Original Paper

Human Heredity

Hum Hered 2010;70:75–91
DOI: 10.1159/000312819

Received: November 3, 2009
Accepted after revision: January 25, 2010
Published online: June 17, 2010

A Latent Class Model for Testing for Linkage and Classifying Families when the Sample May Contain Segregating and Non-Segregating Families

Laurel A. Bastone[a] Richard S. Spielman[b] Xingmei Wang[c, d]
Thomas R. Ten Have[c, d] Mary E. Putt[c, d]

[a]Global Biometrics Science, Bristol-Myers Squibb, Pennington, N.J., [b]Department of Genetics and [c]Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine and [d]Center for Clinical Epidemiology and Biostatistics, Philadelphia, Pa., USA

Key Words
Linkage · Heterogeneity · Haseman-Elston regression · Latent class analysis · Gene expression

Abstract
In a quantitative trait locus (QTL) study, it is usually not feasible to select families with offspring that simultaneously display variability in more than one phenotype. When multiple phenotypes are of interest, the sample will, with high probability, contain 'non-segregating' families, i.e. families with both parents homozygous at the QTL. These families potentially reduce the power of regression-based methods to detect linkage. Moreover, follow-up studies in individual families will be inefficient, and potentially even misleading, if non-segregating families are selected for the study. Our work extends Haseman-Elston regression using a latent class model to account for the mixture of segregating and non-segregating families. We provide theoretical motivation for the method using an additive genetic model with two distinct functions of the phenotypic outcome, squared difference *(SqD)* and mean-corrected product *(MCP)*. A permutation procedure is developed to test for linkage; simulation shows that the test is valid for both phenotypic functions. For rare alleles, the method provides increased power compared to a 'marginal' approach that ignores the two types of families; for more common alleles, the marginal approach has better power. These results appear to reflect the ability of the algorithm to accurately assign families to the two classes and the relative weights of segregating and non-segregating families to the test of linkage. An application of Bayes rule is used to estimate the family-specific probability of segregating. High predictive value positive values for segregating families, particularly for *MCP*, suggest that the method has considerable value for identifying segregating families. The method is illustrated for gene expression phenotypes measured on 27 candidate genes previously demonstrated to show linkage in a sample of 14 families.

Introduction

In conventional quantitative-trait locus (QTL) mapping studies, families are selected into the sample on the basis of phenotypic variation among the offspring. This selection ensures that the offspring, and by definition

Richard S. Spielman, who was integrally involved with this project, passed away suddenly in April 2009 during preparation of this article.

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2010 S. Karger AG, Basel

Accessible online at:
www.karger.com/hhe

Mary E. Putt
Department of Biostatistics and Epidemiology
University of Pennsylvania School of Medicine
Philadelphia, PA 19104 (USA)
Tel. +1 215 573 7020, Fax +1 215 573 4865, E-Mail mputt@mail.med.upenn.edu

their parents, have high probability of displaying allelic variation at the QTL. We define segregating families as those with at least one parent who is heterozygous at the QTL. When multiple phenotypes are of interest, it is difficult to select families that simultaneously segregate for all phenotypes. These studies are likely to include some families where both parents are homozygous at one or more of the QTL of interest. The offspring in these 'non-segregating' families lack genetically determined variation in phenotype. Since non-segregating families do not contribute to the linkage signal, their inclusion in the sample should reduce the power of the analysis to detect linkage compared to a sample of the same size composed of segregating families. Additionally, there is little to be gained from including material from non-segregating families in subsequent follow-up studies. The research described here was motivated by the Genetics of Gene Expression study, a genome-wide linkage study performed initially in 14 pedigrees from the Centre d'Etude du Polymorphisme Humain (CEPH) collection [1]. The phenotypes of interest were mRNA expression levels of several thousand genes measured on each individual using microarray technology [2]. In these analyses we anticipate that many QTL will be characterized by non-segregating families as pedigrees were not selected for simultaneous variation in all of the phenotypes.

Here, we propose to measure linkage with an extension of Haseman-Elston regression based on a latent class model to take into account segregating and non-segregating families. Latent class analysis is a model-based method for clustering data into unobserved, or latent, classes, assuming that the marginal or observed distribution of data is a mixture of two or more distributions [3]. Latent class methods have previously been used in population-based genetic studies to identify subclasses of disease phenotypes, and to infer population substructure [4–12]. However, we are unaware of previous applications of latent class methodology to family-based linkage studies. Our method provides an objective test for linkage in the presence of combinations of segregating and non-segregating families. Through an application of Bayes rule, we estimate the probability that individual families are segregating, and thus, identify families that are likely to be informative in follow-up statistical and molecular studies of putative regulators of gene expression. This is a unique aspect of our approach, one that, to our knowledge is not addressed by current methods.

We briefly introduced a latent class approach to this problem in an earlier study of *trans* regulators of gene expression [13]. Here we describe in detail an improvement on the original testing approach using examples of *cis* regulation of gene expression as motivation. We justify the new method theoretically and explore its operating characteristics in a simulation study. In the Methods section, the latent class extension to Haseman-Elston regression is developed along with a method for fitting the model. The modified hypothesis test is presented along with the approach for identifying subjects from segregating families. In separate Results sections, we first describe analytic findings and a simulation study and then illustrate the approach with an example of the analysis of several selected phenotypes from the motivating study. We conclude with a Discussion of the methods.

Methods

Overview

We demonstrate the potential importance of the problem of non-segregating families in the population and review the additive genetic model and the principles of regression-based linkage analysis. A latent class model extends regression-based linkage to account for heterogeneity in outcome due to segregating and non-segregating families. We explore the properties of two phenotype functions in the context of the two-class model and develop testing and estimating procedures.

Occurrence of Non-Segregating Families

For a single QTL with two alleles, let H and L denote the high- and low-expression alleles, with population frequencies p and $q = 1 - p$ respectively. Under Hardy-Weinberg equilibrium, the proportion of segregating families (π) in the population is

$$\pi = 4p^3q + 4p^2q^2 + 4pq^3 = 4pq(1 - pq) \qquad (2.1)$$

The proportion of non-segregating families is

$$1 - \pi = (1 - 2pq)^2 \qquad (2.2)$$

a value which exceeds 0.25 for all values of p, with larger rates occurring as one allele becomes less common. Even with relatively common alleles, $1 - \pi$ is substantial; for example with $p = 0.10$, close to two thirds of the families in the population are non-segregating. Non-segregating families are unlikely to be included in a sample of families ascertained using a proband. However, samples selected without regard to phenotype should frequently contain non-segregating families.

Genetic Model

We consider each expression phenotype independently. Let W_{ij} denote the phenotypic value for offspring j ($j = 1, ..., n_i$) in the i-th family ($i = 1, ..., N$), where n_i ($n_i \geq 2$) is the number of offspring in family i and N is the number of families in the sample. We use an additive genetic model, without considering dominance, i.e.

$$W_{ij} = \gamma + \nu \cdot G_{ij} + F_i + e_{ij} \qquad (2.3)$$

Table 1. Parameters of the marginal and two-class model for *SqD* and *MCP* phenotype functions for a constant number of siblings per family (*n*)

Model	Phenotype	Expectation
Marginal	SqD	$2\left[Var\left(W_{ij}\right)-Cov\left(W_{ij}W_{ij'}\lvert Z_{ik}\right)\right]$
	MCP	$\frac{1}{n}Var\left(W_{ij}\right)+\frac{\left(n-1\right)^2}{n\left(n+1\right)}Cov\left(W_{ij}W_{ij'}\right)-\frac{n^3-n^2-2n+4}{n^2\left(n+1\right)}Cov\left(W_{ij}W_{ij'}\lvert Z_{ik}\right)$
Two-Class	SqD	$2\left[Var\left(W_{ij}\lvert C_i\right)-Cov\left(W_{ij}W_{ij'}\lvert Z_{ik}C_i\right)\right]$
	MCP	$\frac{1}{n}Var\left(W_{ij}\lvert C_i\right)+\frac{\left(n-1\right)^2}{n\left(n+1\right)}Cov\left(W_{ij}W_{ij'}\lvert C_i\right)-\frac{n^3-n^2-2n+4}{n^2\left(n+1\right)}Cov\left(W_{ij}W_{ij'}\lvert Z_{ik}C_i\right)$

where γ is the average phenotypic value of a heterozygote (genotype *HL*), ν is the additive effect of the *H* allele, G_{ij} is the number, minus one, of *H* alleles carried by individual *j* ($G_{ij} = \{-1, 0, 1\}$), F_i is a family-level random effect, and e_{ij} is an individual error term [14]. We assume the F_i are independent of the G_{ij} and are independent and identically distributed with mean zero and variance σ_F^2 while the e_{ij} are independent of G_{ij} and F_i and are independent and identically distributed with mean zero and variance σ_e^2.

Consider the *k*-th ($k = 1, ..., m_i$, $m_i = \binom{n_i}{2}$) sibling pair, or sibpair, comprised of individuals *j* and *j′*. Let X_{ik} be the number of alleles shared IBD at the QTL by sibpair *k* in family *i*. The covariance of phenotypic values for a sibpair, or phenotypic covariance, is a linear function of their IBD sharing, i.e.

$$Cov[W_{ij}, W_{ij'} \mid X_{ik}] = \sigma_F^2 + \nu^2 pqX_{ik} \tag{2.4}$$

[15]. Equation (2.4) assumes that IBD sharing at the QTL, X_{ik}, is known, whereas in practice IBD sharing is approximated at single nucleotide polymorphism (SNP) markers. Let Z_{ik} denote the number of alleles shared IBD at a SNP marker, where θ is the recombination fraction between the marker and the QTL ($0 \leq \theta \leq 1/2$). Here $\theta = 0$ indicates that the marker and the QTL do not recombine while $\theta = 1/2$ indicates that the marker and the QTL are unlinked. The phenotypic covariance in this case is

$$Cov[W_{ij}, W_{ij'} \mid Z_{ik}] = \sigma_F^2 + 4\nu^2 pq\theta(1-\theta) + \nu^2 pq(1-2\theta)^2 Z_{ik} \tag{2.5}$$

Extension of Haseman-Elston Regression

Let Y_{ik} denote the squared difference (*SqD*) in phenotype for siblings *j* and *j′* in sibpair *k* in family *i*

$$Y_{ik} = (W_{ij} - W_{ij'})^2 \tag{2.6}$$

Haseman and Elston [15] describe a 'marginal' regression-based test of linkage. Here Y_{ik} is regressed on IBD sharing at the QTL, X_{ik}. Subsequent efforts to increase the power of the method led to the use of other functions of phenotypic values, including the family mean-corrected product (*MCP*) and weighted sums of the *SqD* and squared sums [16–22]. In addition to *SqD*, we considered *MCP*, where

$$Y_{ik} = (W_{ij} - \overline{W}_{i\cdot})\,(W_{ij'} - \overline{W}_{i\cdot})$$

and

$$\overline{W}_{i\cdot} = \frac{1}{n_i}\sum\nolimits_{j=1}^{n_i} W_{ij}$$

denotes the mean phenotypic value for the *i*-th family. Table 1 shows that in the marginal model, the expectation of these functions are linear combinations of phenotypic covariance and variance terms.

The expectations in table 1 are evaluated by substituting the appropriate expressions for the variance and covariance terms from table 2. The appendix provides further details of the derivations. The expected value of the phenotype function, either the *SqD* or *MCP*, is a linear function of the number of alleles shared IBD at the marker, Z_{ik} with the general form

$$E[Y_{ik} \mid Z_{ik}] = \alpha + \beta Z_{ik} \tag{2.7}$$

For *SqD* the marginal intercept is $\alpha = 2\sigma_e^2 + 8\nu^2 pq\,[\theta^2 - \theta + 1/2]$; for *MCP* the closed-form solution for the intercept is complicated but is readily computed numerically. The specific values of the 'marginal' slope, β, in terms of the genetic model appear in table 3. For both phenotype functions the terms from the genetic model that contribute to the slope come exclusively from the conditional covariance terms in Equation (2.5). The slope for *SqD* is independent of sample size and the slopes for *SqD* and *MCP* are identical up to the multiplicative term

$$\frac{n^3 - n^2 - 2n + 4}{2n^2\left(n+1\right)}$$

where here, for the purpose of the derivation only, the number of subjects per family, *n*, was set to be constant for all families. Note from table 3 that $\theta = 1/2$ implies $\beta = 0$ under both models.

To extend the model in Equation (2.7), let C_i be a binary variable indicating whether family *i* belongs to a segregating ($C_i = \mathcal{S}$) or non-segregating ($C_i = \mathcal{N}$) 'class' of families. The expectation of the phenotype functions conditional on class appear in table 1. When evaluated in terms of the genetic model using the terms in table 2, the expectations of the phenotype function conditional on class are again linear functions of the number of alleles shared IBD at the marker, Z_{ik}, i.e.

$$E[Y_{ik} \mid Z_{ik}, C_i = \mathcal{S}] = \alpha_{\mathcal{S}} + \beta_{\mathcal{S}} Z_{ik} \text{ or } E[Y_{ik} \mid Z_{ik}, C_i = \mathcal{N}] = \alpha_{\mathcal{N}} + \beta_{\mathcal{N}} Z_{ik} \tag{2.8}$$

Table 2. Covariance and variance terms

Class	Term	Results
Marginal	$Cov\left(W_{ij}W_{ij'}\vert Z_{ik}\right)$	$\sigma_F^2 + 4\nu^2 pq\theta\left(1-\theta\right) + \nu^2 pq\left(1-2\theta\right)^2 Z_{ik}$
	$Cov\left(W_{ij}W_{ij'}\right)$	$\sigma_F^2 + \nu^2 pq$
	$Var\left(W_{ij}\right)$	$\sigma_e^2 + \sigma_F^2 + \nu^2 pq$
Segregating	$Cov\left(W_{ij}W_{ij'}\vert Z_{ik}, C=\mathcal{S}\right)$	$\sigma_F^2 + \frac{\nu^2}{4\left(1-pq\right)^2}\left[2pq\left(pq+1\right) - 1 + 4\theta\left(1-\theta\right)\left(1-pq\right)\right] + \frac{\nu^2\left(1-2\theta\right)^2}{4\left(1-pq\right)} Z_{ik}$
	$Cov\left(W_{ij}W_{ij'}\vert C=\mathcal{S}\right)$	$\sigma_F^2 + \nu^2\left(\frac{2pq\left(1+pq\right)}{4\left(1-pq\right)^2}\right)$
	$Var\left(W_{ij}\vert C=\mathcal{S}\right)$	$\sigma_e^2 + \sigma_F^2 + \nu^2\left(\frac{1+2p^2q^2}{4\left(1-pq\right)^2}\right)$
Non-Segregating	$Cov\left(W_{ij}W_{ij'}\vert Z_{ik}, C=\mathcal{N}\right)$	$\sigma_F^2 + \nu^2\frac{2p^2q^2}{\left(p^2+q^2\right)^2}$
	$Cov\left(W_{ij}W_{ij'}\vert C=\mathcal{N}\right)$	$\sigma_F^2 + \nu^2\left(\frac{2p^2q^2}{\left(p^2+q^2\right)^2}\right)$
	$Var\left(W_{ij}\vert C=\mathcal{N}\right)$	$\sigma_e^2 + \sigma_F^2 + \nu^2\left(\frac{2p^2q^2}{\left(p^2+q^2\right)^2}\right)$

Table 3. Slopes parameterized using the genetic model

Model	Parameter	Phenotype	
		SqD	*MCP*
Marginal	β	$-2\nu^2 pq\left(1-2\theta\right)^2$	$-2\nu^2 pq\left(1-2\theta\right)^2\frac{\left(n^3-n^2-2n+4\right)}{2n^2\left(n+1\right)}$
Two-Class	$\beta_{\mathcal{S}}$	$-\frac{\nu^2\left(1-2\theta\right)^2}{2\left(1-pq\right)}$	$-\frac{\nu^2\left(1-2\theta\right)^2}{2\left(1-pq\right)}\cdot\frac{\left(n^3-n^2-2n+4\right)}{2n^2\left(n+1\right)}$
	$\beta_{\mathcal{N}}$	0	0

For *SqD* the class-specific intercepts are

$$\alpha_{\mathcal{S}} = 2\sigma_e^2 + \frac{\nu^2\left[2\theta^2 - 2\theta + 1\right]}{1-pq}$$

and $\alpha_{\mathcal{N}} = 2\sigma_e^2$; for *MCP* the closed-form intercept terms are more complex but can be computed. Table 3 presents the slopes for the two classes for both *SqD* and *MCP*.

Estimation

Let $f(Y_{ik})$ be the marginal density and let $f(Y_{ik} \mid C_i)$ be the density conditional on class membership of the *i*-th family. To fit the model, we assume independence of sibpairs within a family, conditional on latent class. The conditional densities are assumed to be normal with constant variances, $f(Y_{ik} \mid C_i = \mathcal{S}) \sim N(\mu_{\mathcal{S}ik}, \sigma^2)$, where $\mu_{\mathcal{S}ik} = E[Y_{ik} \mid Z_{ik}, C_i = \mathcal{S}]$ from Equation (2.8) and table 1. Similarly, $f(Y_{ik} \mid C_i = \mathcal{N}) \sim N(\mu_{\mathcal{N}ik}, \sigma^2)$, where $\mu_{\mathcal{N}ik} = E[Y_{ik} \mid Z_{ik}, C_i = \mathcal{N}]$. Thus, the marginal density, $\boldsymbol{f}(\boldsymbol{Y}_i \mid \boldsymbol{Z}_i)$, for a vector of sibpairs with squared phenotypic difference $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{im_i})$ and IBD sharing $\boldsymbol{Z}_i = (Z_{i1}, ..., Z_{im_i})$ is

$$\boldsymbol{f}\left(\boldsymbol{Y}_i \vert \boldsymbol{Z}_i\right) = \pi\prod_{k=1}^{m_i} f\left(Y_{ik} \vert Z_{ik}, C_i = \mathcal{S}\right) + \left(1-\pi\right)\prod_{k=1}^{m_i} f\left(Y_{ik} \vert Z_{ik}, C_i = \mathcal{N}\right) \qquad (2.9)$$

In essence, the latent class approach is a method for fitting two class-specific Haseman-Elston regression models, while accounting for the uncertainty in the latent class designations of individual families.

The likelihood was maximized using a general quasi-Newton procedure implemented in SAS PROC TRAJ (v.9 using the NOVAR option; [23]). There are several issues to consider. First, segregating and non-segregating classes are not explicitly designated by the maximization procedure; this information is inferred by designating the segregating class as the one with the minimum of the two class-specific estimates of the slopes. Second, the two-class model sometimes fails to converge. Even when both types of families are present in the population, a given sample may contain only one type of family, and in this case the algorithm is expected to fail to converge to a two-class solution. In practice, the two-class model also fails to converge at times when both segregating and non-segregating families are present.

Because the two-class model sometimes failed to converge, we incorporated the possibility of fitting the marginal model from Equation (2.7) into the estimation. Let D indicate convergence of the two-class model. Convergence occurs ($D = 1$) if the following three conditions are satisfied: (1) The numerical convergence criteria of the estimation procedure are met. (2) The class-specific parameter estimates are distinct, i.e. the slope and/or intercepts differ by a specified constant, here set to 0.01. (3) The estimate of the proportion of families that are segregating, $\hat{\pi}$ is non-zero and less than 1.0.

Otherwise, $D = 0$ and the marginal model from Equation (2.7) is fit using maximum likelihood, implemented here with PROC TRAJ by specifying a model with one class and with signifiance assessed using permutation. In addition to estimates of the slope and intercept terms, the estimation procedure directly yields $\hat{\pi}$, the maximum likelihood estimate of π which, in turn, yields an estimate of p through simple rearrangement of Equation (2.1).

Lastly, using Bayes rule and assuming normal densities, the family-specific probability of membership in the segregating class is

$$P\left(C_i = \mathcal{S} \middle| \boldsymbol{Y}_i, \boldsymbol{Z}_i\right)$$

$$= \frac{\pi \prod_k f\left(Y_{ik} \middle| Z_{ik}, C_i = \mathcal{S}\right)}{\pi \prod_k f\left(Y_{ik} \middle| Z_{ik}, C_i = \mathcal{S}\right) + (1-\pi) \prod_k f\left(Y_{ik} \middle| Z_{ik}, C_i = \mathcal{N}\right)} \qquad (2.10)$$

$$= \left[1 + \frac{(1-\pi)}{\pi} \exp - \frac{1}{2\sigma^2} \sum_{k=1}^{n_i} \left(r^2_{\mathcal{N}ik} - r^2_{\mathcal{S}ik}\right)\right]^{-1} \qquad (2.11)$$

where $r_{\mathcal{N}ik} = y_{ik} - \mu_{\mathcal{N}ik}$ and $r_{\mathcal{S}ik} = y_{ik} - \mu_{\mathcal{S}ik}$ are class-specific residuals: if $D = 1$, estimates of $\mu_{\mathcal{S}ik}$, $\mu_{\mathcal{N}ik}$, σ^2 and π are used to estimate the family-specific probabilities of segregation in Equation (2.11). If $D = 0$ then this probability is not estimated. For any gene, if the magnitude of the residuals is small relative to σ^2, then individual families show little variation in $P(C_i = \mathcal{S} \mid \boldsymbol{Y}_i, \boldsymbol{Z}_i)$ with values being largely driven by the term

$$\frac{(1-\pi)}{\pi}$$

On the other hand, if the data from a particular family are well fit by one of the conditional means, then the residuals for the other class will be large and this will tend to drive $P(C_i = \mathcal{S} \mid \boldsymbol{Y}_i, \boldsymbol{Z}_i)$ toward zero or one.

Hypothesis Testing

For the marginal Haseman-Elston model in Equation (2.7), the one-sided test for linkage is

$$H_0\text{: } \theta = 1/2 \text{ versus } H_1\text{: } \theta < 1/2 \qquad (2.12)$$

or equivalently

$$H_0\text{: } \beta = 0 \text{ versus } H_1\text{: } \beta < 0 \qquad (2.13)$$

where ν, p, and q are assumed to be fixed.

For the two-class model, the hypothesis regarding θ in Equation (2.12) remains of interest. However since $\beta_{\mathcal{N}} = 0$, the hypothesis regarding the slope is now

$$H_0\text{: } \beta_S = 0 \text{ versus } H_1\text{: } \beta_S < 0 \qquad (2.14)$$

The estimated test statistic, $\hat{\beta}_S$, for the hypothesis tests defined in Equation (2.13) and Equation (2.14) is defined as

$$\hat{\beta}_{\mathcal{S}} = \begin{cases} min\left(\utilde{\hat{\beta}}\right) & \text{if } D = 1 \\ \hat{\beta} & \text{if } D = 0 \end{cases} \qquad (2.15)$$

where $\utilde{\hat{\beta}} = (\hat{\beta}_{\mathcal{N}} \hat{\beta}_S)$ denotes the vector of estimated slopes from the two-class model and $\hat{\beta}$ is the estimated slope for the marginal regression. When $D = 1$ we infer that the minimum of the two slopes, min $(\utilde{\hat{\beta}})$, estimates the slope in the segregating class. When $D = 0$ the marginal slope is our best estimate of the slope in the segregating class, even in light of the possibility that $\hat{\beta}$ estimates the slope in the non-segregating class, or a mixture of segregating and non-segregating families. The test is carried out using a permutation procedure that retains the correlation structure of sibpairs within families. Specifically the observed value of $\hat{\beta}_S$ is determined. Then we re-fit the two-class model to data where the vector of sibpair outcomes for each family, $\boldsymbol{y}_i = (y_{i1}, y_{i2}, ..., y_{im_i})$, is permuted while holding the IBD sharing constant. The permutation test statistic, $\hat{\beta}_S^{(t)}$ is recomputed for the $t = 1, ...,$ T permutations. The one-sided permutation p value is

$$P = \frac{\sum_{t=1}^{T} I\left(\hat{\beta}_{\mathcal{S}}^{(t)} < \hat{\beta}_{\mathcal{S}}\right)}{T}$$

where $I(\cdot)$ is the indicator function.

Results: Analytic Findings and Simulation Study

Analytic Comparison of SqD *and* MCP

To use the two-class model for inference, the Newton-Raphson algorithm must converge to a solution that identifies two distinct classes. While the test for linkage is based on the slope in the segregating class, differences between the intercepts as well as the slopes for the segregating and non-segregating classes potentially contribute to the ability of the algorithm to converge to a two-class model. For both *SqD* and *MCP*, α_S and $\alpha_{\mathcal{N}}$ are independent of IBD sharing with differences between segregating and non-segregating classes even in the absence of linkage ($\theta = 1/2$) that increase with increasing p (result not shown). For the slopes, table 3 shows that the slope for the non-segregating families, $\beta_{\mathcal{N}}$, is zero for both *SqD* and *MCP* while, for segregating families, the β_S are identical, except for the multiplicative term

$$\frac{\left(n^3 - n^2 - 2n + 4\right)}{2n^2\left(n+1\right)}$$

which has an upper bound of 1/2 in very large families. β_S is zero when the marker is unlinked ($\theta = 1/2$) and non-zero under linkage ($0 < \theta < 1/2$). Unlike the intercepts, the slopes contribute no information useful for discerning classes in the absence of linkage. Intuitively, as segregating families become rare, the ratio for the marginal and segregating slopes should become large. Indeed, the ratio of slopes is identical to the proportion of segregating families; using the results in table 1 for either *SqD* or *MCP* and Equation (2.1)

$$\frac{\beta}{\beta_S} = 4pq\left(1 - pq\right) = \pi \tag{3.1}$$

Simulation Study

Parameters of the Simulation Study

Simulation was used to assess the operating characteristics of the proposed permutation test for the two-class model compared to a permutation test based on the marginal model. The simulation study also described the capacity of the approach to discriminate between segregating and non-segregating families. Using the genetic model, parental alleles were drawn from two unlinked (independent) loci, using a Bernoulli(p) distribution and assuming Hardy-Weinberg equilibrium. One of the alleles was arbitrarily designated as the QTL and the other as an 'Unlinked SNP' or null marker. We then simulated the random transmission of alleles from parent to each of eight offspring, or in a limited number of cases to four offspring. We determined the true IBD status of all sib-pairs. Phenotypic values were determined conditional on the QTL at each locus and assuming the additive genetic model from Equation (2.3) and a normal distribution for F_i and e_{ij}, with variances 0.3 and 0.2, respectively. To assess type I error rate, we fit the latent class model to each set of simulated data using IBD sharing at a 'null' marker, i.e. a marker independent of phenotype. We simulated 500 sets of data, varying the number of families, N, the additive allelic effect, ν, and the allele frequency p. Allele frequencies of $p = 0.025$, 0.2, and 0.5 were chosen to represent a range from rare to common alleles. We computed the nominal type I error and power as well as the mean and empirical variance of the estimated slope in the segregating class. We also assessed the discriminative ability of the proposed method conditional on achieving convergence for the two-class model; specifically, we assessed the predictive value positive *(PVP)* (the probability that a member of the nominal segregating class was a segregating family), sensitivity (probability of assignment of a segregating family to the nominal segregating class) and specificity (probability of assignment of a non-segregating family to the nominal non-segregating class) by assigning a family to either the segregating or non-segregating class based on whether the probability in Equation (2.11) exceeded 0.50.

Type I Error and Power

Table 4 shows the results for $N = 14$; similar patterns were observed for a more limited study for $N = 28$ with the same number of sibs per family. The type I error rate is well controlled; for a nominal type I error rate of 0.05, and a simulation error of 0.0097, the empirical error ranged from 0.046 to 0.068 for *SqD* and from 0.036 to 0.056 for *MCP*. Power demonstrates two distinct patterns. For the two-class model, when segregating families are rare ($p = 0.025$; $\pi = 0.095$), *MCP* has better power than *SqD*; when segregating families are common ($p = 0.2$; $\pi = 0.54$ or $p = 0.5$; $\pi = 0.75$), *SqD* has better power than *MCP*. In marginal models, as expected from theoretical results [16, 17], *MCP* has better power than *SqD* across the range of simulations. The question of whether the marginal or the two-class model has better power again depends on the frequency of the segregating families. At low allele frequencies, the two-class model with *MCP* had the highest power; at high allele frequencies, the marginal model, again using *MCP*, had the highest power.

Discrimination of Segregating and Non-Segregating Families

With respect to discriminating between segregating and non-segregating families under the two-class model, *PVP* values for *MCP* consistently exceeded those of *SqD*. For $p = 0.025$, *PVP* values for *MCP* were 35–80%, and while these values are somewhat low, they represent a substantial enrichment in segregating families in the nominal segregating class compared to the 9.5% expected for the entire sample at this allele frequency. In contrast, *PVP* values for *SqD* were 16–25% at $p = 0.025$, indicating that the nominal segregating class was dominated by non-segregating families. At higher allele frequencies discrimination is excellent; *PVP* values for *MCP* were at least 97% for $p \geq 0.2/\pi \geq 0.54$ and at least 87% for *SqD*.

Table 4. Type I error rate, at an unlinked SNP, or null marker, and power, at the QTL, for the hypothesis tests based on the two-class and marginal tests along with *PVP* for the two-class approach

p	*π*	*n*	Power/type I error				*PVP*, %	
			two-class		marginal			
			SqD	*MCP*	*SqD*	*MCP*	*SqD*	*MCP*
Unlinked SNP (null marker)								
0.025	0.095	0.8	0.052	0.056	0.058	0.054	9.7	8.5
0.025	0.095	1.0	0.046	0.052	0.048	0.060	9.4	8.1
0.025	0.095	1.2	0.052	0.056	0.064	0.064	9.6	8.5
0.200	0.54	0.8	0.060	0.060	0.066	0.068	52.6	54.5
0.200	0.54	1.0	0.054	0.056	0.050	0.058	51.5	53.3
0.200	0.54	1.2	0.068	0.036	0.048	0.058	52.5	53.5
0.500	0.75	0.8	0.056	0.054	0.046	0.056	74.0	75.9
0.500	0.75	1.0	0.036	0.050	0.038	0.038	74.9	74.4
0.500	0.75	1.2	0.040	0.046	0.028	0.022	74.7	73.3
QTL								
0.025	0.095	0.8	0.232	0.288	0.200	0.242	15.5	34.8
0.025	0.095	1.0	0.336	0.448	0.312	0.338	22.9	69.5
0.025	0.095	1.2	0.490	0.584	0.440	0.482	24.6	79.6
0.200	0.54	0.8	0.766	0.696	0.894	0.934	87.0	97.0
0.200	0.54	1.0	0.928	0.784	0.980	0.990	93.5	99.0
0.200	0.54	1.2	0.950	0.862	0.990	0.996	97.9	99.7
0.500	0.75	0.8	0.910	0.788	0.988	0.998	96.6	99.4
0.500	0.75	1.0	0.950	0.884	1.00	1.00	98.7	99.5
0.500	0.75	1.2	0.960	0.900	1.00	1.00	99.6	99.7

Results shown are for 14 families, 8 sibs per family, and 500 simulations.

Factors Affecting Relative Power and Detection of Segregating Families

To better understand these patterns, table 5 along with figures 1 and 2 explore the impact of factors including the rate of convergence to a two-class solution, and the composition of the nominal segregating class. This in turn leads to a discussion of how bias in $\hat{\beta}_S$ and its empirical standard error may impact relative power.

Probability of Converging to a Two-Class Solution

First consider convergence rates ($P(D = 1)$) (table 5). At the 'null marker', two classes of families are present and this variation is reflected in the distribution of the phenotype, but there is no association between IBD sharing and phenotypic variation. Here differences in intercepts potentially provide information about class membership, while at the QTL we expect both intercepts and slopes to contribute to discrimination between classes. For 14 families, rates of fit for the two-class model ($P(D = 1)$) at the null marker ranged approximately between 95 and 97% for *SqD* and between 35 and 38% for *MCP*. Convergence rates for both phenotypes were similar to convergence rates in simulations where there was no genetically based heterogeneity and where we generated only a random intercept and held $\nu = 0$ (results not shown). Thus the intercepts in practice provide little discriminatory information. Importantly, the probability of observing families from both classes (both segregating and non-segregating families) is 0.753 when $p = 0.025/\pi = 0.095$, and over 0.98 at the two higher allele frequencies. Focussing on $p = 0.025/\pi = 0.095$ it is clear that *SqD* frequently finds two classes when only one exists. The results for the simulation where $\nu = 0$ indicate *SqD* does so independently of any genetic information. *SqD* may converge more frequently than *MCP* because of the shape of the distribution. For example, *SqD* had skewness/kurtosis values of around –2.8/11 compared to –0.8/8.6 for *MCP* when $\nu = 1.0$ and $p = 0.2$.

At the *QTL*, *SqD* continued to have higher convergence rates than *MCP*. Here, *SqD* converged to a two-class model in over 95% of simulations while *MCP* converged at more reasonable rates of 58–73% at $p = 0.025$ and over

Table 5. Convergence rates ($P(D = 1)$), sensitivity and specificity at an unlinked SNP or null marker, and the QTL

p	π	n	$P(D = 1)$, %		Sensitivity, %		Specificity, %	
			SqD	*MCP*	*SqD*	*MCP*	*SqD*	*MCP*
Unlinked SNP (null marker)								
0.025	0.095	0.8	95.6	36.2	53.9	52.9	48.4	45.2
0.025	0.095	1.0	95.6	35.2	52.0	50.0	49.0	46.2
0.025	0.095	1.2	96.4	37.1	53.1	50.6	49.2	45.9
0.200	0.54	0.8	96.4	35.8	49.4	51.8	48.7	49.0
0.200	0.54	1.0	97.4	35.2	49.0	54.9	46.1	43.3
0.200	0.54	1.2	96.0	36.0	49.8	53.5	47.9	44.8
0.500	0.75	0.8	96.0	38.4	49.8	50.1	48.8	51.9
0.500	0.75	1.0	96.8	37.2	49.7	49.3	51.5	49.2
0.500	0.75	1.2	97.2	34.6	49.8	47.6	50.5	49.7
QTL								
0.025	0.095	0.8	95.6	58.3	69.7	25.6	48.6	93.8
0.025	0.095	1.0	95.6	71.8	74.0	41.4	73.3	97.6
0.025	0.095	1.2	96.4	73.2	80.1	55.3	75.6	98.3
0.200	0.54	0.8	96.4	97.0	48.4	25.7	91.7	99.1
0.200	0.54	1.0	96.0	99.4	50.8	33.5	95.9	99.6
0.200	0.54	1.2	99.8	99.6	54.1	42.4	98.6	99.8
0.500	0.75	0.8	96.0	99.4	40.0	22.1	95.9	99.6
0.500	0.75	1	96.8	99.5	41.9	29.5	98.4	99.6
0.500	0.75	1.2	97.2	99.7	44.4	33.5	99.5	99.7

Results for sensitivity and specificity include only those simulations where the two-class model converged ($D = 1$). Results shown are for 14 families, 8 sibs per family, and 500 families.

97% at higher allele frequencies. Again, we note that *SqD* converged at rates of 95–99% when $p = 0.025$ even when the probability of at least one segregating family in the sample was only 0.753.

Composition of Nominal Segregating Class

With respect to the composition of the nominal segregating class based on the posterior probabilities from Equation (2.11), *SqD* consistently shows higher sensitivity and *MCP* consistently shows higher specificity at the *QTL*. The net result, seen in table 4, is that the *PVP* for *MCP* is higher. However, while the nominal segregating class for *MCP* is highly enriched in segregating families, its low sensitivity means that a large proportion of segregating families in *MCP* are assigned to the non-segregating class. This observation appears important in understanding relative power.

Power of *MCP* versus *SqD* in the Two-Class Model

We first consider the relative power of *MCP* and *SqD* for the two-class model (table 4) and suggest why *MCP* is more powerful at low allele/segregating family frequency. When the algorithm fits the slopes for the two-class model, each family contributes to the estimates from both classes, with the relative contribution weighted by the goodness of fit of the family-level data to each class. With the caveat that the weights in the fitting algorithm are continuous rather than binary, we gain some insight into the relative weighting of families to the slopes by considering the composition of the nominal segregating family class. Figure 1 shows that for *SqD*, min ($\hat{\beta}$), the estimated slope in the segregating class from the two-class fit (Equation (2.15)), underestimates β_S, the true parameter, when $p = 0.025$, a finding consistent with the result that the nominal segregating class for *SqD* at $p = 0.025$ is dominated by non-segregating families ($PVP \leq 25\%$). In contrast for *MCP* $\hat{\beta}_S$, the estimated slope in the segregating class from the two-class fit (Equation (2.15)), substantially overestimates β_S when $p = 0.025$. The bias may reflect the fact that the nominal segregating class is dominated by segregating families while sensitivity for *MCP* at this allele frequency is no greater than 55%, suggesting that the contribution of a few segregating families to the slope estimate, probably those with extreme slopes, is large while those segregating families with less extreme slopes make little contribution to the estimate of β_S. Underestimation of β_S by *SqD* and overestimation by *MCP* is consistent with the finding that *MCP* is more powerful than *SqD* at low allele frequency for the two-class model.

We next suggest an explanation for why *SqD* is more powerful at higher allele/segregating family frequency. Because we do not have closed-form solutions for variance estimates, we focus on relative power using *MCP* at $p = 0.025$ as the baseline. At higher allele frequencies ($p = 0.2$ or $p = 0.5$), the nominal segregating classes for both *SqD* and *MCP* are dominated by segregating families. Sensitivity, while lower than at $p = 0.025$, is higher for *SqD*. For both phenotypes, min ($\hat{\beta}$), the slope estimate in the segregating class, overestimates β_S. Since the magnitude of the relative bias is larger for *MCP* (fig. 1), we suggest that the precision of the estimate, rather than the bias may explain why *SqD* has better power than *MCP* for the two-class model at higher allele frequencies (table 4). We consider increases in the mean slope and the empirical standard deviation (SD) at the higher allele frequency relative to $p = 0.025$. Specifically for each phenotype we computed two ratios, the slope ratio is the mean of the slope estimates at $p = 0.2$ versus $p = 0.025$, and the SD ratio is the empirical SD of the slope estimates at $p = 0.2$ versus $p = 0.025$. Intuitively, we hypothesized that larger slope ratios would tend to yield a larger increase in power, while larger SD ratios would have an opposite effect, yielding a smaller increase or decrease in power. Figure 2 (top row) illustrates the ratios. For the nominal segregating class, the slope ratio was higher and the SD ratio was smaller for *SqD* compared to *MCP*. Slope ratios were 2.5 to 3.1-fold larger for each value of ν for *SqD* and 1.8 to 2.0-fold larger for *MCP*. In contrast, ratios for the empirical SD were 1.3 to 1.6-fold larger for *SqD* and 1.7 to 1.9-fold larger for *MCP*. Similar patterns were observed for $p = 0.500$ versus $p = 0.025$. Thus, while *MCP* is highly specific across allele frequencies, the variance of the slope estimate in the segregating class is relatively higher than *SqD* for more common versus rare alleles. We speculate that compared to *SqD*, *MCP*'s low sensitivity effectively reduces the size of the nominal segregating class when allele frequencies are larger, and this inflates the standard error and reduces power relative to *SqD*.

Power of Marginal versus Two-Class Model

Lastly, we seek to understand why, for each phenotype, the two-class model had better power than its marginal counterpart at low allele frequency while the marginal model had better power at higher allele frequency. We focus on *MCP* as this was the phenotype with the best absolute power. At $p = 0.025$ the ratio of population slopes for the segregating class (β_S) and the marginal model (β) is 10.5, reflecting the fact that only 9.5% of families in the population are segregating (Equation (3.1)); in fact the observed slope ratio for the segregating versus marginal models for *MCP* was even higher, ranging from 11.7 to 21. The power of the two-class model at low allele frequency appears to reflect, at least in part, the large contribution of non-segregating families to the marginal slope along with the enrichment of segregating families in the nominal segregating class of the two-class fit and, for *MCP*, the upward bias of $\hat{\beta}_S$, in cases where the two-class *MCP* model converges.

At higher allele frequency, we speculate that the higher power of the marginal approach, compared to the two-class approach, reflects a smaller contribution of non-segregating families to the marginal slope, along with

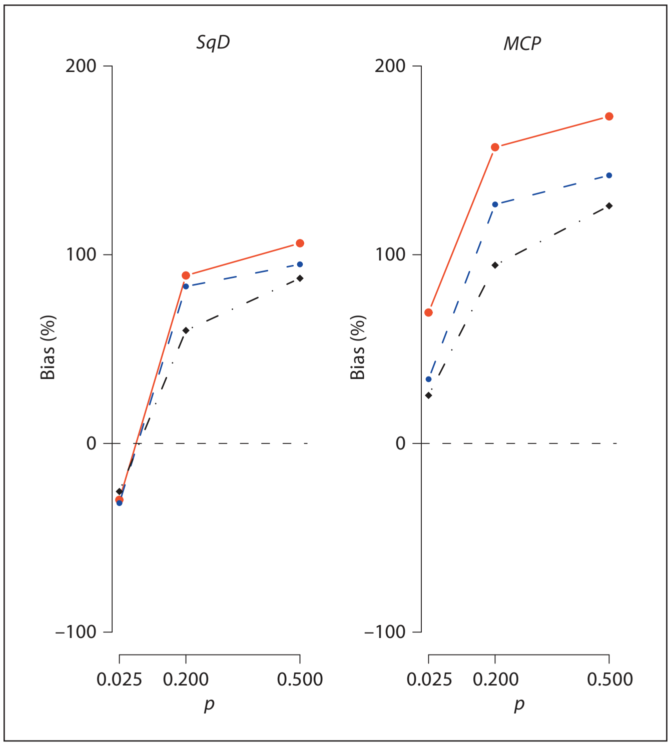


Fig. 1. Relative bias of *SqD* and *MCP* slope estimates as a function of allele frequency *p* for $\nu = 0.8$ (solid line, red), 1.0 (dashed line, blue) and 1.2 (dash-dot line, black) in the nominal segregating class of the two-class model (colors refer to online version; white, grey, and black, respectively, in the print version). Relative bias is the difference between the mean of the estimated slope and its expectation expressed as a percentage of the expectation. Horizontal dashed line indicates absence of bias.

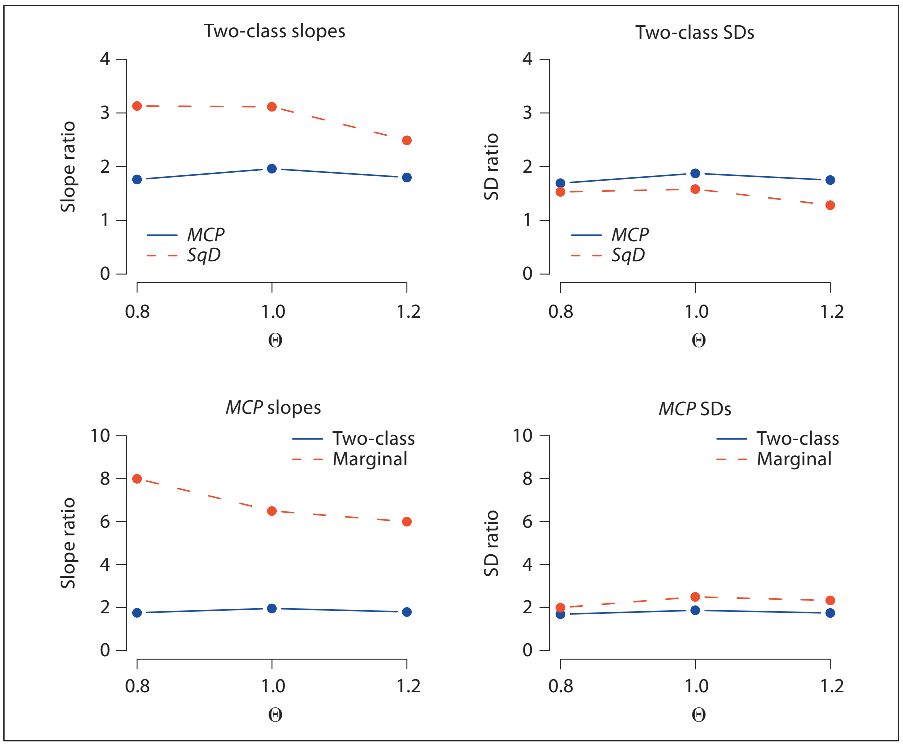


Fig. 2. Observed slope and standard deviation ratios for $p = 0.200$ versus $p = 0.025$ at different values of ν. The top row compares ratios for *MCP* (solid line, blue) and *SqD* (dashed line, red) for the two-class model; the bottom row compares ratios for *MCP* for the two-class (solid line, blue) and the marginal model (dashed line, red). Note that the two-class results (solid line, blue) represent identical data but the y-axis scales differ in the two rows (colors refer to online version; black and white in the print version).

Color version available online

poor sensitivity of the two-class model, i.e. a relatively poor ability to classify segregating families as such. For *MCP* at $p = 0.2$ the ratio of population slopes for the segregating class in the two-class model and the marginal model, β_S/β, is 1.9, a value much smaller than at $p = 0.025$. However, the ratio of observed slopes ranged from 3.5 to 4.6, again reflecting the unbiasedness of the marginal slope and the upward bias of $\hat{\beta}_S$. Since the slope estimates for the segregating class in the two-class model continue to be much higher than for the marginal model, we suggest that the precision of the slope estimate may explain why the marginal model has better power than the two-class model at higher allele frequencies (table 4). Because we do not have closed-form solutions for the variance of the slope estimate, we use $p = 0.025$ as a reference and considered relative values of the slopes and standard deviations. For the marginal model the slope ratio is the mean of the estimated β at $p = 0.2$ versus $p = 0.025$, and the SD ratio is the empirical SD of the slope estimate at $p = 0.2$ versus $p = 0.025$. Figure 2 (bottom row) shows the slope and SD ratios for the marginal and two-class models for *MCP* at different values of ν. As described above, slope ratios for the *MCP* two-class model are 1.8 to 2.0; figure 2 shows that for the marginal model the slope ratios are much larger, on the order of 6.0 to 8.0. In contrast for the two-class model, the SD ratios are 1.7 to 1.9, and for the marginal model, they are only slightly higher, 2.0 to 2.5. Thus the relative increase in the slope for the marginal model compared to the two-class model appears to outpace the relative increase in SD. A comparison of the marginal *MCP* model to the two-class *SqD* model yields similar results. We hypothesize that at larger allele frequencies, the low sensitivity of the two-class model reduces the size of the nominal segregating class, effectively inflating the standard error of the slope estimate. This in turn reduces the power of the two-class model relative to the marginal model. The relative power of the marginal and two-class approaches appears to reflect trade-offs between the relative magnitude of the slope in the segregating class and the overall population, and the inability of the two-class model to assign the majority of segregating families to the nominal segregating class.

Effect of Reduced Number of Sibs per Family

At the suggestion of the reviewers, we briefly explored the impact of a smaller number of sibships. Results of a simulation with 28 families and four sibs per family are shown in table 6 for $\nu = 1$ at the QTL. The type I error rate was controlled and appeared somewhat conservative for both phenotypes (not shown). The results were similar to

Table 6. Power, convergence rates ($P(D = 1)$), sensitivity and specificity at the QTL

p	π	Power				*PVP*, %		$P(D = 1)$, %		Sensitivity, %		Specificity, %	
		two-Class		marginal									
		SqD	*MCP*	*SqD*	*MCP*	*SqD*	*MCP*	*SqD*	*MCP*	*SqD*	*MCP*	*SqD*	*MCP*
0.025	0.095	0.270	0.132	0.252	0.198	14.2	14.7	99.0	56.8	54.9	23.8	66.4	94.9
0.200	0.54	0.444	0.160	0.856	0.846	87.4	93.9	99.8	88.8	28.0	09.0	95.4	99.3
0.500	0.75	0.554	0.326	0.972	0.980	97.8	99.2	99.6	95.8	24.9	10.1	99.2	99.8

Results for sensitivity and specificity include only those simulations where the two-class model converged (D = 1). Results shown are for 28 families and 4 sibs per family across 500 simulations for $\nu = 1$.

those described although both the marginal and two-class models yielded hypothesis tests with lower power. Additionally, for the smaller number of sibships, the marginal model had substantially, better power than the two-class model, except for $p = 0.025$ for the *SqD* function. For the two-class model, *SqD* consistently had better power than *MCP*. As with the larger sibship size, *MCP* consistently had better specificity and worse sensitivity than *SqD*, and this yielded better *PVP* for the *MCP* phenotype. The *PVP* values were consistently lower than for the larger sibship sizes, but still exceeded 90% for *MCP* for $p \geq 0.20$.

We were unable to carry out the permutation test successfully using fewer than four sibs per family. It appeared that because the number of permutations was small (six per family for three sibs per family), that the within-family permutation did not yield sufficient variation to differentiate between the results under the null and alternative hypotheses, leading to a test with literally no power. While the permutation procedure could not be used, we were still able to fit the two-class model, and this information could be used to classify families.

Results: Application to the CEPH Study

Description of the CEPH Study

The motivating study is fully described in Morley et al. [2]. Briefly, 3,554 gene expression phenotypes were analyzed for linkage in a genome-wide study of cell lines from 14 Utah pedigrees with seven to nine offspring. Using a recent modification of Haseman-Elston regression [21], 142 phenotypes yielded sufficient linkage evidence to achieve genome-wide significance (p values $\leq 4.3 \times 10^{-7}$). These expression phenotypes were further classified based on location of the linked SNP marker relative to each expression phenotype. Target genes with expression mapped to within 5 Mb of their genomic location were classified as *cis*-regulated; otherwise genes were classified as *trans*-regulated. Here, we illustrate the two-class approach using 27 expression phenotypes which demonstrated statistically significant evidence of linkage to a *cis* regulator in the original study [2]. We selected the single SNP closest to the start of transcription of each of these 27 genes. While we identified a single SNP for the analysis, we used *Merlin* to estimate the number of alleles shared IBD at the SNP marker using genotype data from all SNPs on the chromosome containing the target gene using [24]. *Merlin* is a multipoint algorithm which uses information from SNPs in the neighborhood of the SNP of interest to infer IBD status. Genotype data from grandparents, parents, and offspring were used in the IBD calculations, although only outcome data from the offspring were used in the latent class analysis. This approach suggests that IBD sharing at the QTL for these data should be well approximated by IBD sharing at a nearby genetic marker.

Hypothesis Tests

Table 7 shows the full set of parameter estimates for the *MCP* function ordered by $\hat{\pi}$, their estimated probability of membership in the segregating class. p values for hypothesis tests for both *SqD* and *MCP* are shown for both the marginal and two-class models. For both functions, the two-class model converged for all genes. The slope estimates followed the expected pattern: $\hat{\beta}_{\mathcal{N}}$ was very close to zero, and $\hat{\beta}$ for the marginal model was intermediate to $\hat{\beta}_S$ and $\hat{\beta}_{\mathcal{N}}$. Estimates of minor allele frequencies, $\hat{p}$, for those genes where the latent class model converged ranged from 0.018 to 0.127; these values correspond to estimated prevalences of segregating families

Table 7. Results of latent class analysis for expression phenotypes

Gene	Estimates					p values			
	$\hat{\beta}$	$\hat{\beta}_{\mathcal{N}}$	$\hat{\beta}_{S}$	$\hat{\pi}$	$\hat{p}$	marginal		two-class	
						MCP	*SqD*	*MCP*	*SqD*
CHI3L2	0.801	0.603	3.741	0.071	0.018	<0.001	<0.001	**0.005**	**0.009**
CTBP1	0.038	0.013	0.253	0.071	0.018	<0.001	<0.001	**0.001**	<0.001
CTSH	0.030	0.021	0.151	0.071	0.018	<0.001	<0.001	<**0.001**	0.011
PPAT	0.044	0.025	0.332	0.071	0.018	<0.001	<0.001	**0.005**	0.044
SMARCB1	0.023	0.008	0.179	0.071	0.018	<0.001	0.003	0.001	<**0.001**
ZNF85	0.106	0.005	0.922	0.071	0.018	<0.001	0.001	**0.001**	0.354
POMZP3	0.216	0.148	1.072	0.071	0.018	<0.001	<0.001	<**0.001**	0.007
VAMP8	0.019	0.009	0.094	0.131	0.033	<0.001	0.011	0.042	**0.038**
CPNE1	0.062	0.023	0.330	0.143	0.036	<0.001	<0.001	0.002	**0.001**
CSTB	0.032	0.014	0.135	0.143	0.036	<0.001	0.001	**0.007**	0.010
ICAP-1A	0.048	0.010	0.179	0.143	0.036	<0.001	<0.001	**0.001**	**0.001**
S100A13	0.053	0.002	0.336	0.143	0.036	<0.001	0.013	**0.006**	0.044
GSTM1	0.098	0.054	0.282	0.180	0.045	<0.001	<0.001	0.114	**0.027**
LOC388796	0.090	0.028	0.289	0.217	0.055	<0.001	<0.001	0.023	<**0.001**
IRF5	0.072	0.032	0.204	0.220	0.055	<0.001	<0.001	0.061	**0.006**
EIF3S8	0.022	0.009	0.057	0.233	0.059	<0.001	<0.001	0.071	**0.036**
RPS26	0.014	0.003	0.051	0.250	0.063	<0.001	<0.001	**0.067**	0.093
PSPHL	0.685	0.056	2.165	0.282	0.071	<0.001	<0.001	0.038	**0.001**
TM7SF3	0.074	0.008	0.282	0.286	0.072	<0.001	0.001	0.048	**0.002**
DDX17	0.091	0.002	0.267	0.294	0.074	<0.001	<0.001	0.154	**0.013**
HSD17B12	0.021	0.007	0.057	0.306	0.077	<0.001	<0.001	0.021	**0.001**
TCEA1	0.010	–0.002	0.037	0.311	0.079	0.005	0.046	**0.130**	0.212
IL16	0.026	0.002	0.083	0.324	0.082	0.010	0.023	0.156	**0.040**
CGI-96	0.074	0.017	0.176	0.360	0.091	<0.001	<0.001	0.046	**0.008**
GSTM2	0.070	0.001	0.165	0.434	0.112	<0.001	<0.001	0.131	**0.035**
LOC64167	1.047	–0.002	2.190	0.490	0.127	<0.001	<0.001	0.050	**0.004**

Complete results are shown for the *MCP* phenotype. p values for the *SqD* phenotype also shown with bold values indicating the smaller of the two p values. Nominal segregating and non-segregating families for genes in bold are illustrated in figure 3.

Table 8. Estimated probability of membership in the segregating class for individual families for genes in figure 3

Gene	CEPH family identifier													
	1333	1340	1341	1345	1346	1347	1362	1408	1416	1418	1421	1423	1424	1454
SMARCB1	0.010	0.006	0.003	0.002	0.002	0.004	0.009	0.006	0.004	0.004	0.018	0.793	0.000	0.002
ZNF85	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
POMZP3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000
CGI-96	0.166	0.124	0.175	0.171	0.564	0.689	0.544	0.614	0.285	0.141	0.713	0.164	0.150	0.105
PSPHL	1.000	0.000	0.000	0.006	0.000	0.000	0.000	0.997	0.000	0.000	1.000	0.000	0.000	0.993
TM7SF3	0.035	0.707	0.012	0.127	0.055	0.627	0.008	0.037	0.874	0.711	0.021	0.016	0.038	0.101

of between 0.07 and 0.49. Values of $\hat{p}$ were generally higher for *SqD* ranging up to $\hat{p} = 0.276$ and $\hat{\pi} = 0.64$ (results not shown). As expected for genes that reached strict statistical significance for linkage in the genome-wide study reported in [2], permutation p values in the test based on the marginal model were generally small (<0.001). p values for the two-class hypothesis test in Equation (2.13) were generally larger than for the marginal model, but were less than 0.05 in 19 out of 27 genes for *MCP* and for 24 out of 27 genes for *SqD*. Based on the simulation study, we predicted that p values would be smaller for *MCP* than for *SqD* at the lower allele frequencies with a reversal in this pattern at the higher allele frequencies. Indeed, we observed eight genes with $\hat{p} = 0.018$ and $\hat{\pi} = 0.071$, and of these genes the p value was smaller for *MCP* than for *SqD* in seven cases. Of the 13 genes with the highest allele frequencies, specifically with $\hat{p} > 0.05$ and $\hat{\pi} > 0.20$, the p value was larger for *MCP* than for *SqD* in 11 cases.

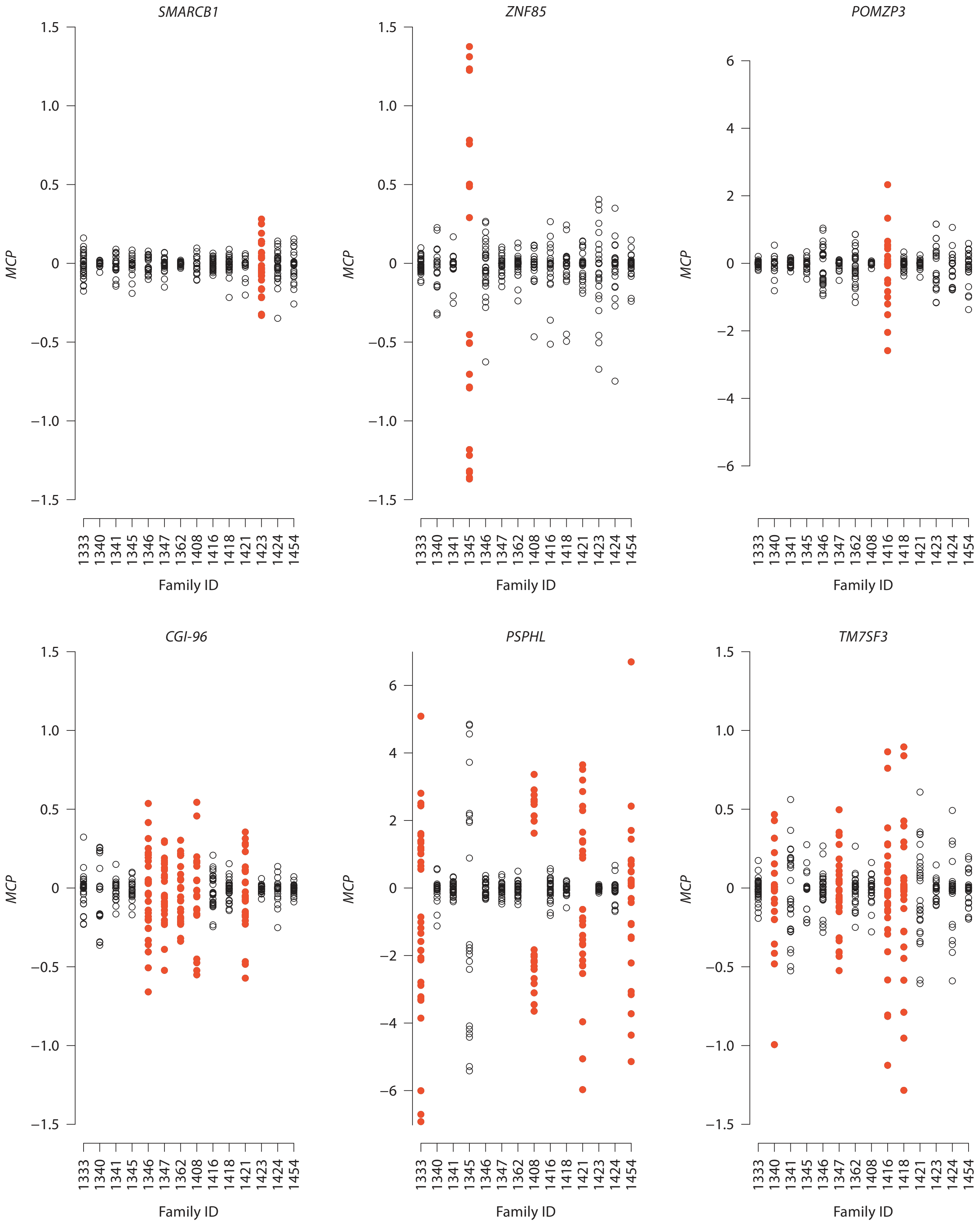
Designation of Segregating Families

Figure 3 illustrates *MCP* for each family in six example genes with estimated allele frequencies, $\hat{p}$, ranging from 0.018 to 0.091. Note that the scale for *MCP* differs among genes. Families were designated as segregating if estimates of the class-specific probabilities $P(C_i = \mathcal{S} \mid \mathbf{Y}_i, \mathbf{Z}_i)$ from Equation (2.10) exceeded 0.50; data for these families appear in red (grey in the print version). Class-specific probabilities of membership in the segregating class appear in table 8. As expected individual families differed in their designation as a segregating family across genes. Typically, families with the largest range in *MCP* values were designated as segregating while families with the smallest range in *MCP* values were designated as non-segregating. For *PSPHL* and *TM7SF3*, the two-class approach also designated several families as segregating that had smaller ranges in *MCP* than families that were designated as non-segregating. For these six genes, the probability of membership in a segregating class tended to approach 1.0 or 0.0 for most families. However, for *CGI-96* and *TM7SF3* there was more uncertainty in the designation of a segregating versus a non-segregating family. In interpreting these results, we keep in mind the simulation study indicating that while we have reasonable confidence that families assigned to the nominal segregating class are indeed segregating, the low sensitivity for *MCP* means that we may have falsely assigned segregating families to the non-segregating class.

Discussion

Using a novel latent class model, we have developed and evaluated an estimation and testing procedure that accounts for heterogeneity in outcome due to segregating and non-segregating families. The approach provides an objective test for linkage which may be valuable when the prevalence of segregating families in the sample is small, as well as information that can be used to identify segregating families, which in turn may inform the selection of families or individuals for future study. In our motivating study, multiple measures of gene expression were the phenotypes of interest. However, the method could also be used in studies where multiple continuous endpoints are collected but families are selected into the study without regard to phenotypic variation in the outcome of interest. One limitation to performing the latent class analysis with a large number of markers is that the permutation procedure proposed here is computationally intensive; to obtain the resolution needed to make multiple comparison adjustments the computational burden would be even larger. For illustrative purposes, we ignored the issue of multiple comparisons but in practice this adjustment would be important, particularly if a large number of candidate loci were examined.

For large sibships, the hypothesis test for *MCP* is somewhat more powerful than *SqD* in marginal, regression-based linkage analyses [17, 22]. For the two-class approach, simulation suggests that *SqD* is more powerful than *MCP* for higher allele frequency whereas *MCP* is more powerful than *SqD* for rare alleles. However, the marginal *MCP* model was more powerful than both latent class procedures at higher allele frequencies, despite ignoring the distinction between information from segregating and non-segregating families. The two-class procedure has several steps, and there are a number of factors at each step that may contribute to this finding. Our results suggest that factors that may contribute to this finding include the skewed distribution of the original phenotypes, which in turn results in higher specificity for *MCP* and higher sensitivity for *SqD*. While the results describing power are complex, the ability of the two-class approach using the *MCP* to identify segregating families was consistent and impressive, with *PVP* values that consistently exceeded 95% at higher allele frequencies. For the marginal model, additional phenotype functions have been proposed. For example, correction by the best linear unbiased predictor (BLUP) of the population mean was shown to perform better than correction by the family mean [22]. However, since the means in the segre-



Color version available online

gating and non-segregating classes differ, the marginal BLUP should be biased for the conditional means. Another phenotype function uses residuals to obtain an optimally weighted function of the squared sum and squared difference and was used in the original analysis of these data [2, 21]. Like the BLUP-based approach, this function relies on marginal residuals and is not well suited for use as the outcome in the two-class model without modification. In limited studies we found that phenotype functions which depend on marginal moments gave unexpected results, including introducing bias or reducing the ability of the algorithm to detect the latent classes.

We used maximum likelihood to fit the two-class model, assuming normality with separate means but equal variances in the two classes. We further assumed that sibpairs within a family are independent, conditional on the latent class. Each of these three assumptions (normality of the phenotype functions, equality of variance and conditional independence) are undoubtedly violated, and it is perhaps not surprising that the slope estimates in the two-class model are highly biased. Early in our study, we found that when the two-class algorithm converged, the maximum likelihood estimates of the variance seriously underestimated the empirical variance of the slope in the segregating class. This finding is consistent with results from the simulation study, essentially ruling out the possibility of using a Wald-type test [Bastone, unpubl. data]. An additional problem in the development of a valid hypothesis test arose as a result, of the algorithm failing to consistently converge to a two-class solution. As a result the test statistic involves both the estimated slope from the two-class model, and in the event of non-convergence, the estimated slope in the marginal model. Permutation was used to address these issues. The mixture distribution of the test statistic was estimated under the null hypothesis using permutation. The procedure yields valid type I error rates, and allows the procedure to be used irrespective of the bias in the estimate, and whether a two-class model can be fit to the observed data. In developing this approach, we initially used a permutation procedure where we accepted the null hypothesis when the two-class model failed to converge [13, 26]. In simulation studies, this approach yielded acceptable type I error rates, but, obviously, failed to detect linkage in any gene where the two-class model did not fit and yielded lower power than the procedure detailed here.

Fig. 3. *MCP* for selected families with $\hat{\pi}$ ranging from 0.07 to 0.29. For each gene, families designated as segregating, based on a posterior probability for membership in the segregating class of at least 0.5, are shown in red (grey in the print version). Note the larger scale for the dependent variable for *PSPHL* and *POMZP3* compared to the other four genes. Posterior probability for membership in the segregating class are shown in table 7.

In practice one of the limitations of this study is the computational burden of the permutation procedure and in practice the computational burden would be even larger when multiple markers are used. Permutation provides a valid test when assumptions about the distribution of the outcome and the conditional independence are violated. Morever, for the method proposed here, the test statistic has a mixture distribution that is not easily modelled parametrically. However, permutation is computationally intensive and is not feasible for either trios or where there is a single sibpair. Within the framework of the two-class model, generalized estimating equations (GEE) provide a framework for incorporating correlation into the covariance matrix for each family while relaxing the normality and homogeneity of variance assumptions [25–27]. GEE requires correct specification of the variance just as ordinary least square (OLS), and so accomodating heterogeneous variances in GEE is explicitly done in the same way with variance weights just as with OLS under a linear model. We have made some initial progress in using GEE to fit the two-class model. By incorporating more reasonable assumptions into the model, estimation using GEE, perhaps in combination with a sequential testing procedure to account for the two sizes of models that are potentially fit, might allow us to develop a semiparametric test and, at least in larger samples, avoid the need for permutation.

While the development of a valid hypothesis test involved a number of complications, the results of the classification procedure are highly encouraging given the importance of accurately identifying segregating families in the presence of heterogeneity. Our simulation results suggest that the method using *MCP* is clearly capable of identifying segregating families with very high *PVP* when these families are relatively common. In cases where segregating families are rare, *PVP* values are lower, but still provide substantial improvements in information compared to randomly selecting families for follow-up. We note that families with larger sibships can be better classified using this approach. One limitation of our method is that the sensitivity for detection of segregating families is low. In our experience, the identification of segregating families is generally of greater interest than the identification of non-segregating families. We caution that the two-class approach described here would not be appropriate for a study where accurate identification of both types of families was needed. As with the hypothesis testing problem, it is possible that improved sensitivity could be achieved by more reasonable assumptions in the model, perhaps by using a GEE framework.

Other family-based linkage methods have been extended to account for heterogeneity using mixture models, including parametric linkage methods for qualitative traits and the variance components method for quantitative traits [28–30]. Our latent class extension of Haseman-Elston regression is not limited to distinguishing between segregating and non-segregating families. A similar approach could be used to model the genetic, or locus-based, heterogeneity of a complex trait such as a quantitative measure of disease, even when families are selected on the basis of phenotype and all families are assumed to be segregating. In contrast to the model-based approaches, ordered subset analysis and recursive partitioning are data-driven methods that have been applied to linkage problems in the presence of heterogeneity [31–34]. These approaches provide a test for linkage and assign families to classes but, unlike the method proposed here, do not include an approach for quantifying the uncertainty of class assignment through the posterior probability of membership in the segregating class.

Lastly, we note that large marker sets are now widely used in human gene-mapping. Incorporating an approach into the methods described here, that would allow us to test for linkage with multiple loci, is a topic of further study.

Acknowledgements

Research supported by NIH grants T32-GM08216 (L.A.B.) and NICHD P30 HD026979 (M.E.P.). The data analyzed here were obtained with support from NIH grants HG 002386 and GM081930 (R.S.S. and Dr. Vivien G. Cheung). Dr. Bobby Jones provided inputs for successful implementation of PROC TRAJ; Drs. Warren Ewens, Nandita Mitra and Vivian Cheung provided comments and support during the evolution of the project. We thank the reviewers for their thoughtful comments and suggestions.

Appendix

Expectation of Squared Difference (SqD)

For the marginal model, table 1 shows the expectation of the *SqD* function conditional on IBD sharing as [15]:

$$E(W_{ij} - W_{ij'} \mid Z_{ik})^2 = 2[E(W_{ij} \mid Z_{ik}) - E(W_{ij}W_{ij'} \mid Z_{ik})]$$
$$= 2[Var(W_{ij}) - Cov(W_{ij}W_{ij'} \mid Z_{ik})]$$

The terms from table 2 needed to derive the slope in table 3 and intercept in terms of the genetic model are the variance of W_{ij}, which follows directly from the genetic model (Equation (2.3)), as well as the conditional covariance. Using the definition of W_{ij} and results previously derived by Haseman and Elston [15], the covariance of W_{ij} and $W_{ij'}$ at the QTL conditional on membership in sibpair k is:

$$\begin{aligned} Cov(W_{ij}\ W_{ij'} \mid X_{ik}) &= \sigma_F^2 + \nu^2 Cov(G_{ij}G_{ij'} \mid X_{ik}) \\ &= \sigma_F^2 + \nu^2 E(G_{ij}G_{ij'} \mid X_{ik}) - E(G_{ij})^2 \\ &= \sigma_F^2 + \nu^2 pqX_{ik} \end{aligned}$$

Considering sharing at the SNP marker leads directly to the term in Equation (2.5). Similarly, conditioning on class in addition to IBD sharing yields:

$$E(W_{ij} - W_{ij'} \mid Z_{ik}C_i)^2 = 2[Var(W_{ij} \mid C_i) - Cov(W_{ij}W_{ij'} \mid Z_{ik}C_i)]$$

For each class, $Cov(W_{ij}W_{ij} \mid Z_{ik}C_i)$ is derived in Appendix B of Bastone [26].

Expectation of Mean Corrected Product (MCP)

For the marginal model, table 1 shows the expectation of the *MCP* function conditional IBD on sharing as:

$$E[-(W_{ij} - \overline{W}_i)(W_{ij'} - \overline{W}_i) \mid Z_{ik}] = Var(\overline{W}_i \mid Z_{ik}) + Cov(W_{ij}W_{ij'} \mid Z_{ik}) - 2Cov(W_{ij}\overline{W}_i \mid Z_{ik}) \quad \text{(A.1)}$$

Note that

$$Var(\overline{W}_i | Z_{ik}) = \frac{1}{n} Var(W_{ij}) + \frac{2}{n^2} Cov(W_{ij}W_{ij'} | Z_{ik}) + \frac{n(n-1)-2}{n^2} Cov(W_{ij}W_{ij'} | jj' \in k') \quad \text{(A.2)}$$

where $jj' \in k'$ denotes a sibpair belonging to a sibpair other than k. To derive $Cov(W_{ij}W_{ij'} \mid jj' \in k')$ the marginal covariance of a sibpair can be written as:

$$Cov(W_{ij}W_{ij'}) = Cov(W_{ij}W_{ij'} | Z_{ik}) \frac{2}{n(n-1)} + Cov(W_{ij}W_{ij'} | jj' \in k') \frac{n(n-1)-2}{n(n-1)}$$

and thus:

$$Cov(W_{ij}W_{ij'} | jj' \in k') = Cov(W_{ij}W_{ij'}) \frac{n(n-1)}{n(n-1)-2} - Cov(W_{ij}W_{ij'} | Z_{ik}) \frac{2}{n(n-1)-2} \quad \text{(A.3)}$$

Also

$$\begin{aligned} Cov(W_{ij}\overline{P}_i | Z_{ik}) &= \frac{1}{n} E\left[(W_{ij} - \mu) \left(\sum_j W_{ij} - \mu \right) | Z_{ik} \right] \\ &= \frac{1}{n} Var(W_{ij}) + \frac{1}{n} Cov(W_{ij}W_{ij'} | Z_{ik}) \\ &\quad + \frac{n-2}{n} Cov(W_{ij}W_{ij'} | jj' \in k') \end{aligned} \quad \text{(A.4)}$$

Using the information from Equations (A.3), (A.2) and (A.4) with Equation (A.1) yields the following result:

$$E\left[-\left(W_{ij}-\overline{P}_i\right)\left(W_{ij}-\overline{P}_i\right)|Z_{ik}\right]=\frac{1}{n}Var\left(W_{ij}\right)$$

$$+\frac{\left(n-1\right)^2}{n\left(n+1\right)}Cov\left(W_{ij}W_{ij'}\right)$$

$$-\frac{n^3-n^2-2n+4}{n^2\left(n+1\right)}Cov\left(W_{ij}W_{ij'}|Z_{ik}\right)$$

In this form the terms from table 2 can be substituted to yield the slope terms in table 3, and a computer program to numerically determine the intercepts. Similarly,

$$E\left[-\left(W_{ij}-\overline{P}_i|Z_{ik}C_i\right)\left(W_{ij}-\overline{P}_i|Z_{ik}C_i\right)\right]$$

$$=\frac{1}{n}Var\left(W_{ij}|C_i\right)-\frac{n^3-n^2-2n+4}{n^2\left(n+1\right)}Cov\left(W_{ij}W_{ij'}|Z_{ik}C_i\right)$$

$$+\frac{\left(n-1\right)^2}{n\left(n+1\right)}Cov\left(W_{ij}W_{ij'}|C_i\right)$$

For each class, $Cov(W_{ij}W_{ij'} \mid C_i)$ is derived in Appendix B of Bastone [26].

References

1 Dausset J, Cann H, Cohen D, Lathrop M, Lalouel JM, White R: Centre d'Etude du Polymorphisme Humain (CEPH): collaborative genetic mapping of the human genome. Genomics 1990;6:575–577.

2 Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG: Genetic analysis of genome-wide variation in human gene expression. Nature 2004;430: 743–747.

3 Clogg CC: Latent Class Models; in Arminger G, Clogg CC, Sobel ME (eds): Handbook of Statistical Modeling for the Social Behavioral Sciences. New York, Plenum Press, 1995.

4 Pritchard JK, Rosenberg NA: Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 1999;65:220–228.

5 Pritchard JK, Stephens M, Donnelly P: Inference of population structure using multilocus genotype data. Genetics 2000a;155:945–959.

6 Pritchard JK, Stephens M, Rosenberg NA, Donnelly P: Association mapping in structured populations. Am J Hum Genet 2000b; 67:170–181.

7 Koch R, Julius U, Jaross W, Schroeder HE: Estimation of the heritability of latent variables which are included in a structural model for metabolic syndrome. Hum Hered 2001; 52:171–176.

8 Satten GA, Flers D, Yang Q: Accounting for unmeasured population substructure in case- control studies of genetic association using a novel latent-class model. Am J Hum Genet 2001;68:466–477.

9 Todd RD, Rasmussen ER, Neuman RJ, Reich W, Hudziak JJ, Bucholz KK, Madden PA, Heath A: Familiality and heritability of subtypes of attention deficit hyperactivity disorder in a population sample of adolescent female twins. Am J Psychiatry 2001;158: 1891–1898.

10 Keel PK, Fichter M, Quadflieg N, Bulik CM, Baxter MG, Thornton L, Halmi KA, Kaplan AS, Strober M, Woodside DB, Crow SJ, Mitchell JE, Rotondo A, Mauri M, Cassano G, Treasure J, Goldman D, Berrettini WH, Kaye WH: Application of a latent class analysis to empirically define eating disorder phenotypes. Arch Gen Psychiatry 2004;61:192–200.

11 Nyholt DR, Gillespie NG, Heath AC, Merikangas KR, Duffy DL, Martin NG: Latent class and genetic analysis does not support migraine with aura and migraine without aura as separate entities. Genet Epidemiol 2004;26:231–244.

12 Purcell S, Sham P: Properties of structured association approaches to detecting population stratification. Hum Hered 2004;58:93–107.

13 Bastone LA, Putt ME, TenHave TR, Chueng VG, Spielman RS: Genetic heterogeneity and trans regulators of gene expression. BMC Proc 2007;1(suppl 1):S80.

14 Falconer DS, Mackay TFC: Quantitative Genetics, fourth edition, Harlow. Prentice Hall, 1996.

15 Haseman JK, Elston RC: The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 1972;2:3–19.

16 Wright FA: The phenotypic difference discards sib-pair QTL linkage information. Am J Hum Genet 1997;60:740–742.

17 Wright FA: Information perspectives of the Haseman-Elston method. Hum Hered 2003; 55:132–142.

18 Drigalenko E: How sib pairs reveal linkage. Am J Hum Genet 1998;63:1242–1245.

19 Elston RC, Buxbaum S, Jacobs KB, Olson JM: Haseman Elston revisited. Genet Epidemiol 2000;19:1–17.

20 Forrest WF: Weighting improves the 'New Haseman-Elston' method. Hum Hered 2001; 52:47–54.

21 Shete S, Jacobs KB, Elston RC: Adding further power to the Haseman-Elston method for detecting linkage in larger sibships: weighting sums differences. Hum Hered 2003;55:79–85.

22 Wang T, Elston RC: A modified revisited Haseman-Elston method to further improve power. Hum Hered 2004;57:109–116.

23 Jones BL, Nagin DS, Roeder K: A SAS procedure based on mixture models for estimating developmental trajectories. Sociol Methods Res 2001;29:374–393.

24 Abecasis GR, Cherny SS, Cookson WO, Cardon LR: Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 2002;30:97–101.

25 Liang KY, Zeger SL: Longitudinal data analysis using generalized linear models. Biometrika 1986;73:3–22.

26 Bastone LA: Methods for the genetic analysis of complex continuous traits, University of Pennsylvania, Philadelphia, PA. PhD Dissertation, 2007.

27 Reboussin BA, Liang KY, Reboussin DM: Estimating equations for a latent transition model with multiple discrete indicators. Biometrics 1999;55:839–845.

28 Hodge SE, Anderson CE, Neiswanger K, Sparkes RS, Rimoin DL: The search for heterogeneity in insulin-dependent diabetes mellitus (IDDM): linkage studies, two-locus models, genetic heterogeneity. Am J Hum Genet 1983;35:1139–1155.

29 Ott J: Linkage analysis and family classification under heterogeneity. Ann Hum Genet 2003;47:311–320.

30 Ekstrom CT, Dalgaard P: Linkage analysis of quantitative trait loci in the presence of heterogeneity. Hum Hered 2003;55:16–26.

31 Hauser ER, Watanabe RM, Duren WL, Bass MP, Langefeld CD, Boehnke M: Ordered subset analysis in genetic linkage mapping of complex traits. Genet Epidemiol 2004;27:53–63.

32 Shannon WE, Province MA, Rao DC: Tree-based recursive partitioning methods for subdividing sibpairs into relatively more homogeneous subgroups. Genet Epidemiol 2001;20:293–306.

33 Costello TJ, Swartz MD, Sabripour M, Gu X, Sharma R, Etzel CJ: Use of tree-based models to identify subgroups increase power to detect linkage to cardiovascular disease traits. BMC Genet 2003;4(suppl 1):S66.

34 Xu W, Schulze TG, De Paulo JR, Bull SB, McMahon FJ, Greenwold CMT: A tree-based model for allele-sharing-based linkage analysis in human complex diseases. Genet Epidemiol 2006;30:155–169.