Am. J. Hum. Genet. 77:768–780, 2005

A Constrained-Likelihood Approach to Marker-Trait Association Studies

Kai Wang[1] and Val C. Sheffield[2,3]

[1]Program in Public Health Genetics, College of Public Health, [2]Department of Pediatrics, and [3]Howard Hughes Medical Institute, University of Iowa, Iowa City

Marker-trait association analysis is an important statistical tool for detecting DNA variants responsible for genetic traits. In such analyses, an analysis model of the mean genetic effects of the genotypes is often specified. For instance, the effect of the disease allele on the trait is often specified to be dominant, recessive, additive, or multiplicative. Although this model-based approach is powerful when the analysis model is correctly specified, it has been found to have low power sometimes when the specified model is incorrect. We introduce an approach that does not require the specification of a particular genetic model. This approach is built upon a constrained maximum likelihood in which the mean genetic effect of the heterozygous genotype is required to not exceed those of the two homozygous genotypes. The asymptotic distribution of the likelihood-ratio statistic is derived for two special cases. A simulation study suggests that this new approach has power comparable to that of the model-based method when the analysis model is correctly specified. This approach uses one marker at a time (i.e., it is a single-marker analysis). However, given the latest findings that powerful inferential procedures for haplotype analyses can be constructed from single-marker analyses, we expect this approach to be useful for haplotype analyses.

Introduction

Marker-trait association analysis is an important statistical tool for detecting DNA variants responsible for genetic traits. It can provide higher mapping resolution than do methods based on closely related meiosis events. In such analyses, it is common to specify an analysis model of the average genetic effects of the genotypes. For instance, the effect of the disease allele on the trait is often specified to be dominant, recessive, additive, or multiplicative. When the specified model is close to the underlying trait model, this model-based approach provides a powerful means of detecting association. However, when the specified model is different from the underlying model, its power may be low (Slager and Schaid 2001; Freidlin et al. 2002; Schaid et al. 2005). In real-data analyses, the underlying genetic model is often unknown. For instance, one promising approach to investigation of the gene-regulation mechanism is to map the expression levels of genes by treating them as quantitative traits (Brem et al. 2002; Schadt et al. 2003; Yvert et al. 2003; Morley et al. 2004). Currently, gene-expression arrays contain thousands of DNA probes, and each probe provides a quantitative measurement of its expression level. Given this large number of traits, the model-based method may miss many true-positive signals. It is desirable to develop methods that have power for a wide range of genetic models.

Analytical methods that have power for a wide range of genetic models are desired not only for single-marker analysis but also for haplotype analysis in which multiple markers are involved. Some recent developments indicate that single-marker tests can be used to construct powerful inference procedures for haplotype analysis. Chapman et al. (2003) found that regression analysis based on a linear combination of tagSNPs is more powerful than the traditional haplotype analysis that is based on haplotype frequencies. Roeder et al. (2005) further found that inferential procedures based on single-marker tests, after correction for multiple testing by permutation or curve fitting, is at least as powerful as the regression method proposed by Chapman et al. (2003). In a more recent study, Schaid et al. (2005) described a testing procedure for haplotype analysis that requires specification of a "kernel." One promising way of specifying a kernel is to derive it from single-marker tests (Schaid et al. 2005).

There have been some studies, mostly on dichotomous traits, that have investigated methods that do not rely on a particular analysis model. Freidlin et al. (2002) proposed a maximin efficiency robust test and a test (named "MAX") based on the maximum of test statistics under several analysis models. They found that the MAX test is generally more powerful than the other one. Freidlin et al. (2002) assumed an a priori ordering of the mean genetic effects for the three genotypes that are induced from the allele to be tested, by assuming that

Received June 9, 2005; accepted for publication August 24, 2005; electronically published September 14, 2005.

Address for correspondence and reprints: Dr. Kai Wang, Program in Public Health Genetics, 2188 WL, College of Public Health, University of Iowa, Iowa City, IA 52242. E-mail: kai-wang@uiowa.edu

0002-9297/2005/7705-0008$15.00

Table 1

Some Quantiles for Test Statistics Λ_{New} and Λ_{Larger}

Frequency of Allele A and Statistic	Statistic at Nominal Significance Level			
	.1	.05	.01	.001
$p = .01$:				
Λ_{New}	4.197	5.507	8.597	13.084
Λ_{Larger}	3.796	5.000	7.874	12.115
$p = .1$:				
Λ_{New}	4.152	5.457	8.536	13.008
Λ_{Larger}	3.778	4.985	7.867	12.113
$p = .3$:				
Λ_{New}	4.111	5.413	8.485	12.949
Λ_{Larger}	3.752	4.964	7.855	12.109
$p = .5$:				
Λ_{New}	4.100	5.401	8.472	12.934
Λ_{Larger}	3.745	4.957	7.852	12.107

Note.—HWE is assumed, so that $p_0 = (1 - p)^2$, $p_1 = 2p(1 - p)$, and $p_2 = p^2$.

the marker allele associated with the disease allele is known. Such an ordering can be difficult to make—for instance, in the expression-level mapping example mentioned above. To remove this restriction, Zheng (2003) proposed a "max and min scores" approach. Another method that does not require the specification of an analysis model is to simply compare mean genotypic effects by use of standard statistical methods such as analysis of variance. But such an approach may have low power due to increased degrees of freedom.

Instead of deriving tests from several model-based statistics or simply comparing the mean genotypic effects, we adopt a constrained-likelihood analysis under a so-called no-overdominance constraint. This constraint requires that the mean genetic effect of the heterozygous genotype not exceed those of the two homozygous genotypes—that is, it is neither larger than the larger of the mean genetic effects of the two homozygous genotypes nor smaller than the smaller of the two. We note that this constraint is satisfied by the commonly assumed analysis models—dominance, recessive, additive, and multiplicative.

In the following section, we introduce our approach in terms of a generalized linear model. We then apply this approach to quantitative traits and dichotomous traits. The asymptotic distribution of the constrained likelihood-ratio statistic for each kind of trait is introduced. Simulation studies were conducted to assess the performance of our method, under different generating models, in comparison with some popular methods. Technical details are given in appendixes A and B.

Methods

Let A denote the allele being tested for association with a trait. The three genotypes induced from allele A are indexed by j ($j = 0,1,2$), where j is the number of copies of allele A. Denote the population frequency of genotype j by p_j. In some situations, the values of p_j are known. For instance, for the F_2 population, $p_0 = p_2 = 0.25$ and $p_1 = 0.5$. Suppose that there are n_j individuals with genotype j. Given the total number of individuals $n := n_0 + n_1 + n_2$, the triplet (n_0,n_1,n_2) follows a trinomial distribution with a parameter vector (p_0,p_1,p_2). The trait value of the ith individual of genotype j is denoted by y_{ji}. For dichotomous traits, it is defined that $y_{ji} = 1$ for cases and $y_{ji} = 0$ for controls. The sample mean for genotype j is denoted by $\bar{y}_j := n_j^{-1} \sum_{i=1}^{n_j} y_{ji}$.

Consider the following generalized linear model for phenotype y with link function $g(\cdot)$: $E(y) = \mu$ and $g(\mu) = \alpha + \delta_1 x_1 + (\delta_1 + \delta_2)x_2$, where x_j, $j = 1,2$, is an indicator of genotype j satisfying $x_j = 1$ if the individual is of genotype j and $x_j = 0$ otherwise. Depending on the random component of the generalized linear model, there may be another nuisance parameter β (possibly a vector). For instance, for normal data, β is the variance of the random component. In the constrained-likelihood approach introduced here, we require a "no-overdominance" constraint on the three mean genotypic effects. That is, the three genotypic means satisfy either $\alpha \leqslant \alpha + \delta_1 \leqslant \alpha + \delta_1 + \delta_2$ or $\alpha \geqslant \alpha + \delta_1 \geqslant \alpha + \delta_1 + \delta_2$. This constraint implies that δ_1 and δ_2 cannot be of different signs, and it can be equivalently written $\delta_1\delta_2 \geqslant 0$, which corresponds to the first and third quadrants on the δ_1-δ_2 plane. When there is no association for allele A, $\delta_1 = \delta_2 = 0$. Let $\Theta_0 = \{(\delta_1,\delta_2,\alpha,\beta):\delta_1 = \delta_2 = 0\}$ and $\Theta_1 = \{(\delta_1,\delta_2,\alpha,\beta):\delta_1\delta_2 \geqslant 0\}$. The hypotheses of interest are

$$H_0:(\delta_1,\delta_2,\alpha,\beta) \in \Theta_0$$
$$H_1:(\delta_1,\delta_2,\alpha,\beta) \in \Theta_1 - \Theta_0 . \quad (1)$$

The other parameters p_0, p_1, p_2, α, and β are nuisance parameters. The requirement $\delta_1\delta_2 \geqslant 0$ contains many commonly used genetic models as special cases. For instance, when the effect of allele A is dominant, we have $\delta_2 = 0$, and there is no restriction on δ_1. When the effect of allele A is recessive, we have $\delta_1 = 0$, and there is no restriction on δ_2.

The likelihood function of the data $\{y_{ji}\}$ is

$$L(\delta_1,\delta_2,\alpha,\beta,p_0,p_1,p_2) = L_1(p_0,p_1,p_2) \times L_2(\delta_1,\delta_2,\alpha,\beta) ,$$

where $L_1(p_0,p_1,p_2) = \Pr(n_1,n_2,n_3|p_0,p_1,p_2,n)$ is the trinomial probability of (n_1,n_2,n_3), given n, and where $L_2(\delta_1,\delta_2,\alpha,\beta) = \prod_{j=0}^{2} \prod_{i=1}^{n_j} \Pr(y_{ji}|\delta_1,\delta_2,\alpha,\beta)$ is the conditional probability of the trait values $\{y_{ji}\}$, given (n_1,n_2,n_3).

The hypotheses in (1) can be tested using the

Table 2

Simulated Type I Error Rates for Quantitative Traits

Frequency of Allele *A*, Sample Size, and Statistic	Type I Error Rate at Nominal Significance Level			
	.1	.05	.01	.001
$p = .1$:				
$n = 100$:				
Λ_{New}	.0774	.0366	.0082	.0011
Λ_{Dom}	.1024	.0517	.0116	.0010
Λ_{Rec}	.0669	.0334	.0068	.0009
Λ_{Add}	.1008	.0516	.0108	.0011
Λ_{Larger}	.0862	.0435	.0086	.0009
Λ_{Unc}	.0787	.0379	.0091	.0011
$n = 200$:				
Λ_{New}	.0908	.0435	.0097	.0006
Λ_{Dom}	.1075	.0524	.0122	.0011
Λ_{Rec}	.0848	.0418	.0076	.0010
Λ_{Add}	.1095	.0529	.0116	.0009
Λ_{Larger}	.0949	.0481	.0102	.0012
Λ_{Unc}	.0957	.0457	.0096	.0007
$p = .3$:				
$n = 100$:				
Λ_{New}	.1018	.0506	.0116	.0011
Λ_{Dom}	.1064	.0548	.0121	.0015
Λ_{Rec}	.1057	.0518	.0110	.0011
Λ_{Add}	.1062	.0533	.0109	.0013
Λ_{Larger}	.1069	.0544	.0126	.0012
Λ_{Unc}	.1111	.0537	.0128	.0011
$n = 200$:				
Λ_{New}	.0971	.0491	.0126	.0014
Λ_{Dom}	.1028	.0543	.0107	.0019
Λ_{Rec}	.0996	.0511	.0116	.0014
Λ_{Add}	.1038	.0532	.0109	.0010
Λ_{Larger}	.1063	.0535	.0113	.0017
Λ_{Unc}	.1039	.0529	.0124	.0015
$p = .5$:				
$n = 100$:				
Λ_{New}	.0972	.0496	.0108	.0012
Λ_{Dom}	.1057	.0508	.0105	.0012
Λ_{Rec}	.1064	.0548	.0130	.0014
Λ_{Add}	.1034	.0508	.0109	.0017
Λ_{Larger}	.1055	.0515	.0120	.0010
Λ_{Unc}	.1041	.0540	.0109	.0011
$n = 200$:				
Λ_{New}	.0997	.0478	.0100	.0013
Λ_{Dom}	.1024	.0507	.0115	.0018
Λ_{Rec}	.1043	.0514	.0107	.0009
Λ_{Add}	.1072	.0520	.0095	.0008
Λ_{Larger}	.1039	.0531	.0112	.0014
Λ_{Unc}	.1031	.0511	.0099	.0013

likelihood-ratio statistic. Since $L_1(p_0,p_1,p_2)$ and $L_2(\delta_1,\delta_2,\alpha,\beta)$ involve two nonoverlapping sets of parameters and since only $L_2(\delta_1,\delta_2,\alpha,\beta)$ contains the parameters of interest, the likelihood-ratio statistic equals

$$\Lambda_{New}: = 2\left[\max_{\delta_1,\alpha,\beta\,(\text{subject to }\delta_1\delta_2\geq 0)} l_2(\delta_1,\delta_2,\alpha,\beta) - l_0\right],$$

where $l_2(\delta_1,\delta_2,\alpha,\beta) = \log[L_2(\delta_1,\delta_2,\alpha,\beta)]$ and $l_0 = \max_{\alpha,\beta} l_2(\delta_1 = 0,\delta_2 = 0,\alpha,\beta)$. The nuisance parameters p_0, p_1, and p_2 do not appear in the calculation of Λ_{New}. However, we show below that the asymptotic distribution of Λ_{New} can depend on p_0, p_1, and p_2.

To compute Λ_{New}, it is essential to compute the constrained maximum of $l_2(\delta_1,\delta_2,\alpha,\beta)$. For this purpose, define

$$\Lambda_{Dom} = 2\left[\max_{\delta_1,\alpha,\beta} l_2(\delta_1,\delta_2 = 0,\alpha,\beta) - l_0\right]$$

to be the likelihood-ratio statistic for the dominance model and

$$\Lambda_{Rec} = 2\left[\max_{\delta_2,\alpha,\beta} l_2(\delta_1 = 0,\delta_2,\alpha,\beta) - l_0\right]$$

to be the likelihood-ratio statistic for the recessive model. Further define $\Lambda_{Larger} = \max\{\Lambda_{Dom},\Lambda_{Rec}\}$. In principle, the constrained maximum of $l_2(\delta_1,\delta_2,\alpha,\beta)$ is straightforward to compute. According to standard optimization theory, there are two possible situations regarding the optimal values of δ_1 and δ_2. They are either in the interior of the region satisfying $\delta_1\delta_2 > 0$ or on the border of this region. The constrained maximum of

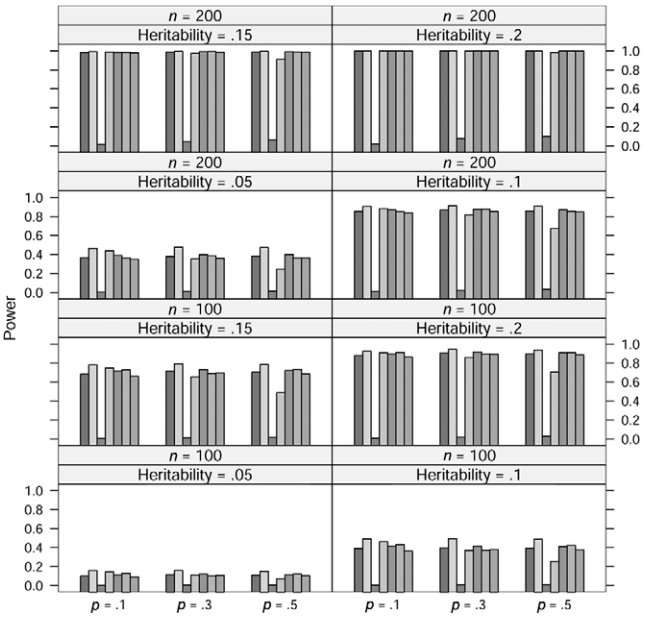

Figure 1 Power comparison for the quantitative trait when the generating model is dominant. The significance level is 0.001. The order of statistics (*shaded bars*) is (from left to right): Λ_{New}, Λ_{Dom}, Λ_{Rec}, Λ_{Add}, Λ_{Larger}, $\Lambda_{Largest}$, and Λ_{Unc}.

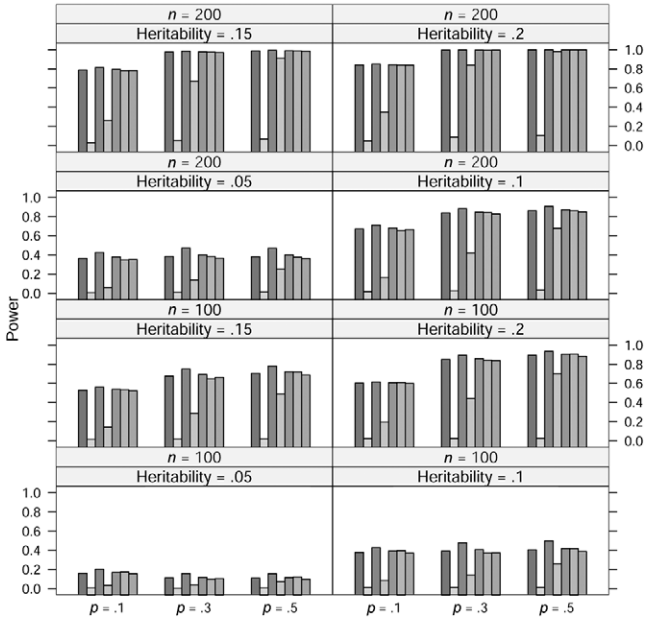

Figure 2 Power comparison for the quantitative trait when the generating model is recessive. The significance level is 0.001. The order of statistics (*shaded bars*) is (from left to right): Λ_{New}, Λ_{Dom}, Λ_{Rec}, Λ_{Add}, Λ_{Larger}, Λ_{Largest}, and Λ_{Unc}.

$l_2(\delta_1,\delta_2,\alpha,\beta)$ equals its unconstrained maximum in the former case and equals Λ_{Larger} in the latter case.

Specifically, the constrained maximum of $l_2(\delta_1,\delta_2,\alpha,\beta)$ can be obtained as follows. Do the unconstrained maximization of $l_2(\delta_1,\delta_2,\alpha,\beta)$, and denote the values of δ_1, δ_2, α, and β for which $l_2(\delta_1,\delta_2,\alpha,\beta)$ is maximized by $\hat{\delta}_1$, $\hat{\delta}_2$, $\hat{\alpha}$, and $\hat{\beta}$, respectively. The constrained maximum of $l_2(\delta_1,\delta_2,\alpha,\beta)$ equals $l_2(\hat{\delta}_1,\hat{\delta}_2,\hat{\alpha},\hat{\beta})$ if $\hat{\delta}_1\hat{\delta}_2 > 0$ and equals Λ_{Larger} otherwise.

Next, we discuss two particular applications of this constrained-likelihood approach, one to quantitative traits and the other to dichotomous traits. For each application, the asymptotic distribution of the likelihood-ratio statistic Λ_{New} is derived. There should be many other possible applications, depending on the specification of the random component and the link function of the generalized linear model.

Quantitative Traits

Quantitative traits are usually modeled through a normal distribution, and the "canonical" link function is the identity function $g(\mu) = \mu$. Assume that the variance of the normal distribution is the same, regardless of the genotype, and denote this common variance by σ^2. The parameter σ^2 corresponds to the nuisance parameter β in our generalized-linear-model setup. Now, the function l_2 becomes

$$l_2(\delta_1,\delta_2,\alpha,\sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{n_0}(y_{0i}-\alpha)^2 + \sum_{i=1}^{n_1}(y_{1i}-\alpha-\delta_1)^2 + \sum_{i=1}^{n_2}(y_{2i}-\alpha-\delta_1-\delta_2)^2\right].$$

The unconstrained maximum-likelihood estimates of α, δ_1, δ_2, and σ^2 are $\hat{\alpha} = \bar{y}_0$, $\hat{\delta}_1 = \bar{y}_1 - \hat{\alpha}$, $\hat{\delta}_2 = \bar{y}_2 - \hat{\delta}_1 - \hat{\alpha}$, and

$$\hat{\sigma}^2 = \frac{1}{n}\left[\sum_{i=1}^{n_0}(y_{0i}-\hat{\alpha})^2 + \sum_{i=1}^{n_1}(y_{1i}-\hat{\alpha}-\hat{\delta}_1)^2 + \sum_{i=1}^{n_2}(y_{2i}-\hat{\alpha}-\hat{\delta}_1-\hat{\delta}_2)^2\right], \quad (2)$$

respectively. For the recessive model ($\delta_1 = 0$), the maximum-likelihood estimates of α and δ_2 are $\hat{\alpha} = (n_0 + n_1)^{-1}(n_0\bar{y}_0 + n_1\bar{y}_1)$ and $\hat{\delta}_2 = \bar{y}_2 - \hat{\alpha}$, respectively. For the dominance model ($\delta_2 = 0$), the maximum-likelihood

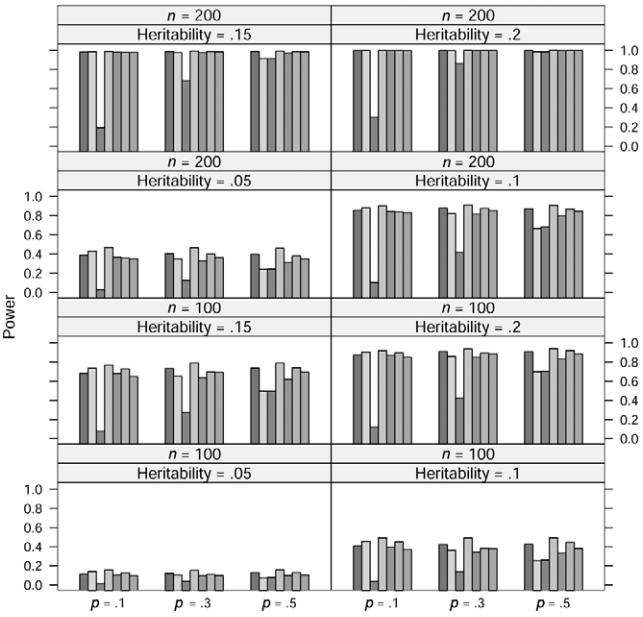

Figure 3 Power comparison for the quantitative trait when the generating model is additive. The significance level is 0.001. The order of statistics (*shaded bars*) is (from left to right): Λ_{New}, Λ_{Dom}, Λ_{Rec}, Λ_{Add}, Λ_{Larger}, Λ_{Largest}, and Λ_{Unc}.

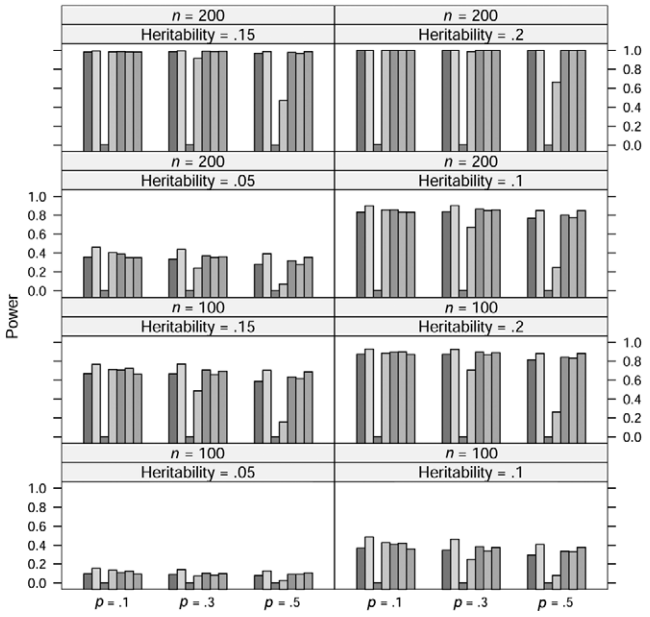

Figure 4 Power comparison for the quantitative trait when the generating model is overdominant. The significance level is 0.001. The order of statistics (*shaded bars*) is (from left to right): Λ_{New}, Λ_{Dom}, Λ_{Rec}, Λ_{Add}, Λ_{Larger}, $\Lambda_{Largest}$, and Λ_{Unc}.

estimates of α and δ_1 are $\hat{\alpha} = \bar{y}_0$ and $\hat{\delta}_1 = (n_1 + n_2)^{-1}(n_1\bar{y}_1 + n_2\bar{y}_2) - \hat{\alpha}$, respectively. The variance σ^2 is estimated by fixing $\hat{\delta}_1 = 0$ (for the recessive model) or $\hat{\delta}_2 = 0$ (for the dominance model) in equation (2). Under the null hypothesis, we have $\delta_1 = \delta_2 = 0$, and the maximum-likelihood estimate of α is $(n_0\bar{y}_0 + n_1\bar{y}_1 + n_2\bar{y}_2)/n$. Given these results, it is straightforward to compute the statistic Λ_{New}.

Let $\gamma = (p_0p_2)^{1/2}[(1 - p_0)(1 - p_2)]^{-1/2}$ and $\kappa = (2\pi)^{-1}\arccos(\gamma)$. In appendix A, it is shown that, as $n \to \infty$,

$$\Pr(\Lambda_{New} \leqslant x) \to (1 - 2\kappa)\Pr(z_1^2 \leqslant x^2, z_2^2 \leqslant x^2) + 2\kappa\Pr(\chi_2^2 < x) ,$$

where $(z_1,z_2)^t$ follows a standard bivariate normal distribution with correlation coefficient γ and where χ_2^2 follows a χ^2 distribution with 2 df. It is also shown in appendix A that, as $n \to \infty$,

$$\Pr(\Lambda_{Larger} \leqslant x) \to \Pr(z_1^2 \leqslant x^2, z_2^2 \leqslant x^2) .$$

The statistic Λ_{Larger} was proposed as a robust test statistic (Freidlin et al. 2002), but its asymptotic distribution was not given. The correlation coefficient γ depends on the genotype frequencies p_0 and p_2. When their true values are unknown, the genotype frequencies p_0 and p_2 can be consistently estimated by their respective sample frequencies.

Dichotomous Traits

Dichotomous traits are usually modeled through binomial distribution, and the "canonical" link function

Table 3

Simulated Type I Error Rates for Dichotomous Traits

Frequency of Allele *A*, Sample Size, and Statistic	Type I Error Rate at Nominal Significance Level			
	.1	.05	.01	.001
$p = .1$:				
$n = 100$:				
Λ_{New}	.0694	.0270	.0047	.0003
Λ_{Dom}	.1094	.0516	.0107	.0007
Λ_{Rec}	.1028	.0164	.0003	.0000
Λ_{Add}	.1041	.0522	.0112	.0008
Λ_{Larger}	.0713	.0271	.0060	.0003
Λ_{Unc}	.0732	.0329	.0059	.0004
$n = 200$:				
Λ_{New}	.1158	.0494	.0075	.0006
Λ_{Dom}	.1064	.0520	.0095	.0013
Λ_{Rec}	.2008	.0577	.0039	.0001
Λ_{Add}	.1045	.0531	.0106	.0011
Λ_{Larger}	.1057	.0424	.0064	.0008
Λ_{Unc}	.1101	.0484	.0078	.0006
$p = .3$:				
$n = 100$:				
Λ_{New}	.1068	.0544	.0118	.0009
Λ_{Dom}	.0869	.0570	.0120	.0009
Λ_{Rec}	.1111	.0621	.0145	.0012
Λ_{Add}	.1033	.0536	.0096	.0005
Λ_{Larger}	.1134	.0553	.0138	.0011
Λ_{Unc}	.1103	.0580	.0119	.0013
$n = 200$:				
Λ_{New}	.0935	.0445	.0099	.0017
Λ_{Dom}	.1039	.0539	.0081	.0009
Λ_{Rec}	.0978	.0493	.0105	.0020
Λ_{Add}	.1035	.0499	.0082	.0014
Λ_{Larger}	.0999	.0499	.0104	.0014
Λ_{Unc}	.0971	.0480	.0113	.0016
$p = .5$:				
$n = 100$:				
Λ_{New}	.0968	.0487	.0100	.0006
Λ_{Dom}	.1017	.0498	.0101	.0011
Λ_{Rec}	.1012	.0509	.0101	.0006
Λ_{Add}	.1014	.0492	.0106	.0007
Λ_{Larger}	.1073	.0546	.0109	.0010
Λ_{Unc}	.1039	.0558	.0115	.0007
$n = 200$:				
Λ_{New}	.0977	.0471	.0084	.0007
Λ_{Dom}	.1008	.0503	.0088	.0011
Λ_{Rec}	.1023	.0539	.0102	.0006
Λ_{Add}	.1053	.0525	.0103	.0009
Λ_{Larger}	.1054	.0508	.0088	.0007
Λ_{Unc}	.1032	.0488	.0083	.0006

NOTE.—Half of the sample is cases, and the other half is controls.

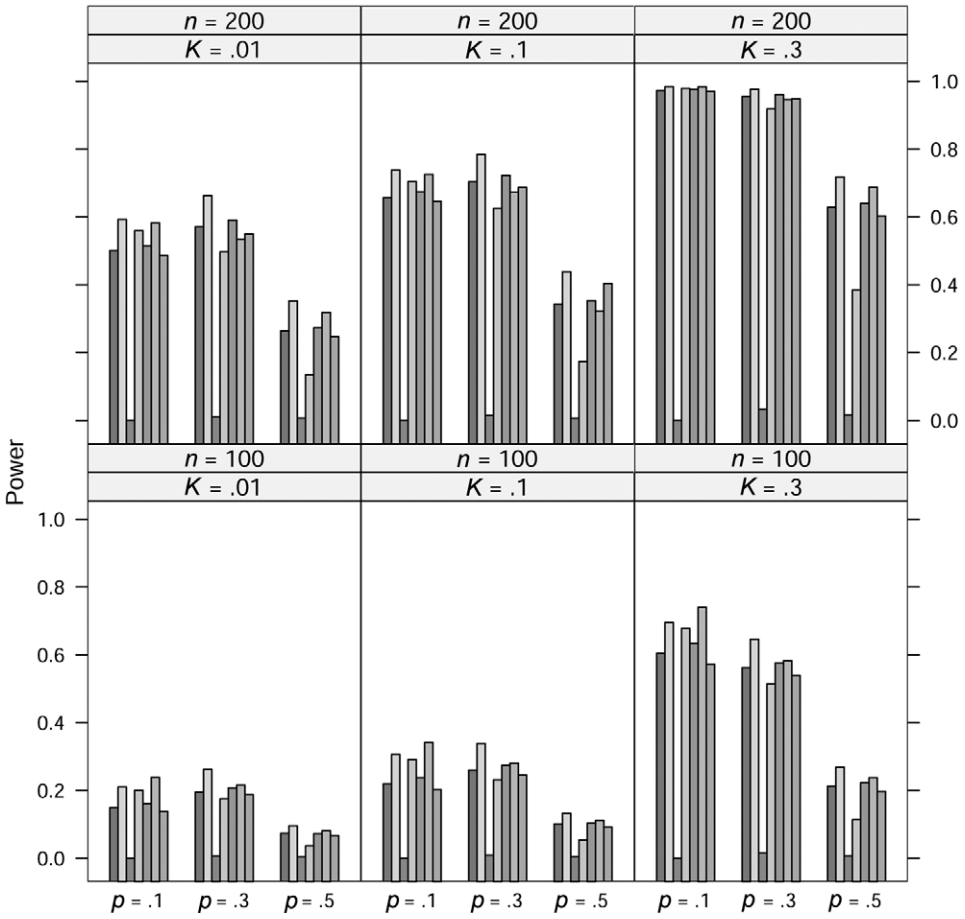

Figure 5 Power comparison for the dichotomous trait when the generating model is dominant. The significance level is 0.001. Half of the sample is cases, and the other half is controls. The order of statistics (*shaded bars*) is (from left to right): Λ_{New}, Λ_{Dom}, Λ_{Rec}, Λ_{Add}, Λ_{Larger}, Λ_{Largest}, and Λ_{Unc}.

is the logit function $g(\mu) = \log[\mu/(1-\mu)]$. In this situation, the log-likelihood function l_2 becomes, up to an additive constant,

$$l_2(\delta_1,\delta_2,\alpha) = \sum_{j=0}^{2} n_j\left[\bar{y}_j \log f_j + (1-\bar{y}_j)\log(1-f_j)\right],$$

where $f_0 := \exp(\alpha)/[1+\exp(\alpha)]$, $f_1 := \exp(\alpha+\delta_1)/[1+\exp(\alpha+\delta_1)]$, and $f_2 := \exp(\alpha+\delta_1+\delta_2)/[1+\exp(\alpha+\delta_1+\delta_2)]$ are the penetrances of the trait for the three genotypes that have 0, 1, and 2 copies of allele A, respectively. We note that the Armitage trend test typically assumes $\delta_1 = \delta_2$ (Sasieni 1997).

The likelihood function $l_2(\delta_1,\delta_2,\alpha)$ depends on δ_1, δ_2, and α through f_0, f_1, and f_2. The constraint $\delta_1\delta_2 \geq 0$ holds if and only if $f_0 \geq f_1 \geq f_2$ or $f_0 \leq f_1 \leq f_2$ holds. It is easy to obtain that the unconstrained maximum-likelihood estimate of f_j is $\bar{y}_j$, $j = 0,1,2$. For the dominance model where $f_1 = f_2$ (equivalent to $\delta_2 = 0$), the maximum-likelihood estimate of f_0 is $\bar{y}_0$, and the maximum-likelihood estimate of $f_1 = f_2$ is $(n_1\bar{y}_1 + n_2\bar{y}_2)/(n_1+n_2)$. For the recessive model where $f_0 = f_1$ (equivalent to $\delta_1 = 0$), the maximum-likelihood estimate of $f_0 = f_1$ is $(n_0\bar{y}_0 + n_1\bar{y}_1)/(n_0+n_1)$, and the maximum-likelihood estimate of f_2 is $\bar{y}_2$. Under the null hypothesis, we have $f_0 = f_1 = f_2$, and the maximum-likelihood estimate is $(n_0\bar{y}_0 + n_1\bar{y}_1 + n_2\bar{y}_2)/n$. Given these results, it is again straightforward to compute the statistic Λ_{New}.

The generalized linear model is appropriate for population samples. For selected samples such as cases and controls, it may not be appropriate. However, it has been shown that simply applying logit regression to case-control data affects only the intercept term α and not the slopes δ_1 and δ_2 (Prentice and Pyke 1979). Following the arguments of Prentice and Pyke (1979), it can be shown that the intercept term depends on δ_1, δ_2, and α and that the constraint $\delta_1\delta_2 \geq 0$ has no impact on the value the intercept term can take (details omitted). So, for a case-control study, one can treat the data as if they were population samples and apply the proposed test.

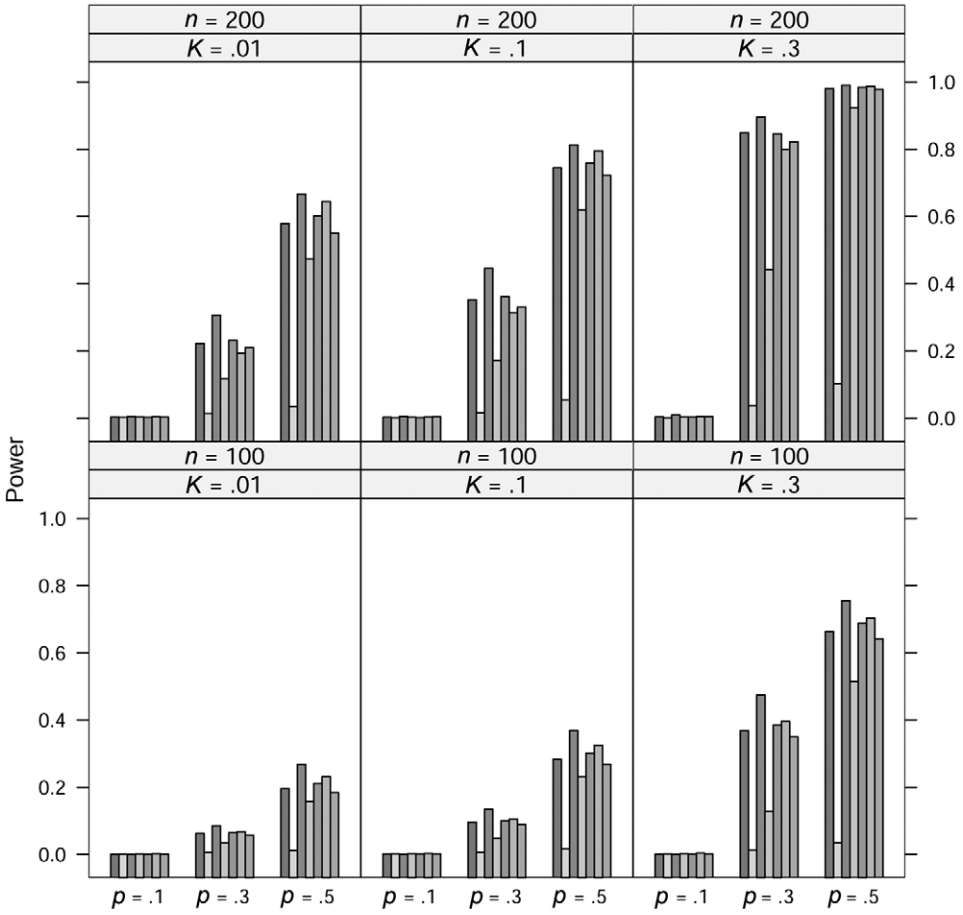

Figure 6 Power comparison for the dichotomous trait when the generating model is recessive. The significance level is 0.001. Half of the sample is cases, and the other half is controls. The order of statistics (*shaded bars*) is (from left to right): Λ_{New}, Λ_{Dom}, Λ_{Rec}, Λ_{Add}, Λ_{Larger}, $\Lambda_{Largest}$, and Λ_{Unc}.

In appendix B, it is shown that the asymptotic distribution of Λ_{New} in this situation is identical to that of the likelihood-ratio statistic Λ_{New} for continuous traits. When true values are unknown, the genotype frequencies p_0 and p_2 can be estimated by their respective sample frequencies in the combined sample of cases and controls.

Let p denote the frequency of allele A. Under the assumption of Hardy-Weinberg equilibrium (HWE), some quantiles of the asymptotic distribution for Λ_{New} are tabulated in table 1 for $p = 0.01$, 0.1, 0.3, and 0.5. We note that, for $p = 0.5$, the genotype frequencies are $p_0 = p_2 = 0.25$ and $p_1 = 0.5$, which are the expected genotype frequencies for the F_2 population.

Simulation

Simulation studies were done for both quantitative traits and dichotomous traits. The type I error rate and the power were computed on the basis of 10,000 replications.

Quantitative Traits

Consider the following data-generating model:

$$y = G + \epsilon \ ,$$

where G is the genotypic value and ϵ is an independent environmental factor. Let $G = -a$ for genotype 0, $G = d$ for genotype 1, and $G = a$ for genotype 2. The distribution of ϵ is taken to be the standard normal distribution whose mean is 0 and whose variance is 1. The heritability for this model is $h^2 := V_G/(V_G + 1)$, where V_G is the variance of the genotypic effect. Under the assumption of HWE, p_0, p_1, and p_2 can be written as $p_0 = q^2, p_1 = 2pq$, and $p_2 = p^2$, where p is the frequency of allele A and $q = 1 - p$. According to Falconer and Mackay (1996), the genotypic variance $V_G = 2pq[a + d(q - p)]^2 + (2pqd)^2$. Four generating models are considered. They are the dominance model ($d = a$), the recessive model ($d = -a$), the additive model ($d = 0$), and an overdominance model ($d = 2a$). Given heritability h^2, the

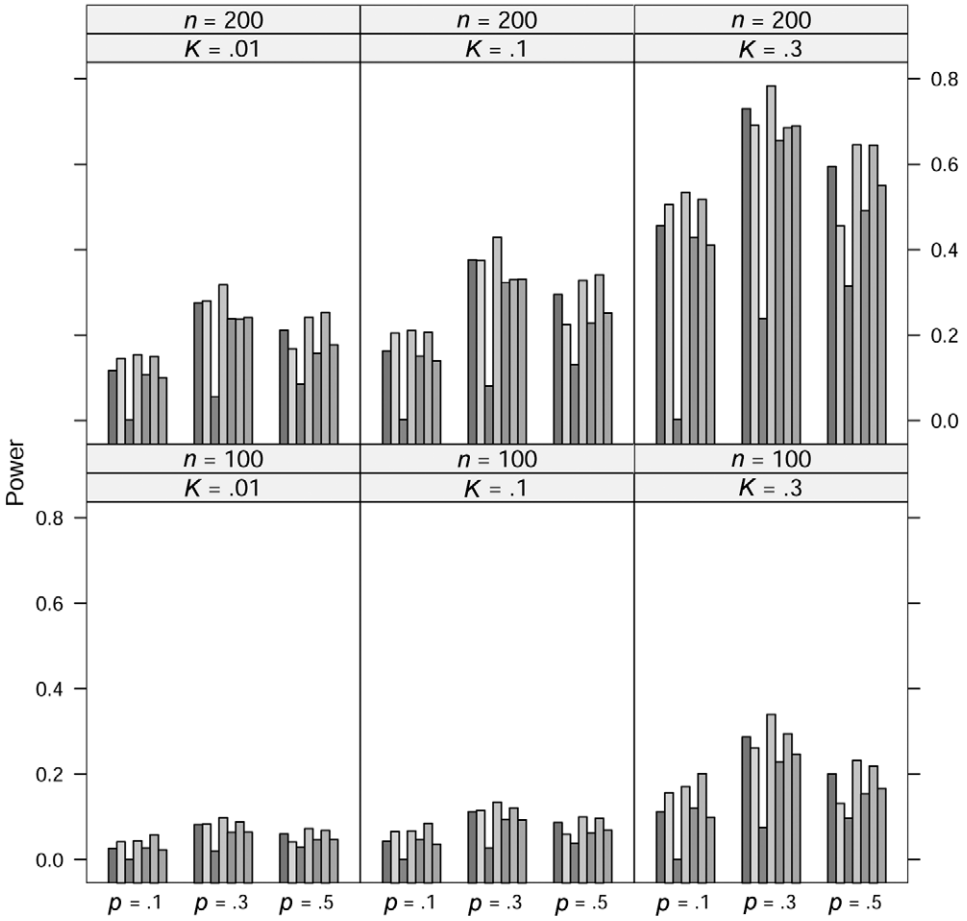

Figure 7 Power comparison for the dichotomous trait when the generating model is additive. The significance level is 0.001. Half of the sample is cases, and the other half is controls. The order of statistics (*shaded bars*) is (from left to right): Λ_{New}, Λ_{Dom}, Λ_{Rec}, Λ_{Add}, Λ_{Larger}, $\Lambda_{Largest}$, and Λ_{Unc}.

value of V_G can be obtained from $V_G = h^2/(1 - h^2)$. For any given allele frequency p, the value of a for each model can be determined as follows: $a = [V_G/4pq^2(1 + q)]^{1/2}$ for the dominance model, $a = [V_G/4p^2q(1 + p)]^{1/2}$ for the recessive model, $a = (V_G/2pq)^{1/2}$ for the additive model, and $a = [V_G/(2pq(4q - 1)^2 + 16p^2q^2)]^{1/2}$ for the overdominance model.

To analyze data from these generating models, seven statistics are computed. They are the proposed statistic Λ_{New}, the likelihood-ratio statistic Λ_{Dom} for the dominant model ($\delta_2 = 0$), the likelihood-ratio statistic Λ_{Rec} for the recessive model ($\delta_1 = 0$), the likelihood-ratio statistic Λ_{Add} for the additive model ($\delta_2 = 2\delta_1$), the statistic $\Lambda_{Larger} = \max\{\Lambda_{Dom},\Lambda_{Rec}\}$, the statistic $\Lambda_{Largest} := \max\{\Lambda_{Dom},\Lambda_{Rec},\Lambda_{Add}\}$, and the unconstrained likelihood-ratio statistic Λ_{Unc} that tests whether the three genotypic means are the same or not. The asymptotic distributions for Λ_{New} and Λ_{Larger} are given in this article. The three statistics Λ_{Dom}, Λ_{Rec}, and Λ_{Add} follow an asymptotic χ^2 distribution with 1 df. The asymptotic distribution for Λ_{Unc} is a χ^2 with 2 df. The statistic $\Lambda_{Largest}$ was introduced by Freidlin et al. (2002). Since its null distribution is unknown, simulated critical values are used in its power study.

The simulated type I error rates for statistics Λ_{New}, Λ_{Dom}, Λ_{Rec}, Λ_{Add}, Λ_{Larger}, and Λ_{Unc} are reported in table 2 for allele A frequency $p = 0.1$, 0.3, and 0.5 and for sample size $n = 100$ and 200. These type I error rates are close to their respective nominal significance levels, which suggests that all these tests have valid size. A similar phenomenon is also observed in situations where HWE does not hold (data not shown).

In the power study, the critical values for statistics Λ_{New}, Λ_{Dom}, Λ_{Rec}, Λ_{Add}, Λ_{Larger}, and Λ_{Unc} are from the respective asymptotic null distributions of these statistics, and the critical values for statistic $\Lambda_{Largest}$ at allele A frequency $p = 0.1$, 0.3, and 0.5 are obtained from a simulation with 10,000 replications. At significance level 0.001, the power of these seven statistics at allele frequency $p = 0.1$, 0.3, and 0.5, heritability $h^2 = 0.05$,

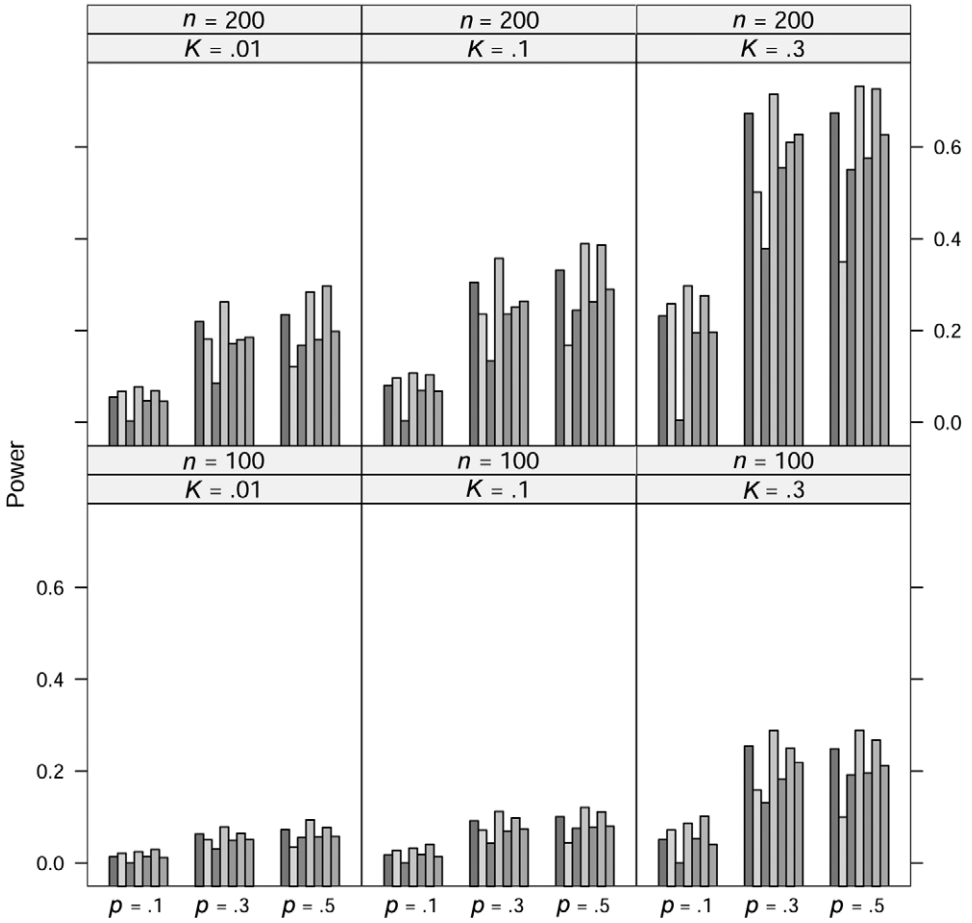

Figure 8 Power comparison for the dichotomous trait when the generating model is multiplicative. The significance level is 0.001. Half of the sample is cases, and the other half is controls. The order of statistics (*shaded bars*) is (from left to right): Λ_{New}, Λ_{Dom}, Λ_{Rec}, Λ_{Add}, Λ_{Larger}, $\Lambda_{Largest}$, and Λ_{Unc}.

0.1, 0.15, and 0.2, and sample size $n = 100$ and 200 is graphed in figures 1, 2, 3, and 4, in which the generating models are dominant, recessive, additive, and overdominant, respectively. It can be seen from these figures that, as expected, the statistic Λ_{Dom} performs best when the generating model is dominant, as do the statistic Λ_{Rec} when the model is recessive and the statistic Λ_{Add} when the model is additive. The statistic Λ_{Dom} performs worst when the generating model is recessive, and the statistic Λ_{Rec} performs worst when it is dominant. Overall, the four statistics Λ_{New}, Λ_{Larger}, $\Lambda_{Largest}$, and Λ_{Unc} seem to have similar power. When the generating model is dominant or recessive, the statistic Λ_{Larger} shows marginally the best power among these four statistics. When the generating model is additive, the statistic $\Lambda_{Largest}$ is slightly better than any of the other three. For the overdominance model, the statistic Λ_{New} has slightly more power than the statistic Λ_{Unc} for $p = 0.1$. For $p = 0.3$ and $p = 0.5$, the statistic Λ_{Unc} shows higher power than the statistic Λ_{New}.

Dichotomous Traits

Let $K = \sum_{i=0}^{2} f_i p_i$ be the prevalence of the trait. The frequency of genotype i would be $f_i p_i / K$ in the cases and $(1 - f_i)p_i/(1 - K)$ in the controls. In the absence of association, $f_0 = f_1 = f_2 = K$, and there is no difference in genotype frequencies between cases and controls. Let $\gamma_i = f_i/f_0$, $i = 1,2$, be the relative risk of genotype i compared with genotype 0. In the simulation, we consider a dominance model ($\gamma_1 = \gamma_2$), a recessive model ($\gamma_1 = 1$), an additive model ($\gamma_1 = (1 + \gamma_2)/2$), a multiplicative model ($\gamma_1 = \gamma_2^{1/2}$), and an overdominance model ($\gamma_1 = 2\gamma_2$). Given population prevalence K and the relative risk γ_2, f_0 can be determined from $f_0 = K/(p_0 + \gamma_1 p_1 + \gamma_2 p_2)$, from which $f_1 = \gamma_1 f_0$ and $f_2 = \gamma_2 f_0$ can be computed for each model. To simplify the calculation, it is assumed again that HWE holds and that the frequency of allele A is denoted by p.

The type I error rates for statistics Λ_{New}, Λ_{Dom}, Λ_{Rec}, Λ_{Add}, Λ_{Larger}, and Λ_{Unc} are reported in table 3 for al-

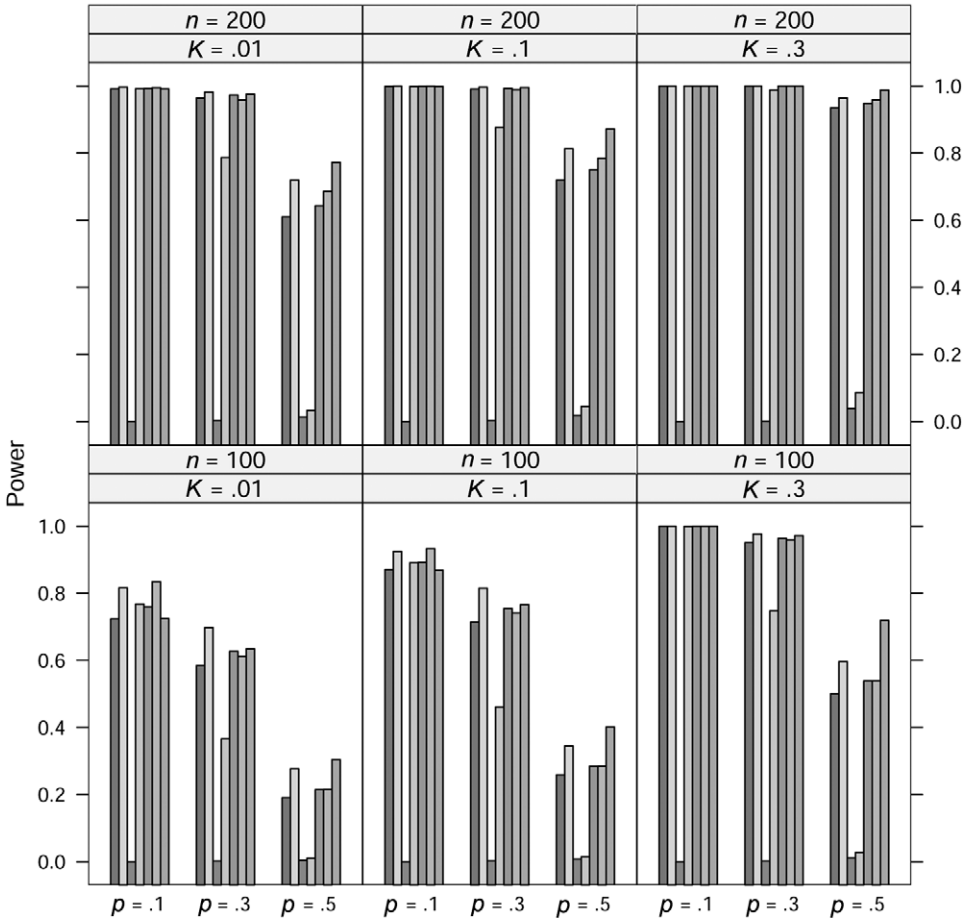

Figure 9 Power comparison for the dichotomous trait when the generating model is overdominant. The significance level is 0.001. Half of the sample is cases, and the other half is controls. The order of statistics (*shaded bars*) is (from left to right): Λ_{New}, Λ_{Dom}, Λ_{Rec}, Λ_{Add}, Λ_{Larger}, $\Lambda_{Largest}$, and Λ_{Unc}.

lele *A* frequency $p = 0.1$, 0.3, and 0.5 and sample size $n = 100$ and 200. These results suggest that all six statistics have valid type I error rates. A similar finding is also found in simulation studies where HWE fails (data not shown).

In the power study, the critical values for the seven statistics, including $\Lambda_{Largest}$, are obtained in the same manner as in the case of quantitative traits. At significance level 0.001, the power of the seven statistics is simulated for allele *A* frequency $p = 0.1$, 0.3, and 0.5, prevalence $K = 0.01$, 0.1, and 0.3, and sample size $n = 100$ and 200 under five models; the results for the dominance model, the recessive model, the additive model, the multiplicative model, and the overdominance model are graphed in figures 5, 6, 7, 8, and 9, respectively. In these simulations, the value of γ_2 is fixed at $\gamma_2 = 3$. Similar simulations were also done for $\gamma_2 = 2$, but the patterns were similar and so the results are not reported here. The pattern of the power for these seven statistics is similar to that in the case of quantitative traits.

For the recessive model, these seven statistics have very similar power when allele *A* frequency $p = 0.1$. We suspect that this may be caused by the rarity of the genotype homozygous for allele *A*, since all statistics would be close to the same likelihood-ratio statistic that tests the equality of the genotype effects between the other two genotypes.

Discussion

We have developed a constrained-likelihood approach to marker-trait association analysis. This approach does not require the specification of an analysis model. We investigated two applications of this approach, one for quantitative traits and the other for dichotomous traits. The asymptotic distribution of the constrained likelihood-ratio statistic was derived for both types of traits. Simulation studies suggest that this approach has power to detect association for a range of genetic models. Simulation results also suggest that the power of this ap-

proach seems to be close to that of the model-based method when the analysis model is correctly specified. It should be noted that the constrained-likelihood approach has been applied to linkage analyses performed using affected siblings (Holmans 1993) and using extreme discordant sib pairs (Knapp 1998; Freidlin et al. 2003) and to association studies using parent-sibs trios (Zheng et al. 2003) but not to the situation considered in the current article.

This approach provides an alternative to the model-based method and to the statistic Λ_{Unc} that does not have any restriction on the genotypic means. It has power for a wider range of underlying models than does the model-based method, and it is more specific than the statistic Λ_{Unc}. It seems to be an appealing alternative method for cases in which specification of an analysis model is inappropriate—for instance, in studies in which numerous phenotypes are analyzed simultaneously, such as expression QTL mapping (Brem et al. 2002; Schadt et al. 2003; Yvert et al. 2003; Morley et al. 2004).

This approach is related to order-restricted inference methods (Robertson et al. 1988). Order-constrained inference methods are suitable for situations in which the allele being tested for association is expected to increase or decrease the trait value—for instance, gene-mapping association studies of knockout mice. In comparison, the proposed approach does not require any ordering of the mean genotypic effects. What it requires is that the mean effect of the heterozygous genotype not exceed those of the two homozygous genotypes. In other words, if one copy of the allele being tested has an increasing or decreasing effect on the trait, then having an additional copy of this allele will not weaken this trend, regardless of its direction. That is, the allele under investigation shows no overdominance effect.

This approach provides a single-marker test. There are some recent findings that single-marker tests can be used to construct inferential procedures for haplotype analysis that are more powerful than some commonly used haplotype analysis methods (Roeder et al. 2005; Schaid et al. 2005). So, our approach has some interesting implications for haplotype analysis as well. For instance, Roeder et al. (2005) studied the performance of, among other inferential procedures, a statistic that is the largest single-marker test statistic over a set of markers. By permuting the affection status among cases and controls, it was found that this statistic is more powerful than a test proposed by Chapman et al. (2003). The single-marker test used by Roeder et al. (2005) is based on allele counts and requires HWE for it to be valid (Sasieni 1997). On the other hand, our proposed approach is based on genotypes and does not require HWE for its validity. It will be interesting to assess the performance of the procedures of Roeder et al. (2005) when the single-marker test is substituted with the test proposed here.

The constrained-likelihood approach presented here is conceptually simple. Such simplicity may be useful for its generalization to situations other than those considered here. For instance, one could include covariates in the analysis or consider markers with more than two alleles.

Acknowledgments

We thank Dr. Christopher Bartlett, Dr. Jian Huang, and three anonymous reviewers, for their useful comments. This work is supported in part by National Institutes of Health grant R01-EY-11298 (to V.C.S. and K.W.). V.C.S. is an Investigator of the Howard Hughes Medical Institute.

Appendix A

Asymptotic Distribution of Λ_{New}: Continuous Traits

It is straightforward to obtain that, under H_0, the Fisher information matrix for $(\delta_1,\delta_2,\alpha,\sigma^2)$ is

$$I_0 = \sigma^{-2}\begin{bmatrix} 1-p_0 & p_2 & 1-p_0 & 0 \\ p_2 & p_2 & p_2 & 0 \\ 1-p_0 & p_2 & 1 & 0 \\ 0 & 0 & 0 & (2\sigma^2)^{-1} \end{bmatrix}.$$

According to theorem 16.7 of van der Vaart (1998), when the null hypothesis is true, the likelihood-ratio statistic Λ_{New} converges to

$$\Gamma := \inf_{h\in\Theta_0}(X-h)^t I_0(X-h) - \inf_{h\in\Theta_1-\Theta_0}(X-h)^t I_0(X-h)$$

as $n\to\infty$, where X is normally distributed with mean vector 0 and covariance matrix I_0^{-1}.

Let $\tilde{I}_0$ be the partial variance matrix of the first two components of X, given the latter two components. That is,

$$\begin{aligned}\tilde{I}_0 &= \sigma^{-2}\left(\begin{bmatrix}1-p_0 & p_2\\ p_2 & p_2\end{bmatrix}-\begin{bmatrix}1-p_0 & 0\\ p_2 & 0\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & (2\sigma^2)^{-1}\end{bmatrix}^{-1}\begin{bmatrix}1-p_0 & p_2\\ 0 & 0\end{bmatrix}\right)\\ &= \sigma^{-2}\begin{bmatrix}p_0(1-p_0) & p_0p_2\\ p_0p_2 & p_2(1-p_2)\end{bmatrix}.\end{aligned}$$

After some matrix calculation, it can be shown that Γ can be written as

$$\Gamma: = \inf_{\tilde{h}\in\tilde{\Theta}_0}(\tilde{X}-\tilde{h})^t\tilde{I}_0(\tilde{X}-\tilde{h}) - \inf_{\tilde{h}\in\tilde{\Theta}_1-\tilde{\Theta}_0}(\tilde{X}-\tilde{h})^t\tilde{I}_0(\tilde{X}-\tilde{h}),$$

where $\tilde{X}$ is a bivariate normal random vector with mean vector 0 and variance matrix $\tilde{I}_0^{-1}$, $\tilde{\Theta}_0 = \{(\delta_1,\delta_2):\delta_1 = \delta_2 = 0\}$, and $\tilde{\Theta}_1 = \{(\delta_1,\delta_2):\delta_1\delta_2 \geqslant 0\}$. For Γ to have 2 df, the two components of $\tilde{X}$ must be either both strictly positive or both strictly negative. Since the correlation coefficient between these two components is $\gamma = (p_0p_2)^{1/2}[(1-p_0)(1-p_2)]^{-1/2}$, the probability that the two components are both strictly positive or both strictly negative is $2 \times (2\pi)^{-1}\arccos(\gamma) = 2\kappa$. When both components of $\tilde{X}$ are strictly positive or negative, the constraint $\delta_1\delta_2 \geqslant 0$ is not binding and Γ has 2 df.

The probability that Γ does not have 2 df is $1-2\kappa$. In this case, we have $\Lambda_{\text{New}} = \Lambda_{\text{Larger}} = \max\{\Lambda_{\text{Dom}},\Lambda_{\text{Rec}}\}$. According to standard asymptotic theory (e.g., that of Cox and Hinkley [1974]), Λ_{Dom} is asymptotically equivalent to the score statistic T^2_{Dom}, where

$$T_{\text{Dom}} = [n_0(n-n_0)]^{1/2}(n\hat{\sigma}^2)^{-1/2}[(n_1\bar{y}_1 + n_2\bar{y}_2)/(n_1+n_2) - \bar{y}_0],$$

and Λ_{Rec} is asymptotically equivalent to the score statistic T^2_{Rec}, where

$$T_{\text{Rec}} = [n_2(n-n_2)]^{1/2}(n\hat{\sigma}^2)^{-1/2}[\bar{y}_2 - (n_0\bar{y}_0 + n_1\bar{y}_1)/(n_0+n_1)].$$

The maximum-likelihood estimate of σ^2 under the null hypothesis is

$$\hat{\sigma}^2 = n^{-1}\sum_{j=0}^{2}\sum_{i=1}^{n_j}(y_{ji}-\bar{y})^2,$$

where $\bar{y} = (n_0\bar{y}_1 + n_1\bar{y}_1 + n_2\bar{y}_2)/n$ is the grand mean of the trait. The correlation coefficient between T_{Dom} and T_{Rec} is $(n_0n_2)^{1/2}[(n-n_0)(n-n_2)]^{-1/2}$, which converges to $\gamma = (p_0p_2)^{1/2}[(1-p_0)(1-p_2)]^{-1/2}$ as $n\to\infty$. Asymptotically, T_{Dom} and T_{Rec} jointly follow bivariate normal distribution $(z_1,z_2)^t$, whose correlation coefficient is γ and for which both z_1 and z_2 follow the standard normal distribution. Hence, for any $x \geqslant 0$, we have

$$\begin{aligned}\Pr(\Lambda_{\text{Larger}} \leqslant x) &= \Pr(\Lambda_{\text{Dom}} \leqslant x, \Lambda_{\text{Rec}} \leqslant x)\\ &\to \Pr(z_1^2 \leqslant x, z_2^2 \leqslant x)\\ &= \Pr(-x^{1/2} \leqslant z_1 \leqslant x^{1/2}, -x^{1/2} \leqslant z_2 \leqslant x^{1/2}).\end{aligned}$$

Appendix B

Asymptotic Distribution of Λ_{New}: Dichotomous Traits

The arguments for dichotomous traits are parallel to those for continuous traits given in appendix A. For dichotomous traits, the Fisher information matrix I_0 becomes

$$f_0(1-f_0)\begin{bmatrix} 1-p_0 & p_2 & 1-p_0 \\ p_2 & p_2 & p_2 \\ 1-p_0 & p_2 & 1 \end{bmatrix},$$

and the matrix $\tilde{I}_0$ becomes

$$f_0(1-f_0)\begin{bmatrix} p_0(1-p_0) & p_0p_2 \\ p_0p_2 & p_2(1-p_2) \end{bmatrix}.$$

Both matrices are proportional to those in appendix A. So, the probability that Λ_{New} has 2 df remains the same.

The calculation of Λ_{Larger} is also parallel. The expressions for T_{Dom} and T_{Rec} remain the same, but the expression for $\hat{\sigma}^2$ is different. In the situation of dichotomous traits, we need to substitute $\hat{\sigma}^2$ with $\bar{y}(1-\bar{y})$, where $\bar{y}$ is the proportion of cases among the total number of subjects. This change in $\hat{\sigma}^2$ does not affect the asymptotic distribution of Λ_{Larger} or, hence, that of Λ_{New}.

Web Resources

The URL for data presented herein is as follows:

K.W.'s Web site, http://arctica.public-health.uiowa.edu/research.html (for a publicly available R program that implements the method described in this article)

References

Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. Science 296:752–755

Chapman JM, Cooper JD, Todd JA, Clayton DG (2003) Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. Hum Hered 56:18–31

Cox DR, Hinkley DV (1974) Theoretical statistics. Chapman and Hall, London

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th ed. Prentice Hall, London

Freidlin B, Zheng G, Li Z, Gastwirth J (2002) Trend tests for case-control studies of genetic markers: power, sample size and robustness. Hum Hered 53:146–152

——— (2003) Efficiency robust tests for mapping quantitative trait loci using extremely discordant sib pairs. Hum Hered 55:117–124

Holmans P (1993) Asymptotic properties of affected–sib-pair linkage analysis. Am J Hum Genet 52:362–374

Knapp M (1998) Evaluation of a restricted likelihood ratio test for mapping quantitative trait loci with extreme discordant sib pairs. Ann Hum Genet 62:75–87

Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman R, Cheung VG (2004) Genetic analysis of genome-wide variation in human gene expression. Nature 430:733–734

Prentice RL, Pyke R (1979) Logistic disease incidence models and case-control studies. Biometrika 66:403–411

Robertson T, Wright FT, Dykstra RL (1988) Order restricted statistical inference. John Wiley & Sons, New York

Roeder K, Bacanu SA, Zhao X, Devlin B (2005) Analysis of single-locus tests to detect gene/disease associations. Genet Epidemiol 28:207–219

Sasieni PD (1997) From genotypes to genes: doubling the sample size. Biometrics 53:1253–1261

Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, Linsley PS, Mao M, Stoughton RB, Friend SH (2003) Genetics of gene expression surveyed in maize, mouse and man. Nature 422: 297–302

Schaid DJ, McDonnell SK, Hebbring SJ, Cunningham JM, Thibodeau SN (2005) Nonparametric tests of association of multiple genes with human disease. Am J Hum Genet 76: 780–793

Slager S, Schaid D (2001) Case-control studies of genetic markers: power and sample size approximations for Armitage's test for trend. Hum Hered 52:149–153

van der Vaart AW (1998) Asymptotic statistics. Cambridge University Press, Cambridge

Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, Mackelprang R, Kruglyak L (2003) *Trans*-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. Nat Genet 35:57–64

Zheng G (2003) Use of max and min scores for trend tests for association when the genetic model is unknown. Stat Med 22:2657–2666

Zheng G, Chen Z, Li Z (2003) Tests for candidate-gene association using case-parents design. Ann Hum Genet 67:589–597