Genetic Epidemiology 33: 740–750 (2009)

Shrinkage Estimation for Robust and Efficient Screening of Single-SNP Association from Case-Control Genome-Wide Association Studies

Sheng Luo,[1] Bhramar Mukherjee,[2] Jinbo Chen,[3] and Nilanjan Chatterjee[4]*

[1]*Division of Biostatistics, University of Texas Health Science Center, Houston, Texas*
[2]*Department of Biostatistics, University of Michigan, Ann Arbor, Michigan*
[3]*Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, Pennsylvania*
[4]*Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS, Rockville, Maryland*

Population-based case-control design has become one of the most popular approaches for conducting genome-wide association scans for rare diseases like cancer. In this article, we propose a novel method for improving the power of the widely used single-single-nucleotide polymorphism (SNP) two-degrees-of-freedom (2 d.f.) association test for case-control studies by exploiting the common assumption of Hardy-Weinberg Equilibrium (HWE) for the underlying population. A key feature of the method is that it can relax the assumed model constraints via a completely data-adaptive shrinkage estimation approach so that the number of false-positive results due to the departure of HWE is controlled. The method is computationally simple and is easily scalable to association tests involving hundreds of thousands or millions of genetic markers. Simulation studies as well as an application involving data from a real genome-wide association study illustrate that the proposed method is very robust for large-scale association studies and can improve the power for detecting susceptibility SNPs with recessive effects, when compared to existing methods. Implications of the general estimation strategy beyond the simple 2 d.f. association test are discussed. *Genet. Epidemiol.* 33:740–750, 2009. Published 2009 Wiley-Liss, Inc.[†]

Key words: association test; case-control studies; genome scan; Hardy-Weinberg Equilibrium; retrospective likelihood

Contract grant sponsor: National Heart Lung and Blood Institute; Contract grant number: R01 HL091172-01; Contract grant sponsors: National Cancer Institute; NSF; Contract grant number: DMS 07-06935; Contract grant sponsor: NIH; Contract grant number: R03 CA130045-01.
*Correspondence to: Nilanjan Chatterjee, Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS, 6120 Executive Blvd., EPS 8038, Rockville, MD 20852. E-mail: chattern@mail.nih.gov
Received 3 October 2008; Revised 18 March 2009; Accepted 25 March 2009
Published online 11 May 2009 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20428

INTRODUCTION

The identification of large numbers of single-nucleotide polymorphisms (SNPs) across the human genome and the development of technologies for massive multiplex genotyping have now made genome-wide association studies (GWAS) involving hundreds of thousands of markers feasible [Hirschhorn and Daly, 2005; Thomas et al., 2005; Wang et al., 2005]. A number of successful studies have now been able to identify novel susceptibility loci for complex diseases like cancer, heart disease, and diabetes [McPherson et al., 2007; Yeager et al., 2007; Ridderstrale and Nilsson, 2008]. In GWAS, the evaluation of the association between a disease trait and an individual SNP often constitutes the initial analytic step. The lack of statistical significance in this first step may lead to the exclusion of an SNP from further scrutiny. Thus, to reduce the chance of false negatives, it is important to use powerful methods for preliminary screening of associations.

Population-based case-control studies are now being increasingly used for conducting genome-wide association scans. A widely used method for testing single-SNP associations in case-control studies is the Cochran-Armitage one-degree-of-freedom trend test [Armitage, 1955; Sasieni, 1997; Slager and Schaid, 2001; Freidlin et al., 2002], which is known to be optimal when the mode of effect for an SNP is multiplicative. An alternative method, which is known to have robust power under alternative modes of effect, is the two-degrees-of-freedom (2 d.f.) χ^2-test for independence between case-control and genotype status. The power of the standard 2 d.f. test, however, can be low for detection of SNPs with recessive effects, often because of the lack of sufficient sample size for homozygous variants among the cases and controls. To resolve this sparse data problem, we recently proposed the use of the assumption of Hardy-Weinberg Equilibrium (HWE) for estimation of the genotype frequencies among the controls, then comparing the resulting distribution with the empirical genotype distribution of the cases to obtain a novel 2 d.f. test of association [Chen and Chatterjee, 2007]. We showed that the proposed methodology can increase the power of 2 d.f. tests in a major way under non-multiplicative genetic effects, with the gain being particularly dramatic under the recessive model. A number of other reports had also previously pointed out

that ''retrospective'' methods for analysis of case-control studies can exploit assumptions of HWE or a related population genetics model to gain major power for both genotype- and haplotype-based tests of association [Epstein and Satten, 2003; Satten and Epstein, 2004; Thompson et al., 2004].

A major limitation of all HWE-based tests of genetic association is that they can lead to serious inflation of type-I error when the underlying assumptions of HWE or other genetic models are violated. In Chen and Chatterjee [2007], we characterized the bias of the 2 d.f. test analytically and showed that even the modest departure of HWE can lead to an unacceptably high increase in the type-I error of the procedures. The main objective of this article is to develop a 2 d.f. test that can gain power by exploiting the model assumptions of HWE for the underlying population and yet be resistant to bias when the model assumptions are violated. The method involves estimation of genotype-specific disease odds ratio parameters by data-adaptive ''shrinkage'' of a ''model-free'' estimator that does not require HWE assumption toward a ''model-based'' estimator that directly exploits the HWE constraints. The amount of ''shrinkage'' is sample-size-adaptive and data-adaptive so that in large samples the method has no bias irrespective of whether the assumptions of HWE hold and yet the method can gain efficiency by shrinking the analysis toward HWE, but only to the extent that the data validate the assumptions. The closed-form expression of the estimator itself and the availability of a simple variance estimator facilitate rapid computation of a corresponding Wald-type 2 d.f. test for GWAS involving hundreds of thousands of SNPs.

We evaluate the performance of the proposed method compared with a number of alternative tests using both simulated and real data. In particular, we use data from the Cancer Genetics Markers of Susceptibility (CGEMS) study to evaluate the ability of the proposed shrinkage estimation procedure to protect against inflated type-I errors due to the departure of HWE that may occur on a genome-wide scale. The study reveals potential problems associated with the application of the so-called ''retrospective'' methods on a genome-wide scale, even though the underlying assumption of HWE overall may be a good assumption for the genome. These studies together suggest that the proposed novel shrinkage estimation procedure is a promising method for testing genetic association in case-control studies. The method can gain major power over standard case-control analysis by exploiting the possible constraint of HWE for the underlying population and yet can adapt itself to protect against inflation of type-I error when the HWE constraints are violated. We also discuss the potential implications of these findings beyond the context of the simple 2 d.f. test considered in this article.

METHODS

The genotype information for an individual SNP in a case-control study can be represented by the 2×3 contingency table presented in Table I. Here D is the indicator of case ($D = 1$) or control ($D = 0$) status and G is the number of minor alleles carried by an individual ($G = 0, 1, 2$). Let $P_{dg} = \mathrm{pr}(G = g|D = d)$, $d = 0$ and 1, denote the population genotype frequencies for the controls and the cases, respectively. The likelihood L for case-control data is given by the product of two sets of multinomial probabilities, $L = L_1 \times L_0 = \prod_{g=0}^{2} p_{1g}^{n_{1g}} \times \prod_{g=0}^{2} p_{0g}^{n_{0g}}$, where n_{1g} and n_{0g} denote numbers of cases and controls with genotype g, respectively. In addition, define $n_{d+} = \sum_{g=0}^{2} n_{dg}$ for $d = 0$ and 1, i.e., n_{1+} for the number of cases and n_{0+} for the number of controls.

TABLE I. SNP genotype frequencies in diseased ($D = 1$) and disease-free ($D = 0$) subjects in the population

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

SNP, single-nucleotide polymorphism.

We consider re-parameterizing the likelihood in terms of alternative parameters of interest. Following Lindley [1988], we define

$$\theta = 0.5 \log \frac{4p_{00}p_{02}}{p_{01}^2} \quad \text{and} \quad \omega = 0.5 \log \frac{p_{00}}{p_{02}}. \qquad (1)$$

Note that θ and ω characterize the genotype frequencies of the controls according to the formulas $p_{00} = e^{2\omega}/(1 + e^{2\omega} + 2e^{\omega-\theta})$, $p_{01} = 2e^{\omega-\theta}/(1 + e^{2\omega} + 2e^{\omega-\theta})$, and $p_{02} = 1/(1 + e^{2\omega} + 2e^{\omega-\theta})$. The Hardy-Weinberg Disequilibrium (HWD) coefficient θ is a measure of the departure from HWE among controls, with $\theta = 0$, $\theta > 0$, and $\theta < 0$ corresponding to HWE, excess homozygosity, and excess heterozygosity, respectively. We note that the HWE assumption is reasonable for the underlying population, which will include both diseased and disease-free subjects. However, for rare diseases like certain cancers, the assumption of HWE is reasonable in the control population, as they approximately represent the underlying whole population. Further, let $\boldsymbol{\psi} = (\psi_0, \psi_1, \psi_2) = (1, P_{11}P_{00}/P_{01}P_{10}, P_{12}P_{00}/P_{02}P_{10})$ be the disease odds ratio parameter vector associated with the genotypes $G = 1$ and 2 relative to the baseline genotype $G = 0$. Let $\boldsymbol{\beta}^{\mathrm{T}} = (\log P_{11}P_{00}/P_{01}P_{10}, \log P_{12}P_{00}/P_{02}P_{10}) = (\log \psi_1, \log \psi_2)$.

Given θ and ω and hence the genotype frequency for the controls, we can characterize the genotype frequencies for the cases by $\boldsymbol{\psi}$ according to the formula

$$p_{1g} = \frac{\psi_g p_{0g}}{\sum_{g=0}^{2} \psi_g p_{0g}} \quad \text{for } g = 0, 1, 2. \qquad (2)$$

Thus, the likelihood for case-control data, $L = L(\boldsymbol{\beta}, \omega, \theta)$, is a function of $\boldsymbol{\psi}$, ω, and θ.

Let $\hat{\boldsymbol{\beta}}(\theta)$ denote the maximum-likelihood (ML) estimate of $\boldsymbol{\beta}$ for a fixed value of θ. When $\theta = 0$, i.e., when HWE holds among the controls, the ML estimate of $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}(\theta = 0)$, which we have shown previously [Chen and Chatterjee, 2007], can be expressed in closed form as

$$(\hat{\boldsymbol{\beta}}(\theta = 0))^{\mathrm{T}} = (\hat{\beta}_1(\theta = 0), \hat{\beta}_2(\theta = 0)) = \left(\log\left(\frac{n_{11}n_{00}^{\mathrm{E}}}{n_{10}n_{01}^{\mathrm{E}}}\right), \log\left(\frac{n_{12}n_{00}^{\mathrm{E}}}{n_{10}n_{02}^{\mathrm{E}}}\right) \right), \qquad (3)$$

where $n_{00}^{E} = n_{0+}(1-\hat{f})^2$, $n_{01}^{E} = n_{0+}2\hat{f}(1-\hat{f})$, and $n_{02}^{E} = n_{0+}\hat{f}^2$ denote the expected genotype counts for the controls computed assuming HWE, with the estimated allele frequency $\hat{f} = (n_{01} + 2n_{02})/2n_{0+}$. If θ is left completely unconstrained, then the ML estimate of $\boldsymbol{\beta}$ is given by the standard case-control estimator

$$(\hat{\boldsymbol{\beta}})^{T} = (\hat{\beta}_1, \hat{\beta}_2) = \left(\log\left(\frac{n_{11}n_{00}}{n_{10}n_{01}}\right), \log\left(\frac{n_{12}n_{00}}{n_{10}n_{02}}\right)\right). \quad (4)$$

The unconstrained ML estimator can also be expressed as $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\hat{\theta})$, where $\hat{\theta} = 0.5\log\{4n_{00}n_{02}/n_{01}^2\}$ denotes the ML estimator of θ.

We propose to combine $\hat{\boldsymbol{\beta}}(\theta = 0)$ and $\hat{\boldsymbol{\beta}}(\hat{\theta})$, the constrained and unconstrained estimators of $\boldsymbol{\beta}$, using an empirical-Bayes-type shrinkage estimation approach that we developed earlier for combining alternative estimates of the gene-environment interaction parameter obtained with or without the assumption of gene-environment independence in the underlying population [Mukherjee and Chatterjee, 2008]. In particular, following a very general formulation of the problem we described in that article, we propose to use the composite estimator (referred to as the vector-based shrinkage estimator, EB1) as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{EB1} =& \hat{\Delta}^{T}\hat{\theta}\hat{\theta}^{T}\hat{\Delta}(\widehat{V}_{\hat{\boldsymbol{\beta}}} + \hat{\Delta}^{T}\hat{\theta}\hat{\theta}^{T}\hat{\Delta})^{-1}\hat{\boldsymbol{\beta}} \\ &+ \widehat{V}_{\hat{\boldsymbol{\beta}}}(\widehat{V}_{\hat{\boldsymbol{\beta}}} + \hat{\Delta}^{T}\hat{\theta}\hat{\theta}^{T}\hat{\Delta})^{-1}\hat{\boldsymbol{\beta}}^{0} \\ =& \hat{\boldsymbol{\beta}} - \widehat{V}_{\hat{\boldsymbol{\beta}}}(\widehat{V}_{\hat{\boldsymbol{\beta}}} + \hat{\theta}^{2}\hat{\Delta}^{T}\hat{\Delta})^{-1}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{0}), \end{aligned} \quad (5)$$

where $\widehat{V}_{\hat{\boldsymbol{\beta}}}$ denotes the estimated asymptotic variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ as in Breslow and Day [1984] and $\hat{\Delta} = \partial\hat{\boldsymbol{\beta}}(\theta)/\partial\theta|_{\theta=0}$. We refer the reader to Mukherjee and Chatterjee [2008] for the detailed rationale for the estimator. Intuitively, we note that $\hat{\boldsymbol{\beta}}_{EB1}$ is an weighted average of the constrained and unconstrained estimators. As the sample size increases, and hence $\widehat{V}_{\hat{\boldsymbol{\beta}}}$ decreases, the composite estimator puts more weight on the robust unconstrained estimator. The weight also depends on $\hat{\theta}$, the data-driven estimate of the HWD coefficient. If the absolute value of $\hat{\theta}$ increases, i.e., if the data suggest departure of HWE, then less weight is given to the constrained estimator. The influence of θ on the weight depends on $\hat{\Delta}$, which determines the rate of change of $\hat{\boldsymbol{\beta}}(\theta)$ as a function of θ at the point $\theta = 0$. In the Appendix, we derive a closed-form expression for $\hat{\Delta}$. In formula (5), the EB estimator for $\boldsymbol{\beta} = (\beta_1, \beta_2)$ is presented where $\boldsymbol{\beta}$ is treated as a whole vector. Alternatively, one could derive an EB estimator for each of the two components of $\boldsymbol{\beta}$ separately. In a vectorized form, we can write the alternative EB estimator (referred to as the component-wise shrinkage estimator, EB2) as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{EB2} =& \hat{\boldsymbol{\beta}} - \mathrm{diag}[\widehat{V}_{\hat{\boldsymbol{\beta}}}][\mathrm{diag}(\widehat{V}_{\hat{\boldsymbol{\beta}}} + \hat{\theta}^{2}\hat{\Delta}^{T}\hat{\Delta})]^{-1}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{0}) \\ =& \hat{\boldsymbol{\beta}} - M(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{0}), \end{aligned} \quad (6)$$

where $\mathrm{diag}(A)$ is the matrix that takes the diagonal of matrix A but sets all the off-diagonal elements to zero and $M = \mathrm{diag}(\widehat{V}_{\hat{\boldsymbol{\beta}}})[\mathrm{diag}(\widehat{V}_{\hat{\boldsymbol{\beta}}} + \hat{\theta}^{2}\hat{\Delta}^{T}\hat{\Delta})]^{-1}$. In the current as well as other applications [see, e.g., Chen et al., 2009], we have found that the component-wise method generally produces more shrinkage compared to its multivariate counterpart. This observation is purely based on extensive empirical studies under several simulation settings. Theoretical justification of the performance advantage of EB2 over EB1 in mean-squared error (MSE) and power are still unknown. In the current HWE context with only two parameters, EB2 and EB1 are noted to have very similar MSE, but EB2 has better power properties across all scenarios. With increase in the dimension of the parameter space, as in the haplotype-based estimation context of Chen et al. [2009], the efficiency advantage of EB2 over EB1 becomes more pronounced. Simulation studies indicate that for a large number of parameters, the off-diagonal elements of $(\widehat{V}_{\hat{\boldsymbol{\beta}}})[(\widehat{V}_{\hat{\boldsymbol{\beta}}} + \hat{\theta}^{2}\hat{\Delta}^{T}\hat{\Delta})]^{-1}$ are quite variable across the samples, which possibly offset the advantages of a full multivariate vector-wise shrinkage. The issue of relative efficiency of EB2 over EB1 merits further theoretical exploration.

In the Appendix, we use the delta method to obtain an estimate of the variance-covariance matrix (Σ_{EB}) for the two EB estimators of $\boldsymbol{\beta} = (\beta_1, \beta_2)$. For each of the methods, a 2 d.f. Wald test can be constructed as $T_i = W_{EBi} = \hat{\boldsymbol{\beta}}_{EBi}^{T}\Sigma_{EBi}^{-1}\hat{\boldsymbol{\beta}}_{EBi}$, for $i = 1, 2$.

We perform simulation studies to compare the type-I error and power for four alternative tests of association: (1) the standard unconstrained 2 d.f. test; (2) the 2 d.f. test assuming HWE in the controls; (3) a two-step method that first tests the HWE constraint (i.e., null hypothesis $\theta = 0$) among the controls at a designated significance level, then uses a constrained test if not rejected, or uses a unconstrained test if rejected; and (4) Wald tests based on the proposed EB estimation procedures. In these simulation studies, we assume that the disease susceptibility allele is the less frequent or minor allele. Given the minor allele frequency (MAF) f and HWD coefficient θ, we calculate the genotype frequencies for the controls, p_{0g}, according to formula (1). Further, given the odds ratio parameters $\psi_0 = 1$ (reference group), ψ_1, and ψ_2, we obtain the genotype frequencies for the cases (i.e., p_{1g}) using formula (2). The genotypes for the cases and the controls are then generated from the respective multinomial distributions.

RESULTS

SIMULATION STUDIES

In the first set of simulations, we examine the type-I errors of various tests under the null hypothesis of no disease-genotype association. We simulate data in the settings that involve two sample sizes (i.e., $n_0 = n_1 = 500$ and $n_0 = n_1 = 2{,}000$) and multiple combinations of coefficients θ (i.e., $\theta = 0$, $0.05\log(1.2)$, $0.5\log(1.6)$, and $0.5\log(2.0)$, referred to as HWE, small, modest, and large deviation from HWE, respectively) and MAFs f. We choose the significance levels α to be 0.05 for the sample size of 500 and 1.0e−5 for the sample size of 2,000. We observe from Table II that when HWE holds, all of the different procedures, except the two-step method, maintain the desired type-I error level very well. The inflation of the type-I error in the two-step method in this setting is probably due to the fact that the procedure ignores the variability associated with uncertainty in the underlying model selection procedure at the first step. When HWE is

TABLE II. The type-I error for alternative tests under the null hypothesis of no disease-genotype association (i.e., $\psi_{Aa} = \psi_{aa} = 1$)

HWD coeff. (θ)	MAF	Unconstrained	Constrained	Two-step	EB1	EB2
500 cases and 500 controls, α = 0.05, 10,000 simulations						
θ = 0	0.1	0.03	0.04	0.04	0.03	0.03
	0.2	0.05	0.05	0.06	0.04	0.04
	0.3	0.05	0.05	0.06	0.04	0.04
θ = 0.5 log(1.2)	0.1	0.03	0.07	0.07	0.04	0.04
	0.2	0.05	0.09	0.09	0.05	0.06
	0.3	0.05	0.11	0.11	0.06	0.07
θ = 0.5 log(1.6)	0.1	0.03	0.15	0.13	0.06	0.07
	0.2	0.05	0.31	0.22	0.09	0.11
	0.3	0.05	0.50	0.23	0.08	0.12
θ = 0.5 log(2.0)	0.1	0.03	0.24	0.18	0.08	0.09
	0.2	0.04	0.57	0.22	0.07	0.11
	0.3	0.05	0.82	0.13	0.06	0.10
2,000 cases and 2,000 controls, α = 1.0e−5, 1 million simulations						
θ = 0	0.1	3.0e−6	8.0e−6	1.0e−5	3.0e−6	3.0e−6
	0.2	1.0e−5	1.4e−5	1.8e−5	1.1e−5	1.3e−5
	0.3	1.6e−5	1.1e−5	2.1e−5	8.0e−6	8.0e−6
θ = 0.5 log(1.2)	0.1	4.0e−6	2.6e−4	2.3e−4	5.5e−5	6.5e−5
	0.2	1.7e−5	8.0e−4	6.1e−4	1.3e−4	1.9e−4
	0.3	1.1e−5	2.1e−3	1.3e−3	2.3e−4	3.6e−4
θ = 0.5 log(1.6)	0.1	1.0e−6	9.0e−3	4.9e−3	6.8e−4	8.1e−4
	0.2	5.0e−6	0.11	1.1e−2	4.7e−4	9.7e−4
	0.3	8.0e−6	0.39	3.5e−3	7.9e−5	3.7e−4
θ = 0.5 log(2.0)	0.1	3.0e−6	5.9e−2	1.5e−2	1.4e−3	1.7e−3
	0.2	5.0e−6	0.57	2.3e−3	5.1e−5	1.8e−4
	0.3	1.1e−5	0.95	2.0e−5	1.0e−5	3.1e−5

Results are obtained based on simulating case-control data sets with either 500 cases and 500 controls (upper panel) or 2,000 cases and 2,000 controls (lower panel). Desired significance level of the tests are assumed to be $\alpha = 0.05$ and 10^{-5} for studies with 500 and 2,000 cases, respectively. Empirical significance levels of the tests are obtained by 10,000 and 1 million simulations, respectively. HWD, Hardy-Weinberg Disequilibrium; MAF, minor allele frequency.

violated, we observe that the type-I error of the constrained test rapidly increases with θ and becomes unacceptably high even under modest deviation from HWE. The two-step method, although it reduces the problem of type-I error inflation to a large extent, can still produce a large inflation of the type-I error. The EB procedures provide much better control of type-I error, compared with both the constrained and the two-step method. In particular, it is encouraging to note that when the departure of HWE is small, say $|\theta| \leq 0.5 \log(1.2)$, a range where the large majority of HWE departures are likely to appear in practice (see, e.g., Fig. 3 in the CGEMS application), the type-I errors of the EB procedures are generally very close to the nominal level. As θ further increases, the type-I errors of the EB procedures initially increase and then eventually again decrease.

In the next set of simulations, we assume HWE in the control population and explore the power of various test procedures under different combinations of MAF and odds ratio parameters. Figure 1 displays the power curves estimated from 10,000 simulated data sets of 500 cases and 500 controls. It is clear that in this setting the constrained test can gain major power over the unconstrained test, especially when the true effect of the genotype is recessive. The EB1 test procedure, although it gives up some efficiency compared with the constrained test, retains a major power advantage over the unconstrained test for detecting recessive genetic effects. The power of EB1 was slightly lower than that of EB2 (not shown in Fig. 1). The power of the two-stage test lies between the unconstrained and constrained tests, as expected.

In Table III, we show the power for various tests of association under the recessive model and different combinations of the MAF f and the HWE coefficient θ. We observe that when there is small departure from HWE, a scenario that is likely to be common in practice, the EB procedures can maintain desired type-I error levels fairly well (as seen in Table II) and yet can gain substantial power over the unconstrained test. Similar comparisons for the dominant model are shown in Table IV. Here we observe that under small departures from HWE, the EB procedures generally perform similarly to the unconstrained test. Under large departures from HWE, however, the EB procedures can sometimes have a substantial loss of power compared with the unconstrained test. Since some of the tests we consider do not strictly maintain type-I error under the departure of HWE, we also provide MSE for the parameter estimates as an alternative way of comparing the performance of the different estimators. The results are similar to those presented in Mukherjee and Chatterjee [2008]. Under HWE, the EB methods produce MSE comparable to the constrained estimator, which has the smallest MSE. Under departures from HWE, the EB methods produce the smallest or close to the smallest MSE among all methods we considered (as shown in Tables III and IV).

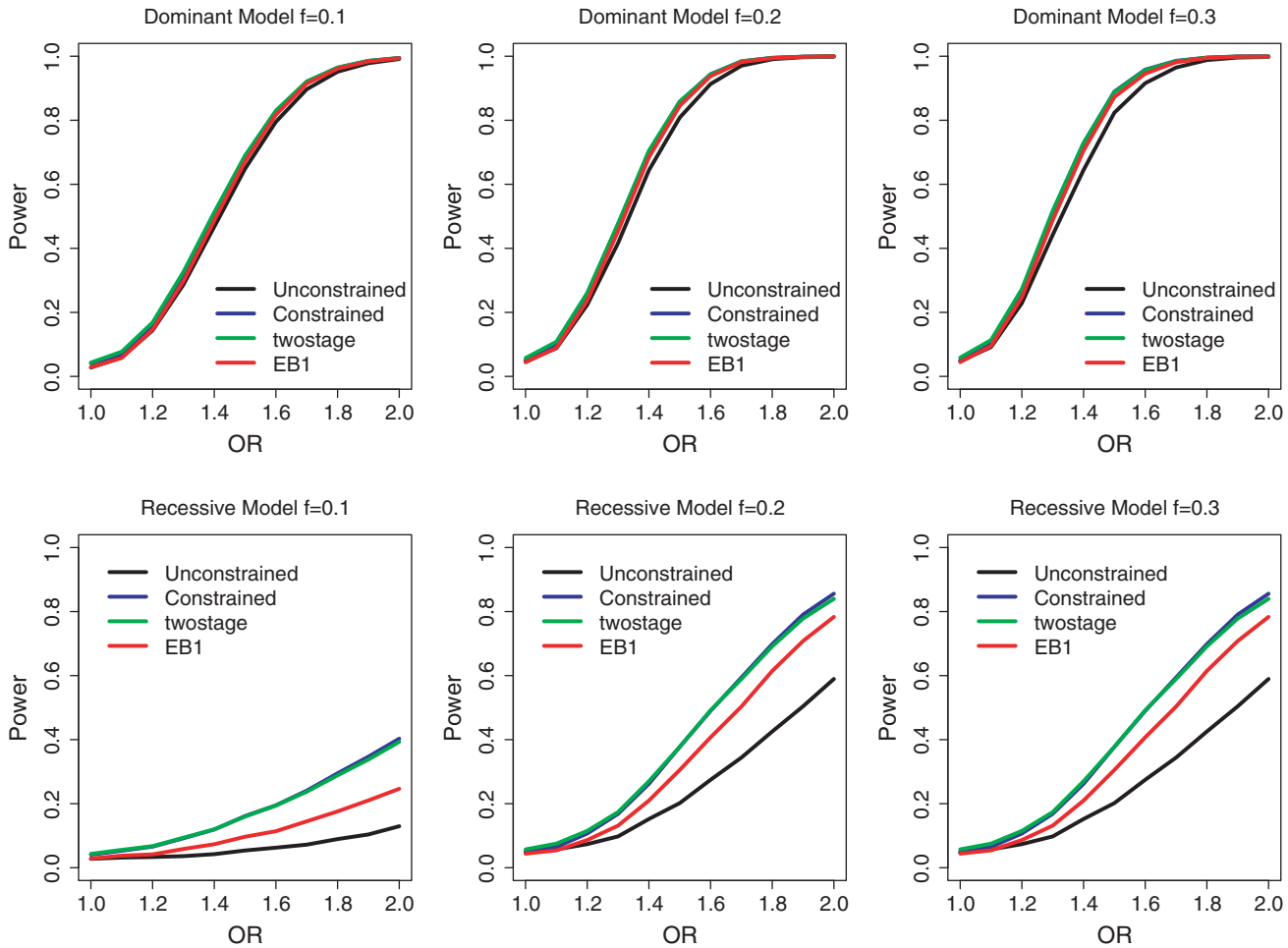

Fig. 1. Power comparison for alternative case-control tests of association: (i) a standard 2 d.f. test (unconstrained), (ii) a 2 d.f. test assuming HWE in controls (constrained), (iii) a two-step test that selects between the constrained and unconstrained tests based on a test of HWE among the controls, and (iv) the proposed EB tests. Data are simulated for a case-control study of 500 cases and 500 controls, assuming that HWE holds for the underlying population. The effect of the SNP on the risk of the disease is assumed to follow either a dominant (upper panel) or a recessive pattern (lower panel). All of the tests are performed at a significance level of $\alpha = 0.05$. HWE, Hardy-Weinberg Equilibrium; 2 d.f., two degrees of freedom; SNP, single-nucleotide polymorphism.

THE CANCER GENETICS MARKERS OF SUSCEPTIBILITY (CGEMS) STUDY

We evaluate the performance of alternative 2 d.f. tests of association using data from the CGEMS study, an NCI enterprise initiative to conduct multi-stage whole-genome association studies to identify genes giving rise to increased risks of prostate and breast cancers. In this article, we will focus on data from the initial scan for the prostate cancer study, involving genotype data on about 550,000 SNPs from 1,172 cases and 1,157 controls. An initial report from the study describing the increased risk of prostate cancer associated with the 8*q*24 region has been published [Yeager et al., 2007]. Sequential replication studies are now ongoing for about 5% of the SNPs that are considered to be promising based on the data from the initial scan. The details of the CGEMS study design and the results from the initial scan can be found at the website https://caintegrator.nci.nih.gov/cgems/.

Figure 2 shows the Q-Q plots associated with 449,698 SNPs from 22 non-sex chromosomes with MAFs larger than 0.05 for the four different tests of association: (i) unconstrained; (ii) constrained; (iii) two-stage; and (iv) EB2. Each plot in the figure displays the empirical percentile of the *P*-values associated with one of the four 2 d.f. tests against the percentiles of the expected null distribution. For a well-designed study and a robust analytic method, Q-Q plots for GWAS are expected to follow the diagonal lines closely, given that at most a handful of the SNPs are likely to be truly associated with the disease. Thus, large-scale departure of the Q-Q plot from the expected diagonal is often considered to be indicative of bias in the underlying study design or/and analytic method.

In Figure 2, we observe that the Q-Q plot for the unconstrained test closely follows the diagonal line except at the extreme tail of the distributions, where $P < 10^{-4}$. This plot suggests that the CGEMS study does not suffer from any large-scale systematic bias such as those due to population stratification or differential genotyping error. Moreover, the standard 2 d.f. test of association is a robust method for analysis of data from this study. In contrast, we observe that the Q-Q plot for the constrained test departs

TABLE III. The power for different tests and mean-squared error corresponding to the estimate of log(ψ_{aa}) (in parentheses)

Sample size	HWD coeff. (θ)	MAF	Unconstrained	Constrained	Two-step	EB1	EB2
$N = 500$	$\theta = 0$	0.1	0.120 (0.383)	0.379 (0.171)	0.371 (0.205)	0.224 (0.232)	0.233 (0.229)
		0.2	0.556 (0.087)	0.834 (0.056)	0.821 (0.065)	0.707 (0.066)	0.756 (0.063)
		0.3	0.883 (0.043)	0.980 (0.035)	0.973 (0.037)	0.940 (0.037)	0.967 (0.036)
	$\theta = 0.5\log(1.2)$	0.1	0.118 (0.378)	0.517 (0.180)	0.484 (0.211)	0.283 (0.235)	0.306 (0.228)
		0.2	0.575 (0.082)	0.945 (0.065)	0.874 (0.071)	0.735 (0.068)	0.809 (0.064)
		0.3	0.893 (0.043)	0.999 (0.042)	0.959 (0.043)	0.930 (0.041)	0.973 (0.040)
	$\theta = 0.5\log(1.6)$	0.1	0.125 (0.367)	0.721 (0.288)	0.598 (0.310)	0.331 (0.276)	0.372 (0.268)
		0.2	0.595 (0.082)	0.993 (0.135)	0.754 (0.117)	0.656 (0.089)	0.776 (0.087)
		0.3	0.905 (0.040)	1.000 (0.071)	0.911 (0.053)	0.897 (0.045)	0.963 (0.046)
	$\theta = 0.5\log(2.0)$	0.1	0.126 (0.365)	0.822 (0.435)	0.581 (0.409)	0.322 (0.319)	0.375 (0.310)
		0.2	0.605 (0.080)	0.999 (0.228)	0.656 (0.126)	0.619 (0.097)	0.755 (0.096)
		0.3	0.916 (0.038)	1.000 (0.111)	0.916 (0.045)	0.884 (0.043)	0.964 (0.045)
$N = 2{,}000$	$\theta = 0$	0.1	0.003 (0.082)	0.228 (0.039)	0.218 (0.051)	0.086 (0.054)	0.098 (0.052)
		0.2	0.520 (0.021)	0.965 (0.014)	0.943 (0.016)	0.782 (0.016)	0.859 (0.015)
		0.3	0.986 (0.011)	1.000 (0.009)	0.997 (0.009)	0.994 (0.010)	0.999 (0.009)
	$\theta = 0.5\log(1.2)$	0.1	0.003 (0.081)	0.500 (0.057)	0.438 (0.068)	0.166 (0.062)	0.190 (0.060)
		0.2	0.548 (0.020)	0.999 (0.026)	0.808 (0.025)	0.654 (0.020)	0.786 (0.020)
		0.3	0.989 (0.010)	1.000 (0.014)	0.990 (0.013)	0.986 (0.011)	0.997 (0.011)
	$\theta = 0.5\log(1.6)$	0.1	0.003 (0.079)	0.862 (0.174)	0.462 (0.128)	0.123 (0.087)	0.149 (0.087)
		0.2	0.585 (0.020)	1.000 (0.093)	0.592 (0.026)	0.505 (0.023)	0.675 (0.024)
		0.3	0.993 (0.010)	1.000 (0.044)	0.993 (0.010)	0.982 (0.011)	0.997 (0.012)
	$\theta = 0.5\log(2.0)$	0.1	0.004 (0.078)	0.968 (0.336)	0.248 (0.134)	0.052 (0.097)	0.071 (0.096)
		0.2	0.613 (0.019)	1.000 (0.184)	0.613 (0.019)	0.501 (0.022)	0.682 (0.022)
		0.3	0.995 (0.010)	1.000 (0.085)	0.995 (0.010)	0.978 (0.011)	0.997 (0.010)

The disease-genotype odds ratios are assumed to follow a ''recessive'' pattern with $\psi_{Aa} = 1$, $\psi_{aa} = 1.4^2$. Results are based on 10,000 simulated case-control data sets, each with 500 cases and 500 controls (upper panel), and on 1,000,000 simulated case-control data sets, each with 2,000 cases and 2,000 controls (lower panel). All tests are performed at a significance level of $\alpha = 0.05$ for the study with 500 cases and $\alpha = 10^{-5}$ for the study with 2,000 cases. HWD, Hardy-Weinberg Disequilibrium; MAF, minor allele frequency.

dramatically from the diagonal line in the range of $P < 10^{-2}$. For example, the constrained test finds 1,716 SNPs to have P-values less than 10^{-3}, while under the null hypothesis of no association, only 450 (i.e., $449{,}698 \times 10^{-3} \simeq 450$) such SNPs would be expected in the study. This indicates a major inflation of the type-I error for the constrained test due to departure of the HWE assumption. The Q-Q plot corresponding to the two-step procedure suggests that although the type-I error inflation is substantially reduced compared with the constrained test, it still remains significantly higher than desired. The two-step method, for example, finds 122 SNPs to have P-values less than 10^{-4}, while under the null hypothesis of no association, only 45 (i.e., $449{,}698 \times 10^{-4} \simeq 45$) such SNPs would be expected. The Q-Q plot for the EB2 procedure strikingly resembles that of the robust unconstrained test. The plot closely follows the diagonal line except at the extreme tail of the distribution. The pattern provides empirical evidence that EB-type procedures perform very well in controlling the type-I error rates for the related tests of association under realistic departures from HWE that may arise in GWAS. We refrain from presenting the Q-Q plot for the EB1 procedure in this example as it appears to be very similar to EB2, and EB2 does have a slight edge over EB1 in terms of power for detecting disease-SNP association.

In Figure 3, we show the histogram of the estimated HWD coefficient θ for the 449,698 SNPs we studied. It is clear that, overall, HWE is a good assumption for the genome, with 69.6 and 96.7% of the estimated coefficients falling between the $\pm 0.5\log(1.2)$ and $\pm 0.5\log(1.6)$ limits, respectively. Nevertheless, a test based on the assumption of HWE can lead to a major inflation of type-I error for large-scale studies.

The CGEMS group has recently reported results from a replication study involving 3,941 cases and 3,964 controls [Thomas et al., 2008]. Based on a ''joint analysis'' of the initial scan and replication study, the report has listed 17 SNPs that have met genome-wide significance for their association with prostate cancer. Given that associations of these SNPs with prostate cancer are now considered to be ''replicated,'' we can use these SNPs to evaluate the power of alternative methods for the analysis of the initial CGEMS scan. From the results shown in Table V, we observe that for 12 out of the 17 SNPs (rows 1–12 of Table V), both EB-based procedures produce smaller P-values than the standard 2 d.f. test, while for 2 other SNPs (rows 13 and 14 of Table V), one of the EB-based procedures produces smaller P-values. The decrease in P-values, however, is quite modest in general. These results are intuitive, given that none of the SNPs shows a genotype odds ratio pattern that resembles a recessive model, under which we would have expected to see a larger gain in power by exploiting the HWE assumption.

DISCUSSION

In this article, we propose a powerful test for genetic association in case-control studies by exploiting the common assumption of HWE for the underlying population. Unlike previous methods that have also aimed to gain

TABLE IV. The power for different tests and sum of mean-squared errors corresponding to the point estimates of $\log(\psi_{Aa})$ and $\log(\psi_{aa})$ (in parentheses)

Sample size	HWD coeff. (θ)	MAF	Unconstrained	Constrained	Two-step	EB1	EB2
$N = 500$	$\theta = 0$	0.1	0.465 (0.479)	0.498 (0.275)	0.507 (0.308)	0.477 (0.329)	0.482 (0.324)
		0.2	0.637 (0.123)	0.693 (0.089)	0.697 (0.098)	0.664 (0.100)	0.677 (0.096)
		0.3	0.656 (0.069)	0.741 (0.057)	0.740 (0.060)	0.700 (0.060)	0.714 (0.058)
	$\theta = 0.5\log(1.2)$	0.1	0.455 (0.471)	0.460 (0.273)	0.478 (0.308)	0.446 (0.323)	0.437 (0.314)
		0.2	0.621 (0.120)	0.600 (0.099)	0.622 (0.107)	0.598 (0.103)	0.593 (0.098)
		0.3	0.651 (0.067)	0.602 (0.066)	0.639 (0.068)	0.610 (0.064)	0.611 (0.062)
	$\theta = 0.5\log(1.6)$	0.1	0.420 (0.472)	0.447 (0.368)	0.473 (0.402)	0.389 (0.362)	0.365 (0.349)
		0.2	0.621 (0.119)	0.614 (0.173)	0.658 (0.154)	0.549 (0.124)	0.530 (0.123)
		0.3	0.659 (0.066)	0.629 (0.112)	0.688 (0.084)	0.565 (0.072)	0.580 (0.075)
	$\theta = 0.5\log(2.0)$	0.1	0.384 (0.468)	0.482 (0.519)	0.485 (0.507)	0.351 (0.408)	0.323 (0.395)
		0.2	0.600 (0.113)	0.714 (0.273)	0.667 (0.165)	0.489 (0.131)	0.478 (0.133)
		0.3	0.658 (0.064)	0.789 (0.176)	0.686 (0.074)	0.516 (0.071)	0.575 (0.076)
$N = 2{,}000$	$\theta = 0$	0.1	0.370 (0.103)	0.428 (0.059)	0.431 (0.071)	0.397 (0.073)	0.417 (0.071)
		0.2	0.690 (0.029)	0.792 (0.021)	0.790 (0.024)	0.744 (0.024)	0.770 (0.023)
		0.3	0.718 (0.017)	0.853 (0.014)	0.847 (0.015)	0.793 (0.015)	0.816 (0.015)
	$\theta = 0.5\log(1.2)$	0.1	0.329 (0.101)	0.314 (0.076)	0.334 (0.087)	0.307 (0.079)	0.297 (0.077)
		0.2	0.674 (0.029)	0.614 (0.034)	0.661 (0.034)	0.624 (0.028)	0.612 (0.028)
		0.3	0.723 (0.017)	0.621 (0.022)	0.702 (0.020)	0.651 (0.017)	0.648 (0.018)
	$\theta = 0.5\log(1.6)$	0.1	0.269 (0.099)	0.285 (0.191)	0.301 (0.148)	0.217 (0.106)	0.178 (0.105)
		0.2	0.642 (0.028)	0.620 (0.108)	0.651 (0.034)	0.515 (0.032)	0.490 (0.034)
		0.3	0.729 (0.016)	0.671 (0.069)	0.731 (0.016)	0.588 (0.017)	0.641 (0.019)
	$\theta = 0.5\log(2.0)$	0.1	0.225 (0.098)	0.374 (0.354)	0.277 (0.154)	0.152 (0.117)	0.112 (0.118)
		0.2	0.613 (0.027)	0.830 (0.210)	0.614 (0.028)	0.400 (0.032)	0.446 (0.032)
		0.3	0.728 (0.016)	0.930 (0.134)	0.728 (0.016)	0.491 (0.018)	0.652 (0.017)

The disease-genotype odds ratios are assumed to follow a "dominant" pattern with $\psi_{Aa} = \psi_{aa} = 1.4$. Results are based on 10,000 simulated case-control data sets, each with 500 cases and 500 controls (upper panel), and on 1,000,000 simulated case-control data sets, each with 2,000 cases and 2,000 controls (lower panel). All tests are performed at a significance level of $\alpha = 0.05$ for the study with 500 cases and $\alpha = 10^{-5}$ for the study with 2,000 cases. HWD, Hardy-Weinberg Disequilibrium; MAF, minor allele frequency.

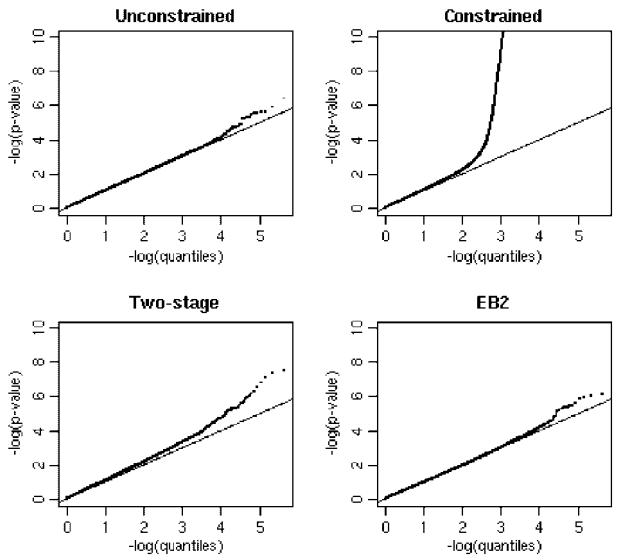

Fig. 2. Q-Q plots for the CGEMS genome-wide association study of prostate cancer. Each panel represents a plot for the percentiles of the observed *P*-values, obtained from a specific test of association, against those expected under the "null" hypothesis of no association. The solid line represents the diagonal $Y = X$. CGEMS, Cancer Genetics Markers of Susceptibility.

efficiency for case-control association testing by exploiting HWE for the underlying population, the proposed EB procedure can data-adaptively relax the underlying constraints and thus can reduce the chance of false-positive results when the HWE assumption is violated. Simulation studies as well as an application involving a GWAS show that the EB procedure can maintain appropriate control over the type-I error rate for large-scale studies that would have natural deviations from HWE of varying degrees across different loci. Further, our studies illustrate that the EB procedure has a major power advantage over standard case-control tests for the detection of susceptibility SNPs with effects resembling a "recessive" pattern. In addition, the closed-form expression of the EB estimators and the simple corresponding variance estimation make the computation cost comparable to that of the unconstrained test in the study setting of GWAS.

The pattern of power seen for different methods under different models for genetic effects is intuitive. The constrained test gains power over its unconstrained counterpart by incorporating additional information from the departure of the observed genotype distribution in the case-control sample from the assumed HWE model for the population. If an SNP is under HWE in the population, its genotype distribution approximately follows HWE in the controls, under the assumption of rare disease. Moreover, when the effect of an SNP is multiplicative (log-additive) per copy of an allele, it can be shown that, again assuming rare disease, the HWE for the population implies HWE for the cases [Sasieni, 1997]. It is expected that when the

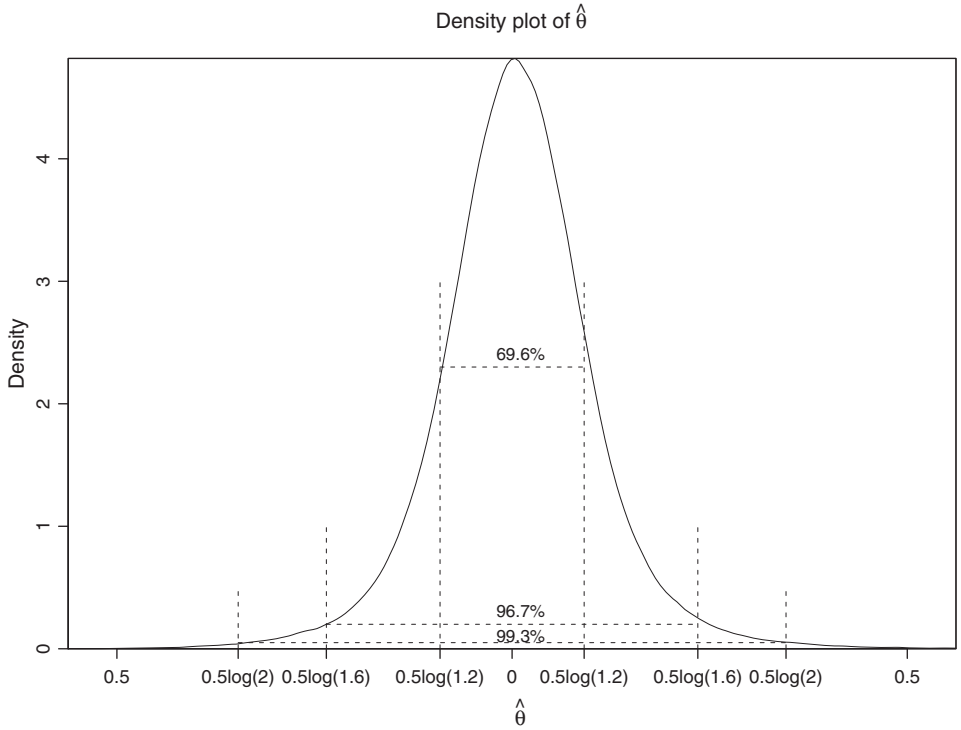

Fig. 3. Histogram of estimates of θ, a log-odds-ratio measure of Hardy-Weinberg Disequilibrium, for the 449,698 SNPs studied in 22 non-sex chromosomes in the CGEMS study with minor allele frequencies larger than 0.05. The values $\theta = 0$, $\theta > 0$, and $\theta < 0$ correspond to HWE, excess homozygosity, and excess heterozygosity, respectively. SNP, single-nucleotide polymorphism; CGEMS, Cancer Genetics Markers of Susceptibility; HWE, Hardy-Weinberg Equilibrium.

TABLE V. The comparison of the *P*-values of various tests of association for specific target SNPs in the CGEMS study

		Overall				
SNP	MAF	Unconstrained	Constrained	Two-step	EB1	EB2
rs4242382	0.100	3.60e−5	2.71e−5	2.71e−5	3.06e−5	3.17e−5
rs4242384	0.099	3.42e−5	2.75e−5	2.75e−5	2.82e−5	2.82e−5
rs1447295	0.102	1.78e−4	1.38e−4	1.38e−4	1.34e−4	1.44e−4
rs7837688	0.097	1.29e−5	5.24e−6	5.24e−6	5.65e−6	5.67e−6
rs11988857	0.117	2.30e−5	2.41e−5	2.41e−5	1.84e−5	1.89e−5
rs10993994	0.365	2.64e−3	1.22e−3	1.22e−3	2.54e−3	1.57e−3
rs9656816	0.088	1.20e−3	5.22e−4	5.22e−4	8.71e−4	9.80e−4
rs6983267	0.489	1.04e−2	7.68e−3	7.68e−3	7.65e−3	7.60e−3
rs4430796	0.489	9.99e−3	3.59e−3	3.59e−3	3.57e−3	3.54e−3
rs7501939	0.421	5.37e−3	4.32e−4	4.32e−4	1.33e−3	9.07e−4
rs7014346	0.350	5.50e−3	4.69e−3	4.69e−3	4.84e−3	4.76e−3
rs7837328	0.394	2.48e−3	2.33e−3	2.33e−3	2.41e−3	2.38e−3
rs1106207	0.428	3.46e−3	3.97e−4	3.46e−4	5.23e−3	2.32e−3
rs7017300	0.131	4.83e−5	6.65e−5	6.65e−5	4.21e−5	5.21e−5
rs4962416	0.263	9.89e−5	6.65e−5	6.65e−5	1.40e−4	1.09e−4
rs1486567	0.231	4.26e−2	5.13e−2	5.13e−2	5.48e−2	4.91e−2
rs10896449	0.490	0.024	0.094	0.094	0.064	0.071

SNP, single-nucleotide polymorphism; CGEMS, Cancer Genetics Markers of Susceptibility; MAF, minor allele frequency.

non-multiplicative effect of an SNP is larger, so is the departure for the distribution of its genotypes from HWE, in the cases and hence in the case-enriched case-control sample. Thus, the efficiency gains for the constrained and the EB-type shrinkage procedures over the unconstrained one are expected to be increasing with the magnitude of the non-multiplicative effect of an SNP. In our simulation, under the multiplicative model for the effect of an SNP, we do not see any difference in efficiency among the methods (results not shown). Under the dominant model, which corresponds to modest departure from the multiplicative model, we observe some gain in efficiency for the

constrained and the EB procedures. Under the recessive model, which corresponds to large departure from the multiplicative effect, we observe the highest gain in efficiency for the constrained and the EB procedures.

In this article, we have focused on the 2 d.f. single-SNP test of genetic association. The proposed shrinkage estimation strategy, however, can be used to improve the power of other types of genetic association tests in case-control studies. For single-SNP association testing with unknown modes of genetic effect, for example, a popular alternative to the 2 d.f. test is the MAX procedure which uses the maximum of the single-SNP Z-statistics for the additive, dominant, and recessive models as the test statistics for detecting association. For case-control studies, the power of the MAX procedure can potentially be improved by deriving the component Z-statistics by exploiting the HWE constraints for the genotype distribution of the controls. In particular, the proposed shrinkage estimation strategy can be used to estimate the disease-genotype odds ratios and their standard errors under alternative modes of genetic effect and hence to derive the corresponding Wald statistics.

The proposed shrinkage estimation strategy can also potentially be used to improve the power of case-control genetic association tests involving loci with more than two alleles. The general strategy would involve first estimating disease-genotype odds ratios, once using the empirical genotype frequency for the controls, once assuming HWE constraints for the controls, then combining the two estimators using the empirical-Bayes-type weighting strategy proposed here. Further research is merited on the development of such multi-allelic tests, especially in the context of haplotype-based association studies, where the additional complexity arises from the fact that haplotype-phase information is typically missing from the observable genotype data.

In conclusion, we believe that the proposed shrinkage estimation strategy, considering its power, robustness, generalizability, and computational simplicity, overall is a promising approach for detecting genetic associations from case-control studies.

ACKNOWLEDGMENTS

The research of Nilanjan Chatterjee was supported by a Gene-Environment Initiative (GEI) grant from the National Heart Lung and Blood Institute (R01 HL091172-01) and by the Intramural research program of the National Cancer Institute. The research of Bhramar Mukherjee was partially supported by NSF DMS 07-06935 and NIH grant R03 CA130045-01.

REFERENCES

Armitage P. 1955. Tests for linear trends in proportions and frequencies. Biometrics 11:375–386.

Breslow NE, Day N. 1984. Statistics of Case-Control Studies. New York: Marcel Dekker.

Chen J, Chatterjee N. 2007. Exploiting Hardy-Weinberg equilibrium for efficient screening of single SNP associations from case-control studies. Hum Hered 63:196–204.

Chen YH, Chatterjee N, Carroll RJ. 2009. Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies. J Am Stat Assoc, to appear.

Epstein MP, Satten GA. 2003. Inference on haplotype effects in case-control studies using unphased genotype data. Am J Hum Genet 73:1316–1329.

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

Hirschhorn JN, Daly MJ. 2005. Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6:95–108.

Lindley DV. 1988. Statistical inference concerning Hardy-Weinberg equilibrium. Bayesian Stat 3:307–326.

McPherson R, Pertsemlidis A, Kavaslar N, Stewart A, Roberts R, Cox DR, Hinds DA, Pennacchio LA, Tybjaerg-Hansen A, Folsom AR, Boerwinkle E, Hobbs HH, Cohen JC. 2007. A common allele on chromosome 9 associated with coronary heart disease. Science 316:1488–1491.

Mukherjee B, Chatterjee N. 2008. Exploiting gene-environment independence for analysis of case-control studies: an empirical Bayes approach to trade off between bias and efficiency. Biometrics 64:685–694.

Ridderstrale M, Nilsson E. 2008. Type 2 diabetes candidate gene CAPN10: first, but not last. Curr Hypertens Rep 10:19–24.

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

Satten GA, Epstein MP. 2004. Comparison of prospective and retrospective methods for haplotype inference in case-control studies. Genet Epidemiol 27:192–201.

Slager SL, Schaid DJ. 2001. Evaluation of candidate genes in case-control studies: a statistical method to account for related subjects. Am J Hum Genet 68:1457–1462.

Thomas DC, Haile RW, Duggan D. 2005. Recent developments in genomewide association scans: a workshop summary and review. Am J Hum Genet 77:337–345.

Thomas G, Jacobs KB, Yeager M, Kraft P, Wacholder S, Orr N, Yu K, Chatterjee N, Welch R, Hutchinson A, Crenshaw A, Cancel-Tassin G, Staats B, Wang Z, Gonzalez-Bosquet J, Fang J, Deng X, Berndt S, Calle E, Feigelson H, Thun M, Rodriguez C, Albanes D, Virtamo J, Weinstein S, Schumacher F, Giovannucci E, Willett W, Cussenot O, Valeri A, Andriole G, Crawford E, Tucker M, Gerhard D, Fraumeni J, Hoover R, Hayes R, Hunter D, Chanock S. 2008. Multiple novel loci identified in a genome-wide association study of prostate cancer. Nat Genet 40:310–315.

Thompson D, Witte JS, Slattery M, Goldgar D. 2004. Increased power for case-control studies of single nucleotide polymorphisms through incorporation of family history and genetic constraints. Genet Epidemiol 27:215–224.

Wang WY, Barratt BJ, Clayton DG, Todd JA. 2005. Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet 6:109–118.

Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, Minichiello MJ, Fearnhead P, Yu K, Chatterjee N, Wang Z, Welch R, Staats B, Calle E, Feigelson H, Thun M, Rodriguez C, Albanes D, Virtamo J, Weinstein S, Schumacher F, Giovannucci E, Willett W, Cancel-Tassin G, Cussenot O, Valeri A, Andriole G, Gelmann E, Tucker M, Gerhard D, Fraumeni J, Hoover R, Hunter D, Chanock S, Thomas G. 2007. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet 39:645–649.

APPENDIX A

A.1. DERIVATION OF $\hat{\Delta}$

The likelihood function for controls, L_0, which is proportional to $p_{00}^{n_{00}} p_{01}^{n_{01}} p_{02}^{n_{02}}$, can be expressed in terms of θ and ω as

$$L_0 = \frac{e^{\theta(n_{00}+n_{02})} e^{\omega(n_{00}-n_{02})}}{(1+e^{\theta}\cosh\omega)^{n_{0+}}},$$

where the hyperbolic cosine $\cosh\omega = (e^{\omega} + e^{-\omega})/2$. Taking the derivative of the logarithm of L_0 with respect to ω, we get

$$\frac{\partial \log L_0}{\partial \omega} = (n_{00} - n_{02}) - \frac{n_{0+}\sinh\omega}{e^{-\theta} + \cosh\omega},$$

where the hyperbolic sine $\sinh\omega = (e^{\omega} - e^{-\omega})/2$. Equating the last equation to zero, taking the derivative on both sides with respect to θ, and letting $\theta = 0$, we get

$$k = \frac{\partial\hat{\omega}}{\partial\theta}(\theta = 0) = \frac{n_{00} - n_{02}}{(n_{00} - n_{02})\sinh\hat{\omega} - n_{0+}\cosh\hat{\omega}}.$$

The unconstrained ML estimator $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\theta = \hat{\theta})$ in (4) can be expressed in terms of $\hat{\theta}$ and $\hat{\omega}$ as

$$\hat{\boldsymbol{\beta}}^{\mathrm{T}} = \left(\hat{\theta} + \hat{\omega} + \log\left(\frac{n_{11}}{2n_{10}}\right), 2\hat{\omega} + \log\left(\frac{n_{12}}{n_{10}}\right)\right).$$

Thus, we have

$$\hat{\Delta} = \frac{\partial\hat{\boldsymbol{\beta}}^{\mathrm{T}}}{\partial\theta}(\theta = 0) = (1 + k, 2k).$$

A.2. VARIANCE CALCULATION FOR THE EB1 ESTIMATOR

We note that the total numbers of cases and controls, n_{0+} and n_{1+}, are fixed by the study design. The cell counts for genotype AA, Aa, and aa both in cases and controls follow a multinomial distribution and the cases are independent of the controls. Then the variance-covariance matrix for the cell count vector $\mathbf{n} = (n_{11}, n_{12}, n_{01}, n_{02})^{\mathrm{T}}$, denoted by B, is given by

$$B = \begin{pmatrix} n_{11}\left(1 - \frac{n_{11}}{n_{1+}}\right) & -\frac{n_{11}n_{12}}{n_{1+}} & 0 & 0 \\ -\frac{n_{11}n_{12}}{n_{1+}} & n_{12}\left(1 - \frac{n_{12}}{n_{1+}}\right) & 0 & 0 \\ 0 & 0 & n_{01}\left(1 - \frac{n_{01}}{n_{0+}}\right) & -\frac{n_{01}n_{02}}{n_{0+}} \\ 0 & 0 & -\frac{n_{01}n_{02}}{n_{0+}} & n_{02}\left(1 - \frac{n_{02}}{n_{0+}}\right) \end{pmatrix}. \tag{A.1}$$

We notice that the matrix $(\widehat{V}_{\hat{\boldsymbol{\beta}}} + \hat{\Delta}^{\mathrm{T}}\hat{\theta}\hat{\theta}^{\mathrm{T}}\hat{\Delta})^{-1}$ in (5) is of the form $(V + uu^{\mathrm{T}})^{-1}$. From matrix algebra, we know that $(V + uu^{\mathrm{T}})^{-1} = V^{-1} - (V^{-1}u)(u^{\mathrm{T}}V^{-1})/(1 + u^{\mathrm{T}}V^{-1}u)$. Then we can simplify (5) as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\mathrm{EB1}} =& \hat{\boldsymbol{\beta}} - \widehat{V}_{\hat{\boldsymbol{\beta}}}\left(\widehat{V}_{\hat{\boldsymbol{\beta}}}^{-1} - \frac{\hat{\theta}^2\widehat{V}_{\hat{\boldsymbol{\beta}}}^{-1}\hat{\Delta}^{\mathrm{T}}\hat{\Delta}\widehat{V}_{\hat{\boldsymbol{\beta}}}^{-1}}{1 + \hat{\theta}^2\hat{\Delta}\widehat{V}_{\hat{\boldsymbol{\beta}}}^{-1}\hat{\Delta}^{\mathrm{T}}}\right)(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^0) \\ =& \hat{\boldsymbol{\beta}}^0 + \frac{\hat{\theta}^2\hat{\Delta}^{\mathrm{T}}\hat{\Delta}\widehat{V}_{\hat{\boldsymbol{\beta}}}^{-1}}{1 + \hat{\theta}^2\hat{\Delta}\widehat{V}_{\hat{\boldsymbol{\beta}}}^{-1}\hat{\Delta}^{\mathrm{T}}}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^0). \end{aligned} \tag{A.2}$$

Since $\widehat{V}_{\hat{\boldsymbol{\beta}}}$ and $\hat{\Delta}$ approach zero at the rate of $O(1/n)$, we may ignore the variation in $\widehat{V}_{\hat{\boldsymbol{\beta}}}$ and $\hat{\Delta}$ and treat them as constants while computing the variance-covariance matrix of the EB1 estimator. Then the EB1 estimator can be viewed as a fixed function of the cell count vector $\mathbf{n}$. Take the derivative of the EB1 estimator with respect to the cell count vector $\mathbf{n}$. The corresponding gradient matrix A_1 is $A_1 = (\partial\hat{\boldsymbol{\beta}}_{\mathrm{EB1}}/\partial n_{11}, \partial\hat{\boldsymbol{\beta}}_{\mathrm{EB1}}/\partial n_{12}, \partial\hat{\boldsymbol{\beta}}_{\mathrm{EB1}}/\partial n_{01}, \partial\hat{\boldsymbol{\beta}}_{\mathrm{EB1}}/\partial n_{02})$. The relevant derivatives are as follows: for $j = 1, 2$,

$$\frac{\partial\hat{\boldsymbol{\beta}}_{\mathrm{EB1}}}{\partial n_{1j}} = \frac{\partial\hat{\boldsymbol{\beta}}^0}{\partial n_{1j}} + \frac{\hat{\theta}^2\hat{\Delta}^{\mathrm{T}}\hat{\Delta}\widehat{V}_{\hat{\boldsymbol{\beta}}}^{-1}}{1 + \hat{\theta}^2\hat{\Delta}\widehat{V}_{\hat{\boldsymbol{\beta}}}^{-1}\hat{\Delta}^{\mathrm{T}}}\left(\frac{\partial\hat{\boldsymbol{\beta}}}{\partial n_{1j}} - \frac{\partial\hat{\boldsymbol{\beta}}^0}{\partial n_{1j}}\right),$$

$$\begin{aligned}\frac{\partial\hat{\boldsymbol{\beta}}_{\mathrm{EB1}}}{\partial n_{0j}} =& \frac{\partial\hat{\boldsymbol{\beta}}^0}{\partial n_{0j}} + \hat{\Delta}^{\mathrm{T}}\hat{\Delta}\widehat{V}_{\hat{\boldsymbol{\beta}}}^{-1}\left[\frac{2\hat{\theta}}{(1 + \hat{\theta}^2\hat{\Delta}\widehat{V}_{\hat{\boldsymbol{\beta}}}^{-1}\hat{\Delta}^{\mathrm{T}})^2}\frac{\partial\hat{\theta}}{\partial n_{0j}}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^0)\right. \\ &\left. + \frac{\hat{\theta}^2}{1 + \hat{\theta}^2\hat{\Delta}\widehat{V}_{\hat{\boldsymbol{\beta}}}^{-1}\hat{\Delta}^{\mathrm{T}}}\left(\frac{\partial\hat{\boldsymbol{\beta}}}{\partial n_{0j}} - \frac{\partial\hat{\boldsymbol{\beta}}^0}{\partial n_{0j}}\right)\right], \end{aligned}$$

$$\frac{\partial\hat{\boldsymbol{\beta}}}{\partial n_{11}} = \frac{\partial\hat{\boldsymbol{\beta}}^0}{\partial n_{11}} = \begin{pmatrix} \frac{1}{n_{11}} + \frac{1}{n_{1+} - n_{11} - n_{12}} \\ \frac{1}{n_{1+} - n_{11} - n_{12}} \end{pmatrix},$$

$$\frac{\partial\hat{\boldsymbol{\beta}}}{\partial n_{12}} = \frac{\partial\hat{\boldsymbol{\beta}}^0}{\partial n_{12}} = \begin{pmatrix} \frac{1}{n_{1+} - n_{11} - n_{12}} \\ \frac{1}{n_{12}} + \frac{1}{n_{1+} - n_{11} - n_{12}} \end{pmatrix},$$

$$\frac{\partial\hat{\boldsymbol{\beta}}}{\partial n_{01}} = \begin{pmatrix} \frac{1}{n_{01}} + \frac{1}{n_{0+} - n_{01} - n_{02}} \\ \frac{1}{n_{0+} - n_{01} - n_{02}} \end{pmatrix},$$

$$\frac{\partial\hat{\boldsymbol{\beta}}}{\partial n_{02}} = \begin{pmatrix} -\frac{1}{n_{0+} - n_{01} - n_{02}} \\ -\frac{1}{n_{02}} - \frac{1}{n_{0+} - n_{01} - n_{02}} \end{pmatrix},$$

$$\frac{\partial\hat{\boldsymbol{\beta}}^0}{\partial n_{01}} = \begin{pmatrix} -\frac{1}{2n_{02} + n_{01}} - \frac{1}{2n_{0+} - n_{01} - 2n_{02}} \\ -\frac{2}{2n_{02} + n_{01}} - \frac{2}{2n_{0+} - n_{01} - 2n_{02}} \end{pmatrix},$$

$$\frac{\partial\hat{\boldsymbol{\beta}}^0}{\partial n_{02}} = \begin{pmatrix} -\frac{2}{2n_{02} + n_{01}} - \frac{2}{2n_{0+} - n_{01} - 2n_{02}} \\ -\frac{4}{2n_{02} + n_{01}} - \frac{4}{2n_{0+} - n_{01} - 2n_{02}} \end{pmatrix},$$

$$\frac{\partial\hat{\theta}}{\partial n_{11}} = \frac{\partial\hat{\theta}}{\partial n_{12}} = 0, \quad \frac{\partial\hat{\theta}}{\partial n_{01}} = -\frac{1}{n_{01}} - \frac{1}{2(n_{0+} - n_{01} - n_{02})},$$

$$\frac{\partial\hat{\theta}}{\partial n_{02}} = \frac{1}{2n_{02}} - \frac{1}{2(n_{0+} - n_{01} - n_{02})}.$$

The variance-covariance matrix of the EB1 estimator, denoted by Σ_{EB1}, is given by $A_1BA_1^{\mathrm{T}}$, where T represents the matrix transpose.

A.3. VARIANCE CALCULATION FOR THE EB2 ESTIMATOR

The derivation of the variance for the EB2 estimator follows that of the EB1 estimator. Note that (6) is also a function of the cell count vector **n**. Take the derivative of the EB2 estimator with respect to the vector **n**. The corresponding gradient matrix $A_2 = (\partial\hat{\boldsymbol{\beta}}_{\mathrm{EB2}}/\partial n_{11}, \partial\hat{\boldsymbol{\beta}}_{\mathrm{EB2}}/\partial n_{12}, \partial\hat{\boldsymbol{\beta}}_{\mathrm{EB2}}/\partial n_{01}, \partial\hat{\boldsymbol{\beta}}_{\mathrm{EB2}}/\partial n_{02})$. The relevant derivatives are

$$\frac{\partial\hat{\boldsymbol{\beta}}_{\mathrm{EB2}}}{\partial n_{1j}} = \frac{\partial\hat{\boldsymbol{\beta}}}{\partial n_{1j}},$$

$$\frac{\partial\hat{\boldsymbol{\beta}}_{\mathrm{EB2}}}{\partial n_{0j}} = \frac{\partial\hat{\boldsymbol{\beta}}}{\partial n_{0j}} - \left[\frac{\partial M}{\partial n_{0j}}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^0) + M\left(\frac{\partial\hat{\boldsymbol{\beta}}}{\partial n_{0j}} - \frac{\partial\hat{\boldsymbol{\beta}}^0}{\partial n_{0j}}\right)\right],$$

and

$$\frac{\partial M}{\partial n_{0j}} = \begin{pmatrix} -\dfrac{2\widehat{V}_{\hat{\boldsymbol{\beta}}_{\mathrm{Aa}}}\hat{\theta}\hat{\Delta}^2_{\mathrm{Aa}}}{(\widehat{V}_{\hat{\boldsymbol{\beta}}_{\mathrm{Aa}}} + \hat{\theta}^2\hat{\Delta}^2_{\mathrm{Aa}})^2}\dfrac{\partial\hat{\theta}}{\partial n_{0j}} & 0 \\ 0 & -\dfrac{2\widehat{V}_{\hat{\boldsymbol{\beta}}_{\mathrm{aa}}}\hat{\theta}\hat{\Delta}^2_{\mathrm{aa}}}{(\widehat{V}_{\hat{\boldsymbol{\beta}}_{\mathrm{aa}}} + \hat{\theta}^2\hat{\Delta}^2_{\mathrm{aa}})^2}\dfrac{\partial\hat{\theta}}{\partial n_{0j}} \end{pmatrix}$$

for $j = 1, 2$,

where $\partial\hat{\boldsymbol{\beta}}/\partial n_{0j}$, $\partial\hat{\boldsymbol{\beta}}^0/\partial n_{0j}$, and $\partial\hat{\theta}/\partial n_{0j}$ are computed in the section "Derivation of $\hat{\Delta}$."

The variance-covariance matrix of the EB2 estimator, denoted by Σ_{EB2}, is computed as $A_2 B A_2^{\mathrm{T}}$, where B is shown in the section "Derivation of $\hat{\Delta}$."