymorphisms by a PCR-based assay simultaneously utilizing two distinct restriction enzymes. *Clin Chem* 43:1657-1659
- Zou G, Pan D, Zhao H (2003) Genotyping error detection through tightly linked markers. *Genetics* 164:1161-1173
- Zou G, Zhao H (2003) Haplotype frequency estimation in the presence of genotyping errors. *Hum Hered* 56:131-138