

Binomial Mixture Model-based Association Tests under Genetic Heterogeneity

HUI ZHOU, WEI PAN

*Division of Biostatistics, School of Public Health, University of Minnesota,
Minneapolis, MN 55455*

April 30, 2009, revised June 24, 2009

Correspondence author: Wei Pan

Telephone: (612) 626-2705

Fax: (612) 626-0660

Email: weip@biostat.umn.edu

Address: Division of Biostatistics, MMC 303,
School of Public Health, University of Minnesota,
Minneapolis, Minnesota 55455-0392, U.S.A.

ABSTRACT

Most of the existing association tests for population-based case-control studies are based on comparing the mean genotype scores between the case and control groups, which may not be efficient under genetic heterogeneity. Given that most common diseases are genetically heterogeneous, caused by mutations in multiple loci, it may be beneficial to fully account for genetic heterogeneity in an association test. Here we first propose a binomial mixture model for such a purpose and develop a corresponding mixture likelihood ratio test (MLRT) for a single locus. **We also consider two methods to combine single-locus-based MLRTs across multiple loci in linkage disequilibrium to boost power when causal SNPs are not genotyped.** We show with a wide spectrum of numerical examples that under genetic heterogeneity the proposed tests are more powerful than some commonly used association tests.

Key Words: Case-control study; EM algorithm; Genome-wide association study; Likelihood ratio test; Linkage disequilibrium (LD); Mixture models; Permutation.

1 Introduction

Common diseases and complex phenotypes are often genetically heterogeneous with different etiologies in different individuals. Here we consider the situation when a disease (or other phenotype) is caused by mutations in multiple unlinked loci, referred to as locus heterogeneity (Ott 1999). Under locus heterogeneity, the population of individuals with disease may be decomposed into various subpopulations, each with disease caused by mutations at different loci (or their combinations). As for admixture mapping in linkage analysis (Smith 1963; Ott 1983), we propose a binomial mixture model for genotype scores to account for possibly heterogeneous subpopulations in the case population for genetic association studies. While most existing association tests aim to detect the mean difference of genotype scores between the case and con-

trol groups, ignoring genetic heterogeneity in the case group fails to utilize differences of other higher moments of genotype scores, leading to power loss, as to be shown. Furthermore, if disease causal SNPs are not genotyped, it may boost statistical power to combine single-locus-based tests across multiple loci in linkage disequilibrium (LD) with any causal SNP; for this purpose, we consider two combining methods for multiple loci.

In the following, we first introduce our statistical models and a mixture likelihood ratio test (MLRT) to detect disease association with a single nucleotide polymorphism (SNP), then we propose two methods to combine such single-locus-based MLRTs across multiple loci in a candidate gene or region. We conducted extensive simulations to demonstrate power gains of our proposed tests over some commonly used association tests. For illustration we applied the tests to a published amyotrophic lateral sclerosis (ALS) dataset (Schymick et al 2007). We end with a summary of our conclusions and a discussion on related issues, limitations and future work.

2 Methods

2.1 Models

We first consider detecting disease association with an individual SNP at a single locus based on a case-control study. Suppose that genotype score X is the number of the minor allele at the locus for a subject. If the subject is in the control group, we assume

$$X \sim Bin(2, \theta_b),$$

where θ_b is the background probability of having the minor allele on a chromosome for the controls. In contrast, for the case group, we assume

$$X \sim \pi Bin(2, \theta) + (1 - \pi) Bin(2, \theta_b), \tag{1}$$

where θ is the probability of having the minor allele on a chromosome for a subpopulation of cases with disease caused by (or associated with) the minor allele, while for the other subpopulation of cases the disease is caused by mutations at other unlinked loci and thus for them the probability of having the minor allele at the locus of interest is the same as that for the controls. The mixture model explicitly accounts for genetic heterogeneity of the case group if the mixture model is not degenerated with $\theta \neq \theta_b$, and $\pi \neq 0$ or 1.

Although for a binomial distribution $X^* \sim Bin(2, \theta^*)$ with $\theta^* = \pi\theta + (1 - \pi)\theta_b$, its mean $E(X^*)$ equals to $E(X)$ of (1), their higher moments are different. For example, for the second-moments,

$$E(X^{*2}) - E(X^2) = 2\pi\theta(1 + \theta) + 2(1 - \pi)\theta_b(1 + \theta_b) - 2\theta^*(1 + \theta^*) = \pi(1 - \pi)(\theta - \theta_b)^2 \geq 0,$$

where the strict inequality holds for the non-degenerated case with $\theta \neq \theta_b$, $\pi \neq 0$ and $\pi \neq 1$. Hence, the binomial mixture model introduces overdispersion as compared to a binomial distribution with the equal mean. While most existing association tests, e.g. Hotelling's T^2 tests (Xiong et al 2002; Fan and Knapp 2003) and those based on logistic regression (Clayton et al 2004; Pan 2009), aim to compare the mean difference of genotype scores between the control and case groups, they ignore the possible genetic heterogeneity in the case group and thus possible difference in high moments of genotype scores between the two groups. The main motivation of this paper is to take advantage of possible differences in high moments as suggested by the genetic heterogeneity and associated mixture model, in addition to the mean genotype score difference between the two groups, to improve power.

The binomial distribution assumption implies that the Hardy-Weinberg equilibrium (HWE) holds for the control group. In contrast, under the mixture model, the HWE does not hold for the case group, as shown below. Let a and A denote the

minor and the other alleles respectively. We have

$$\begin{aligned}
 Pr(a) &= Pr(aa) + \frac{1}{2}Pr(Aa) = Pr(X = 2) + \frac{1}{2}Pr(X = 1) \\
 &= \pi\theta^2 + (1 - \pi)\theta_b^2 + \frac{1}{2}(2\pi\theta(1 - \theta) + 2(1 - \pi)\theta_b(1 - \theta_b)) \\
 &= \pi\theta + (1 - \pi)\theta_b,
 \end{aligned}$$

and

$$Pr(aa) = Pr(X = 2) = \pi\theta^2 + (1 - \pi)\theta_b^2.$$

Thus, we have $Pr(aa) \neq Pr(a)Pr(a)$ unless for the degenerated case with $\pi = 0$ or 1 , or $\theta = \theta_b$.

Although in general the binomial mixture model (1) is not identifiable (McLachlan and Peel 2000, p.164), if θ_b is given (as can be estimated from the control group), we prove in Appendix A.1 that it is indeed identifiable.

In our model, we assume a common background probability θ_b , whereas in reality it is possible that there are two different background probabilities, say θ_{b1} and θ_b , for the case and control groups respectively. Interestingly, under some quite general conditions, it can be shown that there exists another two-component binomial mixture model satisfying i) that it is equivalent to the original mixture model for the case group and ii) that its background probability is θ_b , the same background probability for the control group. The main reason is due to the general non-identifiability of the binomial mixture model (1). Under the situations where the conditions do not hold, we can find another binomial mixture model that is equivalent to the original one such that its background probability θ_{b2} has a minimum difference from θ_b , and the difference is often small. More details are given in Appendix A.2.

2.2 Estimation

We propose a two-step procedure for parameter estimation: first, based on only the control sample, we obtain a maximum likelihood estimate (MLE) of θ_b , say $\hat{\theta}_b$; second,

fixing $\theta_b = \hat{\theta}_b$, we apply an EM algorithm to the case group to obtain a maximum penalized likelihood estimate (MPLE) of other parameters in the mixture model.

Specifically, in the first step, suppose among m controls, there are m_0 , m_1 and m_2 individuals with genotype values equal to 0, 1 and 2 respectively, then

$$\hat{\theta}_b = \frac{m_1 + 2m_2}{2(m_0 + m_1 + m_2)}. \quad (2)$$

In the second step, we fix θ_b at $\hat{\theta}_b$, and use the EM to maximize a penalized log-likelihood for the case group. Suppose x_j is the genotype score for case j for $j = 1, \dots, n$, and let $z_{k,j}$ be the indicator of whether case j is indeed from component k , $k = 1$ or 2 . If we could observe $z_{k,j}$, then a penalized log-likelihood for the complete data is

$$\log L_c = \sum_{k=1}^2 \sum_{j=1}^n z_{k,j} (\log \pi_k + \log f_B(x_j; \theta_k)) + C \log \pi_1, \quad (3)$$

where $\pi_1 = \pi$, $\pi_2 = 1 - \pi$, $\theta_1 = \theta$, $\theta_2 = \hat{\theta}_b$, $f_B(\cdot; \theta)$ is the probability mass function for $Bin(2, \theta)$, and the penalty $C \log \pi_1$ is used to stabilize the estimate of π_1 . Following Fu et al (2006), we used $C = 1$ throughout.

At iteration r , the E-step yields

$$Q = E(\log L_c | Data) = \sum_{k=1}^2 \sum_{j=1}^n \hat{\tau}_{k,j}^{(r)} (\log(\pi_k) + \log(f_B(x_j; \theta_k))) + C \log(\pi_1), \quad (4)$$

where

$$\hat{\tau}_{k,j}^{(r)} = \frac{\hat{\pi}_k^{(r-1)} f_B(x_j; \hat{\theta}_k^{(r-1)})}{\sum_{k=1}^2 \hat{\pi}_k^{(r-1)} f_B(x_j; \hat{\theta}_k^{(r-1)})} \quad (5)$$

is the posterior probability of case j 's coming from component k ; each superscript (r) denotes an estimate at iteration r . In the M-step, we maximize Q with respect to the parameters:

$$\begin{aligned} \frac{\partial Q}{\partial \theta_1} &= \sum_{j=1}^n \tau_{1,j}^{(r)} \left(-\frac{2I(x_j = 0) + I(x_j = 1)}{1 - \theta_1} + \frac{I(x_j = 1) + 2I(x_j = 2)}{\theta_1} \right) = 0 \\ \implies \hat{\theta}_1^{(r)} &= \frac{\sum_{j=1}^n \tau_{1,j}^{(r)} (I(x_j = 1) + 2I(x_j = 2))}{2 \sum_{j=1}^n \tau_{1,j}^{(r)}}, \end{aligned} \quad (6)$$

and

$$\begin{aligned}\frac{\partial Q}{\partial \pi_1} &= \frac{\sum_{j=1}^n \tau_{1,j}^{(r)}}{\pi_1} - \frac{\sum_{j=1}^n \tau_{2,j}^{(r)}}{1 - \pi_1} + \frac{C}{\pi_1} = 0 \\ \implies \hat{\pi}_1^{(r)} &= \frac{\sum_{j=1}^n \tau_{1,j}^{(r)} + C}{n + C}, \quad \hat{\pi}_2^{(r)} = 1 - \hat{\pi}_1^{(r)}.\end{aligned}\tag{7}$$

Then we increase the iteration number r by one and iterate the above E- and M-steps until convergence, obtaining the MPLEs $\hat{\pi}$ and $\hat{\theta}$.

Although we can jointly estimate all the parameters simultaneously by applying the EM algorithm to maximize the sum of the penalized log-likelihood for the case group and the binomial log-likelihood for the control group, we found the above two-step procedure was more stable with better performance, presumably due to the non-identifiability of the mixture model with θ_b not fixed, leading to estimates at local maxima of the joint likelihood and thus degraded performance.

2.3 Tests

To test on disease association with the single SNP, after obtaining the parameter estimates, we use a mixture likelihood ratio test (MLRT) to contrast possible distributional difference of genotype scores between the control and case groups. Our MLRT statistic is

$$MLRT = 2 \left(l_n(\hat{\pi}, \hat{\theta}, \hat{\theta}_b) - l_n(1, \hat{\theta}_b, \hat{\theta}_b) \right),\tag{8}$$

where each log-likelihood l_n is given by

$$l_n(\pi, \theta, \theta_b) = \sum_{j=1}^n \log (\pi f_B(x_j; \theta) + (1 - \pi) f_B(x_j; \theta_b)),\tag{9}$$

calculated based on the mixture model (1) for the case group. To assess statistical significance, we use permutations:

Step 1. For the given data, calculate the MLRT statistic, say $MLRT_0$;

Step 2. Permute the pooled control and case samples by randomly shuffling the disease status for each subject;

Step 3. Calculate the MLRT statistic $MLRT^{(b)}$ based on the permuted data;

Step 4. Repeat Steps 2 and 3 for $b = 1, \dots, B$;

Step 5. Calculate the permutation p-value P as $P = \sum_{b=1}^B I(MLRT^{(b)} > MLRT_0)/B$.

For multiple, say K , loci, possibly in linkage disequilibrium (LD), we first calculate the MLRT statistic $MLRT_j$ for each locus $j = 1, \dots, K$ based on the original data. Second, we define two combined statistics

$$\text{Max-MLRT} = \max_{1 \leq j \leq K} MLRT_j, \quad \text{or} \quad \text{Sum-MLRT} = \sum_{j=1}^K MLRT_j.$$

Then we use permutation to obtain a p-value for each of the two combining methods.

Note that such a combining and permutation procedures can be equally applied to other tests. For example, at each locus j , we can use the two sample Z-test:

$$Z_j = \frac{\bar{X}_{1,j} - \bar{X}_{0,j}}{\sqrt{\frac{\hat{\sigma}_{1,j}^2}{n} + \frac{\hat{\sigma}_{0,j}^2}{m}}},$$

where $\bar{X}_{1,j}$ and $\hat{\sigma}_{1,j}^2$ are the sample mean and variance of genotype scores for the case group, while $\bar{X}_{0,j}$ and $\hat{\sigma}_{0,j}^2$ are for the control group, and n and m are the sample sizes for the two groups respectively. To combine the test statistics across multiple loci, we use

$$\text{Max-Z} = \max_{1 \leq j \leq K} Z_j^2, \quad \text{or} \quad \text{Sum-Z} = \sum_{j=1}^K Z_j^2.$$

Again a permutation procedure is used to obtain p-values for test statistics Max-Z and Sum-Z.

As a comparison, we also consider the Armitage trend (T) test (Armitage 1955; Freidlin et al 2002), and possibly the 2-DF chi-squared (χ^2) test for each locus, and their corresponding combining methods for multiple loci.

Note that the combining method Max is the similar to the commonly used minP method that takes the minimum p-value of single-locus-based tests, while the Sum method is similar to the sum of squared score (SSUw) test of combining multiple individual single-locus-based score tests (Pan 2009). The goal for combining multiple loci is to account for LD and thus boost power. There is no best combining method: in general, the performance of any combining method depends on the unknown data distribution.

3 Results

3.1 Simulated LD Patterns

3.1.1 Simulation set-ups

We fixed the number of SNPs to be 7, including the disease causing SNP_0 at the first locus. The genotype for the causal SNP_0 was directly generated from the mixture model (1). We generated genotype values for SNP_1 through SNP_6 by a latent variable model for haplotype $(Y_0, Y_1, \dots, Y_6)'$, which was assumed to be multivariate normal $N(0, \Sigma)$. We used $Var(Y_j) = 1$ and one of two correlation structures: 1) the compound symmetry (CS) structure with $Corr(Y_j, Y_k) = \rho$ for any $j \neq k$, or 2) the AR(1) structure with $Corr(Y_j, Y_k) = \rho^{|j-k|}$. The joint distribution of $(Y_0, Y_1, \dots, Y_6)'$ suggested a conditional distribution of $(Y_1, \dots, Y_6)'$ given Y_0 :

$$(Y_1, \dots, Y_6)' | Y_0 = y_0 \sim N(y_0 \Sigma_{10} \Sigma_{00}^{-1}, \Sigma_{11} - \Sigma_{10} \Sigma_{00}^{-1} \Sigma_{01}) \quad (10)$$

with $\Sigma_{00} = Var(Y_0) = 1$, $\Sigma_{10} = Cov((Y_1, \dots, Y_6)', Y_0)$ and $\Sigma_{11} = Cov((Y_1, \dots, Y_6)')$.

Since we sampled the genotype value of SNP_0 from the mixture model, we knew $SNP_0 = 0$ or 1 in the haplotype. According to a specified minor allele frequency (MAF) for SNP_0 , say MAF_0 , we sampled y_0 from a truncated normal distribution ranging from $-\infty$ to the normal quantile of MAF_0 if $SNP_0 = 1$, or ranging

from the normal quantile of MSF_0 to ∞ if $SNP_0 = 0$. Once y_0 is known, we generated $(y_1, \dots, y_6)'$ from the conditional distribution. Because each Y_j had a marginal distribution of $N(0, 1)$, once y_j was known, we dichotomized y_j with a truncated normal distribution with a specified MAF, which was randomly drawn from a uniform distribution $U(0.1, 0.4)$. After dichotomizing each y_0, y_1, \dots, y_6 , we obtained a simulated haplotype. Similarly we generated another haplotype, and summed up the two haplotypes to obtain the genotype values (X_0, X_1, \dots, X_6) for a case.

The genotype values for the controls were similarly generated with $\pi = 0$. After genotypes for $n = 500$ cases and $m = 500$ controls were generated, for any test, we calculated its test statistics for locus 1 to locus 6 separately, and recorded the Max- and Sum-statistics of the six locus-specific test statistics. The p-values were obtained by permutation with $B = 200$. Note that the causal SNP_0 was not used by any combining method.

We considered ten simulation set-ups with either a CS or an AR(1) correlation structure and a variety of parameter values as shown below.

Set-up	1	2	3	4	5	6	7	8	9	10
CS, ρ	0.4	0.4	0.4	0.4	0.4	0.4	0.2	0.2	0.2	0.2
AR(1), ρ	0.5	0.5	0.5	0.5	0.5	0.5	0.3	0.3	0.3	0.3
θ	0.2	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4
θ_b	0.2	0.1	0.1	0.1	0.3	0.3	0.3	-	0.1	-
π	0	0.1	0.2	0.3	0.3	0.9	0.9	1	0.9	1

For each non-null set-up, we simulated 250 independent datasets. Therefore the empirical power \hat{p} had a Monte Carlo standard error $\sqrt{\hat{p}(1 - \hat{p})/250} \leq 0.032$. For the null set-up, we simulated 500 datasets to improve the accuracy of the estimated Type I error rates.

3.1.2 Simulation result

The results for data with a CS correlation structure are shown in Table 1 and Table 2, while those for an AR(1) correlation structures are shown in Table 3 and Table 4. For set-up 1 (i.e. the null case), the type I error rates for all tests were around the significance level at 0.05 or 0.1 respectively. For the set-ups with a non-zero mixing proportion, there were substantial power gains by the Max- or Sum-MLRT test over the other tests. In particular, it is noted that, although the power differences between various single-locus tests were not large, there could exist dramatic differences between the various combined tests. Between the Max-MLRT and Sum-MLRT tests, the latter seemed to be the winner. It is reassuring that when there was no genetic heterogeneity as for set-ups 8 and 10, the power of the Max-MLRT and Sum-MLRT tests was comparable to that of the Max-Z/T and Sum-Z/T tests. In general, the 2-DF χ^2 test had lower power than other tests.

For comparison, we also applied several other tests comparing the mean genotype scores for multiple loci: Hotelling's T^2 test, a test based on principle component analysis (PCA) of the multilocus genotype scores (Wang and Abbott 2008), two sum of squared score (SSU and SSUw) tests, a test taking the minimum p-value of single-locus score tests (UminP), and a multivariate score test. All the above tests, except Hotelling's T^2 and UminP tests, are based on fitting a multivariate logistic regression model relating the disease probability to genotype scores (or their principle components, PCs). For the PCA test, as recommended by Wang and Abbott (2008), we took the first few PCs such that they explained at least 85% of total variation. Clayton et al (2004) showed that Hotelling's T^2 test is closely related to the multivariate score test. Although Wang and Abbott (2008) showed that the PCA-based test performed well for continuous traits, while Pan (2009) supported the high power of the SSU, SSUw and UminP tests under various conditions (but without genetic heterogeneity), all the above tests did not perform as well as our proposed

Table 1: Empirical power/size of the tests for simulated data with a CS correlation structure.

Set-up	α	Test	SNP_0	SNP_1	SNP_2	SNP_3	SNP_4	SNP_5	SNP_6	Max	Sum
1 (Null)	0.05	MLRT	0.032	0.026	0.026	0.028	0.040	0.022	0.028	0.058	0.054
		Z	0.064	0.056	0.056	0.036	0.060	0.028	0.060	0.052	0.040
		T	0.072	0.044	0.060	0.048	0.072	0.044	0.068	0.052	0.048
		χ^2	0.042	0.044	0.054	0.042	0.048	0.046	0.048	0.054	0.056
	0.1	MLRT	0.068	0.062	0.068	0.048	0.090	0.050	0.068	0.136	0.122
		Z	0.128	0.124	0.116	0.076	0.136	0.068	0.116	0.112	0.112
		T	0.120	0.108	0.120	0.088	0.136	0.076	0.108	0.108	0.132
		χ^2	0.084	0.100	0.108	0.108	0.096	0.088	0.096	0.100	0.108
2	0.05	MLRT	0.728	0.168	0.148	0.140	0.128	0.112	0.124	0.348	0.448
		Z	0.524	0.088	0.080	0.072	0.068	0.068	0.064	0.088	0.084
		T	0.544	0.080	0.080	0.064	0.060	0.064	0.072	0.088	0.108
		χ^2	0.060	0.088	0.060	0.064	0.052	0.040	0.056	0.068	0.064
	0.1	MLRT	0.816	0.264	0.228	0.240	0.244	0.188	0.224	0.516	0.584
		Z	0.676	0.152	0.136	0.136	0.148	0.128	0.120	0.160	0.176
		T	0.676	0.168	0.144	0.136	0.152	0.116	0.132	0.148	0.184
		χ^2	0.104	0.152	0.116	0.112	0.096	0.088	0.100	0.100	0.120
3	0.05	MLRT	0.988	0.228	0.232	0.200	0.228	0.220	0.280	0.524	0.592
		Z	0.984	0.144	0.152	0.136	0.124	0.148	0.180	0.204	0.304
		T	0.984	0.144	0.156	0.120	0.136	0.128	0.184	0.200	0.276
		χ^2	0.072	0.032	0.056	0.060	0.060	0.060	0.052	0.060	0.060
	0.1	MLRT	1.000	0.344	0.332	0.336	0.316	0.320	0.412	0.652	0.732
		Z	0.996	0.220	0.236	0.224	0.216	0.232	0.240	0.292	0.404
		T	0.996	0.216	0.224	0.228	0.212	0.228	0.272	0.292	0.380
		χ^2	0.128	0.064	0.116	0.112	0.112	0.124	0.096	0.100	0.128
4	0.05	MLRT	1.000	0.368	0.348	0.408	0.372	0.412	0.376	0.700	0.784
		Z	1.000	0.264	0.268	0.328	0.256	0.284	0.260	0.404	0.564
		T	1.000	0.268	0.264	0.304	0.232	0.268	0.260	0.404	0.572
		χ^2	0.148	0.056	0.072	0.096	0.072	0.052	0.048	0.044	0.064
	0.1	MLRT	1.000	0.504	0.504	0.524	0.508	0.544	0.488	0.796	0.852
		Z	1.000	0.372	0.396	0.424	0.348	0.412	0.356	0.544	0.668
		T	1.000	0.380	0.384	0.432	0.364	0.408	0.360	0.536	0.688
		χ^2	0.252	0.096	0.116	0.152	0.148	0.092	0.116	0.124	0.128
5	0.05	MLRT	0.304	0.072	0.068	0.064	0.072	0.076	0.056	0.084	0.072
		Z	0.284	0.072	0.072	0.056	0.060	0.076	0.052	0.072	0.072
		T	0.276	0.072	0.052	0.052	0.060	0.068	0.040	0.072	0.056
	0.1	MLRT	0.404	0.152	0.116	0.120	0.124	0.116	0.120	0.148	0.152
		Z	0.388	0.136	0.128	0.124	0.128	0.112	0.100	0.144	0.112
		T	0.392	0.156	0.132	0.112	0.124	0.104	0.100	0.144	0.144
6	0.05	MLRT	0.984	0.312	0.268	0.280	0.324	0.260	0.252	0.348	0.688
		Z	0.980	0.200	0.188	0.200	0.252	0.180	0.172	0.264	0.388
		T	0.980	0.204	0.196	0.192	0.228	0.188	0.160	0.264	0.392
	0.1	MLRT	1.00	0.448	0.392	0.448	0.476	0.372	0.388	0.520	0.820
		Z	0.988	0.284	0.264	0.276	0.352	0.272	0.260	0.396	0.520
		T	0.988	0.280	0.272	0.292	0.352	0.260	0.268	0.392	0.496

Table 2: Empirical power/size of the tests for simulated data with a CS correlation structure (Table 1 continued).

Set-up	α	Test	SNP_0	SNP_1	SNP_2	SNP_3	SNP_4	SNP_5	SNP_6	Max	Sum
7	0.05	MLRT	0.992	0.108	0.096	0.096	0.108	0.064	0.104	0.132	0.128
		Z	0.992	0.100	0.092	0.092	0.096	0.060	0.084	0.136	0.136
		T	0.992	0.092	0.080	0.088	0.096	0.080	0.068	0.132	0.128
	0.1	MLRT	0.992	0.216	0.120	0.152	0.164	0.124	0.156	0.220	0.224
		Z	0.992	0.200	0.116	0.148	0.144	0.128	0.140	0.192	0.228
		T	0.992	0.172	0.120	0.144	0.144	0.124	0.144	0.188	0.208
8	0.05	MLRT	1.000	0.112	0.136	0.080	0.080	0.064	0.104	0.132	0.152
		Z	1.000	0.104	0.124	0.064	0.068	0.064	0.092	0.116	0.136
		T	1.000	0.108	0.116	0.056	0.064	0.068	0.080	0.108	0.140
		χ^2	0.992	0.060	0.068	0.076	0.080	0.080	0.068	0.096	0.100
	0.1	MLRT	1.000	0.184	0.196	0.148	0.156	0.120	0.156	0.204	0.240
		Z	1.000	0.168	0.184	0.116	0.140	0.112	0.144	0.184	0.232
		T	1.000	0.196	0.168	0.108	0.140	0.112	0.140	0.176	0.220
		χ^2	1.000	0.124	0.108	0.116	0.116	0.172	0.128	0.168	0.180
9	0.05	MLRT	1.000	0.420	0.368	0.424	0.408	0.416	0.476	0.704	0.880
		Z	1.000	0.408	0.360	0.408	0.400	0.404	0.460	0.680	0.880
		T	1.000	0.412	0.356	0.388	0.400	0.396	0.460	0.680	0.876
	0.1	MLRT	1.000	0.556	0.516	0.560	0.524	0.552	0.628	0.808	0.928
		Z	1.000	0.536	0.496	0.548	0.528	0.520	0.608	0.792	0.904
		T	1.000	0.536	0.496	0.544	0.524	0.536	0.604	0.792	0.928
10	0.05	MLRT	1.000	0.552	0.492	0.484	0.480	0.508	0.532	0.808	0.940
		Z	1.000	0.516	0.480	0.464	0.468	0.472	0.524	0.804	0.940
		T	1.000	0.512	0.480	0.460	0.456	0.480	0.528	0.804	0.940
		χ^2	1.000	0.480	0.448	0.448	0.480	0.440	0.460	0.756	0.916
	0.1	MLRT	1.000	0.668	0.588	0.596	0.572	0.652	0.640	0.892	0.956
		Z	1.000	0.656	0.588	0.576	0.552	0.632	0.628	0.896	0.960
		T	1.000	0.656	0.588	0.572	0.564	0.612	0.620	0.896	0.952
		χ^2	1.000	0.656	0.528	0.568	0.532	0.608	0.632	0.868	0.936

Table 3: Empirical power/size of the tests for simulated data with an AR(1) correlation structure.

Set-up	α	Test	SNP_0	SNP_1	SNP_2	SNP_3	SNP_4	SNP_5	SNP_6	Max	Sum
1 (Null)	0.05	MLRT	0.038	0.040	0.050	0.052	0.044	0.038	0.058	0.054	0.054
		Z	0.056	0.036	0.042	0.034	0.052	0.022	0.050	0.038	0.028
		T	0.046	0.024	0.038	0.030	0.048	0.028	0.040	0.038	0.036
	0.1	MLRT	0.078	0.110	0.102	0.094	0.108	0.070	0.102	0.112	0.108
		Z	0.088	0.080	0.076	0.066	0.108	0.050	0.078	0.070	0.072
		T	0.082	0.068	0.076	0.066	0.098	0.050	0.066	0.070	0.080
2	0.05	MLRT	0.756	0.148	0.120	0.144	0.092	0.100	0.132	0.348	0.420
		Z	0.572	0.092	0.060	0.084	0.048	0.040	0.052	0.064	0.052
		T	0.584	0.080	0.064	0.096	0.048	0.032	0.064	0.064	0.056
	0.1	MLRT	0.828	0.244	0.216	0.248	0.200	0.184	0.188	0.500	0.572
		Z	0.680	0.156	0.104	0.156	0.108	0.072	0.108	0.100	0.096
		T	0.688	0.148	0.112	0.144	0.092	0.084	0.100	0.096	0.096
3	0.05	MLRT	0.992	0.304	0.168	0.164	0.132	0.116	0.128	0.420	0.540
		Z	0.988	0.204	0.084	0.076	0.072	0.056	0.068	0.112	0.156
		T	0.988	0.212	0.080	0.084	0.060	0.056	0.064	0.108	0.144
	0.1	MLRT	0.996	0.480	0.260	0.252	0.212	0.224	0.252	0.584	0.684
		Z	0.996	0.308	0.168	0.160	0.132	0.120	0.136	0.164	0.236
		T	0.992	0.312	0.152	0.148	0.116	0.124	0.144	0.164	0.236
4	0.05	MLRT	1.000	0.524	0.188	0.092	0.104	0.120	0.104	0.530	0.628
		Z	1.000	0.448	0.116	0.064	0.056	0.080	0.052	0.268	0.204
		T	1.000	0.432	0.116	0.056	0.048	0.080	0.056	0.268	0.248
	0.1	MLRT	1.000	0.656	0.336	0.168	0.184	0.196	0.184	0.660	0.740
		Z	1.000	0.540	0.216	0.100	0.108	0.116	0.100	0.388	0.308
		T	1.000	0.536	0.216	0.104	0.104	0.116	0.100	0.388	0.364
5	0.05	MLRT	0.308	0.072	0.068	0.064	0.068	0.076	0.036	0.116	0.112
		Z	0.300	0.080	0.044	0.056	0.068	0.076	0.052	0.084	0.072
		T	0.288	0.060	0.040	0.052	0.060	0.092	0.036	0.080	0.080
	0.1	MLRT	0.412	0.156	0.144	0.104	0.108	0.124	0.068	0.188	0.152
		Z	0.392	0.164	0.096	0.108	0.128	0.112	0.064	0.144	0.108
		T	0.396	0.136	0.104	0.096	0.124	0.120	0.064	0.136	0.120
6	0.05	MLRT	0.996	0.356	0.180	0.140	0.112	0.088	0.132	0.200	0.444
		Z	0.988	0.244	0.108	0.058	0.064	0.028	0.056	0.152	0.128
		T	0.988	0.240	0.112	0.062	0.056	0.028	0.052	0.152	0.144
	0.1	MLRT	1.000	0.488	0.276	0.220	0.164	0.128	0.184	0.320	0.640
		Z	1.000	0.400	0.164	0.144	0.088	0.068	0.100	0.220	0.244
		T	0.996	0.356	0.164	0.112	0.100	0.064	0.100	0.212	0.268

Table 4: Empirical power/size of the tests for simulated data with an AR(1) correlation structure (Table 3 continued).

Set-up	α	Test	SNP_0	SNP_1	SNP_2	SNP_3	SNP_4	SNP_5	SNP_6	Max	Sum	
7	0.05	MLRT	1.000	0.088	0.048	0.080	0.060	0.056	0.044	0.032	0.040	
		Z	1.000	0.072	0.048	0.084	0.052	0.056	0.040	0.036	0.048	
		T	1.000	0.076	0.048	0.088	0.032	0.044	0.040	0.032	0.032	
	0.1	MLRT	1.000	0.132	0.124	0.124	0.096	0.116	0.112	0.112	0.080	0.088
		Z	1.000	0.120	0.112	0.136	0.084	0.144	0.112	0.112	0.088	0.104
		T	1.000	0.128	0.108	0.104	0.072	0.116	0.096	0.080	0.064	
8	0.05	MLRT	1.000	0.164	0.068	0.080	0.040	0.068	0.040	0.092	0.096	
		Z	1.000	0.152	0.056	0.084	0.040	0.064	0.040	0.104	0.076	
		T	0.996	0.144	0.068	0.072	0.044	0.060	0.044	0.100	0.104	
	0.1	MLRT	1.000	0.248	0.132	0.136	0.084	0.108	0.084	0.164	0.168	
		Z	1.000	0.244	0.116	0.132	0.092	0.104	0.088	0.164	0.152	
		T	1.000	0.232	0.112	0.132	0.080	0.104	0.084	0.164	0.172	
9	0.05	MLRT	1.000	0.840	0.132	0.056	0.024	0.052	0.032	0.580	0.468	
		Z	1.000	0.820	0.116	0.052	0.032	0.056	0.020	0.560	0.516	
		T	1.000	0.820	0.096	0.048	0.028	0.048	0.032	0.552	0.524	
	0.1	MLRT	1.000	0.912	0.228	0.104	0.060	0.096	0.092	0.712	0.636	
		Z	1.000	0.908	0.224	0.100	0.060	0.080	0.096	0.704	0.636	
		T	1.000	0.900	0.208	0.092	0.064	0.080	0.088	0.704	0.652	
10	0.05	MLRT	1.000	0.860	0.136	0.076	0.052	0.088	0.068	0.660	0.592	
		Z	1.000	0.836	0.136	0.060	0.052	0.084	0.068	0.652	0.572	
		T	1.000	0.848	0.128	0.060	0.048	0.072	0.056	0.652	0.628	
	0.1	MLRT	1.000	0.916	0.204	0.108	0.092	0.148	0.116	0.760	0.716	
		Z	1.000	0.912	0.208	0.104	0.104	0.132	0.104	0.756	0.716	
		T	1.000	0.916	0.196	0.100	0.096	0.136	0.108	0.748	0.748	

Table 5: Empirical power/size of some multi-locus tests for simulated data with a CS or AR(1) correlation structure.

Corr	Set-up	α	T^2	PCA	SSU	SSUw	UminP	Sco	
CS	1	0.05	0.032	0.028	0.036	0.052	0.036	0.036	
		(Null) 0.1	0.096	0.068	0.076	0.076	0.080	0.092	
	2	0.05	0.056	0.060	0.088	0.096	0.096	0.080	0.068
		0.1	0.140	0.144	0.148	0.144	0.144	0.144	0.144
	3	0.05	0.148	0.152	0.248	0.240	0.240	0.160	0.144
		0.1	0.248	0.236	0.356	0.368	0.368	0.244	0.248
4	0.05	0.384	0.400	0.584	0.588	0.588	0.424	0.388	
	0.1	0.496	0.528	0.696	0.708	0.708	0.596	0.492	
AR(1)	1	0.05	0.024	0.024	0.044	0.028	0.032	0.020	
		(Null) 0.1	0.092	0.092	0.100	0.100	0.092	0.080	
	2	0.05	0.068	0.080	0.080	0.076	0.092	0.072	
		0.1	0.124	0.136	0.140	0.128	0.132	0.124	
	3	0.05	0.124	0.148	0.140	0.136	0.132	0.132	
		0.1	0.212	0.224	0.196	0.196	0.208	0.220	
	4	0.05	0.248	0.240	0.248	0.256	0.288	0.240	
		0.1	0.356	0.372	0.364	0.384	0.388	0.368	

MLRT in the current situation with genetic heterogeneity, as shown in Table 5. The main reason is that all the above tests, as the Z- and T-tests, are based on contrasting the mean genotype scores between the case and control groups while ignoring between-group differences in higher moments that are present under the genetic heterogeneity assumption.

3.2 Robustness to mis-specified models

In the previous simulations, we had a correct modeling assumption: the genotype value at any locus had a binomial distribution for the control group, and a mixture of two binomials (possibly degenerated with $\pi = 1$) for the case group. We investigated the robustness of the proposed tests to the violation of the above modeling assumption. Specifically, we consider data generated from various mixture models and from multiple causal SNPs respectively.

3.2.1 Various mixture models

We considered four scenarios: (1) for both the case and control groups, the genotype had the same mixture distribution of two binomials; (2) it was a binomial for the control group, but a mixture of three binomials for the case group; (3) for the control group it was a mixture of two binomials with mixing proportions π_b 's, while for the case group it was a mixture of three binomials with mixing proportions π 's; (4) the same as (3) except with different parameter values. For each set-up, simulated data were generated as in the previous section, and the correlation structure (for the latent variables) was always CS with the pair-wise correlation $\rho = 0.4$. The parameter values for the four scenarios are summarized in the following table.

Set-up	(1)	(2)	(3)	(4)
θ 's	(.4,.1)	(.4,.25,.1)	(.4,.25,.1)	(.4,.25,.1)
π 's	(0.2,0.8)	(0.1,0.1,0.8)	(0.16,0.2,0.64)	(0.2,0.16,0.64)
θ_b 's	(.4,.1)	0.1	(.4,.1)	(.25, .1)
π_b 's	(0.2,0.8)	1	(0.2,0.8)	(0.2,0.8)

As before, the MLRT test was applied under the assumption of having one and two binomial components for the control and case groups respectively, which was incorrect. As a comparison, we also considered the ideal MLRT, denoted as $MLRT_T$, under the (ideal but impractical) assumption that the true numbers of the mixture components for the control and case groups were known, and the corresponding mixture models were fitted for the two groups. For each set-up, the results based on 250 replicates are shown in Table 6. For set-up 1, the null case, the Type I error rates of MLRT and other tests were close to the nominal levels. For other three set-ups, as expected, $MLRT_T$ was most powerful (with its correct and strongest modeling assumption). Interestingly, our proposed MLRT could still be either more powerful than or as powerful as the Z- and T-tests.

Table 6: Empirical power/size of the tests under various mixture models.

Set-up	α	Test	SNP_0	SNP_1	SNP_2	SNP_3	SNP_4	SNP_5	SNP_6	Max	Sum
(1)	0.05	MLRT	0.048	0.052	0.040	0.056	0.052	0.044	0.060	0.076	0.040
		$MLRT_K$	0.044	0.064	0.056	0.064	0.068	0.056	0.060	0.060	0.072
		Z	0.044	0.048	0.040	0.060	0.056	0.064	0.080	0.076	0.048
		T	0.032	0.052	0.044	0.052	0.064	0.052	0.064	0.056	0.040
	0.1	MLRT	0.108	0.096	0.084	0.116	0.104	0.072	0.112	0.128	0.100
		$MLRT_K$	0.064	0.112	0.104	0.104	0.124	0.116	0.112	0.124	0.116
		Z	0.116	0.088	0.080	0.112	0.104	0.088	0.116	0.116	0.108
		T	0.096	0.080	0.108	0.112	0.108	0.112	0.112	0.108	0.088
(2)	0.05	MLRT	0.900	0.104	0.120	0.112	0.100	0.116	0.124	0.168	0.192
		$MLRT_K$	0.920	0.144	0.136	0.136	0.112	0.096	0.132	0.180	0.236
		Z	0.868	0.096	0.108	0.080	0.084	0.108	0.088	0.116	0.152
		T	0.860	0.084	0.112	0.080	0.080	0.104	0.084	0.112	0.144
	0.1	MLRT	0.940	0.176	0.200	0.168	0.172	0.192	0.192	0.240	0.252
		$MLRT_K$	0.952	0.256	0.204	0.196	0.180	0.168	0.196	0.260	0.324
		Z	0.916	0.160	0.172	0.144	0.152	0.176	0.160	0.196	0.204
		T	0.920	0.168	0.156	0.148	0.152	0.188	0.156	0.192	0.200
(3)	0.05	MLRT	0.128	0.032	0.052	0.068	0.088	0.072	0.040	0.060	0.068
		$MLRT_K$	0.208	0.116	0.080	0.076	0.080	0.104	0.112	0.096	0.104
		Z	0.160	0.032	0.060	0.068	0.080	0.072	0.048	0.056	0.080
		T	0.152	0.036	0.056	0.056	0.076	0.076	0.040	0.056	0.080
	0.1	MLRT	0.204	0.104	0.092	0.104	0.148	0.104	0.060	0.092	0.132
		$MLRT_K$	0.340	0.192	0.132	0.136	0.108	0.148	0.140	0.172	0.180
		Z	0.244	0.104	0.096	0.112	0.152	0.112	0.064	0.096	0.140
		T	0.248	0.112	0.092	0.096	0.136	0.116	0.068	0.096	0.132
(4)	0.05	MLRT	0.920	0.124	0.128	0.132	0.116	0.144	0.132	0.148	0.252
		$MLRT_K$	0.932	0.144	0.188	0.136	0.136	0.128	0.144	0.192	0.260
		Z	0.892	0.096	0.068	0.112	0.108	0.124	0.116	0.132	0.212
		T	0.900	0.092	0.064	0.108	0.116	0.120	0.114	0.128	0.212
	0.1	MLRT	0.996	0.204	0.184	0.212	0.196	0.208	0.220	0.288	0.376
		$MLRT_K$	0.984	0.240	0.236	0.218	0.212	0.196	0.236	0.348	0.432
		Z	0.936	0.172	0.152	0.200	0.200	0.196	0.184	0.240	0.324
		T	0.940	0.164	0.148	0.200	0.204	0.192	0.184	0.244	0.308

We note that, although in general the binomial mixture model with two or three components, each with a binomial total $N = 2$, is not identifiable (McLachlan and Peel 2000, p.164), for our purpose of testing, it is not an issue: because we used the (modified) likelihood ratio test, any of the equivalent models should give the same maximized likelihood value, and thus the equal MLRT statistics and resulting p-values.

3.2.2 Multiple causal SNPs

Here we consider that the disease is caused by multiple SNPs, possibly in LD, explicitly accounting for genetic heterogeneity. For simplicity, we assumed that there were two causal SNPs, say SNP_{01} and SNP_{02} with genotype scores of X_{01} and X_{02} . The disease status $Y = 1$ or 0 was generated from a logistic regression model:

$$\text{LogitPr}(Y = 1) = \beta_0 + \beta_1 X_{01} + \beta_2 X_{02}. \quad (11)$$

As before, we assumed that there were six genotyped SNPs, say SNP_1 to SNP_6 , in LD with SNP_{01} ; their latent variables followed a multivariate Normal distribution with a CS(0.4) correlation matrix. Note that neither SNP_{02} nor any of its neighboring SNPs was genotyped. The latent variables for the two causal SNPs also had a Normal distribution with a correlation ρ , and the disease-causing minor allele frequencies (MAFs) were MAF_{01} and MAF_{02} respectively. We considered the following five set-ups:

Set-up	1	2	3	4	5
ρ	0.2	0.5	0.8	0	0.5
β_0	-1.38	-1.38	-1.38	-1.38	-1.38
β_1	$\log(1.6)$	$\log(1.6)$	$\log(1.6)$	$\log(1.6)$	$\log(1.6)$
β_2	$\log(1.6)$	$\log(1.6)$	$\log(1.6)$	$-\log(1.6)$	$-\log(1.6)$
MAF_{01}	0.2	0.2	0.2	0.2	0.2
MAF_{02}	0.1	0.1	0.1	0.2	0.2

Note β_0 was chosen to yield a disease probability of 0.20 for individuals carrying no causal alleles. By default we adopted the dosage coding for X_{01} and X_{02} , implying the additive (ADD) mode of inheritance (MOI); for set-ups 4 and 5, the dominant (DOM) and recessive (REC) MOI were also considered. Note that our assumed mixture model for genetic heterogeneity model does not imply an additive MOI, or

even a corresponding logistic regression model unless under the special cases of rare disease or of no genetic heterogeneity; see Appendix A.3 for details.

The empirical powers of the tests are shown in Table 7. To save space, we only give the results for significance level $\alpha = 0.05$, but similar conclusions can be drawn for $\alpha = 0.1$. It is clear that, as before, our proposed MLRT had the highest power across all the set-ups. It is also interesting to note that, even for non-additive MOI with genetic heterogeneity, the 2-DF χ^2 test (based on permutations) did not work well.

3.3 HapMap data for gene CHI3L2

To mimic real LD patterns, we extracted the SNP data for gene CHI3L2 from the 90 CEU (Utah residents with ancestry from northern and western Europe) samples from the HapMap web site. We excluded SNPs with MAFs less than 0.2, imputed for missing genotypes by randomly drawing an observed genotype of the same allele from other samples, and removed those perfectly correlated SNPs, leaving to 17 SNPs. To generate simulated data while maintaining the LD structure, first, we fixed the disease causing SNP_0 to be SNP rs2182114, and grouped the samples into $group_0$, $group_1$ and $group_2$ based on whether the number of minor allele at SNP_0 was 0, 1 or 2, respectively. Next, we generated a random variate X_0 from the mixture distribution (1) for the case group, or from a binomial distribution for the control group; if $X_0 = k$, we randomly drew a sample from $group_k$. We repeated the above steps to generate a simulated dataset with $m = n = 500$ cases and controls respectively. The causal SNP was always removed from the data supplied to calculate Max and Sum statistics, and we used $B = 200$ permutations to calculate p-values. The parameter values used for three set-ups were $(\pi, \theta, \theta_b) = (0.3, 0.4, 0.4)$, $(0.3, 0.4, 0.3)$, $(0.3, 0.4, 0.25)$ and $(0.3, 0.4, 0.2)$ respectively.

The empirical test size and power are shown in Table 8. Due to limited space, only

Table 7: Empirical power of the tests with two-causal SNPs.

Set-up	MOI	α	Method	SNP_{01}	SNP_1	SNP_2	SNP_3	SNP_4	SNP_5	SNP_6	Max	Sum
1	ADD	0.05	MLRT	1.000	0.240	0.264	0.268	0.256	0.256	0.264	0.440	0.504
			Z	1.000	0.164	0.188	0.172	0.180	0.184	0.172	0.344	0.420
			T	1.000	0.164	0.180	0.156	0.176	0.176	0.188	0.332	0.428
			χ^2	0.996	0.132	0.128	0.156	0.128	0.136	0.152	0.224	0.328
2	ADD	0.05	MLRT	1.000	0.312	0.360	0.356	0.328	0.332	0.356	0.536	0.624
			Z	1.000	0.264	0.308	0.292	0.276	0.288	0.284	0.472	0.536
			T	1.000	0.276	0.300	0.308	0.288	0.292	0.272	0.468	0.532
			χ^2	1.000	0.224	0.216	0.212	0.212	0.196	0.208	0.288	0.428
3	ADD	0.05	MLRT	1.000	0.472	0.468	0.464	0.496	0.476	0.480	0.648	0.712
			Z	1.000	0.396	0.412	0.404	0.424	0.420	0.416	0.556	0.644
			T	1.000	0.392	0.400	0.396	0.416	0.408	0.404	0.552	0.648
			χ^2	1.000	0.244	0.264	0.248	0.256	0.272	0.276	0.396	0.568
4	REC	0.05	MLRT	0.436	0.200	0.184	0.196	0.188	0.212	0.212	0.332	0.384
			Z	0.288	0.168	0.148	0.156	0.144	0.160	0.152	0.284	0.320
			T	0.304	0.152	0.156	0.148	0.144	0.164	0.168	0.276	0.312
			χ^2	0.272	0.064	0.072	0.068	0.096	0.052	0.056	0.076	0.064
	ADD	0.05	MLRT	1.000	0.264	0.256	0.276	0.244	0.252	0.264	0.432	0.480
			Z	0.984	0.224	0.212	0.216	0.192	0.204	0.208	0.324	0.376
			T	0.976	0.212	0.216	0.208	0.196	0.212	0.212	0.336	0.392
			χ^2	0.976	0.116	0.128	0.128	0.152	0.136	0.136	0.220	0.276
	DOM	0.05	MLRT	0.944	0.248	0.244	0.252	0.236	0.240	0.232	0.392	0.452
			Z	0.924	0.184	0.176	0.196	0.168	0.168	0.148	0.304	0.356
			T	0.908	0.180	0.172	0.184	0.176	0.172	0.152	0.312	0.364
			χ^2	0.932	0.116	0.088	0.084	0.092	0.116	0.092	0.112	0.188
5	REC	0.05	MLRT	0.184	0.112	0.128	0.132	0.124	0.132	0.120	0.192	0.208
			Z	0.148	0.064	0.084	0.104	0.084	0.092	0.084	0.128	0.144
			T	0.152	0.080	0.076	0.084	0.084	0.096	0.104	0.132	0.148
			χ^2	0.180	0.060	0.028	0.044	0.036	0.032	0.056	0.012	0.036
	ADD	0.05	MLRT	0.924	0.172	0.152	0.164	0.156	0.168	0.156	0.244	0.276
			Z	0.868	0.104	0.108	0.112	0.108	0.124	0.108	0.156	0.184
			T	0.852	0.112	0.096	0.108	0.116	0.112	0.092	0.160	0.176
			χ^2	0.768	0.068	0.068	0.060	0.104	0.088	0.084	0.116	0.136
	DOM	0.05	MLRT	0.804	0.152	0.136	0.124	0.132	0.144	0.136	0.224	0.252
			Z	0.756	0.088	0.096	0.108	0.100	0.092	0.100	0.140	0.164
			T	0.724	0.096	0.088	0.096	0.092	0.104	0.108	0.144	0.168
			χ^2	0.728	0.072	0.104	0.080	0.072	0.060	0.080	0.080	0.088

some representative loci were shown. Note that the fifth SNP is the causal SNP_0 , and a test based on it would be powerful. It is clear that all the tests had correct Type I error rates, and the Sum-MLRT was the most powerful. Interestingly, the power of the Sum-MLRT test, though without the use of the disease causing SNP_0 (under the assumption that SNP_0 was not genotyped), was close to that of the single-locus MLRT test based on SNP_0 .

For comparison, we also used the MAF at 0.01 as the cut-off, selecting 27 SNPs. The ninth SNP was causal. We obtained the similar results (Table 8) and drew the same conclusions as before.

3.4 HapMap data for gene IL21R

For the same HapMap CEU samples, we also considered the region of gene IL21R. We processed the data by following the same steps as for gene CHI3L2, resulting in 28 SNPs. In contrast to using a fixed disease-causing SNP for gene CHI3L2, we randomly chose an SNP as disease-causing in each simulated dataset; other aspects of data generation were the same as for gene CHI3L2. Again the causal SNP was always unused in calculating any Max or Sum statistic. We considered two set-ups with the parameter values $(\pi, \theta, \theta_b) = (0.3, 0.4, 0.25)$ and $(0.4, 0.4, 0.25)$ respectively.

For each set-up, the sample sizes were $m = n = 500$ with permutation number $B = 200$, and the empirical power was estimated based on 250 replicated datasets. The results are shown in Table 9 with some representative loci/SNPs included. Clearly the Sum-MLRT test was most powerful. In particular, the Sum-MLRT test was much more powerful than any single-SNP-based tests, showing power gains by combining multiple SNPs in linkage disequilibrium. In addition, Sum-MLRT was always more powerful than Max-MLRT, whereas Max-Z or Max-T test could be more powerful than Sum-Z or Sum-T test.

Table 8: Empirical power/size for gene CHI3L2 with 17 or 25 SNPs based on the HapMap data.

#SNPs	Set-up	α	Test	SNP_4	SNP_0	SNP_7	SNP_{10}	SNP_{13}	SNP_{15}	SNP_{17}	Max	Sum	
17	1 (Null)	0.05	MLRT	0.048	0.056	0.064	0.028	0.036	0.044	0.052	0.056	0.052	
			Z	0.044	0.048	0.044	0.032	0.044	0.036	0.036	0.048	0.048	
			T	0.044	0.036	0.048	0.04	0.028	0.044	0.044	0.048	0.052	
		2	0.05	MLRT	0.096	0.128	0.104	0.072	0.08	0.104	0.104	0.092	0.088
				Z	0.092	0.112	0.092	0.064	0.072	0.104	0.092	0.084	0.076
				T	0.104	0.104	0.088	0.096	0.096	0.096	0.088	0.088	0.084
	3		0.05	MLRT	0.268	0.476	0.252	0.360	0.140	0.276	0.132	0.300	0.484
				Z	0.208	0.336	0.208	0.240	0.088	0.160	0.060	0.228	0.276
				T	0.204	0.320	0.208	0.236	0.064	0.176	0.056	0.228	0.272
		4	0.05	MLRT	0.408	0.608	0.376	0.512	0.236	0.412	0.232	0.448	0.636
				Z	0.288	0.484	0.300	0.348	0.144	0.264	0.116	0.336	0.380
				T	0.284	0.492	0.292	0.364	0.140	0.292	0.112	0.336	0.380
	5		0.05	MLRT	0.444	0.712	0.392	0.596	0.176	0.504	0.180	0.536	0.752
				Z	0.400	0.644	0.376	0.460	0.144	0.348	0.096	0.480	0.552
				T	0.416	0.632	0.384	0.484	0.136	0.368	0.088	0.500	0.572
		6	0.05	MLRT	0.560	0.864	0.496	0.732	0.304	0.676	0.272	0.660	0.856
				Z	0.500	0.764	0.528	0.604	0.220	0.484	0.148	0.644	0.676
				T	0.508	0.780	0.552	0.612	0.220	0.520	0.148	0.648	0.688
	7		0.05	MLRT	0.636	0.932	0.540	0.844	0.232	0.752	0.184	0.720	0.928
				Z	0.652	0.908	0.628	0.788	0.224	0.636	0.112	0.832	0.820
				T	0.644	0.904	0.620	0.780	0.212	0.656	0.112	0.832	0.848
		8	0.05	MLRT	0.748	0.980	0.688	0.908	0.368	0.848	0.356	0.804	0.972
				Z	0.752	0.948	0.728	0.864	0.312	0.760	0.228	0.880	0.896
				T	0.768	0.952	0.736	0.860	0.300	0.752	0.220	0.876	0.916
9	0.05		MLRT	0.636	0.932	0.540	0.844	0.232	0.752	0.184	0.720	0.928	
			Z	0.652	0.908	0.628	0.788	0.224	0.636	0.112	0.832	0.820	
			T	0.644	0.904	0.620	0.780	0.212	0.656	0.112	0.832	0.848	
	10	0.05	MLRT	0.748	0.980	0.688	0.908	0.368	0.848	0.356	0.804	0.972	
			Z	0.752	0.948	0.728	0.864	0.312	0.760	0.228	0.880	0.896	
			T	0.768	0.952	0.736	0.860	0.300	0.752	0.220	0.876	0.916	
11		0.05	MLRT	0.636	0.932	0.540	0.844	0.232	0.752	0.184	0.720	0.928	
			Z	0.652	0.908	0.628	0.788	0.224	0.636	0.112	0.832	0.820	
			T	0.644	0.904	0.620	0.780	0.212	0.656	0.112	0.832	0.848	
	12	0.05	MLRT	0.748	0.980	0.688	0.908	0.368	0.848	0.356	0.804	0.972	
			Z	0.752	0.948	0.728	0.864	0.312	0.760	0.228	0.880	0.896	
			T	0.768	0.952	0.736	0.860	0.300	0.752	0.220	0.876	0.916	
13		0.05	MLRT	0.636	0.932	0.540	0.844	0.232	0.752	0.184	0.720	0.928	
			Z	0.652	0.908	0.628	0.788	0.224	0.636	0.112	0.832	0.820	
			T	0.644	0.904	0.620	0.780	0.212	0.656	0.112	0.832	0.848	
	14	0.05	MLRT	0.748	0.980	0.688	0.908	0.368	0.848	0.356	0.804	0.972	
			Z	0.752	0.948	0.728	0.864	0.312	0.760	0.228	0.880	0.896	
			T	0.768	0.952	0.736	0.860	0.300	0.752	0.220	0.876	0.916	
15		0.05	MLRT	0.636	0.932	0.540	0.844	0.232	0.752	0.184	0.720	0.928	
			Z	0.652	0.908	0.628	0.788	0.224	0.636	0.112	0.832	0.820	
			T	0.644	0.904	0.620	0.780	0.212	0.656	0.112	0.832	0.848	
	16	0.05	MLRT	0.748	0.980	0.688	0.908	0.368	0.848	0.356	0.804	0.972	
			Z	0.752	0.948	0.728	0.864	0.312	0.760	0.228	0.880	0.896	
			T	0.768	0.952	0.736	0.860	0.300	0.752	0.220	0.876	0.916	
17		0.05	MLRT	0.636	0.932	0.540	0.844	0.232	0.752	0.184	0.720	0.928	
			Z	0.652	0.908	0.628	0.788	0.224	0.636	0.112	0.832	0.820	
			T	0.644	0.904	0.620	0.780	0.212	0.656	0.112	0.832	0.848	
	18	0.05	MLRT	0.748	0.980	0.688	0.908	0.368	0.848	0.356	0.804	0.972	
			Z	0.752	0.948	0.728	0.864	0.312	0.760	0.228	0.880	0.896	
			T	0.768	0.952	0.736	0.860	0.300	0.752	0.220	0.876	0.916	
19		0.05	MLRT	0.636	0.932	0.540	0.844	0.232	0.752	0.184	0.720	0.928	
			Z	0.652	0.908	0.628	0.788	0.224	0.636	0.112	0.832	0.820	
			T	0.644	0.904	0.620	0.780	0.212	0.656	0.112	0.832	0.848	
	20	0.05	MLRT	0.748	0.980	0.688	0.908	0.368	0.848	0.356	0.804	0.972	
			Z	0.752	0.948	0.728	0.864	0.312	0.760	0.228	0.880	0.896	
			T	0.768	0.952	0.736	0.860	0.300	0.752	0.220	0.876	0.916	
21		0.05	MLRT	0.636	0.932	0.540	0.844	0.232	0.752	0.184	0.720	0.928	
			Z	0.652	0.908	0.628	0.788	0.224	0.636	0.112	0.832	0.820	
			T	0.644	0.904	0.620	0.780	0.212	0.656	0.112	0.832	0.848	
	22	0.05	MLRT	0.748	0.980	0.688	0.908	0.368	0.848	0.356	0.804	0.972	
			Z	0.752	0.948	0.728	0.864	0.312	0.760	0.228	0.880	0.896	
			T	0.768	0.952	0.736	0.860	0.300	0.752	0.220	0.876	0.916	
23		0.05	MLRT	0.636	0.932	0.540	0.844	0.232	0.752	0.184	0.720	0.928	
			Z	0.652	0.908	0.628	0.788	0.224	0.636	0.112	0.832	0.820	
			T	0.644	0.904	0.620	0.780	0.212	0.656	0.112	0.832	0.848	
	24	0.05	MLRT	0.748	0.980	0.688	0.908	0.368	0.848	0.356	0.804	0.972	
			Z	0.752	0.948	0.728	0.864	0.312	0.760	0.228	0.880	0.896	
			T	0.768	0.952	0.736	0.860	0.300	0.752	0.220	0.876	0.916	
25		0.05	MLRT	0.636	0.932	0.540	0.844	0.232	0.752	0.184	0.720	0.928	
			Z	0.652	0.908	0.628	0.788	0.224	0.636	0.112	0.832	0.820	
			T	0.644	0.904	0.620	0.780	0.212	0.656	0.112	0.832	0.848	
	26	0.05	MLRT	0.748	0.980	0.688	0.908	0.368	0.848	0.356	0.804	0.972	
			Z	0.752	0.948	0.728	0.864	0.312	0.760	0.228	0.880	0.896	
			T	0.768	0.952	0.736	0.860	0.300	0.752	0.220	0.876	0.916	
27		0.05	MLRT	0.636	0.932	0.540	0.844	0.232	0.752	0.184	0.720	0.928	
			Z	0.652	0.908	0.628	0.788	0.224	0.636	0.112	0.832	0.820	
			T	0.644	0.904	0.620	0.780	0.212	0.656	0.112	0.832	0.848	
	28	0.05	MLRT	0.748	0.980	0.688	0.908	0.368	0.848	0.356	0.804	0.972	
			Z	0.752	0.948	0.728	0.864	0.312	0.760	0.228	0.880	0.896	
			T	0.768	0.952	0.736	0.860	0.300	0.752	0.220	0.876	0.916	
29		0.05	MLRT	0.636	0.932	0.540	0.844	0.232	0.752	0.184	0.720	0.928	
			Z	0.652	0.908	0.628	0.788	0.224	0.636	0.112	0.832	0.820	
			T	0.644	0.904	0.620	0.780	0.212	0.656	0.112	0.832	0.848	
	30	0.05	MLRT	0.748	0.980	0.688	0.908	0.368	0.848	0.356	0.804	0.972	
			Z	0.752	0.948	0.728	0.864	0.312	0.760	0.228	0.880	0.896	
			T	0.768	0.952	0.736	0.860	0.300	0.752	0.220	0.876	0.916	
31		0.05	MLRT	0.636	0.932	0.540	0.844	0.232	0.752	0.184	0.720	0.928	
			Z	0.652	0.908	0.628	0.788	0.224	0.636	0.112	0.832	0.820	
			T	0.644	0.904	0.620	0.780	0.212	0.656	0.112	0.832	0.848	
	32	0.05	MLRT	0.748	0.980	0.688	0.908	0.368	0.848	0.356	0.804	0.972	
			Z	0.752	0.948	0.728	0.864	0.312	0.760	0.228	0.880	0.896	
			T	0.768	0.952	0.736	0.860	0.300	0.752	0.220	0.876	0.916	
33		0.05	MLRT	0.636	0.932	0.540	0.844	0.232	0.752	0.184	0.720	0.928	
			Z	0.652	0.908	0.628	0.788	0.224	0.636	0.112	0.832	0.820	
			T	0.644	0.904	0.620	0.780	0.212	0.656	0.112	0.832	0.848	
	34	0.05	MLRT	0.748	0.980	0.688	0.908	0.368	0.848	0.356	0.804	0.972	
			Z	0.752									

Table 9: Empirical test power for gene IL21R with 28 SNPs based on the for HapMap data.

Set-up	α	Test	SNP_1	SNP_7	SNP_8	SNP_{13}	SNP_{19}	SNP_{22}	SNP_{25}	SNP_{28}	Max	Sum
1	0.05	MLRT	0.220	0.132	0.368	0.220	0.128	0.340	0.204	0.316	0.420	0.712
		Z	0.124	0.084	0.288	0.176	0.088	0.204	0.136	0.288	0.344	0.356
		T	0.124	0.080	0.276	0.172	0.096	0.184	0.120	0.296	0.344	0.392
	0.1	MLRT	0.328	0.232	0.496	0.336	0.220	0.436	0.316	0.436	0.576	0.840
		Z	0.204	0.140	0.364	0.232	0.172	0.304	0.212	0.364	0.488	0.480
		T	0.216	0.140	0.372	0.228	0.156	0.296	0.212	0.356	0.488	0.496
2	0.05	MLRT	0.240	0.204	0.456	0.276	0.196	0.400	0.208	0.396	0.572	0.840
		Z	0.196	0.152	0.392	0.192	0.160	0.332	0.160	0.348	0.580	0.436
		T	0.196	0.148	0.368	0.196	0.160	0.332	0.164	0.340	0.576	0.484
	0.1	MLRT	0.356	0.292	0.536	0.392	0.260	0.504	0.336	0.452	0.684	0.904
		Z	0.240	0.212	0.444	0.284	0.204	0.388	0.216	0.396	0.704	0.544
		T	0.236	0.200	0.448	0.296	0.204	0.384	0.224	0.400	0.708	0.608

3.5 Example

For illustration we used the data from a genome-wide association study on sporadic amyotrophic lateral Sclerosis (ALS) (Schymick et al 2007). The original study assayed 555352 unique SNPs for each of 276 patients with sporadic ALS and 268 controls. Using the PLINK software (Purcell et al 2007), we applied the HWE test to the case group to locate SNPs with low p-values, suggesting possible genetic heterogeneity. Note that under our mixture model, the HWE is violated. We looked at a few SNPs with low p-values from the HWE test. Here for illustration we consider such a SNP, rs6596835, on chromosome 6, with a small p-value of $5.999 * 10^{-5}$ against the HWE. We applied the LD blocking algorithm implemented in Haploview (v4.1) (Barrett et al 2005) to 31 SNPs within 20Kb of SNP rs6596835 for the control group; an LD block including SNP rs6596835 and other 4 SNPs were identified (Fig 1).

We applied the tests to the LD block with $B = 2000$ permutations. The results are shown in Table 10. We observe that for several individual SNPs, the p-values from the MLRT were smaller than that of the Z-, T- and χ^2 -tests. Consequently, by combining the five SNPs, the Max-MLRT and Sum-MLRT gave more significant results than their counterparts based on the Z-, T- and χ^2 -tests.

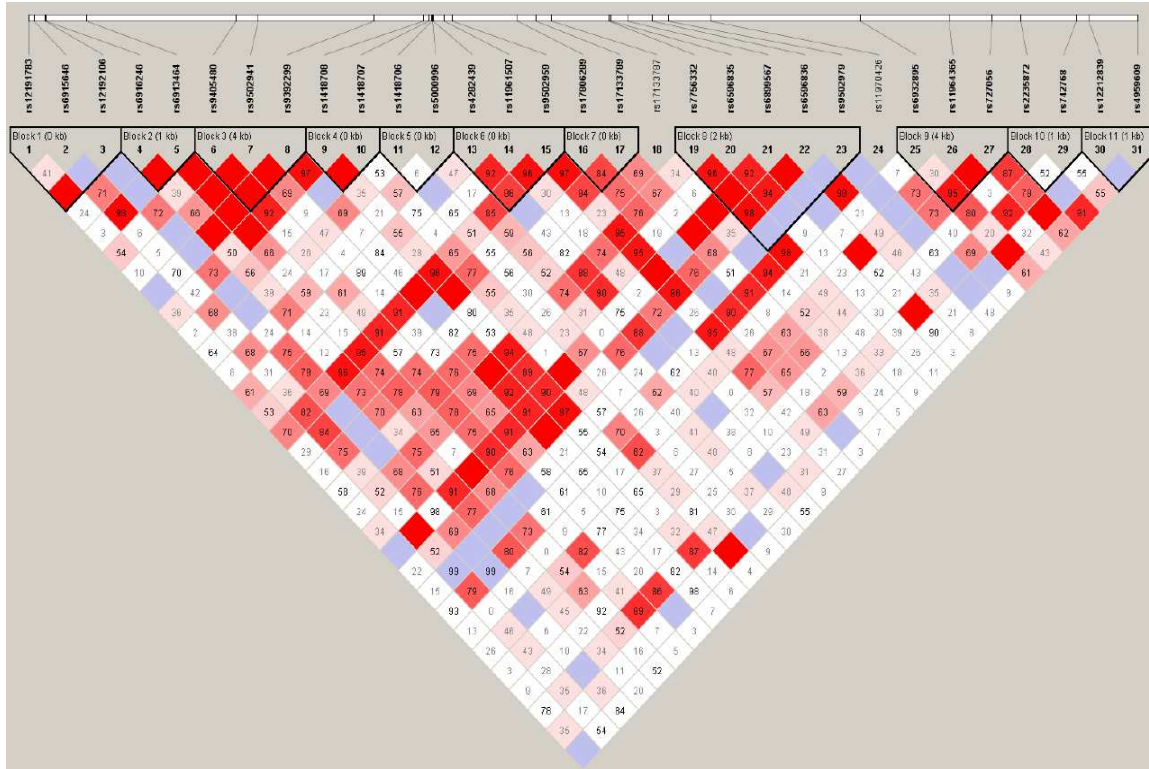


Figure 1: LD blocks around SNP rs6596835 based on the ALS data.

4 Discussion

We have proposed a binomial mixture model for genotype scores to take account of genetic heterogeneity of a disease or phenotype. A mixture likelihood ratio test (MLRT) is applied to contrast the distributional difference of genotype scores between the control and case groups, which utilizes their differences not only in means, but also in higher-moments. As a consequence, the proposed test gains power over those only comparing the mean difference of the genotype scores. The latter group of the tests includes many commonly used ones, such as the Hotelling's T^2 test (Xiong et al 2002; Fan and Knapp 2003) and those based on logistic regression (Clayton et al 2004; Pan 2009). We have also investigated two methods to combine single-locus-based MLRTs for multiple loci, which may further gain power if the causal SNPs

Table 10: P-values for the ALS data with 7 SNPs in an LD block. The 5 SNPs from SNP_1 to SNP_5 are rs7756332, rs6596835, rs6899567, rs6596836 and rs9502979 respectively.

Test	SNP_1	SNP_2	SNP_3	SNP_4	SNP_5	Max	Sum
MLRT	0.0150	0.0065	0.5775	0.0100	0.6835	0.0115	0.0465
Z	0.0305	0.0755	0.4950	0.0195	0.7200	0.0725	0.0875
T	0.0275	0.0725	0.3325	0.0115	0.5350	0.0825	0.0750
χ^2	0.0335	0.1205	0.4715	0.0105	1.0000	0.0915	0.0960

are not genotyped and the causal SNPs are in high linkage disequilibrium (LD) with some genotyped SNPs (Clayton et al 2004). Between the two combining methods, we found that the Max (similar to minP) method consistently performed better than the Sum method when combining multiple MLRTs, whereas in situations without genetic heterogeneity either combining method could be the winner (Pan 2009).

We note that the motivation behind our proposed tests is similar to that for admixture mapping in linkage analysis (Ott 1999). However, in spite of their appearing similarities related to heterogeneous groups and mixture modeling, the issue here is quite different from population stratification in association studies (e.g. Devlin and Roeder 1999; Pritchard et al 2000; Satten et al 2001; Zhang et al 2002; Zhu et al 2002; Chen et al 2003). The core problem in population stratification is the distributional difference of genotype-varying subpopulations between the control and case groups, whereas here we assume that the control group and case group are both from the same general population with possible differences in only disease-associated genotypes. Of course, our assumption might be incorrect, in which case, as usual, we need to correct for population stratification before applying our proposed tests. **For example, we can apply the same idea of corrections by the principle components of some representative**

genome-wide marker genotypes (Price et al 2006): first, we can cluster the cases and controls based on their principle components of the marker genotypes; second, within each cluster, that is, conditional on similar marker genotypes (as manifested by their principle components), we apply our MLRT to detect any distributional difference of genotype scores between the case and control groups; third, we summarize our MLRT statistics across multiple clusters, e.g. by a weighted sum of cluster-specific test statistics with a weight proportional to cluster sample size; finally, as before, we use permutations to obtain a p-value.

In theory, our mixture model can be extended to more than two binomial components for the case group, or to the control group. Although the non-identifiability perhaps is not an issue for a likelihood ratio test, there are other complicating issues, such as data-dependent choice of the number of components (McLachlan 1988). We tried this more general approach (and the ideal case with the true numbers of components known), but did not find substantial power gains, and hence we have skipped a detailed discussion. There are some potential limitations with our current approach. The first is its use of the dosage coding for genotype scores, corresponding to the additive mode of inheritance for rare diseases, though in general there is no corresponding simple mode of inheritance (Appendix A.3). It is not clear how to extend the proposed approach to other coding schemes or inheritance modes. The second limitation is the dependence on the use of permutation to assess statistical significance for multiple loci; it would be desirable to have an asymptotic or other approximate distributional result to facilitate inference. Because of some non-regularities associated with the likelihood for a mixture model (e.g. Hartigan 1985), the standard asymptotic results do not apply even for a single locus (Chen et al 2004); a theoretical development for multiple loci seems challenging, though useful.

R software and a tutorial on its use will be posted on <http://www.biostat.umn.edu/~weip>, and are available upon request.

A Appendix: Theory of Our Assumed Model

A.1 Identifiability

We claim that the two-component mixture model (1) is identifiable for any given θ_b if it is not degenerated (i.e., if $\theta \neq \theta_b$, and $\pi \neq 0$ or 1).

Proof. Suppose that (1) is not identifiable; that is, there exists another

$$f_1(X) = \pi_1 \text{Bin}(2, \theta_1) + (1 - \pi_1) \text{Bin}(2, \theta_b),$$

which is the same as $f(X)$ for the mixture model (1). It is easy to see that, if $\pi_1 = \pi$, then we must have $\theta_1 = \theta$. Hence, we have $\pi_1 \neq \pi$. Denote $\Delta\pi = \pi_1 - \pi \neq 0$.

By the equality of the means $E(X)$ from the two distributions, we have

$$\pi\theta + (1 - \pi)\theta_b = \pi_1\theta_1 + (1 - \pi_1)\theta_b,$$

and thus

$$\pi\theta = \pi\theta_1 + \Delta\pi\theta_1 - \Delta\pi\theta_b. \tag{12}$$

On the other hand, by the equality of the second moments $E(X^2)$ from the two distributions, we have

$$\pi\theta(1 + \theta) + (1 - \pi)\theta_b(1 + \theta_b) = \pi_1\theta_1(1 + \theta_1) + (1 - \pi_1)\theta_b(1 + \theta_b),$$

which, with the use of equation (12), yields

$$\pi\theta^2 = \pi\theta_1^2 + \Delta\pi\theta_1^2 - \Delta\pi\theta_b^2.$$

Hence,

$$\frac{\pi}{\Delta\pi} = \frac{\theta_1^2 - \theta_b^2}{\theta^2 - \theta_1^2},$$

and by equation (12), we have

$$\frac{\pi}{\Delta\pi} = \frac{\theta_1 - \theta_b}{\theta - \theta_1}.$$

Thus, we must have $\theta = \theta_b$, a contradiction to our assumption.

A.2 Equivalent mixture models

We consider the following problems: given any non-degenerated mixture model M1

$$\pi_1 \text{Bin}(2, \theta_1) + (1 - \pi_1) \text{Bin}(2, \theta_{b1}),$$

is it possible to exist a non-degenerated mixture model M2

$$\pi_2 \text{Bin}(2, \theta_2) + (1 - \pi_2) \text{Bin}(2, \theta_{b2})$$

that is equivalent to M1 with $\theta_{b2} \neq \theta_{b1}$? If so, under what conditions? Otherwise, what is θ_{b2} closest to the specified θ_{b1} ? These issues are related to our proposed model: although we assume a common background disease probability θ_b for both the case and control groups, it is possible for our model to allow differing background disease probabilities for the two groups.

We assume throughout that the mixture models are non-degenerated, and without loss of generality that $\theta_{b1} < \theta_1$.

If X is a random variable with distribution M1, and equivalently with M2, by the equality of the corresponding probabilities for $X = 0, 1$ and 2 from the two models, we have

$$\pi_1 \theta_1 + (1 - \pi_1) \theta_{b1} = \pi_2 \theta_2 + (1 - \pi_2) \theta_{b2}, \quad (13)$$

$$\pi_1 \theta_1^2 + (1 - \pi_1) \theta_{b1}^2 = \pi_2 \theta_2^2 + (1 - \pi_2) \theta_{b2}^2, \quad (14)$$

by which we have

$$\frac{\pi_1 \theta_1^2 + (1 - \pi_1) \theta_{b1}^2 - \theta_{b2}^2}{\pi_1 \theta_1 + (1 - \pi_1) \theta_{b1} - \theta_{b2}} = \theta_2 + \theta_{b2},$$

and thus

$$\theta_2 = \frac{\pi_1 \theta_1 (\theta_1 - \theta_{b2}) + (1 - \pi_1) \theta_{b1} (\theta_{b1} - \theta_{b2})}{\pi_1 (\theta_1 - \theta_{b2}) + (1 - \pi_1) (\theta_{b1} - \theta_{b2})}. \quad (15)$$

If $\theta_{b2} < \theta_{b1}$ or $\theta_{b2} > \theta_1$, then $\theta_2 \geq 0$. In this situation, since $\theta_1, \theta_{b1} \leq 1$, from (15), we have $\theta_2 \leq 1$. Now suppose $\theta_{b1} \leq \theta_{b2} \leq \theta_1$, in order for $\theta_2 \geq 0$, we need either

$$\pi_1(\theta_1 - \theta_{b2}) \geq (1 - \pi_1)(\theta_{b2} - \theta_{b1}), \quad (16)$$

$$\pi_1\theta_1(\theta_1 - \theta_{b2}) \geq (1 - \pi_1)\theta_{b1}(\theta_{b2} - \theta_{b1}), \quad (17)$$

or

$$\pi_1(\theta_1 - \theta_{b2}) \leq (1 - \pi_1)(\theta_{b2} - \theta_{b1}), \quad (18)$$

$$\pi_1\theta_1(\theta_1 - \theta_{b2}) \leq (1 - \pi_1)\theta_{b1}(\theta_{b2} - \theta_{b1}). \quad (19)$$

Notice that (16) implies (17) and (19) implies (18) since we assume $\theta_{b1} \leq \theta_1$. Under condition (16), in order for $\theta_2 \leq 1$, we still need numerator of (15) less than its denominator:

$$\begin{aligned} & \pi_1\theta_1(\theta_1 - \theta_{b2}) + (1 - \pi_1)\theta_{b1}(\theta_{b1} - \theta_{b2}) \leq \pi_1(\theta_1 - \theta_{b2}) + (1 - \pi_1)(\theta_{b1} - \theta_{b2}) \\ \Rightarrow \theta_{b2} & \leq \frac{\pi_1\theta_1 - \pi_1\theta_1^2 + (1 - \pi_1)\theta_{b1} - (1 - \pi_1)\theta_{b1}^2}{1 - \pi_1\theta_1 - (1 - \pi_1)\theta_{b1}}. \end{aligned} \quad (20)$$

Combine (16) and (20), we have:

$$\begin{aligned} \theta_{b2} & \leq \min \left(\pi_1\theta_1 + (1 - \pi_1)\theta_{b1}, \frac{\pi_1\theta_1 - \pi_1\theta_1^2 + (1 - \pi_1)\theta_{b1} - (1 - \pi_1)\theta_{b1}^2}{1 - \pi_1\theta_1 - (1 - \pi_1)\theta_{b1}} \right) \\ & = \frac{\pi_1\theta_1 - \pi_1\theta_1^2 + (1 - \pi_1)\theta_{b1} - (1 - \pi_1)\theta_{b1}^2}{1 - \pi_1\theta_1 - (1 - \pi_1)\theta_{b1}}. \end{aligned} \quad (21)$$

We can show

$$\theta_{b1} \leq \frac{\pi_1\theta_1 - \pi_1\theta_1^2 + (1 - \pi_1)\theta_{b1} - (1 - \pi_1)\theta_{b1}^2}{1 - \pi_1\theta_1 - (1 - \pi_1)\theta_{b1}} \leq \theta_1. \quad (22)$$

Under condition (19), rewrite (15) as

$$\frac{(1 - \pi_1)\theta_{b1}(\theta_{b2} - \theta_{b1}) - \pi_1\theta_1(\theta_1 - \theta_{b2})}{(1 - \pi_1)(\theta_{b2} - \theta_{b1}) - \pi_1(\theta_1 - \theta_{b2})} \leq \frac{(1 - \pi_1)\theta_1(\theta_{b2} - \theta_{b1}) - \pi_1\theta_1(\theta_1 - \theta_{b2})}{(1 - \pi_1)(\theta_{b2} - \theta_{b1}) - \pi_1(\theta_1 - \theta_{b2})} = \theta_1 \leq 1$$

Therefore, condition (19) alone guarantees $0 \leq \theta_2 \leq 1$. In fact, we can rewrite condition (19) as

$$\theta_{b2} \geq \frac{\pi_1\theta_1^2 + (1 - \pi_1)\theta_{b1}^2}{\pi_1\theta_1 + (1 - \pi_1)\theta_{b1}}. \quad (23)$$

We can also show

$$\theta_{b1} \leq \frac{\pi_1 \theta_1^2 + (1 - \pi_1) \theta_{b1}^2}{\pi_1 \theta_1 + (1 - \pi_1) \theta_{b1}} \leq \theta_1. \quad (24)$$

As a summary, we have $0 \leq \theta_2 \leq 1$ if one of the following two conditions is satisfied:

$$\text{Condition C1: } \theta_{b2} \leq \frac{\pi_1 \theta_1 - \pi_1 \theta_1^2 + (1 - \pi_1) \theta_{b1} - (1 - \pi_1) \theta_{b1}^2}{1 - \pi_1 \theta_1 - (1 - \pi_1) \theta_{b1}};$$

$$\text{Condition C2: } \theta_{b2} \geq \frac{\pi_1 \theta_1^2 + (1 - \pi_1) \theta_{b1}^2}{\pi_1 \theta_1 + (1 - \pi_1) \theta_{b1}}.$$

Note that the above two conditions can never be satisfied at the same time because one of the right hand sides is strictly larger than the other (unless M1 is degenerated into a single component).

On the other hand, by (15) and (13), we have

$$\pi_2 = \frac{(\pi_1(\theta_1 - \theta_{b2}) + (1 - \pi_1)(\theta_{b1} - \theta_{b2}))^2}{\pi_1(\theta_1 - \theta_{b2})^2 + (1 - \pi_1)(\theta_{b1} - \theta_{b2})^2}. \quad (25)$$

With some algebra, it can be shown that we always have $0 \leq \pi_2 \leq 1$. In summary, under either C1 or C2, we have an equivalent M2.

If an equivalent M2 does not exist, we must have

$$\frac{\pi_1 \theta_1 - \pi_1 \theta_1^2 + (1 - \pi_1) \theta_{b1} - (1 - \pi_1) \theta_{b1}^2}{1 - \pi_1 \theta_1 - (1 - \pi_1) \theta_{b1}} \leq \theta_{b2} \leq \frac{\pi_1 \theta_1^2 + (1 - \pi_1) \theta_{b1}^2}{\pi_1 \theta_1 + (1 - \pi_1) \theta_{b1}}. \quad (26)$$

The θ_{b2} of M2 closest to the specified θ_b would be either the lower or upper end of (26). Thus,

$$\begin{aligned} |\theta_{b2} - \theta_b| &\leq \frac{1}{2} \left(\frac{\pi_1 \theta_1^2 + (1 - \pi_1) \theta_{b1}^2}{\pi_1 \theta_1 + (1 - \pi_1) \theta_{b1}} - \frac{\pi_1 \theta_1 - \pi_1 \theta_1^2 + (1 - \pi_1) \theta_{b1} - (1 - \pi_1) \theta_{b1}^2}{1 - \pi_1 \theta_1 - (1 - \pi_1) \theta_{b1}} \right) \\ &= \frac{\pi_1(1 - \pi_1)(\theta_1 - \theta_{b1})^2}{2(\pi_1 \theta_1 + (1 - \pi_1) \theta_{b1})(1 - \pi_1 \theta_1 - (1 - \pi_1) \theta_{b1})}. \end{aligned} \quad (27)$$

For example, if $\theta_1 = 0.4$, $\theta_{b1} = 0.1$, $\pi_1 = 0.3$, $\theta_{b2} = 0.05$, by (15) and (25), we have $\theta_2 = 0.325$ and $\pi_2 = 0.509$. In this situation, $\theta_{b2} < \theta_{b1}$, so a valid solution is guaranteed. If we change $\theta_{b2} = 0.2$, then from (15) and (25), we have $\theta_2 = -1.7$ and $\pi_2 = 0.00526$. In this situation, the lower and upper bounds in (26) become 0.167 and 0.289 respectively, hence no valid solution since θ_{b2} falls within the two bounds.

We did a grid search on $(\pi_1, \theta_{b1}, \theta_1) \in (0, 1) \times (0, 1/2) \times (\theta_{b1}, 1/2)$. In any case, the resulting π_2 lies between 0 and 1, as proven above, and the range of θ_{b2} leading to invalid solutions matches constraint (26). Furthermore, a summary of the distribution for the values of the right hand side of (27) is the following: the maximum is 0.20, and the 99th, 95th and 90th percentiles are 0.14, 0.089 and 0.064 respectively.

A.3 Relationship to logistic regression

We show how our assumed model is related to logistic regression. Suppose that Y and X are disease indicator and genotype score (i.e. number of copies of an allele) in one locus. Based on our model, we have $X|Y = 1 \sim \alpha Bin(2, \theta) + (1 - \alpha) Bin(2, \theta_b)$ and $X|Y = 0 \sim Bin(2, \theta_b)$. Denote $P(Y = 1) = \tau$. Then by the Bayes theorem, it is easy to write down $Pr(Y|X)$, and thus

$$\begin{aligned} \text{Logit}P(Y = 1|X = x) &= \log \left(\frac{\alpha\tau Bin(2, \theta) + (1 - \alpha)\tau Bin(2, \theta_b)}{(1 - \tau) Bin(2, \theta_b)} \right) \\ &= \log \left(\frac{(1 - \alpha)\tau}{1 - \tau} + \frac{\alpha\tau}{1 - \tau} \left(\frac{\theta}{\theta_b} \right)^x \left(\frac{1 - \theta}{1 - \theta_b} \right)^{2-x} \right). \end{aligned}$$

Hence, in general it can not be written as the form of logistic regression. However, under the special case of either α is close to 1 (i.e. no genetic heterogeneity) or τ is close to 0 (i.e. rare disease), we could ignore the term $(1 - \alpha)\tau/(1 - \tau)$, and thus have a logistic regression model:

$$\text{Logit}P(Y = 1|X = x) = \log \left(\frac{\alpha\tau}{1 - \tau} \right) + 2 \log \left(\frac{1 - \theta}{1 - \theta_b} \right) + \left(\log \left(\frac{\theta}{\theta_b} \right) - \log \left(\frac{1 - \theta}{1 - \theta_b} \right) \right) x.$$

Acknowledgment

This research was partially supported by NIH grants GM081535 and HL65462. We thank the reviewers for many helpful and constructive comments that led to a substantially improved version.

References

- [1] ARMITAGE P (1955). Tests for linear trends in proportions and frequencies. *Biometrics* **11**, 375-386.
- [2] BARRETT JC, FRY B, MALLER J, DALY MJ (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263-265.
- [3] CHEN HS, ZHU X, ZHAO H, ZHANG S (2003). Qualitative semi-parametric test for genetic associations in case-control designs under structured populations. *Ann Hum Genet* **67**, 250-264.
- [4] CHEN H, CHEN J, KALBFLEISCH JD (2004). Testing for a finite mixture model with two components. *Journal of Royal Statistical Society, B* **66**, 95-115.
- [5] CLAYTON D, CHAPMAN J, COOPER J (2004). Use of unphased multilocus genotype data in indirect association studies. *Genet Epidemiol* 27:415-428.
- [6] DEVLIN B, ROEDER K (1999). Genomic control for association studies. *Biometrics* **55**, 997-1004.
- [7] FU Y, CHEN J, KALBFLEISCH JD (2006). Testing for homogeneity in genetic linkage analysis. *Statistica Sinica* **16**, 805-8243.
- [8] FAN R, KNAPP M (2003) Genome association studies of complex diseases by case-control designs. *Am J Hum Genet* 72:850-868.
- [9] FREIDLIN B, ZHENG G, LI Z AND GASTWIRTH JL (2002). Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Human Hered* **53**, 146-152.
- [10] HARTIGAN, J.A. 1985 A failure of likelihood asymptotics for normal mixtures. In *Proc. Berk. Conf. in Honor of J. Neyman and J. Kiefer.* **2** 807-810. Edited by L. LeCam and R.A. Olshen.

- [11] MCLACHLAN, G. (1987). On bootstrapping likelihood ratio test statistics for the number of components in a normal mixture. *Applied Statistics* **36**, 318-324.
- [12] MCLACHLAN G., PEEL, D. (2000). *Finite Mixture Models*. Wiley, New York.
- [13] OTT J. (1983). Linkage analysis and family classification under heterogeneity. *Ann Hum Genet* 47:311-320.
- [14] OTT J. (1999). *Analysis of Human Genetic Linkage*. Third Edition. The John Hopkins University Press, Baltimore.
- [15] PAN W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. To appear in *Genetic Epidemiology*. Published Online on Jan 23 2009. DOI: 10.1002/gepi.20402.
- [16] PRICE AL, PATTERSON NJ, PLENGE RM, WEINBLATT ME, SHADICK NA, REICH D (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* **38**, 904-909.
- [17] PRITCHARD JK, STEPHENS M, DONNELLY P (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.
- [18] PURCELL S, NEALE B, TODD-BROWN K, THOMAS L, FERREIRA MAR, BENDER D, MALLER J, SKLAR P, DE BAKKER PIW, DALY MJ, SHAM PC (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics* **81**, 559-575.
- [19] SATTEN GA, FLANDERS WD, YANG Q (2001). Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am J Hum Genet* **68**, 466-477.

- [20] SCHYMICK JC, SCHOLZ SW, FUNG HC, ET AL. (2007). Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol*, **6**, 322-328.
- [21] SMITH C.A.B. (1963). Testing for heterogeneity of recombination fraction values in human genetics. *Ann Hum Genet* 27:175-182.
- [22] WANG K, ABBOTT D (2008). A principle components regression approach to multilocus genetic association studies. *Genetic Epidemiology* 32:108-118.
- [23] XIONG M, ZHAO J, BOERWINKLE E (2002). Generalized T^2 test for genome association studies. *Am J Hum Genet* 70:1257-1268.
- [24] ZHANG S, KIDD KK, ZHAO H (2002). Detecting genetic association in case-control studies using similarity-based association tests. *Statistica Sinica* **12**, 337-359.
- [25] ZHU X, ZHANG S, ZHAO H, COOPER RS (2002). Association mapping, using a mixture model for complex traits. *Genet Epidemiol* **23**, 181-196.