

Exploiting Hardy-Weinberg Equilibrium for Efficient Screening of Single SNP Associations from Case-Control Studies

Jinbo Chen^a Nilanjan Chatterjee^b

^aDepartment of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, Pa., and ^bBiostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Md., USA

Key Words

Association test · Case-control study · Genome scan · Hardy-Weinberg equilibrium · Retrospective likelihood

Abstract

In case-control studies, the assessment of the association between a binary disease outcome and a single nucleotide polymorphism (SNP) is often based on comparing the observed genotype distribution for the cases against that for the controls. In this article, we investigate an alternative analytic strategy in which the observed genotype frequencies of cases are compared against the expected genotype frequencies of controls assuming Hardy-Weinberg Equilibrium (HWE). Assuming HWE for controls, we derive closed-form expressions for maximum likelihood estimates of the genotype-specific disease odds ratio (OR) parameters and related variance-covariances. Based on these estimates and their variance-covariance structure, we then propose a two-degree-of-freedom test for disease-SNP association. We show that the proposed test can have substantially higher power than a variety of existing methods, especially when the true effect of the SNP is recessive. We also obtain analytic expressions for the bias of the OR estimates when the underlying HWE assumption is violated. We conclude that the novel test would be particularly useful for analyzing data from the initial 'screening' stages of contemporary multi-stage association studies.

Copyright © 2007 S. Karger AG, Basel

Introduction

The evaluation of the association between a disease trait and individual single nucleotide polymorphisms (SNPs) often constitutes the initial step of analysis for association studies. The lack of statistical significance in this first step may lead to the exclusion of a SNP from further scrutiny. Thus, to reduce the chance of false negatives, it is important to use powerful methods for preliminary screening of associations. A variety of methods are available for testing the association of a binary disease outcome with the three-category genotype variable at a bi-allelic locus based on case-control data. A widely used method is the Cochran-Armitage (CA) one-degree-of-freedom test of trend [1–4], which is known to be optimal when the mode-of-effect for a SNP is multiplicative. Alternatively, one can use a two-degree-of-freedom χ^2 test for independence in the 2×3 contingency table defined by the cross-classification of subjects by their case-control and genotype status. The advantage of the latter method is that it does not rely on any model assumption for the 'mode-of-effect' of the SNP.

In this article, we propose to enhance the power of the two-degree-of-freedom test of association by exploiting an assumption of Hardy-Weinberg Equilibrium (HWE) for the controls. The central idea is that if the underlying population is expected to be in HWE and the disease is rare, then the genotype distribution of the controls is ex-

pected to follow the Hardy-Weinberg constraints. We show how to exploit these constraints to obtain highly efficient maximum likelihood estimates (MLEs) of the genotype-specific disease odds ratio (OR) parameters. Based on these estimates and their variance-covariance structure, we then propose a two degree-of-freedom Wald statistic for testing the global null hypothesis of no association of the disease with any of the genotypes. We compare the power of the proposed method against a variety of alternatives including the standard two-degree-of-freedom test, the CA trend test, and a recent test proposed by Song and Elston [6].

Material and Methods

The Standard Two-Degree-of-Freedom Test

Table 1 shows the general form of a 2×3 contingency table that represents the genotype frequencies for a bi-allelic locus in a case-control study of n_{1+} cases ($D = 1$) and n_{0+} controls ($D = 0$). The major and the minor alleles are denoted by A and a , respectively, and the corresponding genotype values are AA , Aa and aa . Without assuming any further model, the maximum likelihood estimates of the OR parameters for genotypes Aa and aa , say denoted by ψ_{Aa} and ψ_{aa} , both defined in reference to the baseline genotype AA , can be obtained by the cross-ratios:

$$\hat{\psi}_{Aa}^{\circ} = \frac{n_{11}n_{00}}{n_{01}n_{10}} \quad \text{and} \quad \hat{\psi}_{aa}^{\circ} = \frac{n_{12}n_{00}}{n_{02}n_{10}}.$$

The variance-covariance matrix of the logarithm of the estimated OR parameters, $(\hat{\beta}_{Aa}^{\circ}, \hat{\beta}_{aa}^{\circ}) \equiv (\log \hat{\psi}_{Aa}^{\circ}, \log \hat{\psi}_{aa}^{\circ})$, is given by the well known formula [7]

$$\hat{V}_{\hat{\beta}_{Aa}^{\circ}, \hat{\beta}_{aa}^{\circ}} = \begin{bmatrix} \frac{1}{n_{11}} + \frac{1}{n_{00}} + \frac{1}{n_{01}} + \frac{1}{n_{10}} & \frac{1}{n_{00}} + \frac{1}{n_{10}} \\ \frac{1}{n_{00}} + \frac{1}{n_{10}} & \frac{1}{n_{12}} + \frac{1}{n_{00}} + \frac{1}{n_{02}} + \frac{1}{n_{10}} \end{bmatrix}.$$

A test for the global null hypothesis that none of the genotypes are associated with the disease risk, i.e. $\psi_{Aa} = \psi_{aa} = 1$ or equivalently $\beta_{Aa} = \beta_{aa} = 0$, can be constructed based on the Wald statistic

$$W_{\beta_{Aa}\beta_{aa}}^{\circ} = (\hat{\beta}_{Aa}^{\circ}, \hat{\beta}_{aa}^{\circ}) \left(\hat{V}_{\hat{\beta}_{Aa}^{\circ}, \hat{\beta}_{aa}^{\circ}} \right)^{-1} (\hat{\beta}_{Aa}^{\circ}, \hat{\beta}_{aa}^{\circ})^T,$$

where the superscript T denotes the matrix transpose. Asymptotically, $W_{\beta_{Aa}\beta_{aa}}^{\circ}$ is distributed as a χ^2 variable with two degrees of freedom under the null hypothesis of no association.

The Two-Degree-of-Freedom Test Exploiting HWE for Controls

The null hypothesis of $\beta_{Aa} = \beta_{aa} = 0$ corresponds to the equality of genotype distribution for the cases and controls in the underlying population. The test statistic $W_{\beta_{Aa}\beta_{aa}}^{\circ}$ attempts to detect difference between these two distributions by comparing the observed genotype counts for the cases against those of the controls. If, however, the distribution of the genotypes for the controls is

Table 1. Observed counts

	D = 0	D = 1	Total
G = AA	n_{00}	n_{10}	n_{+0}
G = Aa	n_{01}	n_{11}	n_{+1}
G = aa	n_{02}	n_{12}	n_{+2}
Total	n_{0+}	n_{1+}	n_{++}

Table 2. Expected counts under HWE in controls

	D = 0	D = 1	Total
G = AA	$\hat{n}_{00} = n_{0+}(1 - \hat{f})^2$	n_{10}	\hat{n}_{+0}
G = Aa	$\hat{n}_{01} = 2n_{0+}\hat{f}(1 - \hat{f})$	n_{11}	\hat{n}_{+1}
G = aa	$\hat{n}_{02} = n_{0+}\hat{f}^2$	n_{12}	\hat{n}_{+2}
Total	n_{0+}	n_{1+}	n_{++}

expected to follow the HW constraints, then a more efficient test for the equality of the two distributions can be obtained by comparing the observed counts among the cases against the expected counts among the controls under HWE. Based on the frequencies shown in table 2, the minor allele frequency (MAF) for the controls can be estimated as $\hat{f} = (2n_{02} + n_{01})/2n_{0+}$. The expected genotype counts for the controls are then given by $n_{00}^E = n_{0+}(1 - \hat{f})^2$, $n_{01}^E = 2n_{0+}\hat{f}(1 - \hat{f})$, and $n_{02}^E = n_{0+}\hat{f}^2$. The odds-ratio parameters ψ_{Aa} and ψ_{aa} can then be estimated based on the expected counts as

$$\hat{\psi}_{Aa}^E = \frac{n_{11}n_{00}^E}{n_{01}^En_{10}} \quad \text{and} \quad \hat{\psi}_{aa}^E = \frac{n_{12}n_{00}^E}{n_{02}^En_{10}}.$$

The asymptotic variance-covariance matrix of the logarithm of the OR estimates, $(\hat{\beta}_{Aa}^E, \hat{\beta}_{aa}^E) \equiv (\log \hat{\psi}_{Aa}^E, \log \hat{\psi}_{aa}^E)$ can be estimated by (details are shown in the Appendix):

$$\hat{V}_{\hat{\beta}_{Aa}^E, \hat{\beta}_{aa}^E} = \begin{bmatrix} \frac{1}{n_{10}} + \frac{1}{n_{11}} + \frac{1}{2n_{00}^E + n_{01}^E} + \frac{1}{2n_{02}^E + n_{01}^E} & \frac{1}{n_{10}} + \frac{1}{n_{0+}} \frac{1}{\hat{f}(1 - \hat{f})} \\ \frac{1}{n_{10}} + \frac{1}{n_{0+}} \frac{1}{\hat{f}(1 - \hat{f})} & \frac{1}{n_{10}} + \frac{1}{n_{12}} + \frac{4}{2n_{02}^E + n_{01}^E} + \frac{4}{2n_{00}^E + n_{01}^E} \end{bmatrix}.$$

The fact that $(\hat{\beta}_{Aa}^E, \hat{\beta}_{aa}^E)$ is more precise than $(\hat{\beta}_{Aa}^{\circ}, \hat{\beta}_{aa}^{\circ})$ can be established by noting that $\hat{V}_{\hat{\beta}_{Aa}^E, \hat{\beta}_{aa}^E} - \hat{V}_{\hat{\beta}_{Aa}^{\circ}, \hat{\beta}_{aa}^{\circ}}$ is asymptotically negative definite.

In the Appendix, we show that the estimates $\hat{\psi}_{Aa}^E$ and $\hat{\psi}_{aa}^E$ maximize the retrospective likelihood [8] for case-control data under the assumption of HWE for controls. We propose use of the Wald statistic, $W_{\beta_{Aa}\beta_{aa}}^E = (\hat{\beta}_{Aa}^E, \hat{\beta}_{aa}^E) \left(\hat{V}_{\hat{\beta}_{Aa}^E, \hat{\beta}_{aa}^E} \right)^{-1} (\hat{\beta}_{Aa}^E, \hat{\beta}_{aa}^E)^T$, for testing the null hypothesis $\beta_{Aa} = \beta_{aa} = 0$. Asymptotically, $W_{\beta_{Aa}\beta_{aa}}^E$ is distributed as central χ^2 distribution with two degrees of freedom under the null hypothesis of no association. Under the alternative hypothesis, it follows a noncentral χ^2 with a non-centrality pa-

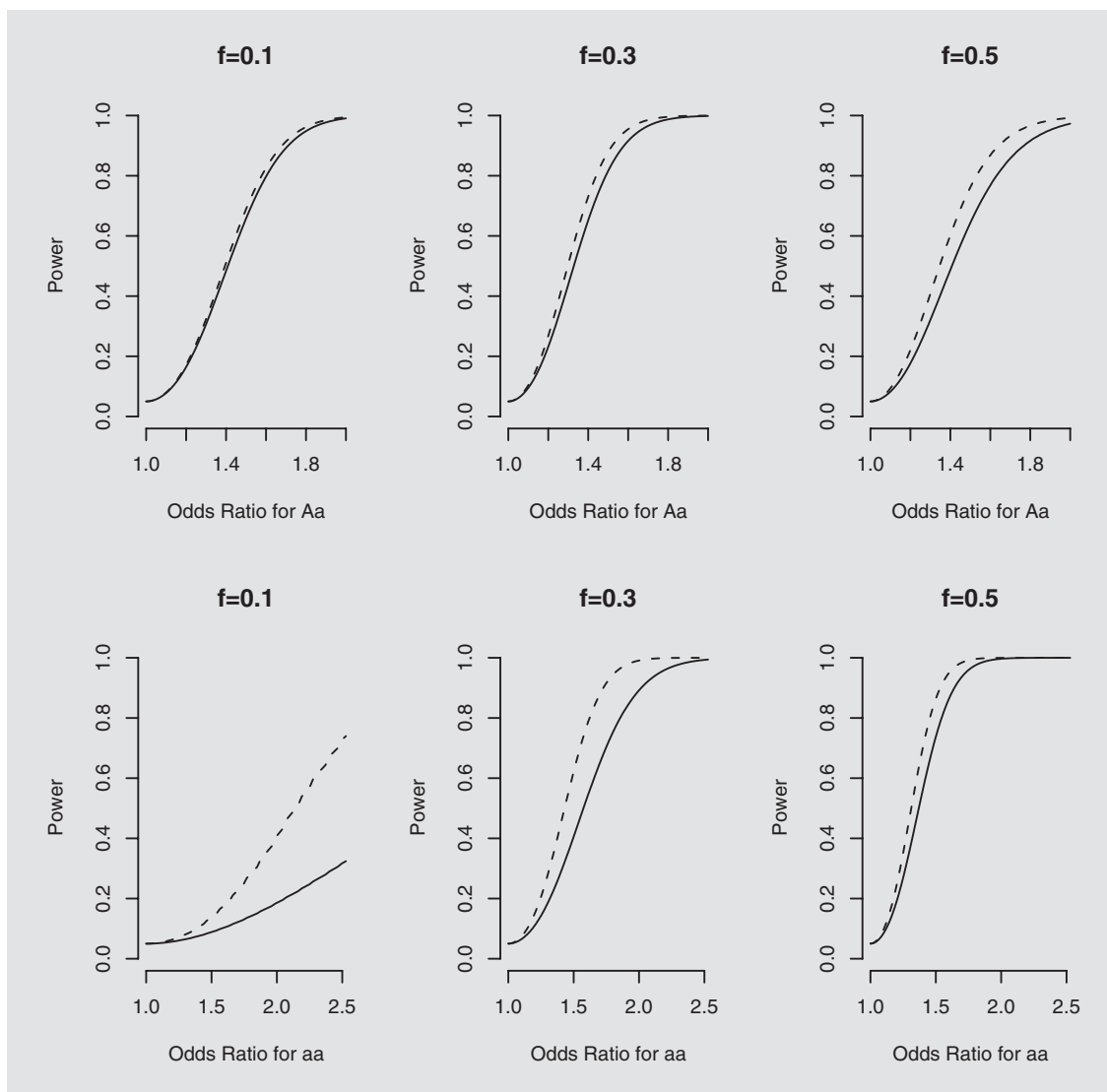


Fig. 1. Power comparison between the proposed and the standard 2 d.f. tests. The upper panels are for the dominant model, and the lower panels are for the recessive model. The solid line corresponds to the standard method, and the dashed line corresponds to the new method. The size of the test was set to be 0.05. 500 cases and controls are used.

parameter being the limit of the statistic itself. The power of the test can be easily computed from standard tables. Because

$$\hat{V}_{\hat{\beta}_{Aa}^o, \hat{\beta}_{aa}^o} - \hat{V}_{\hat{\beta}_{Aa}^E, \hat{\beta}_{aa}^E}$$

is positive definite, and because both $(\hat{\beta}_{Aa}^o, \hat{\beta}_{aa}^o)$ and $(\hat{\beta}_{Aa}^E, \hat{\beta}_{aa}^E)$ converge to the limit (β_{Aa}, β_{aa}) , the non-centrality parameter $W_{\hat{\beta}_{Aa}^o, \hat{\beta}_{aa}^o}^E$ is greater than that of the standard test $W_{\hat{\beta}_{Aa}^E, \hat{\beta}_{aa}^E}^o$, implying higher power of the novel test.

Performance Studies

We performed simulation studies to compare the power of the proposed test against a variety of alternatives. In these simulations, given the minor allele frequency f , the genotype frequencies

for the controls, $p(G = g | D = 0)$, were calculated according to the HWE law. Further, given the odds-ratio parameters $\psi_{AA} = 1$ (reference), ψ_{Aa} and ψ_{aa} , the genotype frequencies for the cases were obtained using the formula

$$p(G = g | D = 1) = \frac{\psi_g p(G = g | D = 0)}{\sum_{g \in \{AA, Aa, aa\}} \psi_g p(G = g | D = 0)}$$

The genotypes for the cases and the controls were then generated from the respective multinomial distributions.

Figure 1 compares the power of the standard and the proposed χ^2 two-degree-of-freedom tests at a significance level of 0.05 for different combinations of minor allele frequencies, OR parameters

Table 3. Number of cases required (assuming equal number of controls) by the standard and new tests to achieve 80% power

	OR	Minor allele frequency in controls							
		0.1		0.2		0.3		0.4	
		standard	new	standard	new	standard	new	standard	new
Multiplicative	1.2	2,853	2,853	1,726	1,726	1,324	1,324	1,145	1,145
	1.4	826	825	482	482	372	372	343	343
	1.6	400	400	235	235	190	190	175	175
	1.8	243	243	147	147	118	118	109	109
	2.0	169	169	103	103	84	84	78	78
Dominant	1.2	3,435	3,410	2,425	2,213	2,337	1,997	2,536	2,012
	1.4	997	956	693	631	679	583	778	616
	1.6	487	467	352	322	349	298	403	319
	1.8	302	292	224	205	226	193	269	211
	2.0	210	203	158	145	165	141	196	154
Recessive	1.2	– ^a	–	–	7,855	6,750	4,475	4,026	2,874
	1.4	–	8,734	3,825	2,523	1,758	1,218	1,102	804
	1.6	6,783	4,433	1,900	1,284	863	612	546	404
	1.8	4,102	2,794	1,095	769	538	390	343	256
	2.0	2,809	1,956	762	549	372	273	237	178

^a Sample size is larger than 2×10^4 .

ters, and different modes of effects for the SNP. The gain in efficiency for the new test is prominent under the recessive model ($\psi_{AA} = \psi_{Aa} = 1$), with the power differences generally larger for larger OR values and smaller minor allele frequencies. Some modest gain in power is also observed under the dominant model ($\psi_{Aa} = \psi_{aa}$). In contrast, under the multiplicative model ($\psi_{Aa} = \psi_{AA}^2$), no difference in power existed between the two tests (data not shown).

Table 3 presents the number of cases required (assuming equal number of controls) to achieve 80% power using the new and the standard two degree-of-freedom tests. In agreement with figure 1, the sample size required when the new approach is adopted can be much smaller under the recessive disease risk model. For example, when the minor allele frequency for the controls was 0.2 and $\psi_{Aa} = 1.6$, the number of cases required by the standard and the proposed tests were 1,900 and 1,284, respectively. The new method could also lead to substantial decrease in the sample size under the dominant model when the minor allele frequency is large.

Recently, Song and Elston [6] proposed a novel test of association that combines the CA-trend statistic and the HWE-trend statistic (HWE-Trend) comparing the Hardy-Weinberg disequilibrium (HWD) coefficients between the cases and the controls. They defined the HWD coefficient to be the difference between the observed and the expected frequencies under HWE for the homozygous variant genotype *aa*. They showed that the combined test was more powerful than the CA-trend or the HWE-trend test alone. Table 4 compares the power of five different tests, including the proposed method and the method of Song and Elston, in a setting that involves a sample size of 500 cases and 500 controls, $\psi_{Aa} = 1.4$ under the multiplicative and dominant models or $\psi_{Aa} = 1.96$ under the recessive model. The critical value for the

test-statistic proposed by Song and Elston [6] was obtained by permutation-based resampling. It appears that the proposed test generally outperforms all the other methods under the dominant and recessive models. Under the multiplicative model, for which the CA-trend test is known to be the most powerful, the proposed test remains a close competitor and loses only 4–8% of power.

Following the suggestion of an anonymous reviewer, next we studied the performance of the proposed test compared to a novel case-only test of association that assumes *f*, the MAF among the control population, is known a priori. We observe that when *f* is known, then assuming HWE in controls, one can estimate the odds-ratio parameters ψ_{Aa} and ψ_{aa} as

$$\hat{\psi}_{Aa}^f = \frac{n_{11}(1-f)^2}{n_{10}[2f(1-f)]} \quad \text{and} \quad \hat{\psi}_{aa}^f = \frac{n_{12}(1-f)^2}{n_{10}f^2},$$

with an asymptotic variance-covariance matrix given by

$$\hat{V}_f = \begin{bmatrix} \frac{1}{n_{11}} + \frac{1}{n_{10}} & \frac{1}{n_{10}} \\ \frac{1}{n_{10}} & \frac{1}{n_{12}} + \frac{1}{n_{10}} \end{bmatrix}.$$

Table 5 compares the power of the corresponding 2 d.f. ‘case only’ test with those of the standard and the proposed 2 d.f. ‘case-control’ tests. We considered different sampling control-to-case ratios, assuming 500 cases and $\psi_{Aa} = 1.4$ for multiplicative and dominant models and $\psi_{Aa} = 1.96$ for the recessive model. Clearly, if *f* is known, then the case-only test can be far superior than either of the two case-control tests when the disease-risk follows the multiplicative or the dominant model. In contrast, under a recessive

Table 4. Size and power of the proposed test with 500 cases and 500 controls

Model ^a	<i>f</i>	New ^b	Standard	CA-trend ^c	HWE-trend ^d	Song & Elston ^e
Under the null hypothesis $\psi_{Aa} = \psi_{aa} = 1$						
Multiplicative	0.1	0.04	0.05	0.04	0.04	0.05
	0.2	0.04	0.04	0.06	0.03	0.04
	0.3	0.05	0.06	0.04	0.06	0.04
	0.4	0.04	0.05	0.06	0.05	0.05
Dominant	0.1	0.03	0.04	0.04	0.06	0.05
	0.2	0.06	0.05	0.05	0.05	0.06
	0.3	0.05	0.04	0.04	0.05	0.06
	0.4	0.06	0.06	0.07	0.05	0.06
Recessive	0.1	0.02	0.04	0.04	0.05	0.05
	0.2	0.04	0.04	0.05	0.04	0.05
	0.3	0.04	0.04	0.04	0.04	0.05
	0.4	0.04	0.05	0.05	0.05	0.06
Under the alternative hypothesis						
Multiplicative ($\psi_{Aa} = 1.4$)	0.1	59.0	56.0	67.0	4.0	66.0
	0.2	82.0	82.0	88.0	6.0	86.0
	0.3	91.0	91.0	95.0	7.0	94.0
	0.4	92.0	92.0	96.0	5.0	95.0
Dominant ($\psi_{Aa} = 1.4$)	0.1	51.0	47.0	56.0	11.0	54.0
	0.2	71.0	64.0	69.0	20.0	65.0
	0.3	73.0	64.0	65.0	23.0	63.0
	0.4	68.0	58.0	56.0	25.0	53.0
Recessive ($\psi_{aa} = 1.96$)	0.1	34.0	10.0	7.0	18.0	7.0
	0.2	84.0	58.0	34.0	56.0	38.0
	0.3	98.0	89.0	72.0	74.0	76.0
	0.4	100.0	98.0	93.0	77.0	94.0

^a Model for the SNP-disease association; ^b new test; ^c Cochran-Armitage test of trend; ^d Hardy-Weinberg equilibrium trend test [7]; ^e test by a weighted average of CA-trend and HWD-trend [7].

sive model, the power of the case-only approach was only modestly higher than the proposed case-control test assuming HWE. As the control-to-case ratio decreased, the power for both the case-control tests, as expected, decreased rapidly under both the multiplicative and the dominant models. In contrast, under the recessive model, decreasing the sampling ratio for the controls had very little impact on the power for case-control test assuming HWE, although the impact gets slightly larger with larger MAF.

Discussion

The proposed method can be easily extended if one wishes to adjust for covariate strata, such as those defined by ethnicity. In this setting, the data can be represented by a series of stratum-specific two-by-three contingency tables. The MLE estimates of the OR parameters for each stratum can be easily obtained in closed form using the proposed method, assuming that HWE holds in controls within each stratum. The stratum-specific ORs then can

be combined over different strata using a method similar to that of Mantel and Haenszel [9]. Details of this estimation method and the corresponding two-degree-of-freedom χ^2 test are provided in the Appendix.

Our work is closely related to that of Epstein and Satten [8], Satten and Epstein [10], and Thompson et al. [11], all of whom described likelihood-based genetic association tests by exploiting the HWE assumption. Thompson et al. assumed HWE holds in the general population from which the cases and the controls have been sampled. This assumption, although more natural, can complicate the computation of maximum likelihood estimates unless the marginal disease prevalence in the underlying population is known [12, 13]. Epstein and Satten [8] and Satten and Epstein [10] focused on haplotype-based analysis of multi-locus genetic data assuming HWE holds in the control population. Our calculations show that for analysis of the data on a single SNP, the retrospective maximum-likelihood estimator proposed by Satten and Ep-

Table 5. The power of the proposed test for different control-to-case ratio assuming OR^a = 1.4 and 300 cases

MAF (f)	Ratio ^b	Models					
		Multiplicative		Dominant		Recessive	
0.1	1.00	0.71 ^c	0.37 ^d	0.58 ^c	0.29 ^d	0.30 ^c	0.27 ^d
	0.75	0.71	0.32	0.61	0.27	0.29	0.24
	0.50	0.68	0.22	0.62	0.19	0.30	0.23
	0.25	0.71	0.09	0.62	0.10	0.32	0.23
0.2	1.00	0.90	0.57	0.74	0.46	0.75	0.64
	0.75	0.89	0.50	0.76	0.40	0.73	0.63
	0.50	0.89	0.42	0.76	0.34	0.73	0.58
	0.25	0.89	0.21	0.74	0.25	0.74	0.57
0.3	1.00	0.95	0.67	0.72	0.49	0.96	0.88
	0.75	0.95	0.63	0.76	0.47	0.95	0.83
	0.50	0.96	0.51	0.73	0.39	0.94	0.80
	0.25	0.95	0.29	0.75	0.31	0.95	0.76
0.4	1.00	0.96	0.73	0.65	0.45	1.00	0.95
	0.75	0.95	0.66	0.67	0.45	0.99	0.93
	0.50	0.96	0.56	0.68	0.40	1.00	0.92
	0.25	0.96	0.35	0.63	0.33	1.00	0.86

^a OR = ψ_{Aa} under multiplicative/dominant model and OR = $\psi_{aa} = 1.96$ under the recessive model; ^b control-to-case ratio; ^c the 'case-only' test with known MAF f ; ^d the proposed test with f estimated from controls.

stein can be obtained in a simple closed form. Furthermore, in this article, we focused on a two-degree-of-freedom 'model-free' test of association that has robust power under unknown modes of genetic effects. Both Epstein and Satten [10] and Thompson et al. [11] studied the effect of HWE assumption on the power of tests assuming specific models for the genetic effect, such as dominant, recessive or multiplicative.

The closed-form expressions for $\hat{\psi}_{Aa}$ and $\hat{\psi}_{aa}$ also facilitate the characterization of their potential bias due to the violation of the HWE assumption. Consider a setting where there are more homozygous controls than expected under HWE. In this situation, the deviation of the genotype frequencies from HWE can be characterized by a fixation parameter ρ [14]. The genotype frequencies for the controls in terms of f and ρ can be described as $p(AA) = (1 - \rho)(1 - f)^2 + \rho(1 - f)$, $p(Aa) = 2(1 - \rho)f(1 - f)$ and $p(aa) = (1 - \rho)f^2 + \rho f$. The respective asymptotic biases of the OR estimates $\hat{\psi}_{Aa}^E$ and $\hat{\psi}_{aa}^E$ can be derived as

$$-\psi_{Aa} \left\{ 2 + \frac{p(Aa)}{p(AA)} \right\} \frac{\rho}{2} \quad \text{and} \quad \psi_{aa} \left\{ 1 - \frac{p(aa)}{p(AA)} \right\} \rho \left(\frac{1}{f} - 1 \right).$$

The magnitude of the bias thus depends on the true OR parameters ψ_{Aa} and ψ_{aa} , the minor allele frequency f , and

the fixation parameter ρ . It is important to note that even under the null hypothesis $\psi_{Aa} = \psi_{aa} = 1$, violation of HWE can cause the OR estimates to be biased. This in turn would lead to an inflated type-I error rate for the proposed testing procedure. For example, with a minor allele frequency of 0.2 and fixation parameter $\rho = 0.05$, the true significance level of the proposed test would be 0.15 in comparison with the nominal level of 0.05. A more detailed study of the bias and type-I error rate of the proposed method when HWE is violated is provided in figure 2 in the appendix.

To check the validity of the HWE assumption, one could conduct certain diagnostics using the control sample itself in a given study. Structural problems in a study, such as hidden population structure or genotyping error, are expected to give rise to violation of HWE across many different loci. Thus, evidence for such structural problems should be investigated using some global measures of HWE departures. For example, the measure of 'tail strength of p-values' defined by Taylor and Tibshirani [15] can be adopted for this purpose. One can also try to detect site-specific violation of HWE using standard tests for HW disequilibrium (HWD). For any SNP that shows marked HW disequilibrium, any evidence of association found based on the proposed method should be re-evalu-

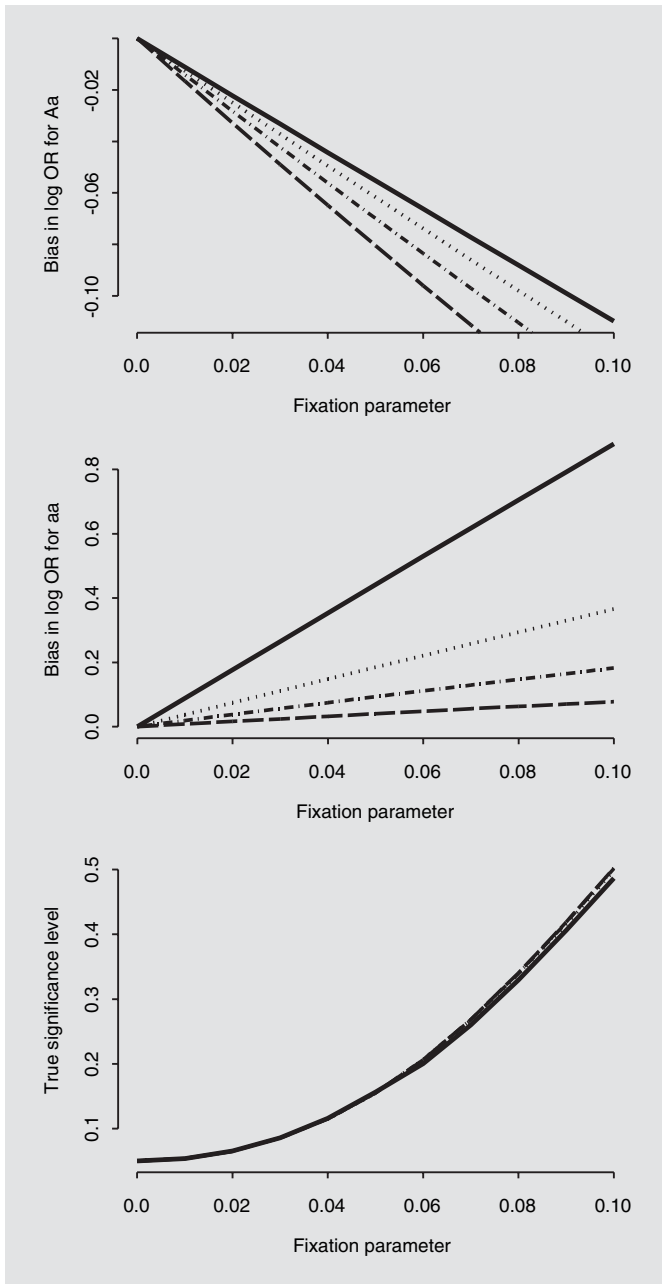


Fig. 2. Bias and type-I error rate for the proposed method as a function of minor allele frequencies and fixation parameters. Solid line corresponds to a minor allele frequency of 0.1, dotted line of 0.2, dash-dotted line of 0.3, and dashed line of 0.4. The size of the test was set to be 0.05.

ated using standard tests that do not require the assumption of HWE. For a study with modest sample size, however, powers for standard HWD tests to detect modest departures from HWE could be quite low. As a result, association tests assuming HWE could remain substantially

biased even if the corresponding HWD test is not statistically significant [16]. In light of this concern, we believe the proposed testing strategy would be most useful for analysis of data from initial ‘screening’ studies the goal of which is to identify promising SNPs that should be pursued with high priority in subsequent larger studies. Once larger samples are available on these promising SNPs, the confirmatory tests of association should be based on more robust methods that do not rely on the HWE assumption.

Appendix

To simplify notations, let $\psi_1 = \psi_{Aa}$ and $\psi_2 = \psi_{aa}$.

Proof that $\hat{\beta}_1$ and $\hat{\beta}_2$ are Maximum Likelihood Estimators
Write the genotypes distribution for the cases as

$$p(G = g | D = 1) = \frac{\psi_g p^0(g)}{p^0(AA) + \psi_1 p^0(Aa) + \psi_2 p^0(aa)},$$

where $p^0(g)$ is the probability of genotype $G = g$ in controls, and ψ_g is the odds ratio parameter for genotype g in reference to the baseline genotype AA . Under the HWE assumption, $p^0(AA) = (1 - f)^2$, $p^0(Aa) = 2f(1 - f)$, and $p^0(aa) = f^2$. The retrospective likelihood under the case-control sampling is $L(\psi_1, \psi_2, f) = \prod_i p(g_i | D = 1)^{D_i} p(g_i | D = 0)^{1 - D_i}$. The score functions for ψ_1 , ψ_2 , and f , obtained by taking the derivatives of $\log L$ with respect to each of these three parameters, are given by

$$U_{\psi_1} = \frac{\partial \log L}{\partial \psi_1} = \frac{n_{11}}{\psi_1} - \frac{n_{1+} 2f(1-f)}{(1-f)^2 + 2f(1-f)\psi_1 + f^2\psi_2}$$

$$U_{\psi_2} = \frac{\partial \log L}{\partial \psi_2} = \frac{n_{12}}{\psi_2} - \frac{n_{1+} f^2}{(1-f)^2 + 2f(1-f)\psi_1 + f^2\psi_2}$$

$$U_f = \frac{\partial \log L}{\partial f} = \frac{-(2n_{+0} + n_{+1})}{1-f} + \frac{2n_{+2} + n_{+1}}{f} - \frac{n_{1+} \{-2(1-f) + 2\psi_1(1-2f) + 2f\psi_2\}}{(1-f)^2 + 2f(1-f)\psi_1 + f^2\psi_2}.$$

By jointly solving the equations $U_{\psi_1} = 0$ and $U_{\psi_2} = 0$, we obtain $\psi_1 = \{n_{11}(1 - f)\}/(2n_{10}f)$ and $\psi_2 = \{n_{12}(1 - f)^2\}/(n_{10}f^2)$. Substituting them into the equation $U_f = 0$, we obtain $\hat{f} = (2n_{02} + n_{01})/2n_{0+}$.

The Asymptotic Distributions of $(\hat{\psi}_1, \hat{\psi}_2)$

The asymptotic distributions of $(\hat{\psi}_1, \hat{\psi}_2)$ can be obtained by applying the Delta method. We write

$$\hat{\psi}_1 = \frac{1}{2} \frac{n_{11}}{n_{1+}} \left(\frac{n_{10}}{n_{1+}} \right)^{-1} \frac{1 - \hat{f}}{\hat{f}} \quad \text{and} \quad \hat{\psi}_2 = \frac{n_{12}}{n_{1+}} \left(\frac{n_{10}}{n_{1+}} \right)^{-1} \left(\frac{1 - \hat{f}}{\hat{f}} \right)^2.$$

We note that n_{1+} and n_{0+} , the total numbers of cases and controls, are fixed by study design. We further observe that n_{10} , n_{11} , and n_{12} , the genotype frequencies for the cases, are independent of \hat{f} , which is obtained using data from only the controls. Let $p_{1g} = p(G = g | D = 1)$ and $\hat{p}_{1g} = n_{1g}/n_{1+}$. The variance-covariance matrix for \hat{p}_{11} , \hat{p}_{12} , and \hat{f} , denote by B , is given by

$$\begin{bmatrix} \frac{1}{n_{1+}} \frac{n_{11}}{n_{1+}} \left(1 - \frac{n_{11}}{n_{1+}}\right) & \frac{-1}{n_{1+}} \frac{n_{11}}{n_{1+}} \frac{n_{12}}{n_{1+}} & 0 \\ \frac{-1}{n_{1+}} \frac{n_{11}}{n_{1+}} \frac{n_{12}}{n_{1+}} & \frac{1}{n_{1+}} \frac{n_{12}}{n_{1+}} \left(1 - \frac{n_{12}}{n_{1+}}\right) & 0 \\ 0 & 0 & \frac{1}{2n_{0+}} \hat{f}(1-\hat{f}) \end{bmatrix}$$

Now $\hat{\beta}_1 = \log(\psi_1)$ and $\hat{\beta}_2 = \log(\psi_2)$ are fixed functions of \hat{p}_{11} , \hat{p}_{12} , and \hat{f} , with the corresponding gradient matrix given by

$$\begin{bmatrix} \frac{\partial \log \psi_1}{\partial p^{11}} & \frac{\partial \log \psi_1}{\partial p^{12}} & \frac{\partial \log \psi_1}{\partial f} \\ \frac{\partial \log \psi_2}{\partial p^{11}} & \frac{\partial \log \psi_2}{\partial p^{12}} & \frac{\partial \log \psi_2}{\partial f} \end{bmatrix}_{p^{11} = \frac{n_{11}}{n_{1+}}, p^{12} = \frac{n_{12}}{n_{2+}}, f = \hat{f}}$$

$$= \begin{bmatrix} \frac{n_{1+} + n_{1+}}{n_{10}} \frac{n_{1+}}{n_{11}} & \frac{n_{1+}}{n_{10}} & \frac{-1}{\hat{f}(1-\hat{f})} \\ \frac{n_{1+}}{n_{10}} & \frac{n_{1+} + n_{1+}}{n_{12}} & \frac{-2}{\hat{f}(1-\hat{f})} \end{bmatrix}$$

The desired variance-covariance matrix of $\hat{\beta}_1$ and $\hat{\beta}_2$ is then given by ABA^T , where T denotes 'matrix transpose'. Simple algebra leads to the variance-covariance matrix given in the text.

The Mantel-Haenszel Estimators and Their Asymptotic Properties

Table A1 and A2 below show the observed and expected genotype frequencies for the cases and the controls in the k th covariate stratum. Applying an idea similar to Mantel and Haenszel [9], we propose estimating the common (ψ_1, ψ_2)

Table A1. Observed counts

	D = 0	D = 1	Total
G = AA	n_{00}^k	n_{10}^k	n_{+0}^k
G = Aa	n_{01}^k	n_{11}^k	n_{+1}^k
G = aa	n_{02}^k	n_{12}^k	n_{+2}^k
Total	n_{0+}^k	n_{1+}^k	n_{++}^k

Table A2. Expected counts under HWE in controls

	D = 0	D = 1	Total
G = AA	$\hat{n}_{00}^k = n_{0+}(1-\hat{f}_k)^2$	\hat{n}_{10}^k	\hat{n}_{+0}^k
G = Aa	$\hat{n}_{01}^k = 2n_{0+}\hat{f}_k(1-\hat{f}_k)$	\hat{n}_{11}^k	\hat{n}_{+1}^k
G = aa	$\hat{n}_{02}^k = n_{0+}\hat{f}_k^2$	\hat{n}_{12}^k	\hat{n}_{+2}^k
Total	\hat{n}_{0+}^k	\hat{n}_{1+}^k	\hat{n}_{++}^k

underlying the K strata as

$$\hat{\psi}_1^m = \frac{\sum_k n_{11}^k \hat{n}_{00}^k / \hat{N}_{1+}^k}{\sum_k \hat{n}_{01}^k \hat{n}_{10}^k / \hat{N}_{1+}^k} \quad \text{and} \quad \hat{\psi}_2^m = \frac{\sum_k n_{12}^k \hat{n}_{00}^k / \hat{N}_{2+}^k}{\sum_k \hat{n}_{02}^k \hat{n}_{10}^k / \hat{N}_{2+}^k},$$

where $N_{1+}^k = \hat{n}_{00}^k + n_{10}^k = \hat{n}_{01}^k + n_{11}^k$ and $N_{2+}^k = \hat{n}_{00}^k + n_{10}^k + \hat{n}_{02}^k + n_{12}^k$. We assume that the total number of covariate strata K is fixed and small, and the number of subjects within each stratum is large. The asymptotic variance of $\log \hat{\psi}_1^m$ and $\log \hat{\psi}_2^m$ can be obtained via Delta method as

$$\sum_k \begin{bmatrix} \frac{\partial \log \hat{\psi}_1^m}{\partial p_{11}^k} & \frac{\partial \log \hat{\psi}_1^m}{\partial p_{12}^k} & \frac{\partial \log \hat{\psi}_1^m}{\partial f} \\ \frac{\partial \log \hat{\psi}_2^m}{\partial p_{11}^k} & \frac{\partial \log \hat{\psi}_2^m}{\partial p_{12}^k} & \frac{\partial \log \hat{\psi}_2^m}{\partial f} \end{bmatrix} A_k \begin{bmatrix} \frac{\partial \log \hat{\psi}_1^m}{\partial p_{11}^k} & \frac{\partial \log \hat{\psi}_1^m}{\partial p_{12}^k} & \frac{\partial \log \hat{\psi}_1^m}{\partial f} \\ \frac{\partial \log \hat{\psi}_2^m}{\partial p_{11}^k} & \frac{\partial \log \hat{\psi}_2^m}{\partial p_{12}^k} & \frac{\partial \log \hat{\psi}_2^m}{\partial f} \end{bmatrix}^T$$

where A_k is the variance-covariance matrix of \hat{p}_{11}^k , \hat{p}_{12}^k and \hat{f}^k given above. The relevant derivatives are as follows:

$$\begin{aligned} \frac{\partial \log \hat{\psi}_1^m}{\partial p_{11}^k} &= \frac{1}{\sum_k n_{11}^k \hat{n}_{00}^k / \hat{N}_{1+}^k} \frac{n_{1+}^k \hat{n}_{00}^k}{\hat{N}_{1+}^k} + \frac{1}{\sum_k \hat{n}_{01}^k \hat{n}_{10}^k / \hat{N}_{1+}^k} \frac{n_{0+}^k \hat{n}_{01}^k}{\hat{N}_{1+}^k} \\ \frac{\partial \log \hat{\psi}_1^m}{\partial p_{12}^k} &= \frac{1}{\sum_k n_{11}^k \hat{n}_{00}^k / \hat{N}_{1+}^k} \frac{n_{1+}^k \hat{n}_{00}^k}{(\hat{N}_{1+}^k)^2} + \frac{1}{\sum_k \hat{n}_{01}^k \hat{n}_{10}^k / \hat{N}_{1+}^k} \frac{n_{0+}^k \hat{n}_{01}^k (N_{1+}^k - n_{10}^k)}{(\hat{N}_{1+}^k)^2} \\ \frac{\partial \log \hat{\psi}_1^m}{\partial f^k} &= \frac{1}{\sum_k n_{11}^k \hat{n}_{00}^k / \hat{N}_{1+}^k} \frac{-2n_{11}^k n_{1+}^k (1-\hat{f}^k) \{ \hat{N}_{1+}^k - n_{0+}^k (1-\hat{f}^k) \hat{f}^k \}}{(\hat{N}_{1+}^k)^2} \\ &\quad - \frac{1}{\sum_k \hat{n}_{01}^k \hat{n}_{10}^k / \hat{N}_{1+}^k} \frac{2n_{0+}^k \hat{p}_{10}^k n_{0+}^k (1-2\hat{f}^k) \{ \hat{N}_{1+}^k + 4n_{0+}^k (\hat{f}^k)^2 (1-\hat{f}^k) \}}{(\hat{N}_{1+}^k)^2} \\ \frac{\partial \log \hat{\psi}_2^m}{\partial p_{11}^k} &= \frac{1}{\sum_k n_{12}^k \hat{n}_{00}^k / \hat{N}_{2+}^k} \frac{n_{12}^k \hat{n}_{00}^k}{(\hat{N}_{2+}^k)^2} + \frac{1}{\sum_k \hat{n}_{02}^k \hat{n}_{10}^k / \hat{N}_{2+}^k} \frac{n_{0+}^k \hat{n}_{02}^k (\hat{N}_{2+}^k - n_{10}^k)}{(\hat{N}_{2+}^k)^2} \\ \frac{\partial \log \hat{\psi}_2^m}{\partial p_{12}^k} &= \frac{1}{\sum_k n_{12}^k \hat{n}_{00}^k / \hat{N}_{2+}^k} \frac{n_{1+}^k \hat{n}_{00}^k}{\hat{N}_{2+}^k} + \frac{1}{\sum_k \hat{n}_{02}^k \hat{n}_{10}^k / \hat{N}_{2+}^k} \frac{n_{0+}^k \hat{n}_{02}^k}{\hat{N}_{2+}^k} \\ \frac{\partial \log \hat{\psi}_2^m}{\partial f^k} &= \frac{1}{\sum_k n_{12}^k \hat{n}_{00}^k / \hat{N}_{2+}^k} \frac{-2n_{12}^k n_{0+}^k (1-\hat{f}^k) \{ \hat{N}_{2+}^k - n_{0+}^k (1-\hat{f}^k) (1-2\hat{f}^k) \}}{(\hat{N}_{2+}^k)^2} \\ &\quad - \frac{1}{\sum_k \hat{n}_{02}^k \hat{n}_{10}^k / \hat{N}_{2+}^k} \frac{2n_{0+}^k \hat{p}_{10}^k \hat{f}_k \{ \hat{N}_{2+}^k + 2n_{0+}^k \hat{f}_k (1-2\hat{f}_k) \}}{(\hat{N}_{2+}^k)^2}. \end{aligned}$$

References

- 1 Armitage P: Tests for linear trends in proportions and frequencies. *Biometrics* 1955; 11:375–386.
- 2 Freidlin B, Zheng G, Li Z, Gastwirth, JL: Trend Tests for Case-Control Studies of Genetic Markers: Power, Sample Size and Robustness. *Hum Hered* 2002;53:146–152.
- 3 Sasieni PD: From genotypes to genes: doubling the sample size. *Biometrics* 1997;53: 1253–1261.
- 4 Slager SL, Schaid DJ: Evaluation of candidate genes in case-control studies: A statistical method to account for related subjects. *Am J Hum Genet* 2001;68:1457–1462.
- 5 Agresti A: *Categorical Data Analysis*. New York, John Wiley and Sons, 1990.
- 6 Song KA, Elston RC: A powerful method of combining measures of association and Hardy-Weinberg disequilibrium for fine-mapping in case-control studies. *Stat Med* 2006; 25:105–126.
- 7 Breslow NE, Day N: *Statistics of case-control studies*. New York, Marcel Dekker, 1984.
- 8 Epstein MP, Satten GA: Inference on haplotype effects in case-control studies using unphased genotype data. *Am J Hum Genet* 2003;73:1316–1329.
- 9 Mantel N, Haenszel W: Statistical aspects of the analysis of data from retrospective studies of diseases. *J Natn Cancer Inst* 1959;22: 719–748.
- 10 Satten GA, Epstein MP: Comparison of prospective and retrospective methods for haplotype inference in case-control studies. *Genet Epidemiol* 2004;27:192–201.
- 11 Thompson D, Witte JS, Slattery M, Goldgar D: Increased power for case-control studies of single nucleotide polymorphism through incorporation of family history and genetic constraints. *Genet Epidemiol* 2004;27:215–224.
- 12 Chatterjee N, Carroll RJ: Semiparametric maximum-likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika* 2005;92:399–418.
- 13 Spinka C, Carroll R, Chatterjee N: Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity. *Genet Epidemiol* 2005;29:108–127.
- 14 Hartl DL, Clark AG: *Principles of population genetics*. Sunderland, Sinauer Associates, 1997.
- 15 Taylor J, Tibshirani R: A tail strength measure for assessing the overall uni-variate significance in a dataset. *Biostatistics* 2006;7: 167–181.
- 16 Zou G, Donner A: The merits of testing Hardy-Weinberg Equilibrium in the analysis of unmatched case-control data: A cautionary note. *Ann Hum Genet* 2006; published online (<http://www.blackwell-synergy.com/doi/abs/10.1111/j.1469-1809.2006.00267.x>).