Biostatistics (2012), **13**, 4, *pp.* 776–790
doi:10.1093/biostatistics/kxs015
Advance Access publication on June 25, 2012

Identifying genetic marker sets associated with phenotypes via an efficient adaptive score test

TIANXI CAI*, XIHONG LIN
tcai@hsph.harvard.edu
Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA
RAYMOND J. CARROLL
Department of Statistics, Texas A&M University, College Station, TX 77843, USA

SUMMARY

In recent years, genome-wide association studies (GWAS) and gene-expression profiling have generated a large number of valuable datasets for assessing how genetic variations are related to disease outcomes. With such datasets, it is often of interest to assess the overall effect of a set of genetic markers, assembled based on biological knowledge. Genetic marker-set analyses have been advocated as more reliable and powerful approaches compared with the traditional marginal approaches (Curtis *and others*, 2005. Pathways to the analysis of microarray data. *TRENDS in Biotechnology* **23**, 429–435; Efroni *and others*, 2007. Identification of key processes underlying cancer phenotypes using biologic pathway analysis. *PLoS One* **2**, 425). Procedures for testing the overall effect of a marker-set have been actively studied in recent years. For example, score tests derived under an Empirical Bayes (EB) framework (Liu *and others*, 2007. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics* **63**, 1079–1088; Liu *and others*, 2008. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC bioinformatics* **9**, 292–2; Wu *and others*, 2010. Powerful SNP-set analysis for case-control genome-wide association studies. *American Journal of Human Genetics* **86**, 929) have been proposed as powerful alternatives to the standard Rao score test (Rao, 1948. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, **44,** 50–57). The advantages of these EB-based tests are most apparent when the markers are correlated, due to the reduction in the degrees of freedom. In this paper, we propose an adaptive score test which up- or down-weights the contributions from each member of the marker-set based on the Z-scores of their effects. Such an adaptive procedure gains power over the existing procedures when the signal is sparse and the correlation among the markers is weak. By combining evidence from both the EB-based score test and the adaptive test, we further construct an omnibus test that attains good power in most settings. The null distributions of the proposed test statistics can be approximated well either via simple perturbation procedures or via distributional approximations. Through extensive simulation studies, we demonstrate

*To whom correspondence should be addressed.

that the proposed procedures perform well in finite samples. We apply the tests to a breast cancer genetic study to assess the overall effect of the FGFR2 gene on breast cancer risk.

Keywords: Adaptive procedures; Empirical Bayes; GWAS; Pathway analysis; Score test, SNP sets.

1. Introduction

With rapid advances in high throughput technology, modern genetic studies have provided datasets that can be used to identify genetic variants associated with various diseases such as cancer, autoimmune diseases, cardiovascular diseases, and psychiatric disorders (e.g. Baum *and others* 2007; Frayling 2007; Hunter *and others* 2007; Rioux *and others* 2007; Yeager *and others* 2007; Sullivan *and others* 2008; Wallace *and others* 2008). These studies, while providing valuable resources for investigating the genetic risk of diseases, impose a grand challenge in identifying important genetic variants, due to the large number of genetic markers under investigation.

The standard approach to screening for important genetic markers is based on single-marker marginal analyses, which may suffer from low power and poor reproducibility (Vo *and others*, 2007). The lack of power is in part attributed to the fact that multiple genetic markers may relate to the phenotype simultaneously, and most have weak or no effects. To overcome such difficulties, biological knowledge-based methods have been advocated to integrate prior information into statistical learning (Brown *and others*, 2000). One useful strategy is through marker-set analysis, where a set of genetic markers are assembled based on prior knowledge such as multiple variants in a gene or a pathway. The results from marker-set analysis are often more reliable, reproducible, and powerful than the results from individual marker analysis (Curtis *and others*, 2005; Efroni *and others*, 2007; Wu *and others*, 2010), and have attractive interpretations.

To identify marker-sets associated with disease outcomes, one may test for the overall effect of a marker-set, of dimension p, on the phenotype of interest. A convenient way to do this is via a standard p-degrees of freedom (DF) Rao score test (Rao, 1948) for the significance of the global effect. However, when p is not small, such a test may have little power, especially when the markers are correlated. One may also consider combining p-values of univariate tests to obtain an overall assessment (Zaykin *and others*, 2002; Lin, 2005; Nyholt, 2004; Moskvina and Schmidt, 2008). Such an approach often works well when the signal is extremely sparse, but may not be powerful when multiple markers jointly affect the outcome. To improve the power of the standard score test, modified score tests have been derived under an empirical bayes (EB) framework (Carlin and Louis, 1997) for various types of models (Thomas *and others*, 1992; Commenges, 1994; Goeman *and others*, 2005; Liu *and others*, 2007, 2008; Wu *and others*, 2010). Such modifications achieve the power gain by taking advantage of the between-marker correlation and thus reducing the effective DF of the test. However, when the true signals are sparse and/or when the markers within a set have weak correlations, these tests also suffer from a lack of power. For example, in the case–control National Cancer Institute Cancer Genetic Markers of Susceptibility (CGEMs) study (Hunter *and others*, 2007) to be discussed in Section 4, one is interested in studying the association between the FGFR2 gene and breast cancer risk. There are 41 typed single-nucleotide polymorphisms (SNP) in FGFR2, including those within a 30-kb region of the gene. As the number of SNPs is not small and many of these SNPs have weak linkage disequilibrium (LD), i.e. week correlation, the standard score test and the EB-based test will suffer loss of power.

To overcome these difficulties, we propose an adaptive score test procedure that incorporates the strength of the signals from the individual markers within a set. The adaptive procedure, defined in Section 2, rescales the marker values by the Z-score of an initial estimator of their association with the outcome and thus down-weights the non-informative markers. We study its theoretical distribution and provide simple procedures for approximating the null distribution of the test statistic. Our numerical studies

suggest that the adaptive test performs well with respect to both empirical size and power when the true signals are relatively sparse and the markers are weakly correlated. However, when the true signals are not sparse and the correlation among the markers is high, the adaptive procedure suffers from the variability in the initial estimator and thus may perform worse compared with the EB score test. In Section 3, we derive an automatic omnibus procedure that combines information from both types of score tests. In Section 4, we examine the overall effect of the FGFR2 gene on the risk of breast cancer using the data from the CGEMS study. Findings from our numerical studies in Section 5 indicate that the omnibus test pays little price for selecting between the EB and adaptive tests. Some concluding remarks are given in Section 6.

2. METHODS

2.1 *Data structure and standard score tests*

Suppose data consist of n independent and identically distributed random vectors, $\{(Y_i, \mathbf{V}_i, \mathbf{U}_i), i = 1, \ldots, n\}$, where Y_i is the response, $\mathbf{V}_i = (V_{i1}, \ldots, V_{ip})^{\mathrm{T}}$ is the set of new markers under investigation, and $\mathbf{U}_i = (U_{i1}, \ldots, U_{ip_u})$ represents other covariates such as confounders with $U_{i1} = 1$, for the ith subject. To test the overall effect of the marker set $\mathbf{V}$ on the response Y conditional on $\mathbf{U}$, e.g. the overall effect of the 41 SNPs in the FGFR2 gene on breast cancer risk, we consider the quasilikelihood model (McCullagh and Nelder, 1989) with

$$E(Y_i \mid \mathbf{W}_i) = \mu(\boldsymbol{\theta}_0^{\mathrm{T}}\mathbf{W}_i) = \mu(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{U}_i + \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{V}_i), \quad \operatorname{var}(Y_i \mid \mathbf{W}_i) = \sigma^2\mathcal{V}(\boldsymbol{\theta}_0^{\mathrm{T}}\mathbf{W}_i), \tag{2.1}$$

where $\mathbf{W}_i = (\mathbf{U}_i^{\mathrm{T}}, \mathbf{V}_i^{\mathrm{T}})^{\mathrm{T}}$, $\mu(\cdot)$, and $\mathcal{V}(\cdot)$ are pre-specified mean and variance functions, and $\boldsymbol{\theta}_0 = (\boldsymbol{\alpha}_0^{\mathrm{T}}, \boldsymbol{\beta}_0^{\mathrm{T}})^{\mathrm{T}}$ is the vector of unknown effects for $\mathbf{U}$ and $\mathbf{V}$. Common examples of model (2.1) includes linear, logistic, Gamma, and Poisson regression.

Our primary interest lies in testing whether the set of genetic markers $\mathbf{V}$ is associated with Y after adjusting for confounders $\mathbf{U}$. That is, we aim to test the null hypothesis,

$$H_0: \quad \boldsymbol{\beta}_0 = 0. \tag{2.2}$$

In standard statistical theory, score-type testing first fits the model under H_0, producing an estimate $\tilde{\boldsymbol{\alpha}}_{\mathrm{QL}}$ that is the solution to the quasilikelihood score equation under H_0, defined as follows. Let $\dot{\mu}(x)$ be the derivative of $\mu(x)$, $\mathcal{G}_1(x) = \dot{\mu}(x)/\mathcal{V}(x)$ and $\mathcal{G}_2(x) = \{\dot{\mu}(x)\}^2/\mathcal{V}(x)$. Then $\tilde{\boldsymbol{\alpha}}_{\mathrm{QL}}$ solves

$$n^{-1}\sum_{i=1}^{n}\mathcal{G}_1(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{U}_i)\{Y_i - \mu(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{U}_i)\}\mathbf{U}_i = 0. \tag{2.3}$$

An estimate of σ^2 under H_0 is

$$\hat{\sigma}^2 = (n - p_u)^{-1}\sum_{i=1}^{n}\{Y_i - \mu(\tilde{\boldsymbol{\alpha}}_{\mathrm{QL}}^{\mathrm{T}}\mathbf{U}_i)\}^2/\mathcal{V}(\tilde{\boldsymbol{\alpha}}_{\mathrm{QL}}^{\mathrm{T}}\mathbf{U}_i). \tag{2.4}$$

The standardized score statistic is then defined as

$$\tilde{\mathbf{S}}_v = n^{-1/2}\sum_{i=1}^{n}\mathcal{G}_1(\tilde{\boldsymbol{\alpha}}_{\mathrm{QL}}^{\mathrm{T}}\mathbf{U}_i)\{Y_i - \mu(\tilde{\boldsymbol{\alpha}}_{\mathrm{QL}}^{\mathrm{T}}\mathbf{U}_i)\}\mathbf{V}_i. \tag{2.5}$$

There are two ways to form a test statistic based on (2.5). Let $\hat{\mathbb{C}}_{ww} = n^{-1}\sum_{i=1}^{n}\mathcal{G}_2(\tilde{\boldsymbol{\alpha}}_{\mathrm{QL}}^{\mathrm{T}}\mathbf{U}_i)\mathbf{W}_i\mathbf{W}_i^{\mathrm{T}}$, partition $\hat{\mathbb{C}}_{ww}$ with top left block $p_u \times p_u$ matrix $\hat{\mathbb{C}}_{uu}$, bottom right block $p \times p$ matrix $\hat{\mathbb{C}}_{vv}$, and non-diagonal blocks

$\hat{\mathbb{C}}_{uv}$ and $\hat{\mathbb{C}}_{vu}$, and $\hat{\mathbb{C}}_{v|u} = \hat{\mathbb{C}}_{vv} - \hat{\mathbb{C}}_{vu}\hat{\mathbb{C}}_{uu}^{-1}\hat{\mathbb{C}}_{uv}$. The standard score test statistic is

$$\tilde{Q} = \tilde{\mathbf{S}}_v^{\mathrm{T}}(\hat{\sigma}^2\hat{\mathbb{C}}_{v|u})^{-1}\tilde{\mathbf{S}}_v, \tag{2.6}$$

and under H_0, it is asymptotically central χ_p^2. The local power of this score test is also well known (Harris and Peers, 1980; Rao, 2005). See Martinez *and others* (2010) for a recent explication.

The simplicity of the standard score test is convenient, but comes at potentially considerable cost because of the p-DF. In genomic applications, p is typically not small and some markers in a set have no effects. For example, there are $p = 41$ SNPs in the FGFR2 gene. Such a large DF carries with it a potential loss of power, as we see in our simulations (Section 5).

Another way to use the score statistic (2.5) is to employ an EB framework by imposing a working assumption that the $\{\beta_{0j}, j = 1, \ldots, p\}$ are independent and follow an arbitrary distribution with mean 0 and variance τ. We hence test the null hypothesis (2.2) by testing whether $\tau = 0$. The working distributional assumption, while useful for deriving a globally valid testing procedure, is not required to hold. This leads to the score test

$$\hat{Q} = \|\tilde{\mathbf{S}}_v\|_2^2 = \sum_{\ell=1}^{p} \tilde{S}_{v\ell}^2, \tag{2.7}$$

where for any vector $\mathbf{a}$, a_ℓ denotes the ℓth element of $\mathbf{a}$, and $\|\mathbf{a}\|_q$ denotes the L_q vector norm. These types of score test statistics have been shown previously to be powerful alternatives to the standard score test under various settings (Goeman *and others*, 2005; Liu *and others*, 2007; Kwee *and others*, 2008; Wu *and others*, 2010). The distribution of (2.7) is given in Section 2.3.

2.2 *An efficient adaptive score test*

As shown in the simulation Section 5, the EB-based score test (2.7) is quite powerful when the signal is not sparse. However, for settings when the signal is sparse and the markers are weakly correlated with each other, this test has limited power due to the high DF paid for the non-informative markers. To overcome such a difficulty, we propose an adaptive score statistic that incorporates information on the signal strength of each marker. Specifically, we first obtain an initial root-n consistent estimator of $\boldsymbol{\beta}_0$, denoted by $\hat{\boldsymbol{\beta}}$. Without loss of generality, we assume that the vector of element-wise variances of $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$, $\boldsymbol{\kappa}^2$, can be consistently estimated by $\hat{\boldsymbol{\kappa}}^2$. Our proposed rescaling factor is $\hat{\mathbf{Z}} = n^{1/2}\hat{\boldsymbol{\beta}} \oslash \hat{\boldsymbol{\kappa}}$, where $\oslash$ represents element-wise division. Here, $\hat{Z}_l$ is the Z-statistic for testing $\beta_{0\ell} = 0$, which reflects the strength of the signal $\beta_{0\ell}$ relative to the noise. The adaptive score test is then constructed by rescaling $\tilde{\mathbf{S}}_v$ in $\hat{Q}$ element-wise by $\hat{\mathbf{Z}}$ such that non-informative markers will be down-weighted towards zero, thus effectively eliminating these markers from the score statistic. We define our adaptive score test statistic as

$$\hat{Q}_{\mathcal{A}} = \hat{Q}_{\mathcal{A}}(\hat{\mathbf{Z}}) = \|\tilde{\mathbf{S}}_v \odot \hat{\mathbf{Z}}\|_2^2 = \sum_{\ell=1}^{p} (\tilde{S}_{v\ell}\hat{Z}_\ell)^2, \tag{2.8}$$

where $\odot$ denotes element-wise product. If the ℓth component $\hat{\beta}_\ell$ of $\hat{\boldsymbol{\beta}}$ is not significantly different from zero, this clearly shows that the ℓth marker plays little role in the test statistic.

For the initial estimator, one may obtain $\hat{\boldsymbol{\beta}}$ as the standard quasilikelihood estimator. Specifically, let $\hat{\boldsymbol{\theta}}_{\mathrm{QL}} = (\hat{\boldsymbol{\alpha}}_{\mathrm{QL}}^{\mathrm{T}}, \hat{\boldsymbol{\beta}}_{\mathrm{QL}}^{\mathrm{T}})^{\mathrm{T}}$ be the solution to $\bar{\mathbf{S}}(\boldsymbol{\theta}) = 0$, where

$$\bar{\mathbf{S}}(\boldsymbol{\theta}) = \begin{bmatrix} \bar{\mathbf{S}}_u(\boldsymbol{\theta})_{p_u \times 1} \\ \bar{\mathbf{S}}_v(\boldsymbol{\theta})_{p \times 1} \end{bmatrix} = n^{-1} \sum_{i=1}^{n} \mathcal{G}_1(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{W}_i)\{Y_i - \mu(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{W}_i)\}\mathbf{W}_i \tag{2.9}$$

is the quasilikelihood score equation under the alternative. A simple choice for the variance estimator of $n^{1/2}(\hat{\beta}_\ell - \beta_{0\ell})$ is

$$\hat{\kappa}_\ell^2 = \hat{\sigma}^2 \hat{\mathbf{B}}_{v\ell}^{\mathrm{T}} \hat{\mathbb{C}}_{ww} \hat{\mathbf{B}}_{v\ell}, \tag{2.10}$$

where $\hat{\sigma}^2$ could be estimated under H_0 as in (2.4), $\hat{\mathbf{B}}_{v\ell}$ is the ℓth row vector of $[-\hat{\mathbb{C}}_{v|u}^{-1}\hat{\mathbb{C}}_{vu}\hat{\mathbb{C}}_{uu}^{-1}, \hat{\mathbb{C}}_{v|u}^{-1}]$ and $\hat{\mathbb{C}}_{v|u} = \hat{\mathbb{C}}_{vv} - \hat{\mathbb{C}}_{vu}\hat{\mathbb{C}}_{uu}^{-1}\hat{\mathbb{C}}_{uv}$. Then one may let $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\mathrm{QL}}$ and $\hat{\mathbf{Z}} = \hat{\boldsymbol{\beta}}_\ell \oslash \hat{\boldsymbol{\kappa}}$ in the adaptive score test (2.8), where $\hat{\boldsymbol{\kappa}} = (\hat{\kappa}_1, \ldots, \hat{\kappa}_p)^{\mathrm{T}}$.

When p is not small and the signals are moderate or weak, $\hat{\boldsymbol{\beta}}_{\mathrm{QL}}$ may have large variation and thus may lead to power loss for the test rescaled with $\hat{\boldsymbol{\beta}}_{\mathrm{QL}}$. To overcome such difficulties, we propose to improve the power by considering a ridge penalized quasilikelihood (PQL) estimator $\hat{\boldsymbol{\theta}}_{\mathrm{RQL}}(\lambda) = \{\hat{\boldsymbol{\alpha}}_{\mathrm{RQL}}(\lambda)^{\mathrm{T}}, \hat{\boldsymbol{\beta}}_{\mathrm{RQL}}(\lambda)^{\mathrm{T}}\}^{\mathrm{T}}$, where $\hat{\boldsymbol{\theta}}_{\mathrm{RQL}}(\lambda)$ is the solution to $\bar{\mathbf{S}}(\boldsymbol{\theta}) - \lambda(\mathbf{0}_{1\times p_u}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}} = 0$, and λ is a tuning parameter with $\lambda \to \lambda_0 \geqslant 0$. The estimator $\hat{\boldsymbol{\beta}}_{\mathrm{QL}}$ corresponds to the maximum PQL estimator under the random effects model $\boldsymbol{\beta}_0 \sim N(\mathbf{0}, (n\lambda)^{-1}\mathbb{I}_{p\times p})$ (Green, 1987; Breslow and Clayton, 1993). In practice, one may choose an optimal $\hat{\lambda}$ based on procedures such as generalized cross-validation. The adaptive score test (2.8) may also be constructed based on $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\mathrm{RQL}}(\hat{\lambda})$ along with the variance estimators $\hat{\kappa}_\ell^2 = \hat{\sigma}^2\{\hat{\mathbf{B}}_{v\ell}^{(\hat{\lambda})}\}^{\mathrm{T}}\hat{\mathbb{C}}_{ww}\hat{\mathbf{B}}_{v\ell}^{(\hat{\lambda})}$, where $\hat{\mathbf{B}}_{v\ell}^{(\hat{\lambda})}$ is the ℓth row vector of $[-\{\hat{\mathbb{C}}_{v|u} + \hat{\lambda}\mathbb{I}_{p\times p}\}^{-1}\hat{\mathbb{C}}_{vu}\hat{\mathbb{C}}_{uu}^{-1}, \{\hat{\mathbb{C}}_{v|u} + \hat{\lambda}\mathbb{I}_{p\times p}\}^{-1}]$.

2.3 *Distribution of the test statistic under the null and under the local alternative*

In the Appendix of the supplementary material available at *Biostatistics* online, we derive the distributions of the test statistics (2.7)–(2.8) under H_0 and more generally under the local alternative $H_{1n} : \boldsymbol{\beta}_0 = n^{-1/2}\mathbf{b}_0$. Define

$$\mathbb{C}_{ww} = \begin{bmatrix} \mathbb{C}_{uu} & \mathbb{C}_{uv} \\ \mathbb{C}_{vu} & \mathbb{C}_{vv} \end{bmatrix} = \begin{bmatrix} E\{\mathcal{G}_2(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{U})\mathbf{U}\mathbf{U}^{\mathrm{T}}\} & E\{\mathcal{G}_2(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{U})\mathbf{U}\mathbf{V}^{\mathrm{T}}\} \\ E\{\mathcal{G}_2(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{U})\mathbf{V}\mathbf{U}^{\mathrm{T}}\} & E\{\mathcal{G}_2(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{U})\mathbf{V}\mathbf{V}^{\mathrm{T}}\} \end{bmatrix}. \tag{2.11}$$

$\mathbb{C}_{v|u} = \mathbb{C}_{vv} - \mathbb{C}_{vu}\mathbb{C}_{uu}^{-1}\mathbb{C}_{uv}$, $\mathbf{z}_0 = \mathbf{b}_0 \oslash \boldsymbol{\kappa}$, and $\mathbf{s}_0 = \mathbb{C}_{v|u}\mathbf{b}_0$. Furthermore, let $\boldsymbol{\varepsilon}_w$ denote a $N(\mathbf{0}, \sigma^2\mathbb{C}_{ww})$ random vector, $\mathbf{B}_{v\ell}^{(\lambda_0)}$ and $\mathbf{A}_{v\ell}$, respectively, denote the ℓ^{th} row of $\mathbb{B}_v^{(\lambda_0)} = [-\{\mathbb{C}_{v|u} + \lambda_0\mathbb{I}_{p\times p}\}^{-1}\mathbb{C}_{vu}\mathbb{C}_{uu}^{-1}, \{\mathbb{C}_{v|u} + \lambda_0\mathbb{I}_{p\times p}\}^{-1}]$ and $\mathbb{A}_v = [-(\mathbb{C}_{uv}^{\mathrm{T}}\mathbb{C}_{uu}^{-1})^{\mathrm{T}}, \mathbb{I}_{p\times p}]$. Then $\hat{Q}_{\mathcal{A}}$ converges to

$$\begin{aligned} \mathcal{Q}_{\mathcal{A}}(\mathbf{b}_0) &= \|(\mathbf{s}_0 + \mathbb{A}_v\boldsymbol{\varepsilon}_w) \odot (\mathbf{b}_0 + \mathbb{B}_v^{(\lambda_0)}\boldsymbol{\varepsilon}_w) \oslash \boldsymbol{\kappa}\|_2^2 \\ &= \sum_{\ell=1}^{p} (s_{0\ell} + \boldsymbol{\varepsilon}_w^{\mathrm{T}}\mathbf{A}_{v\ell})^2 (z_{0\ell} + \kappa_\ell^{-1}\boldsymbol{\varepsilon}_w^{\mathrm{T}}\mathbf{B}_{v\ell}^{(\lambda_0)})^2 \end{aligned} \tag{2.12}$$

in distribution. The same arguments as given in the Appendix of the supplementary material available at *Biostatistics* online can be used to show that under H_{1n}, the EB-score test statistic (2.7) converges in distribution to $\mathcal{Q}(\mathbf{b}_0) = \|\mathbf{s}_0 + \mathbb{A}_v\boldsymbol{\varepsilon}_w\|_2^2$. Under H_0, $\mathbf{b}_0 = \mathbf{0}$ and thus $\hat{Q}_{\mathcal{A}}$ in (2.8) and $\hat{Q}$ in (2.7) converge in distribution to $\mathcal{Q}_{\mathcal{A}}(\mathbf{0}) = \sum_{l=1}^{p} \mathcal{Z}_l^2(\mathbf{A}_{vl}^{\mathrm{T}}\boldsymbol{\varepsilon}_w)^2$ and $\mathcal{Q}(\mathbf{0}) = \sum_{l=1}^{p}(\mathbf{A}_{vl}^{\mathrm{T}}\boldsymbol{\varepsilon}_w)^2$, respectively, where $\mathcal{Z}_l = \kappa_l^{-1}\boldsymbol{\varepsilon}_w^{\mathrm{T}}\mathbf{B}_{vl}^{(\lambda_0)}$.

2.4 *Implementation*

In general, it is straightforward to approximate the null distribution of $\hat{Q}_{\mathcal{A}}$ in (2.8) via perturbation by repeatedly generating realizations of $\hat{Q}_{\mathcal{A}}^* = \|(\hat{\mathbb{A}}_v\hat{\boldsymbol{\varepsilon}}_w^*) \odot (\hat{\mathbb{B}}_v^{(\hat{\lambda})}\hat{\boldsymbol{\varepsilon}}_w^*) \oslash \hat{\boldsymbol{\kappa}}\|_2^2$, where $\hat{\mathbb{A}}_v$ and $\hat{\mathbb{B}}_v^{(\hat{\lambda})}$ are defined by replacing $\mathbb{C}_{ww}^{(\lambda_0)}$ with its empirical counterpart, and where $\hat{\boldsymbol{\varepsilon}}_w^* = n^{-1/2}\sum_{i=1}^{n}\mathcal{G}_1(\tilde{\boldsymbol{\alpha}}_{\mathrm{QL}}^{\mathrm{T}}\mathbf{U}_i)\{Y_i - \mu(\tilde{\boldsymbol{\alpha}}_{\mathrm{QL}}^{\mathrm{T}}\mathbf{U}_i)\}\mathbf{W}_i\mathcal{N}_i$ and $\underline{\mathcal{N}} = (\mathcal{N}_1, \cdots, \mathcal{N}_n)^{\mathrm{T}}$ is a vector of independent $N(0, 1)$ random variables. This is because asymptotically, the distribution of $\hat{\boldsymbol{\varepsilon}}_w^*$ given the observed data and the unconditional distribution of $\boldsymbol{\varepsilon}_w$ are asymptotically the same. Mimicking the Satterthwaite approximation, we find that in our

numerical studies, the null distribution of $\sqrt{\hat{Q}_{\mathcal{A}}}$ can be well approximated by a rescaled χ^2 distribution, i.e. $\sqrt{\hat{Q}_{\mathcal{A}}} \sim c_0 \chi^2_{d_0}$. The scale parameter c_0 and the degrees of freedom d_0 can be estimated by matching the first two moments of $\sqrt{\hat{Q}^*_{\mathcal{A}}}$. Similarly, the null distribution of the standard score test $\hat{Q}$ in (2.7) can be approximated by $\hat{Q}^* = \|\hat{\mathbb{A}}_v \hat{\boldsymbol{\epsilon}}^*_w\|_2^2$.

2.5 *Theoretical differences between the tests*

Some insight as to the difference between our adaptive test and the ordinary score test can be gained by considering the case that there are no additional confounders, so that $\mathbf{U} = 1$ and $\lambda_0 = 0$. In this case, $\mathbb{A}_v = [\mathbf{0}_{p\times 1}, \mathbb{I}_{p\times p}]$, $\mathbb{B}_v^{(\lambda_0)} = [\mathbf{0}_{p\times 1}, a_0^{-2}\boldsymbol{\Sigma}_v \mathbb{I}_{p\times p})^{-1}]$, $\mathbf{s}_0 = a_0^2 \boldsymbol{\Sigma}_v \mathbf{b}_0$ and $a_0^2 = \mathcal{G}_2[\mu^{-1}\{E(Y)\}]$, where $\boldsymbol{\Sigma}_v = \mathrm{var}(\mathbf{V})$. Since $E(\mathbf{V}) = 0$, $\mathbb{C}_{vu} = \mathbb{C}_{uv}^{\mathrm{T}} = \mathbf{0}_{p\times 1}$. Thus $\mathbb{B}_v^{(\lambda_0)}\boldsymbol{\epsilon}_{\mathbf{w}} \sim N(0, \sigma^2 a_0^{-2}\boldsymbol{\Sigma}_v^{-1})$ and $\mathbb{A}_v\boldsymbol{\epsilon}_{\mathbf{w}} \sim N(0, \sigma^2 a_0^2 \boldsymbol{\Sigma}_v)$. Now, let ξ_l^2 denote the lth diagonal element of $\boldsymbol{\Sigma}_v^{-1}$, $\mathbb{D}_v = \mathrm{diag}(\xi_1, \ldots, \xi_p)$, write $\boldsymbol{\Sigma}_v^{-1} = \mathbb{D}_v \mathbb{R}_v \mathbb{D}_v$ and let $\mathcal{Z}(\mathbf{z}_0) \sim N(\mathbf{z}_0, \mathbb{R}_v)$. It follows that $\kappa_l = \sigma a_0^{-1}\xi_l$ and $(\mathbf{b}_0 + \mathbb{B}_v^{(\lambda_0)}\boldsymbol{\epsilon}_{\mathbf{w}}) \oslash \boldsymbol{\kappa} \sim \mathcal{Z}(\mathbf{z}_0)$. Therefore, we may simplify the distribution of $\mathcal{Q}_{\mathcal{A}}(\mathbf{b}_0)$ as

$$(a_0\sigma)^{-2}\mathcal{Q}_{\mathcal{A}}(\mathbf{b}_0) = \|\{\mathbb{D}_v^{-1}\mathbb{R}_v^{-1}\mathcal{Z}(\mathbf{z}_0)\} \odot \mathcal{Z}(\mathbf{z}_0)\|_2^2.$$

Furthermore, if $\mathbf{V}$ is uncorrelated, then $\mathbb{R}_v = \mathbb{I}_{p\times p}$, $\xi_\ell^2 = \mathrm{var}(V_\ell)^{-1}$ and $(a_0\sigma)^{-2}\mathcal{Q}_{\mathcal{A}}(\mathbf{b}_0)$ are equivalent to $\sum_{\ell=1}^p \mathrm{var}(V_\ell)\mathcal{Z}_\ell(\mathbf{z}_0)^4$. On the other hand, the distribution of (2.7) scaled by $(a_0\sigma)^{-2}$ can be written as $\sum_{\ell=1}^p \mathrm{var}(V_\ell)\mathcal{Z}_\ell(\mathbf{z}_0)^2$. Thus, for the orthogonal case, asymptotically, the EB score statistic $\hat{Q}$ is a weighted sum of p independent 1-DF χ^2 random variables with non-centrality parameters $\mathbf{z}_0^2$; whereas the adaptive score statistic is a weighted sum of p-independent *squared* 1-DF χ^2 random variables with non-centrality parameters $\mathbf{z}_0^2$.

3. An efficient adaptive omnibus test

3.1 *Theoretical local power calculations*

To compare the performance of the EB score test and the adaptive test, we consider the simple setting of linear regression with $\sigma^2 = 1$, $\mathbf{U} = 1$ for an intercept and $\mathbf{V}$ is multivariate normal with mean zero, unit variance, and a common correlation ρ. In Figure 1, we present the power curves under the local alternative for $\rho = 0.0$, 0.2 and 0.5 for two extreme settings: (i) when signals are sparse with $\mathbf{b}_0 = (b, 0, \ldots, 0)^{\mathrm{T}}$ and (ii) when all covariates contribute equally with $\mathbf{b}_0 = (b, \ldots, b)^{\mathrm{T}}/\sqrt{p}$. It appears that the adaptive procedure outperforms the EB score test under the sparse setting, but the phenomenon is reversed for the setting where the signals are equally contributed from all p covariates. The relative performance of the adaptive and EB score test procedures also varies with the correlation ρ. The lower the correlation is, the more advantage the adaptive procedure has.

3.2 *Omnibus test and implementation*

In general, the relative performance of the EB-based score test and the adaptive procedure depends on the sparsity of $\mathbf{b}_0$ and the between-marker correlation. In practice, without prior information on these factors, it is unclear which procedure should be chosen for a given dataset. To overcome this difficulty, we propose to automatically combine evidence between the EB score test and the adaptive test by taking the minimum p-value and comparing to its null counterpart. Specifically, let $\hat{P} = \hat{S}(\hat{Q})$ and $\hat{P}_{\mathcal{A}} = \hat{S}_{\mathcal{A}}(\hat{Q}_{\mathcal{A}})$ be the respective p-values based on the EB score test and the adaptive score test, where

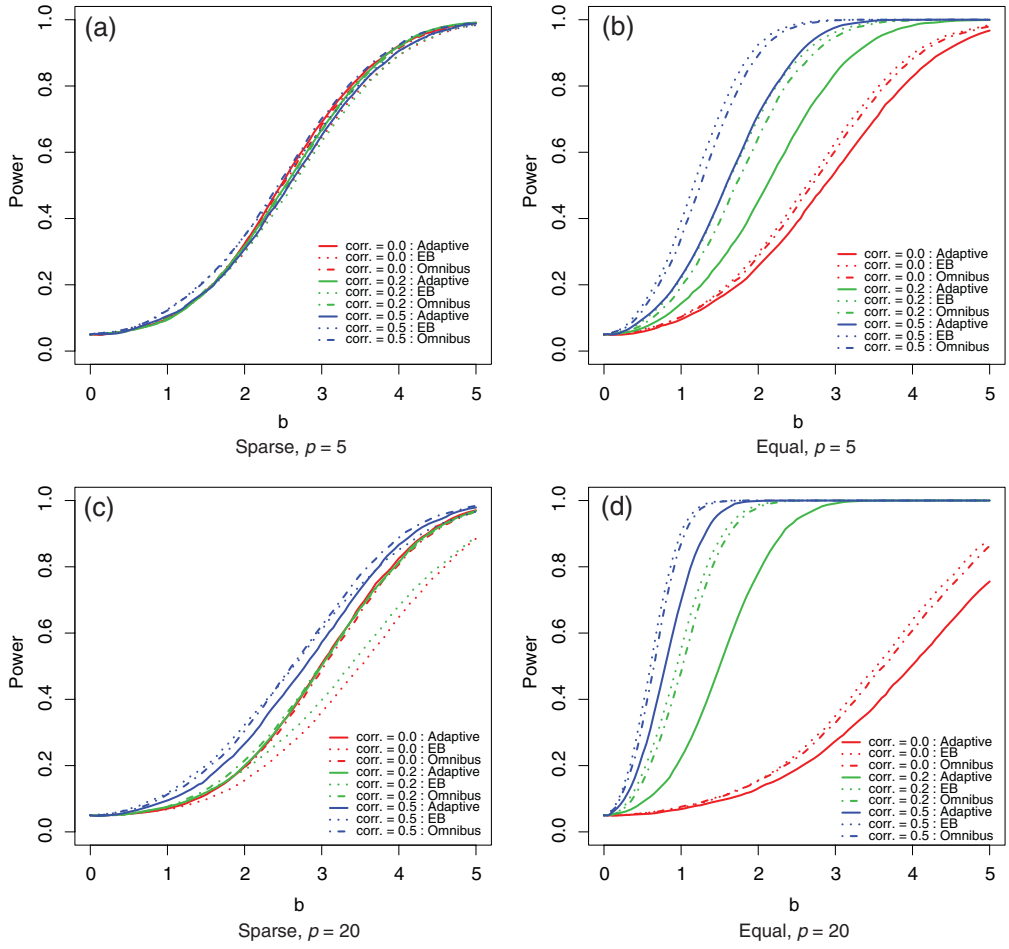

Fig. 1. Theoretical power curve for the adaptive (solid curves), EB-score (dotted curves), and omnibus combining $\hat{Q}$ and $\hat{Q}_{\mathcal{A}}$ (dot dashed curves) tests under local alternatives with various levels of correlations (corr.): 0.0 (thin gray curves); 0.2 (black curves); and 0.5 (thick gray curves).

$\hat{S}(q)$ and $\hat{S}_{\mathcal{A}}(q)$ are estimators of $S(q) = \text{pr}\{\mathcal{Q}(0) > q\}$ and $S_{\mathcal{A}}(q) = \text{pr}\{\mathcal{Q}_{\mathcal{A}}(0) > q\}$, respectively. Then the omnibus test is based on the minimum p-value, $\hat{P}_{\min} = \min(\hat{P}, \hat{P}_{\mathcal{A}})$ which converges in distribution to $\mathcal{P}_{\min} = \min[S\{\mathcal{Q}(0)\}, S_{\mathcal{A}}\{\mathcal{Q}_{\mathcal{A}}(0)\}]$ under H_0. Our simulation studies in Section 5 indicate that the omnibus test pays a relatively low price with respect to power for utilizing two tests. Our theoretical power analysis under the aforementioned two settings also supports this finding as shown in Figure 1.

In practice, the null distribution of $\hat{P}_{\min}$ can be approximated easily via perturbation methods. Let $\{(\hat{Q}^{(b)}, \hat{Q}_{\mathcal{A}}^{(b)}), b = 1, \ldots, B\}$ denote B perturbed realization of $\{(\hat{Q}^*, \hat{Q}_{\mathcal{A}}^*), b = 1, \ldots, B\}$, where for each b, $\hat{Q}^{(b)}$ and $\hat{Q}_{\mathcal{A}}^{(b)}$ are perturbed with the same set of independent normal vector $\underline{\mathcal{N}}^{(b)}$. Then the null distribution of $\hat{P}_{\min}$ can be approximated by the empirical distribution of $\hat{P}_{\min}^{(b)} = \min\{\hat{S}(\hat{Q}^{(b)}), \hat{S}_{\mathcal{A}}(\hat{Q}_{\mathcal{A}}^{(b)})\}$ across $b = 1, \ldots, B$, where $\hat{S}(\cdot)$ and $\hat{S}_{\mathcal{A}}(\cdot)$ are obtained by fitting rescaled χ^2 distributions to $\{\hat{Q}^{(b)}\}$ and $\{\sqrt{\hat{Q}_{\mathcal{A}}^{(b)}}\}$. When $\hat{P}_{\min}$ is very small, it may be challenging to obtain its p-value using resampling, because a large B would be required to ensure adequate approximation. For such settings, we propose to

approximate the null distribution of $\Phi^{-1}(\hat{P}_{\min})$ using a normal mixture. Specifically, we fit a ν_0-population normal mixture, $\sum_{\nu=1}^{\nu_0} \pi_\nu N(\mu_\nu, \sigma_\nu^2)$, to $\{\hat{P}_{\min}^{(b)}, b = 1, \ldots, B\}$ and obtain maximum likelihood estimates for $\{(\pi_\nu, \mu_\nu, \sigma_\nu), \nu = 1, \ldots, \nu_0\}$, denoted by $\{(\hat{\pi}_\nu, \hat{\mu}_\nu, \hat{\sigma}_\nu), \nu = 1, \ldots, \nu_0\}$. Then the p-value can be estimated by $1 - \sum_{\nu}^{\nu_0} \hat{\pi}_\nu \Phi[\hat{\sigma}_\nu^{-1}\{\Phi^{-1}(\hat{P}_{\min}) - \hat{\mu}_\nu\}]$. Through our empirical studies, we find that this approximation works well and hence could be useful when aiming to control for low type I error rates. Similar strategies could be used to approximate the distribution of $\Phi^{-1}(\hat{P}_{\mathcal{A}})$ to improve the χ^2 approximation, especially for the tail probabilities. Here, the number of components ν_0 can be either pre-specified or chosen adaptively using criteria such as BIC. In practice, we find that $\nu_0 = 3$ works well for approximating the distribution of $\Phi^{-1}(\hat{P}_{\min})$.

4. Example: the FGFR2 gene for the risk of breast cancer

We applied our proposed procedures to assess the association between the FGFR2 gene and the risk of sporadic postmenopausal breast cancer using the genome-wide association studies (GWAS) data of the Nurses' Health Study, which was part of the aforementioned CGEMS study (Hunter *and others*, 2007). Using the Illumina HumanHap500 array, this study initially genotyped 1183 women with postmenopausal invasive breast cancer and 1185 individually matched controls. Data for analysis consist of 1091 cases and 1110 controls with complete information. Among the loci reported as potentially associated with breast cancer in Hunter *and others* (2007) include several SNPs in FGFR2 or its intron 2. Here, we examine the overall effect of the FGFR2 gene, consisting of 41 typed SNPs, including those within the 30-kb region of the gene, on the risk of breast cancer. The analysis adjusted for age group, hormone usage, age at menarche, and the first 4 eigenvectors generated from EIGENSTRAT principal components analysis (Price *and others*, 2006) to account for population stratification.

We first fit the data with marginal logistic regression models with one SNP at a time adjusting for these covariates. The log odds ratio estimates along with their 95% confidence intervals obtained from the 41 marginal models are shown in Supplementary material, Figure S1 (see supplementary material available at *Biostatistics* online). Out of these 41 SNPs, 14 SNPs have marginal p-value < 0.05 and 3 SNPs, rs2420946, rs1219648, rs2981579, with p-value $< 10^{-5}$. The SNP rs1219648 has been previously shown to be highly associated with increased risk of breast cancer while both rs2420946 and rs2981579 are in high LD with rs1219648 (Hunter *and others*, 2007). An experimental rationale was presented in indicating that this SNP is part of a haplotype that increases risk for ER+ breast cancer by increasing FGFR2 transcription.

To assess the overall effect of the gene, we employed the aforementioned procedures including the univariate test using the minimum of these p-values. Since this gene is highly associated with breast cancer risk, we used 100 000 perturbation samples along with the normal mixture approach to approximate the p-values. The univariate test gives an overall p-value of 4.0×10^{-5}. On the other hand, our adaptive test gives a p-value of 4.3×10^{-8}, the EB score test a similar p-value of 7.2×10^{-7} while the the standard p-DF score test has no power in detecting the signal with a p-value of 0.12. The omnibus test combining $\hat{Q}$ and $\hat{Q}_{\mathcal{A}}$ yielded a p-value of 1.8×10^{-7} based on the normal mixture approximation. However, a larger number of perturbation samples would be required to obtain more accurate estimate of the p-value for these tests.

5. Simulation studies

5.1 *Setup and null case*

We conducted simulation studies to assess the performance of the proposed score test. For simplicity, we considered the setting that $U = 1$. To mimic the GWAS setting, we generated $\mathbf{V}$ based on the LD structure of two genes: (i) the ASAH1 gene with high LD and (ii) the FGFR2 gene with moderate to low LD. Based on

Table 1. *Empirical sizes* (*in* %) *at target sizes of* 1% *and* 5% *for the score tests when* $\mathbf{V}$ *is simulated based on the* 17 *tagSNPs of the ASAH1 gene and* 31 *tagSNPs of the FGFR2 gene on the illumina chip. Here* $\tilde{Q}$ *is the p-DF score test in* (2.6), $\hat{Q}$ *is the test* (2.7), $\hat{Q}_{\mathcal{A}}$ *is the test at* (2.8) *with initial ridge estimate*, "*Uni*" *is the test that is based on the minimum p-value across p univariate tests, and* $\min_{A_1,\ldots,A_k}$ *represents the omnibus test that takes the minimum p-value of the tests based on* $A_1, \ldots, A_k$. *The p-values are obtained via perturbation (ptb), chi-square approximation* (χ^2), *and normal mixture approximation (Mix).*

		$\tilde{Q}$		$\hat{Q}$			$\hat{Q}_{\mathcal{A}}$			$\min_{\hat{P}_{\mathcal{A}},\hat{P}}$		$\min_{\hat{P}_{\mathcal{A}},\hat{P},\hat{P}_{\text{uni}}}$		$\min_{\hat{P},\hat{P}_{\text{uni}}}$	
n	Size	ptb	χ^2	ptb	χ^2	Uni	ptb	χ^2	Mix	ptb	Mix	ptb	Mix	ptb	Mix
ASAH1 gene															
200	1	0.3	0.2	1.1	1.3	0.8	1.1	1.6	1.3	1.1	1.2	0.9	0.8	0.9	0.8
500	1	0.4	0.3	1.0	1.4	0.8	1.1	1.5	1.3	1.0	1.1	0.7	0.7	0.8	0.8
1000	1	0.7	0.2	0.9	1.1	0.7	1.0	1.3	1.1	1.0	1.1	0.8	0.8	0.8	0.8
200	5	3.1	1.2	5.5	5.4	4.4	5.7	5.6	5.4	5.8	5.6	4.5	4.5	4.4	4.3
500	5	4.3	1.7	5.2	5.1	4.7	5.2	5.2	5.0	5.5	5.2	4.7	4.5	4.6	4.4
1000	5	3.7	1.5	4.6	4.5	4.5	5.2	5.1	4.9	5.0	4.7	4.6	4.5	4.5	4.3
FGFR2 gene															
200	1	0.1	0.1	0.7	0.9	0.7	0.6	0.9	0.8	0.6	0.6	0.7	0.7	0.7	0.7
500	1	0.3	0.5	0.9	1.1	1.0	0.9	1.2	1.0	0.9	1.0	1.1	1.0	1.1	1.1
1000	1	0.6	0.6	0.8	1.0	1.0	0.9	1.3	1.0	0.7	0.7	0.9	0.9	0.9	1.0
200	5	1.7	1.8	5.1	5.2	5.2	4.3	4.4	4.2	4.4	4.2	5.1	5.0	5.1	4.8
500	5	2.8	3.1	5.0	5.1	5.2	4.9	5.0	4.7	4.7	4.5	5.2	5.0	5.2	4.8
1000	5	3.8	3.4	4.6	4.5	5.3	5.1	5.2	4.9	5.0	4.7	5.2	5.0	5.2	4.9

the Illumina 500 K platform, we included $p = 14$ SNPs of the ASAH1 gene and $p = 31$ SNPs of the FGFR2 gene, whose LD heat maps are shown in Supplementary material, Figure S2 (see supplementary material available at *Biostatistics* online). The response variable Y is generated from the linear regression model $Y = \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{V} + \epsilon$ with $\epsilon \sim N(0, 4)$ generated independent of $\mathbf{V}$. For each configuration, we generated 4000 datasets to calculate the empirical size when the null hypothesis is true and 1000 datasets to calculate the empirical power when it is not. For each dataset, the resampling procedure was carried out with $B = 5000$ and $B = 1000$, for the null and alternative settings, respectively. We considered $n = 200$, 500, and 1000.

As a benchmark, we also report results on the univariate test, whose significance is determined by comparing the observed minimum p-value of p univariate tests to its corresponding null distribution. For each simulated dataset, we carried out the following test procedures: (i) the p-DF score test $\tilde{Q}$; (ii) the EB based score test $\hat{Q}$ in (2.7); (iii) adaptive score test in (2.8) rescaled with ridge initial estimator $\hat{Q}_{\mathcal{A}}$; and (iv) univariate test (Uni). In addition, we consider various omnibus tests based on the minimum p-value among the two or three p-values from (ii), (iii), and (iv). Note that we considered only ridge estimators as our initial estimator, since the standard quasilikelihood estimator is unstable due to the high collinearity between the SNPs. To examine how well the χ^2 distribution approximates the null distribution of $\sqrt{\hat{Q}_{\mathcal{A}}}$ and $\hat{Q}$, we also provided the empirical size and power based on the approximation. We also examined the performance of the normal mixture approximation to the distribution of the omnibus test statistics as well as the distribution of $\hat{Q}_{\mathcal{A}}$. Unless noted otherwise, p-values for all other test statistics are based on the perturbation procedure, which can conveniently account for various types of correlations.

First, to examine the validity of our proposed testing procedure in finite samples, we generated data under H_0 model with $\boldsymbol{\beta}_0 = 0$ to assess the size of the score test. As shown in Table 1, the empirical sizes of

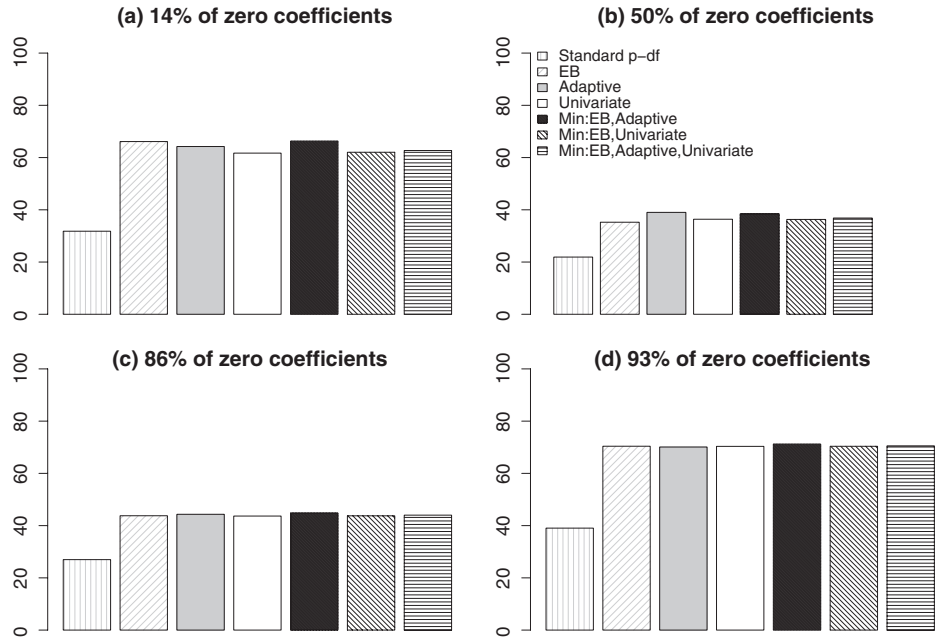

Fig. 2. Empirical power (in %) for various tests using the ASAH1 gene under local alternatives averaged over all the choices of ι's. (a) $q = 12$, $b_{\text{ASAH1}} = 3.5$ (14% sparsity); (b) $q = 7$, $b_{\text{ASAH1}} = 4.1$ (50% sparsity); (c) $q = 2$, $b_{\text{ASAH1}} = 7.1$ (86% sparsity); and (d) $q = 1$, $b_{\text{ASAH1}} = 10.6$ (93% sparsity).

the aforementioned tests at type I error rate of 1% and 5% are summarized in Table 1. Across all the configurations, the empirical sizes are close to the nominal levels for all procedures except for the standard p-DF test, which is often overly conservative due to the correlation among the $\mathbf{V}$. Furthermore, it appears that the χ^2-based approximation works reasonably well in practice for approximating the distribution of $\sqrt{\hat{Q}_{\mathcal{A}}}$ at type I error rate 5%. However, it appears that at a lower error rate of 1%, the χ^2 approximation is slightly anti-conservative, while the normal mixture approximation appears to provide a better approximation and works well for approximating the distribution of other minimum p-value test statistics.

5.2 *Power comparisons*

For empirical power analyses, we let $\boldsymbol{\beta}_0 = n^{-1/2} b_{\mathcal{G}} [\mathbf{0}_{1\times\iota}, \mathbf{1}_{1\times q}, \mathbf{0}_{1\times(p-q-\iota)}]^{\mathrm{T}}$, for $\iota = 0, \ldots, p-q$, and $\mathcal{G}$ indexes either ASAH1 or FGFR2. Hence $\iota + 1$ represents the starting position the causal variants, q determines the sparsity of the signal, and $b_{\mathcal{G}}$ reflects the strength of the signal. We consider 4 choices of q and $b_{\mathcal{G}}$: (i) dense signal $q = \lceil 0.8p \rceil$, $b_{\text{ASAH1}} = 3.5$, $b_{\text{FGFR2}} = 2.8$; (ii) moderately sparse signal $q = \lceil 0.5p \rceil$, $b_{\text{ASAH1}} = 4.1$, $b_{\text{FGFR2}} = 3.5$; (iii) sparse signal $q = \lceil 0.1p \rceil$, $b_{\text{ASAH1}} = 7.1$, $b_{\text{FGFR2}} = 5.4$; and (iv) single causal variant $q = 1$, $b_{\text{ASAH1}} = b_{\text{FGFR2}} = 10.6$. The pattern of the results is similar across the three sample sizes and hence we present only results for $n = 500$.

Since the SNPs in the ASAH1 gene are generally in high LD with each other, we summarize the power of the tests averaged over the entire range of ι in Figure 2. As we expect from the theoretical analysis, the EB test is the most powerful under the dense signal setting with 14% sparsity. For the sparse settings, the adaptive test is at least as powerful as other procedures. Across all settings, the standard p-DF test is the

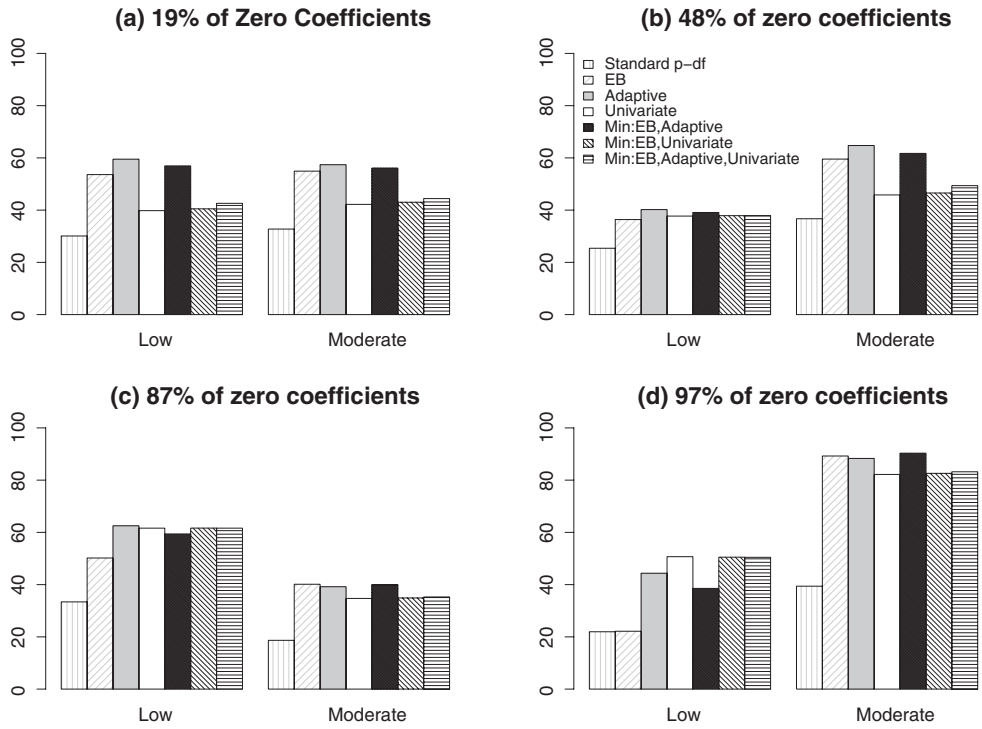

Fig. 3. Empirical Power (in %) for various tests using the FGFR2 gene under local alternatives averaged over the set of ι with low $\wp(\cdot)$ (low) and the set of ι with moderate $\wp(\cdot)$ (moderate). For settings of q and b were considered: (a) $q = 25$, $b_{\text{FGFR2}} = 3.5$ (19% sparsity); (b) $q = 16$, $b_{\text{FGFR2}} = 2.8$ (48% sparsity); (c) $q = 4$, $b_{\text{FGFR2}} = 5.4$ (87% sparsity); and (d) $q = 1$, $b_{\text{FGFR2}} = 10.6$ (97% sparsity).

least powerful. The most robust test is the omnibus test based on $\min_{\hat{P},\hat{P}_{\mathcal{A}}}$, the minimum p-value from $\hat{Q}$ and $\hat{Q}_{\mathcal{A}}$, which always has power similar to the best among (a)–(d).

For the FGFR2 gene, the LD is generally weak among the SNPs, but the correlation structure changes over different regions. To gauge the general pattern of how the correlation might affect the test performances, we let $\wp(\iota)$ denote the *average* correlation between the causal and non-causal SNPs for a given ι and summarize the power by averaging over different levels of $\wp(\iota)$. In Figure 3, we present the power averaged over the set of ι with low $\wp(\cdot)$ and the set of ι with moderate $\wp(\cdot)$. When $\wp(\cdot)$ is low and the signal is moderately sparse, the adaptive test is more powerful than its competitors. For example, when the sparsity is 19%, the average power is 54% for the EB test and 60% for the adaptive test while the univariate test has a power of 40%. The univariate test is generally less powerful except when there is only a single causal variant and $\wp(\cdot)$ is low. When $\wp(\cdot)$ is moderate, the EB test and the adaptive test have more similar performances and the univariate test generally is less powerful. Similar to the results for the ASAH1 gene, the test based on $\min_{\hat{P},\hat{P}_{\mathcal{A}}}$ almost always achieves power close to the best among (a)–(d), except for when $\wp(\cdot)$ is low with extreme sparsity. Other omnibus tests based on $\min_{\hat{P},\hat{P}_{\mathcal{A}},\hat{P}_{\text{uni}}}$ and $\min_{\hat{P},\hat{P}_{\text{uni}}}$ can be less powerful whenever the univariate test does not work well, where $\hat{P}_{\text{uni}}$ denotes the p-value based on the univariate test. For example, for the FGFR2 gene with 19% sparsity and low $\wp(\cdot)$, the powers of the univariate test and the omnibus test that includes the univariate test is only $\sim 43\%$, while the power of $\min_{\hat{P},\hat{P}_{\mathcal{A}}}$ is close to 57%.

6. Discussions

In this paper, we proposed an adaptive score test procedure to test for the effect of a set of genetic markers, by rescaling the design matrix of the genetic markers with an initial estimator of the marker effects. When compared with the EB score test in (2.7), the adaptive test in (2.8) has higher power when the signal is sparse and the between marker correlation is low. The null distribution of $\hat{Q}_{\mathcal{A}}$ can be estimated via a simple Monte-Carlo procedure. In practice, we find that the null distributions of $\hat{Q}$ and $\sqrt{\hat{Q}_{\mathcal{A}}}$ can also be approximated well by skewed χ^2 distributions with DF, $d_{\hat{Q}}$ and $d_{\hat{Q}_{\mathcal{A}}}$, respectively. Furthermore, under the local alternative, the distributions of $\hat{Q}$ and $\sqrt{\hat{Q}_{\mathcal{A}}}$ can be approximated by non-central χ^2 distributions with non-centrality parameters, $\gamma_{\hat{Q}}$ and $\gamma_{\hat{Q}_{\mathcal{A}}}$, respectively. However, providing theoretical justification for why such an approximation works well for $\sqrt{\hat{Q}_{\mathcal{A}}}$ is challenging and warrants further research.

The χ^2 approximations allow us to assess the relative performance of these two testing procedures by comparing $(d_{\hat{Q}}, \gamma_{\hat{Q}})$ to $(d_{\hat{Q}_{\mathcal{A}}}, \gamma_{\hat{Q}_{\mathcal{A}}})$ under various settings. In simulation studies (results not reported) with a compound symmetry correlation structure for $\mathbf{V}$ with correlation $\wp$, we compared how $\wp$ and signal sparsity affect the results. The adaptive test is most effective when $\wp$ is low and sparsity is high, since under such settings $d_{\hat{Q}} \approx d_{\hat{Q}_{\mathcal{A}}}$ but $\gamma_{\hat{Q}_{\mathcal{A}}} > \gamma_{\hat{Q}}$. As $\wp$ increases, $d_{\hat{Q}}$ decreases quickly, but $d_{\hat{Q}_{\mathcal{A}}}$ remains almost constant. As the sparsity of the signal increases, $\gamma_{\hat{Q}}$ decreases while $\gamma_{\hat{Q}_{\mathcal{A}}}$ increases but the magnitude of change in the non-centrality parameters appears to be slower with larger $\wp$, particularly for $\hat{Q}$. Thus, when the correlation increases, the EB-based test gains power by maintaining low DF while the adaptive test pays the price for having higher DF, in part due to the increased difficulty in estimating $\boldsymbol{\beta}_0$.

To get more intuition behind these relative performances, we consider the local alternative $\boldsymbol{\beta}_0 = n^{-1/2}\mathbf{b}_0$ and the setting with orthogonal normal design, $U = 1$, and $\sigma^2 = 1$. Due to the complexity of the power functions, we focus on the setting when p is not small for the ease of approximation. One can show that $\sqrt{\hat{Q}_{\mathcal{A}}}/c_0$ is approximately χ^2 with

$$d_{\hat{Q}_{\mathcal{A}}} \approx \frac{3}{4}p \quad \text{and} \quad \gamma_{\hat{Q}_{\mathcal{A}}} \approx \sqrt{d_{\hat{Q}_{\mathcal{A}}}^2 + \frac{3p}{16}(6\|\mathbf{b}_0\|_2^2 + \|\mathbf{b}_0\|_4^4)} - d_{\hat{Q}_{\mathcal{A}}} \approx \frac{3}{4}\|\mathbf{b}_0\|_2^2 + \frac{1}{8}\|\mathbf{b}_0\|_4^4,$$

while $\hat{Q} \sim \chi_p^2$ and $\gamma_{\hat{Q}} = \|\mathbf{b}_0^2\|_2^2$, where $c_0 \approx 4/\sqrt{3p}$. Comparing the non-sparse case with $\mathbf{b}_0 = (b, \ldots, b)^{\mathrm{T}}/\sqrt{p}$ to the sparse case with $\mathbf{b}_0 = (b, 0, \ldots, 0)^{\mathrm{T}}$, one finds that the EB test has the same power at these two local alternatives since $\|\mathbf{b}_0\|_2^2 = b^2$ in both cases. On the other hand, the adaptive test has much greater power in the sparse setting since $\gamma_{\hat{Q}_{\mathcal{A}}} \approx b^2 + \frac{1}{8}(b^4 - 2b^2)$ for the sparse case and $b^2 + \frac{1}{8p}(b^4 - 2pb^2)$ in the non-sparse case. Hence when signal is sparse and b is not small, the adaptive test gains power by amplifying the strong signals, which is reflected in the increased non-centrality parameter $\gamma_{\hat{Q}_{\mathcal{A}}}$. On the other hand, when signal is not sparse, $b^4 - 2pb^2$ could be substantially smaller than 0 and hence leads to a power loss when compared with the EB test.

In general, the ridge-rescaled adaptive test has more power than the test rescaled by the standard quasi-likelihood estimator, especially when the correlation among the $\mathbf{V}$ is high. The omnibus test which combines information from both the EB-based score test and the adaptive test appears to pick out the winner with relatively little price paid for selecting the better one, at least for the settings we have examined. It will be interesting to extend the proposed procedures to accommodate the rare variants from next generation sequence studies. When the minor allele frequencies of the rare variants are too low, the proposed weight vector $\hat{\mathbf{Z}}$ based on simple initial estimators may be unstable. Alternative weights that account for rare variants and may increase power warrants further research.

As shown in the data example section, when the p-value is extremely small, it remains numerically difficult to obtain a good estimate of the tail probability for the omnibus test due to the requirement of a large number of perturbations. On the other hand, our proposed perturbation procedure would enable us to easily obtain the overall type I error-adjusted p-values when multiple marker sets are under investigation. By generating the same set of $\mathcal{N}$ for all the marker sets, one can obtain the null distribution of the minimum p-value across all marker sets and compare the observed p-value to this null distribution to estimate the adjusted p-value. For approximating the tail probabilities, we find that a normal mixture works well for approximating the distribution of the minimum p-value, both under H_0 and under the alternative.

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

Conflict of Interest: None declared.

FUNDING

Research was supported by grants from the National Institute of Health (R01-GM079330 to T.C.) and the National Science Foundation (DMS-0854970 to T.C.); the National Cancer Institute (R37-CA076404 and P01-CA134294 to X.L.); the National Cancer Institute (R37-CA057030 to R.J.C.) and Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology (KAUST) to R.J.C.

REFERENCES

BAUM, A. E., AKULA, N., CABANERO, M., CARDONA, I., CORONA, W., KLEMENS, B., SCHULZE, T. G., CICHON, S., RIETSCHEL, M., NÖTHEN, M. M. *and others*. (2007). A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder. *Molecular Psychiatry* **13**, 197–207.

BRESLOW, N. E. AND CLAYTON, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

BROWN, M. P. S., GRUNDY, W. N., LIN, D., CRISTIANINI, N., SUGNET, C. W., FUREY, T. S., ARES, M. AND HAUSSLER, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences* **97**, 262–267.

CARLIN, B. P. AND LOUIS, T. A. (1997). Bayes and empirical Bayes methods for data analysis. *Statistics and Computing* **7**, 153–154.

COMMENGES, D. (1994). Robust genetic linkage analysis based on a score test of homogeneity: the weighted pairwise correlation statistic. *Genetic Epidemiology* **11**, 189–200.

CURTIS, R. K., ORESIC, M. AND VIDAL-PUIG, A. (2005). Pathways to the analysis of microarray data. *TRENDS in Biotechnology* **23**, 429–435.

EFRONI, S., SCHAEFER, C. F. AND BUETOW, K. H. (2007). Identification of key processes underlying cancer phenotypes using biologic pathway analysis. *PLoS One* **2**, 425.

Frayling, T. M. (2007). Genome-wide association studies provide new insights into type 2 diabetes aetiology. *Nature Reviews Genetics* **8**, 657–662.

Goeman, J. J., Oosting, J., Cleton-Jansen, A. M., Anninga, J. K. and van Houwelingen, H. C. (2005). Testing association of a pathway with survival using gene expression data. *Bioinformatics* **21**, 1950–1957.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review/Revue Internationale de Statistique*, 245–259.

Harris, P. and Peers, H. W. (1980). The local power of the efficient scores test statistic. *Biometrika* **67**, 525.

Hunter, D. J., Kraft, P., Jacobs, K. B., Cox, D. G., Yeager, M., Hankinson, S. E., Wacholder, S., Wang, Z., Welch, R., Hutchinson, A. *and others*. (2007). A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nature Genetics* **39**, 870–874.

Kwee, L. C., Liu, D., Lin, X., Ghosh, D. and Epstein, M. P. (2008). A powerful and flexible multilocus association test for quantitative traits. *The American Journal of Human Genetics* **82**, 386–397.

Lin, D. Y. (2005). An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics* **21**, 781–7.

Liu, D., Ghosh, D. and Lin, X. (2008). Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC bioinformatics* **9**, 292.

Liu, D., Lin, X. and Ghosh, D. (2007). Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics* **63**, 1079–1088.

Martinez, J. G., Carroll, R. J., Muller, S., Sampson, J. N. and Chatterjee, N. (2010). A note on the effect on power of score tests via dimension reduction by penalized regression under the null. *The International Journal of Biostatistics* **6**, 12.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. New York: Chapman & Hall/CRC.

Moskvina, V. and Schmidt, K. M. (2008). On multiple-testing correction in genome-wide association studies. *Genetic Epidemiology* **32**, 567–573.

Nyholt, D. R. (2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *The American Journal of Human Genetics* **74**, 765–769.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904–909.

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, **44**, 50–57.

Rao, C. R. (2005). Score test: Historical review and recent developments. *Advances in Ranking and Selection, Multiple Comparisons, and Reliability*, 3–20.

Rioux, J. D., Xavier, R. J., Taylor, K. D., Silverberg, M. S., Goyette, P., Huett, A., Green, T., Kuballa, P., Barmada, M. M., Datta, L. W. *and others*. (2007). Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nature Genetics* **39**, 596–604.

Sullivan, P. F., de Geus, E. J. C., Willemsen, G., James, M. R., Smit, J. H., Zandbelt, T., Arolt, V., Baune, B. T., Blackwood, D., Cichon, S. *and others*. (2008). Genome-wide association for major depressive disorder: a possible role for the presynaptic protein piccolo. *Molecular Psychiatry* **14**, 359–375.

THOMAS, D., LANGHOLZ, B., CLAYTON, D., PITKÄNIEMI, J., TUOMILEHTO-WOLF, E. AND TUOMILEHTO, J. (1992). Empirical Bayes methods for testing associations with large numbers of candidate genes in the presence of environmental risk factors, with applications to HLA associations in IDDM. *Annals of Medicine* **24**, 387–92.

VO, T. M., PHAN, J. H., HUYNH, K. N. T. AND WANG, M. D. (2007). Reproducibility of differential gene detection across multiple microarray studies. In *Engineering in Medicine and Biology Society, 2007. 29th Annual International Conference of the IEEE*, pp. 4231–4234.

WALLACE, C., NEWHOUSE, S. J., BRAUND, P., ZHANG, F., TOBIN, M., FALCHI, M., AHMADI, K., DOBSON, R. J., MARÇANO, A. C. B., HAJAT, C. *and others*. (2008). Genome-wide association study identifies genes for biomarkers of cardiovascular disease: serum urate and dyslipidemia. *The American Journal of Human Genetics* **82**, 139–149.

WU, M. C., KRAFT, P., EPSTEIN, M. P., TAYLOR, D. M., CHANOCK, S. J., HUNTER, D. J. AND LIN, X. (2010). Powerful SNP-set analysis for case-control genome-wide association studies. *American Journal of Human Genetics* **86**, 929.

YEAGER, M., ORR, N., HAYES, R. B., JACOBS, K. B., KRAFT, P., WACHOLDER, S., MINICHIELLO, M. J., FEARNHEAD, P., YU, K., CHATTERJEE, N. *and others*. (2007). Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nature genetics* **39**, 645–649.

ZAYKIN, D. V., ZHIVOTOVSKY, L. A., WESTFALL, P. H. AND WEIR, B. S. (2002). Truncated product method for combining *p*-values. *Genetic Epidemiology* **22**, 170–185.

[*Received March 2, 2011; revised April 24, 2012; accepted for publication April 25, 2012*]