Original Paper

Human Heredity

Hum Hered 2011;72:98–109
DOI: 10.1159/000330632

Received: March 10, 2011
Accepted after revision: July 2, 2011
Published online: September 16, 2011

Adaptive Tests for Detecting Gene-Gene and Gene-Environment Interactions

Wei Pan[a] Saonli Basu[a] Xiaotong Shen[b]

[a]Division of Biostatistics, School of Public Health, [b]School of Statistics, University of Minnesota, Minneapolis, Minn., USA

Key Words
Complex traits · Epistasis · Logistic regression · Adaptive Neyman test · Simulation · SSU test · Sum test · UminP test

Abstract
There has been an increasing interest in detecting gene-gene and gene-environment interactions in genetic association studies. A major statistical challenge is how to deal with a large number of parameters measuring possible interaction effects, which leads to reduced power of any statistical test due to a large number of degrees of freedom or high cost of adjustment for multiple testing. Hence, a popular idea is to first apply some dimension reduction techniques before testing, while another is to apply only statistical tests that are developed for and robust to high-dimensional data. To combine both ideas, we propose applying an adaptive sum of squared score (SSU) test and several other adaptive tests. These adaptive tests are extensions of the adaptive Neyman test [Fan, 1996], which was originally proposed for high-dimensional data, providing a simple and effective way for dimension reduction. On the other hand, the original SSU test coincides with a version of a test specifically developed for high-dimensional data. We apply these adaptive tests and their original nonadaptive versions to simulated data to detect interactions between two groups of SNPs (e.g. multiple SNPs in two candidate regions). We found that for sparse models (i.e. with only few non-zero interaction parameters), the adaptive SSU test and its close variant, an adaptive version of the weighted sum of squared score (SSUw) test, improved the power over their non-adaptive versions, and performed consistently well across various scenarios. The proposed adaptive tests are built in the general framework of regression analysis, and can thus be applied to various types of traits in the presence of covariates.

Introduction

Genome-wide association studies (GWASs) have been successful in identifying thousands of genetic variants associated with complex human diseases and traits [Hindorff et al., 2010]. Most of these identified variants are single nucleotide polymorphisms (SNPs). However, these identified variants can only explain a small fraction of heritability [Maher, 2008]. Since almost all existing GWASs are based on univariate analysis of main effects to identify individual SNPs that are marginally associated with a trait, one possible reason for the missing heritability could be missed gene-gene and gene-environment interactions. It has been argued that such interactions are commonly present [e.g. Moore, 2003]. Furthermore, discovery of gene-gene and gene-environment interactions can advance our understanding of the underlying biology. Hence, there have been tremendous research interests and efforts in identifying such interactions. There is a huge body of literature in detecting gene-

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

Accessible online at:
www.karger.com/hhe

Wei Pan
MMC 303, A460 Mayo, Division of Biostatistics
School of Public Health, University of Minnesota
Minneapolis, MN 55455-0392, USA
Tel. +1 612 626 2705, E-Mail weip@biostat.umn.edu

gene and gene-environment interactions, and we will not provide a comprehensive review here; see recent reviews [Musani et al., 2007; Cordell, 2009; Kooperberg et al., 2009; Thomas, 2010]. Existing approaches can be divided into two categories: one is to conduct multilocus or even genome-wide searches [Zhang and Liu, 2007; Wu et al., 2010], while others focus on two or a few genes or candidate regions [Zhao et al., 2006; Zheng et al., 2006]. In each category, some are more exploratory in nature [e.g. Ritchie et al., 2001; Lou et al., 2007; He et al., 2010; Chen et al., 2011], while others are statistically more rigorous in quantifying statistical uncertainty or significance level for any discovery. Here we focus on formal hypothesis testing for possible interactions between two groups of SNPs, or between a group of SNPs and a group of environmental variables. The SNPs in a group, such as neighboring SNPs in a candidate gene or region, can be in linkage disequilibrium (LD). Some authors have considered testing for the significance of a group of SNPs in the presence of its interaction with other SNPs or environmental variables, including both main effects and interactions [e.g. Chatterjee et al., 2006]. In contrast, like other authors [e.g. Kooperberg and LeBlanc, 2008], here we are interested in testing interactions only. This could arise, for example, after an SNP-by-SNP analysis has identified some significant SNPs or loci in a GWAS. We explored whether any of these SNPs or their loci are interacting with each other, or with other loci or environmental variables.

A major statistical challenge is how to deal with a large number of parameters associated with interaction effects, which inevitably lead to loss of power of any statistical test due to increased degrees of freedom (DF) and/or high cost for multiple test adjustment. Naturally, many approaches are based on dimension reduction. They can be based on modeling interaction effects [Chatterjee et al., 2006], restricting the parameter space [Wang, 2008; Song and Nicolae, 2009], or more direct dimensional reduction through principal components or partial least squares [Wang et al., 2009]. Exploring gene-gene or gene-environment independence [Chatterjee and Carroll, 2005; Mukherjee and Chatterjee, 2008] can be also regarded as a way of reducing the dimension of the parameter space. On the other hand, one can also employ some tests that are more robust and thus more suitable to high-dimensional data. For example, Pan [2009] developed two such tests, called sum of squared score (SSU) test and its weighted version (SSUw). The SSU test coincides with the permutation-based test of Goeman et al. [2006], which was specifically developed for high-dimensional data, such as microarray data. Another example is the so-called Sum test, which reduces to testing on only one parameter, no matter how many parameters are in the original model [Chapman and Whittaker, 2008; Pan, 2009]; some existing methods, such as a dimension-reduction approach to testing for multi-locus joint effects of Basu et al. [2011] and a one-parameter interaction model for a pair of SNPs of VanderWeele and Laird [2011], follow the same line. Here, we combine the above two ideas: we extend the adaptive Neyman test [Fan, 1996] to the SSU, SSUw, Sum tests, the usual score test and a univariate test (denoted UminP). The adaptive Neyman test was motivated by the failure of the usual score test (or its asymptotically equivalent Wald test or likelihood ratio test, LRT) in a simple model for high-dimensional data: we observe $Z \sim N(\theta, I)$ and wish to test the null hypothesis $H_0 : \theta = 0$; the usual score test (or Wald's test or LRT) has power tending to the specified type I error rate (i.e. minimal power) if the magnitude of the non-zero components of θ (i.e. its L_2 -norm $\| \theta \|$) is small compared to the dimension of θ while $\| \theta \| \rightarrow \bullet\bullet$. Fan [1996] proposed an adaptive Neyman test to overcome this problem. The basic idea is that, rather than using all the components of Z, one selects an informative subset of the components of Z to maximize the power. Adopting the idea of the adaptive Neyman test, Pan and Shen [2011] proposed an adaptive SSU, adaptive SSUw and adaptive Sum tests for association analysis of main effects of rare variants; here we apply the tests to detect gene-gene interactions. Importantly, in addition to developing a new adaptive test (called aUminP) to combine univariate analyses of individual SNP-SNP pairs and a new version of the adaptive Sum test (called aSum2), we extend the tests to the current context with nuisance parameters. In particular, in the presence of nuisance parameters, the permutation method as proposed in Pan and Shen [2011] is not applicable to inference [Buzkova et al., 2011]; instead, we propose using a general simulation-based method by taking advantage of the asymptotic Normal distribution of the score vector. We also note that the SSU and SSUw tests have never been applied to detect interactions only, though they have previously been shown to perform well in detecting both main and interaction effects [Pan, 2010]. Here we use simulated data to show that our proposed adaptive SSU, SSUw and Sum tests could substantially improve the power over the original SSU, SSUw and Sum tests for sparse models (e.g. when there are only a few non-zero interaction effects among many possible interactions), and the adaptive SSU and SSUw tests consistently performed well across a wide range of scenarios.

Methods

We consider a binary trait, e.g. a disease indicator, as arising in a case-control design in GWASs. Since all the methods discussed here are based on multiple logistic regression, they can be extended to generalized linear models or Cox regression for other types of traits. In addition, we can also incorporate other covariates, such as environmental variables and principal components (PCs) to adjust for population stratification. We focus on testing whether there is any interaction between two groups of SNPs; the methods can be equally applied to test gene-environment interactions. The SNPs within each group may be in LD with each group being an LD block, a part of candidate region, or nearby SNPs inside a sliding window in a genome-wide scan.

Denote the binary trait $Y_i = 0$ for n_0 controls, and $Y_i = 1$ for $n_1 = n - n_0$ cases. Suppose that there are k_1 and k_2 SNPs in the two groups. By default we use an additive genetic model to code the SNPs: $X_{g,ij} = 0$, 1 or 2 for SNP j in group $g = 1$ or 2 for subject i. Other coding schemes can be equally adopted.

Logistic Regression Model and Some Existing Tests

For a binary trait, most existing association tests are based on a logistic regression model:

$$\text{Logit Pr}\left(Y_i = 1\right) = \beta_0 + \sum_{j_1=1}^{k_1} X_{1,ij_1}\beta_{1,j_1} + \sum_{j_2=1}^{k_2} X_{2,ij_2}\beta_{2,j_2} + \sum_{j_1=1}^{k_1}\sum_{j_2=1}^{k_2} X_{1,ij_1}X_{2,ij_2}\beta_{12,j_1j_2}. \quad (1)$$

The null hypothesis to be tested is H_0: $\beta_{12} = (\beta_{12,11}, ..., \beta_{12,k_1k_2})' = 0$. As in Kooperberg and LeBlanc [2008], we test whether there is any interaction between two groups of SNPs, while ignoring possible main effects. A possible scenario is that after detecting significant main effects of one or two groups of SNPs, we would like to test for their interactions. Of course, similarly we can also test both main effects and interactions simultaneously, as done in other studies [e.g. Chatterjee et al., 2006; Pan, 2010].

The most straightforward method is to take one of the three asymptotically equivalent score test, Wald's test and LRT. Since the score test is computationally fastest without the need to iteratively obtain the maximum likelihood estimate of β_{12}, we will focus exclusively on the score test. The score test can be implemented directly based on the score vector for β_{12} while accounting for the nuisance parameters β_1 and β_2 with an appropriate information matrix [Cox and Hinkley, 1974], which, however, we found might have inflated type I error rates for a high-dimensional β_{12}. Alternatively, we take the approach as discussed in Kooperberg and LeBlanc [2008]: the efficient score vector for β_{12} is $U = (U_{1,1}, ..., U_{k_1,k_2})'$ with

$$U_{j_1,j_2} = \sum_{i=1}^{n} U_{i,j_1,j_2} = \sum_{i=1}^{n}\left(Y_i - \hat{p}_i\right)\left(X_{i,j_1}X_{i,j_2} - \hat{\mu}_{i,j_1,j_2}\right),$$

where

$$\hat{p}_i = \widehat{\Pr}\left(Y_i = 1\right)$$

is the fitted probability of $Y_i = 1$ from a main-effects logistic regression model

$$\text{Logit Pr}\left(Y_i = 1\right) = \beta_0 + \sum_{j_1=1}^{k_1} X_{1,ij_1}\beta_{1,j_1} + \sum_{j_2=1}^{k_2} X_{2,ij_2}\beta_{2,j_2},$$

which is the null model for (1) under H_0, and $\hat{\mu}_{i,j_1,j_2} = \hat{\text{E}}(X_{i,j_1}X_{i,j_2})$ is the fitted response value in a linear regression model

$$E\left(X_{i,j_1}X_{i,j_2}\right) = \beta_0 + \sum_{j_1=1}^{k_1} X_{1,ij_1}\beta_{1,j_1} + \sum_{j_2=1}^{k_2} X_{2,ij_2}\beta_{2,j_2}.$$

Under H_0, U has an asymptotically Normal distribution with mean 0 and covariance matrix V that can be estimated as

$$V = \sum_{i=1}^{n} U_{i..}U_{i..}',$$

with $U_{i..} = (U_{i,1,1}, ..., U_{i,k_1,k_2})'$.

Based on the asymptotic Normality of U, we can construct the multivariate score test as

$$T_{Score} = T_{Score}(U) = U'V^{-1}U,$$

which has an asymptotic χ^2 distribution with DF $k = rank(V)$ under H_0. While accounting for the correlation structure of the score vector U, the score test may lose power for high-dimensional data with a large DF k; see section Adaptative Tests for more discussions.

As an alternative to the score test, Pan [2009] proposed two tests, called the sum of squared score (SSU) and the weighted sum of squared score (SSUw) tests:

$$T_{SSU} = T_{SSU}(U) = U'U, \qquad T_{SSU_w} = T_{SSU_w}(U) = U'V_d^{-1}U,$$

where $V_d = \text{Diag}(V)$ is a diagonal matrix with the same diagonal elements of V. Under H_0, each of the two test statistics has an asymptotic distribution of a mixture of χ_1^2's, which can be approximated by a scaled and shifted χ^2 distribution [Pan, 2009]. Each of the two tests can be regarded as a modified score test by ignoring the nondiagonal elements of V, i.e. correlations among the components of U, which is known to be advantageous for high-dimensional data [Chen and Qin, 2010]. Importantly, as shown by Pan [2009], the SSU test is equivalent to the permutation-based version of Goeman's test [2006], which is derived as a variance component score test for a random-effects logistic regression model specifically developed for high-dimensional data. In addition, Han and Pan [2010b] showed a close connection between the SSU test and a genomic distance-based regression approach of Wessel and Schork [2006], which was also successfully applied to high-dimensional microarray gene expression data [Zapala and Schork, 2006]. These connections provide a partial justification for the utility of the SSU and SSUw tests in the current context, in which the dimension of the parameters to be tested is $k = k_1k_2$, which can be large. Empirically, Goeman's test and the SSU test have been shown to perform well with high power in many situations for SNP data [Chapman and Whittaker, 2008; Pan, 2009, 2010].

Another association test with high power under some situations is the so-called Sum test, as noted by Chapman and Whittaker [2008] and Pan [2009]. The Sum test aims to utilize multiple SNPs while avoiding testing on multiple parameters, for which it imposes a generally incorrect working assumption that the parameters to be tested are all equal; in the current context, assuming a common interaction parameter, we have

$$\text{Logit Pr}\left(Y_i = 1\right) = \beta_0 + \sum_{j_1=1}^{k_1} X_{1,ij_1}\beta_{1,j_1} + \sum_{j_2=1}^{k_2} X_{2,ij_2}\beta_{2,j_2} + \sum_{j_1=1}^{k_1}\sum_{j_2=1}^{k_2} X_{1,ij_1}X_{2,ij_2}\beta_{12,c} \quad (2)$$

where $\beta_{12,c}$ represents a common (or more precisely, average) interaction effect. The Sum test has a minimum DF = 1 and thus possibly improved power since one only needs to test on a single parameter with H_0': $\beta_{12,c} = 0$. The Sum test based on the score vector is

$$T_{Sum} = T_{Sum}(U) = \frac{1'U}{\sqrt{1'V1}}$$

which has an asymptotic null distribution of $N(0, 1)$. However, as clearly shown by the Sum statistic, if some association parameters in β_{12} are positive while some are negative, leading to both positive and negative components of U, we may end up with a small T_{Sum} and thus fail to reject H_0'.

Finally, the most commonly used test in GWASs is the univariate or marginal minP (UminP) test on each individual SNPs. In the current context, the UminP test can be formulated as

$$T_{U\,min\,P} = T_{U\,min\,P}(U) = \max_{1 \le j \le k_1 k_2} U_j^2 / V_{jj},$$

where U_j is the j-th component of U, and V_{jj} is the (j, j)th element of V. Equivalently, since each univariate test statistic U_j^2/V_{jj} has an asymptotic χ_1^2 null distribution, T_{UminP} takes the minimum of the p values of the univariate tests. Rather than taking a conservative Bonferroni adjustment for multiple testing, we can adopt either numerical integration [Conneely and Boehnke, 2007] or permutation/simulation methods to calculate its p value. Intuitively, the UminP test is powerful if there is only one or few *large* components of $|\beta_{12}|$; on the other hand, if there are many and yet small non-zero effects, it may have low power.

Adaptive Tests

To motivate the adaptive Neyman test [Fan, 1996] and other adaptive tests, we consider a simple situation with a k-dimensional observation $Z \sim N(\theta, I)$. The goal is to test H_0: $\theta = 0$ versus H_1: $\theta \neq 0$. The score test, Wald's test and LRT all share the same test statistic $\| Z \|^2 = \Sigma_{j=1}^k Z_j^2$, where $\| . \|$ is the L_2 norm. The above three conventional tests suffer from low power for high-dimensional data with a large k. Under H_1 with $\theta = \theta_0 \neq 0$, if $\| \theta_0 \|^2 = o(k)$, then the power of the tests is approximately [Fan, 1996]

$$1 - \Phi(z_{1-\alpha} - \| \theta_0 \|^2 / \sqrt{2k}),$$

which tends to the nominal type I error rate α as $\| \theta_0 \| \to \infty$ with $\| \theta_0 \|^2 = o(k^{1/2})$. That is, under H_1 with $\theta_0 \neq 0$, if the squared L_2-norm of θ_0 increases to ∞ in a slower rate than $k^{1/2}$, then the power of the score (or Wald or LRT) test will diminish; simply put, no matter how big $\| \theta_0 \|^2$ is, if $\| \theta_0 \|^2$ is small compared to dimension k, the power of the test will be minimal.

To overcome the shortcoming of the score test, as a simple way for dimension reduction, Neyman [1937] proposed using only the first few dimensions with a test statistic $\Sigma_{j=1}^m Z_j^2$ with $m \le k$. The power of Neyman's test is approximately

$$1 - \Phi\left(z_{1-\alpha} - \sum_{j=1}^m \theta_{0,j}^2 / \sqrt{2m} \right),$$

which critically depends on the choice of m. To maximize the power, observing $E(Z_j^2 - 1) = \theta_{0,j}^2$, Fan [1996] proposed an adaptive Neyman test with test statistic

$$T_{AN}(Z) = \max_{1 \le m \le k} \left\{ \sum_{j=1}^m \left(Z_j^2 - 1 \right) / \sqrt{2m} \right\}.$$

It is evident that both Neyman's test and its adaptive version depend on the ordering of the components of Z; that is, if we change the order of the components of Z, we may end up with different test results. For the score test, we first take a transformation to de-correlate the components of U: $U^s = V^{-1/2}U$, such that U^s behaves similarly to Z and thus the adaptive Neyman test can be directly applied. Although we do not have a rigorous justification, with $U^s \sim N(\theta, I)$, it seems intuitively reasonable to order the independent components of U^s based on $|U_j^s|$ in a descending order; in this way, for any m, $\Sigma_{j=1}^m (U_j^s)^2$ will be the largest sum among any m components of U^s. Comparing to the expression of T_{AN}, we see that the resulting adaptive test would have high power of rejecting H_0. Note that the larger $|U_j^s|$, the more likely do we have to reject the null hypothesis of $\theta_j = 0$. We call the resulting test with statistic $T_{AN}(U^s)$ an adaptive score (aScore) test.

Pan and Shen [2011] extended the idea of the adaptive Neyman test to other score-based tests. Since it is generally difficult to obtain a closed form of the power function for a test, they proposed using its p value as a surrogate for (one minus) its power. This basic idea has been used in the literature, e.g. by Yu et al. [2009]. A heuristic justification is the following. To be concrete, consider a test statistic with its null and alternative distributions as a central and noncentral χ_d^2, respectively, where d is fixed. For a given dataset, a small p value of the test might result by chance with a small probability, or with a larger probability from the fact that the true distribution of the test statistic is from a noncentral $\chi_d^2(c)$ with $c > 0$; the smaller the p value, the more likely the noncentrality parameter c is large, which in turn implies the increasing power of the test in rejecting the null hypothesis. Of course, the p value is only a rough estimate of (one minus) power.

Specifically, suppose that $U = (U_1, ..., U_k)'$ is the score vector, and denote $U_{(m)} = (U_1, ..., U_m)'$ the subvector containing the first m components with $m \le k$. For any test statistic $T = T(U)$, one can equally apply it to only the first few components of U as $T(U_{(m)})$, and obtain its p value as $P_{T(U_{(m)})}$. Then one constructs an adaptive version of the test T as

$$aT = aT(U) = \min_{1 \le m \le k} P_{T\left(U_{(m)}\right)}.$$

Generally, the distribution of aT is complex; Pan and Shen [2011] proposed a permutation method to obtain its p value. Since permutation methods may not be applicable to testing interactions in the presence of the nuisance parameters for the main effects in model (1) under H_0 [Buzkova et al., 2011], here we propose a simulation method similar to that of Seaman and Muller-Myhsok [2005] and that of Lin [2005]. Since $U \sim N(0, V)$ under H_0, we generate $U^{(1)}$, ..., $U^{(B)}$ iid from $N(0, V)$, then apply the adaptive test to yield $aT^{(b)} = aT(U^{(b)})$ for each $b = 1$, ..., B. The p value for the adaptive test is $\Sigma_{b=1}^B I(aT^{(b)} < aT)/B$.

If we substitute the SSU, SSUw, Sum and UminP test statistics as T respectively, then we have the corresponding adaptive tests called aSSU, aSSUw, aSum and aUminP tests. Note that Pan and Shen [2011] did not consider the aUminP test since a univariate analysis is not expected to perform well for rare variants (due to limited information content in any single rare variant). In addition, the aSum test here differs from, and in our experience, is often more powerful than the adaptive Sum test of Han and Pan [2010a].

The performance of an adaptive test depends on the order of the components of the score vector U. Recall that, as a simple di-

mension reduction strategy to overcome the curse of high dimensionality with the usual score test, an adaptive test aims to use only the first few components of its test statistic while capturing its major departures from H_0, and it therefore makes sense to order the components of the test statistic based on the chance of rejecting H_0. Pan and Shen [2011] proposed a few strategies, some of which can be based on prior knowledge. A general one is based on some (standardized) magnitudes of the components of a test statistic: for the aSSU, aSSUw, aSum and aUminP tests, using a *simple and heuristic* argument of increasing the chance of rejecting H_0, we order the components of U based on the magnitudes of $|U_j|$, $|U_j|/\sqrt{V_{jj}}$, $U_j/\sqrt{V_{jj}}$ and $|U_j|/\sqrt{V_{jj}}$, respectively.

Since the Sum test is sensitive to the signs of U_j's, ordering the components of U by $U_j/\sqrt{V_{jj}}$ works for the aSum test if most of the non-zero components of U are positive (or negative). For example, since most rare causal mutations are expected to be deleterious, the above aSum test can be directly applied to analysis of rare variants [Pan and Shen, 2011]. On the other hand, if the non-zero components of U are mostly negative, we have to change the search direction to increase its power: rather than using the *first* few components of U to construct the Sum statistic and its p value, say $P_{Sum(U_{(m)})}$, we use the last few components of U, say $P_{Sum(U_{(k-m)})}$ with $k = dim(U)$. More generally, we can search the ordered components of U in both directions, yielding a 2-dimensional aSum test:

$$aSum2 = aSum2(U) = \min_{1 \le m \le k} \left\{ P_{T\left(Sum_{(m)}\right)}, P_{T\left(Sum_{(k-m)}\right)} \right\}.$$

Pan and Shen [2011] also discussed accommodating weights on SNPs, e.g. based on the minor allele frequencies (MAFs) of the SNPs, in the above adaptive tests, which is straightforward, though we do not pursue it here.

Simulation Set-Ups

We conducted simulation studies to evaluate and compare the performance of various tests. The simulated data were generated as in Wang and Elston [2008]. Specifically, we simulated two SNP groups with k_1 and k_2 SNPs, respectively; the MAFs of the SNPs were uniformly distributed as $U(0.1, 0.5)$. Each group of the SNPs was independently generated in the following way. First, we generated a latent vector $Z = (Z_1, ..., Z_{k_1})'$ from a multivariate Normal distribution with a first-order auto-regressive (AR1) covariance structure: $Corr(Z_i, Z_j) = \rho_1^{|i-j|}$ between any latent components i and j; we used $\rho_1 = 0$ and $\rho_1 = 0.8$ to generate (neighboring) SNPs in linkage equilibrium and in LD, respectively. Second, the latent vector was dichotomized to yield a haplotype with MAFs each randomly selected between 0.1 and 0.5. Third, we combined two independent haplotypes and obtained genotype data for group 1: $(X_{1,i1}, ..., X_{1,ik_1})$ for subject i. Similarly we generated genotype data for group 2 with a correlation parameter ρ_2 and k_2 SNPs. Fourth, the disease status Y_i of subject i was generated from the logistic regression model (1). We used $\beta_0 = -\log(4)$ for a 20% background disease prevalence, and fixed $\beta_1 = \beta_2 = 0$. For the null case, we used $\beta_{12} = 0$; for non-null cases, we randomly chose a fraction (ZF) of the components of β_{12} as 0, while the remaining ones were randomly chosen from $U(-b, b)$. Fifth, as in a case-control study, we sampled $n_1 = 400$ cases and $n_0 = 400$ controls in each dataset.

We considered several set-ups with various values of SNP group sizes (k_1 and k_2), within-group correlations or LD strengths among the SNPs (ρ_1 and ρ_2), sparsity (ZF) and strength (b) of true interaction effects. By default we assumed no LD between the two SNP groups except for one case, in which case the data were simulated similarly as before with one modification: we generated a latent multivariate Normal variate Z of length $k_1 + k_2 + 2$ with an AR1 correlation structure with parameter ρ_1; the first k_1 and last k_2 components of Z were dichotomized to generate two haplotypes for the two SNP blocks respectively; all other aspects remained the same.

Throughout the simulations, we fixed the test significance level at $\alpha = 0.05$, and used $B = 200$ for each simulation-based method. The results were based on 1,000 independent replicates for each set-up. Each bold print (in each result table) indicates the test with the highest power for a given set-up.

Results

Simulations: No LD between the Two SNP Groups

We first considered a set-up with $k_1 = k_2 = 6$ SNPs in each of the two SNP groups; there was LD among the SNPs within the same group, but no LD between the two groups (table 1). Under the null case, all the tests had satisfactory type I error rates. For the non-null cases, (1) among the nonadaptive tests, if the true model was sparse with high fractions of zero-coefficients (ZF) (i.e. only a small proportion of non-zero interaction effects between the SNPs), the UminP test was most powerful; if the true model was not sparse at all and with strong signals (i.e. $ZF = 0$ and $b = 0.3$), the score test was most powerful; on the other hand, for most cases, the SSU and SSUw performed similarly and were most powerful. (2) Comparing the nonadaptive and adaptive tests, we see that the UminP and aUminP tests always had similar power, while the aSSUw and aSSU tended to more or less improve the power of their corresponding nonadaptive tests; in contrast, the adaptive score test occasionally improved over the score test, but more often was less powerful. (3) Since the true interaction effects could be in either direction, as expected, the Sum test was low powered; however, the aSum2 test had dramatically improved power, which was close to that of the aSSU and aSSUw tests. (4) Overall, the adaptive SSUw and SSU tests were winners.

To explore whether our results were sensitive to B and the total number of simulation replicates for each scenario, we doubled the two numbers: we increased B from 200 to 400, and the number of replicates from 1,000 to 2,000. The results are listed in table 2, which are essentially the same as those in table 1. Hence, we concluded that it was reliable to use $B = 200$ and 1,000 replicates.

When the size of the second SNP group was increased from $k_2 = 6$ to $k_2 = 12$, we would draw similar conclu-

Table 1. Empirical type I error rate (if $ZF = 1$) and power (if $ZF < 1$) of testing for only 2-way interactions with various values of ZF and number of non-zero interaction terms in simulation set-up I: $\rho_1 = 0.8$ and $\rho_2 = 0.8$; $k_1 = 6$ and $k_2 = 6$ SNPs

ZF	NZ	*b*	Score	SSU	SSUw	Sum	UminP	aScore	aSSU	aSSUw	aSum2	aUminP
1	0	–	0.042	0.042	0.039	0.045	0.045	0.038	0.054	0.058	0.055	0.047
0.95	2	0.8	0.338	0.414	0.425	0.299	0.509	0.381	0.508	0.503	0.470	**0.510**
0.90	4	0.8	0.605	0.617	0.638	0.449	0.680	0.586	0.701	**0.713**	0.655	0.674
0.85	6	0.8	0.748	0.742	0.761	0.490	0.796	0.720	0.809	**0.826**	0.779	0.785
0.75	9	0.3	0.244	0.332	0.337	0.257	0.329	0.195	**0.358**	0.347	0.346	0.320
0.75	9	0.4	0.405	0.484	0.497	0.341	0.478	0.345	0.518	**0.520**	0.498	0.472
0.50	18	0.3	0.434	0.487	0.489	0.354	0.448	0.323	0.502	**0.510**	0.472	0.445
0.50	18	0.4	0.682	0.681	0.691	0.442	0.631	0.537	0.713	**0.724**	0.656	0.627
0.25	27	0.2	0.296	0.374	0.382	0.294	0.319	0.201	0.375	**0.383**	0.378	0.326
0.25	27	0.3	0.588	0.626	0.639	0.455	0.571	0.422	0.639	**0.665**	0.615	0.563
0	36	0.2	0.388	0.456	0.458	0.345	0.400	0.270	0.462	**0.471**	0.446	0.395
0	36	0.3	**0.752**	0.706	0.720	0.449	0.655	0.560	0.732	0.735	0.704	0.646

NZ = Non-zero.

Table 2. Empirical type I error rate (if $ZF = 1$) and power (if $ZF < 1$) of testing for only 2-way interactions as in table 1, but with a larger $B = 400$ and 2,000 simulation replicates for each case

ZF	NZ	*b*	Score	SSU	SSUw	Sum	UminP	aScore	aSSU	aSSUw	aSum2	aUminP
1	0	–	0.0460	0.0505	0.0435	0.0470	0.0530	0.0475	0.0530	0.0550	0.0550	0.0520
0.95	2	0.8	0.3470	0.4170	0.4255	0.3080	0.4940	0.3840	**0.4980**	0.4910	0.4650	0.4965
0.90	4	0.8	0.5950	0.6200	0.6375	0.4375	0.6845	0.5850	0.7130	**0.7170**	0.6575	0.6830
0.85	6	0.8	0.7595	0.7400	0.7570	0.4785	0.7950	0.7325	0.8150	**0.8300**	0.7790	0.7905
0.75	9	0.3	0.2450	0.3270	0.3275	0.2405	0.3230	0.1990	**0.3550**	0.3495	0.3300	0.3220
0.75	9	0.4	0.4000	0.4790	0.4825	0.3375	0.4675	0.3475	**0.5185**	**0.5185**	0.4960	0.4675
0.50	18	0.3	0.4220	0.4875	0.4940	0.3535	0.4540	0.3225	0.5080	**0.5185**	0.4770	0.4510
0.50	18	0.4	0.6795	0.6730	0.6790	0.4475	0.6360	0.5335	0.7105	**0.7245**	0.6625	0.6350
0.25	27	0.2	0.2930	0.3805	0.3820	0.2920	0.3265	0.2140	0.3895	**0.3985**	0.3830	0.3265
0.25	27	0.3	0.5995	0.6315	0.6375	0.4360	0.5715	0.4485	0.6480	**0.6640**	0.6170	0.5720
0	36	0.2	0.3960	0.4690	0.4695	0.3530	0.4060	0.2765	0.4825	**0.4970**	0.4615	0.4065
0	36	0.3	0.7475	0.7150	0.7245	0.4570	0.6625	0.5565	0.7380	**0.7515**	0.7095	0.6545

sions, though some of the observed trends were more evident (table 3). For example, the aSSU and aSSUw were winners more consistently, even for the nonsparse model with $ZF = 0$. In particular, the individual SNP-SNP analysis as shown by the UminP or aUminP test was not powerful for moderate to nonsparse models.

We also considered a case where there was LD in one group but not in the other (table 4), and one where there was no LD in any of the two groups (table 5). Most of the previous conclusions still hold. One exception is that, with nonsparse models (i.e. small *ZF*), the aSSU and aSSUw could lose power as compared to their nonadaptive versions, though the loss was small. Importantly, we note that the single SNP-SNP based UminP and aUminP tests were low powered for nonsparse models (e.g. with $ZF \leq 0.75$), especially so for independent SNPs (or environmental variables) as shown in table 5.

Simulations: Two SNP Groups in LD

We next considered the case when the SNPs in one group were correlated with those in the other (or environmental variables), while all other simulation aspects remained the same as those shown in table 1. The results are shown in table 6, which are very close to those in ta-

Table 3. Empirical type I error rate (if $ZF = 1$) and power (if $ZF < 1$) of testing for only 2-way interactions with various values of ZF and number of non-zero NZ interaction terms in simulation set-up II: $\rho_1 = 0.8$ and $\rho_2 = 0.8$; $k_1 = 6$ and $k_2 = 12$ SNPs

ZF	NZ	*b*	Score	SSU	SSUw	Sum	UminP	aScore	aSSU	aSSUw	aSum2	aUminP
1	0	–	0.065	0.058	0.060	0.058	0.055	0.062	0.053	0.050	0.065	0.054
0.98	2	0.8	0.230	0.352	0.370	0.246	0.468	0.285	0.443	0.433	0.389	**0.473**
0.95	4	0.8	0.418	0.576	0.580	0.356	0.654	0.479	0.674	**0.684**	0.597	0.656
0.93	6	0.8	0.570	0.701	0.718	0.461	0.761	0.603	0.790	**0.793**	0.713	0.755
0.88	9	0.8	0.756	0.835	0.841	0.472	0.844	0.761	0.886	**0.902**	0.828	0.839
0.75	18	0.3	0.277	0.431	0.440	0.292	0.375	0.206	0.451	**0.461**	0.416	0.368
0.75	18	0.4	0.470	0.615	0.617	0.375	0.536	0.366	0.648	**0.658**	0.579	0.524
0.50	36	0.3	0.539	0.663	0.663	0.376	0.541	0.410	0.670	**0.693**	0.607	0.539
0.50	36	0.4	0.778	0.833	0.832	0.447	0.735	0.610	0.847	**0.865**	0.769	0.719
0.25	54	0.2	0.390	0.552	0.546	0.351	0.444	0.277	0.563	**0.583**	0.517	0.414
0.25	54	0.3	0.749	0.807	0.802	0.436	0.695	0.556	0.835	**0.848**	0.754	0.674
0	72	0.2	0.498	0.626	0.633	0.361	0.508	0.337	0.624	**0.638**	0.573	0.494
0	72	0.3	0.81	0.863	0.859	0.459	0.760	0.629	0.870	**0.883**	0.815	0.751

NZ = Non-zero.

Table 4. Empirical type I error rate (if $ZF = 1$) and power (if $ZF < 1$) with various values of ZF and number of non-zero interaction terms in simulation set-up III: $\rho_1 = 0.8$ and $\rho_2 = 0$; $k_1 = 6$ and $k_2 = 6$ SNPs

ZF	NZ	*b*	Score	SSU	SSUw	Sum	UminP	aScore	aSSU	aSSUw	aSum2	aUminP
1	0	–	0.048	0.046	0.041	0.052	0.050	0.039	0.039	0.047	0.049	0.049
0.95	2	0.8	0.374	0.417	0.442	0.168	0.487	0.425	0.494	**0.505**	0.391	0.496
0.90	4	0.8	0.599	0.629	0.660	0.244	0.678	0.658	0.695	**0.701**	0.605	0.675
0.85	6	0.8	0.791	0.804	0.828	0.272	0.796	0.803	0.841	**0.856**	0.759	0.780
0.75	9	0.4	0.423	0.483	0.494	0.180	0.409	0.368	**0.512**	0.501	0.432	0.403
0.50	18	0.4	0.700	0.726	**0.733**	0.253	0.570	0.567	0.722	0.718	0.648	0.569
0.25	27	0.3	0.615	0.676	**0.681**	0.243	0.484	0.460	0.641	0.653	0.567	0.475
0	36	0.3	0.761	0.780	**0.786**	0.238	0.572	0.602	0.751	0.764	0.704	0.563

NZ = Non-zero.

Table 5. Empirical type I error rate (if $ZF = 1$) and power (if $ZF < 1$) with various values of ZF and number of non-zero interaction terms in simulation set-up IV: $\rho_1 = 0$ and $\rho_2 = 0$; $k_1 = 6$ and $k_2 = 6$ SNPs

ZF	NZ	*b*	Score	SSU	SSUw	Sum	UminP	aScore	aSSU	aSSUw	aSum2	aUminP
1	0	–	0.047	0.049	0.045	0.063	0.043	0.042	0.039	0.043	0.051	0.040
0.95	2	0.8	0.330	0.335	0.342	0.078	**0.465**	0.455	0.428	0.421	0.272	0.458
0.90	4	0.8	0.578	0.563	0.601	0.114	0.659	**0.675**	0.658	0.672	0.465	0.648
0.85	6	0.8	0.794	0.780	0.805	0.133	0.788	0.824	0.830	**0.835**	0.670	0.784
0.75	9	0.4	0.406	0.435	0.421	0.110	0.294	0.377	**0.444**	0.428	0.333	0.279
0.50	18	0.4	0.717	**0.724**	0.717	0.120	0.456	0.633	0.703	0.706	0.604	0.447
0.25	27	0.3	0.657	**0.670**	0.664	0.113	0.320	0.513	0.620	0.622	0.558	0.312
0	36	0.3	0.785	**0.827**	0.808	0.149	0.368	0.595	0.738	0.752	0.695	0.361

NZ = Non-zero.

Table 6. Empirical type I error rate (if $ZF = 1$) and power (if $ZF < 1$) of testing for only 2-way interactions with various values of ZF and number of non-zero interaction terms in simulation set-up V: $\rho_1 = 0.8$; $k_1 = 6$ and $k_2 = 6$ SNPs, and the two SNP blocks are in LD

ZF	NZ	*b*	Score	SSU	SSUw	Sum	UminP	aScore	aSSU	aSSUw	aSum2	aUminP
1	0	–	0.052	0.044	0.040	0.049	0.042	0.045	0.045	0.045	0.048	0.041
0.95	2	0.8	0.308	0.426	0.415	0.323	**0.494**	0.363	0.480	0.481	0.451	0.491
0.90	4	0.8	0.580	0.625	0.644	0.429	0.696	0.594	0.708	**0.715**	0.670	0.692
0.85	6	0.8	0.733	0.721	0.739	0.460	0.779	0.700	0.811	**0.812**	0.747	0.770
0.75	9	0.3	0.249	0.331	0.337	0.261	0.315	0.201	0.346	**0.356**	0.346	0.311
0.75	9	0.4	0.385	0.474	0.488	0.359	0.447	0.312	0.514	**0.520**	0.486	0.449
0.50	18	0.3	0.427	0.475	0.469	0.339	0.431	0.293	0.490	**0.501**	0.458	0.428
0.50	18	0.4	0.646	0.640	0.640	0.418	0.592	0.484	0.662	**0.670**	0.622	0.586
0.25	27	0.2	0.302	0.400	0.410	0.310	0.348	0.237	0.407	**0.415**	0.401	0.334
0.25	27	0.3	0.606	0.632	0.633	0.437	0.576	0.456	0.646	**0.656**	0.617	0.564
0	36	0.2	0.393	0.468	0.467	0.339	0.406	0.256	0.460	**0.477**	0.461	0.400
0	36	0.3	0.726	0.706	0.727	0.461	0.679	0.546	0.750	**0.763**	0.716	0.674

NZ = Non-zero.

ble 1, suggesting that the tests were not sensitive to LD between the two SNP blocks.

Simulations: Comparison with BEAM and BEAM2

As suggested by a reviewer, we compared the performance of the proposed tests with BEAM (Bayesian Epistasis Association Mapping) of Zhang and Liu [2007] and its improved version BEAM2, which also infers LD blocks [Zhang et al., 2011]. BEAM and BEAM2 aim to detect associated SNPs, with or without interactions, for GWASs, and hence are ideal for SNP selection. In particular, BEAM and BEAM2 do not test for interactions separately; for example, if two SNPs are both marginally associated with the phenotype, then their interaction will not be tested (since the two SNPs have already been selected). Hence, it would not be fair to compare BEAM or BEAM2 with our tests in testing for interactions only, as done in our previous simulations. As an illustration, we applied BEAM and, counted only detected interactions, its power was much lower: in table 1, the maximum power was only 0.262 for ZF = 0.85, compared to the power of 0.826 of the aSSUw test.

For a fair comparison between BEAM/BEAM2 and our proposed tests, we conducted more simulations. The simulation set-ups remained the same as in the first two cases as shown in tables 1 and 3. However, when applying our tests, instead of testing for interactions only, we test for both main effects and 2-way interactions; that is, in model (1), we tested H_0: $\beta_1 = 0$, $\beta_2 = 0$ and $\beta_{12} = 0$. For BEAM2, we used its significance test in detecting any associated SNPs, either marginally or interactively, at the significance level of 0.05 with the Bonferroni adjustment. The results are shown in tables 7 and 8. It is clear that all the other tests except the Sum test were more powerful than BEAM2. In particular, as before, the aSSUw test was the overall winner (closely followed by the aSSU test); its power could be substantially higher than that of BEAM2. Note that BEAM2 is not a global test as the other tests: it does not simply test whether there is any association between the SNPs and the phenotype but aims to identify which SNPs are indeed associated with the phenotype, in which sense BEAM2 targets a much more difficult problem. In addition, BEAM2 uses the more conservative Bonferroni adjustment, while other tests are asymptotically exact.

We also note that, by comparing tables 7 and 8 with tables 1 and 3, we see that any of the tests was more powerful when testing on both main effects and interactions than testing on interactions only, though in truth there were no main effects with $\beta_1 = 0$ and $\beta_2 = 0$ in model (1).

Example

We consider an application of the tests to an amyotrophic lateral sclerosis (ALS) dataset of Schymick et al. [2007]. A GWAS was conducted to identify genetic variants predisposing to developing ALS in a cohort of 276 American sporadic cases and 268 neurologically normal controls. By the univariate test on each SNP one by one, Schymick et al. [2007] identified a list of 34 most significant SNPs with a p value less than 0.0001 (without mul-

Table 7. Empirical type I error rate (if $ZF = 1$) and power (if $ZF < 1$) in testing for both main effects and interactions with various values of ZF and number of non-zero interaction terms in simulation set-up I: $\rho_1 = 0.8$ and $\rho_2 = 0.8$; $k_1 = 6$ and $k_2 = 6$ SNPs, as in table 1

ZF	NZ	*b*	Score	SSU	SSUw	Sum	UminP	aScore	aSSU	aSSUw	aSum2	aUminP	BEAM2
1	0	–	0.042	0.046	0.039	0.046	0.038	0.034	0.036	0.031	0.033	0.038	0.048
0.95	2	0.8	0.612	0.735	0.744	0.620	**0.785**	0.620	0.775	0.781	0.768	0.781	0.589
0.90	4	0.8	0.855	0.862	0.869	0.698	**0.918**	0.841	0.910	0.917	0.902	0.917	0.759
0.85	6	0.8	0.944	0.944	0.957	0.752	**0.970**	0.934	0.965	0.968	0.961	**0.970**	0.858
0.75	9	0.3	0.530	0.662	0.671	0.552	0.686	0.449	**0.705**	0.699	0.678	0.676	0.468
0.75	9	0.4	0.757	0.810	0.805	0.646	0.833	0.662	0.836	**0.851**	0.830	0.840	0.632
0.50	18	0.3	0.794	0.832	0.831	0.626	0.854	0.713	0.858	**0.869**	0.841	0.840	0.635
0.50	18	0.4	0.955	0.946	0.940	0.724	0.944	0.904	0.950	**0.958**	0.842	0.940	0.797
0.25	27	0.2	0.653	0.741	0.742	0.608	0.760	0.541	0.778	**0.786**	0.761	0.756	0.555
0.25	27	0.3	0.905	0.904	0.912	0.705	0.918	0.835	0.917	**0.929**	0.913	0.912	0.761
0	36	0.2	0.778	0.808	0.824	0.646	0.832	0.644	0.833	**0.854**	0.832	0.822	0.644
0	36	0.3	0.967	0.945	0.949	0.747	0.957	0.919	0.956	**0.969**	0.956	0.954	0.846

NZ = Non-zero.

Table 8. Empirical type I error rate (if $ZF = 1$) and power (if $ZF < 1$) of testing for both main effects and interactions with various values of ZF and number of non-zero interaction terms in simulation set-up II: $\rho_1 = 0.8$ and $\rho_2 = 0.8$; $k_1 = 6$ and $k_2 = 12$ SNPs, as in table 3

ZF	NZ	*b*	Score	SSU	SSUw	Sum	UminP	aScore	aSSU	aSSUw	aSum2	aUminP	BEAM2
1	0	–	0.040	0.045	0.050	0.046	0.050	0.053	0.046	0.047	0.044	0.044	0.043
0.98	2	0.8	0.503	0.691	0.724	0.579	**0.780**	0.517	0.765	0.765	0.748	0.776	0.571
0.95	4	0.8	0.794	0.900	0.904	0.707	0.933	0.808	0.935	**0.939**	0.928	0.930	0.797
0.93	6	0.8	0.926	0.945	0.954	0.743	0.970	0.918	0.972	**0.978**	0.970	0.973	0.887
0.88	9	0.8	0.970	0.976	0.981	0.774	0.988	0.962	0.987	**0.991**	0.982	0.987	0.931
0.75	18	0.3	0.663	0.820	0.819	0.620	0.820	0.570	0.842	**0.859**	0.818	0.814	0.625
0.75	18	0.4	0.869	0.922	0.929	0.691	0.938	0.801	0.949	**0.956**	0.935	0.933	0.794
0.50	36	0.3	0.914	0.950	0.957	0.705	0.959	0.838	0.963	**0.970**	0.951	0.953	0.832
0.50	36	0.4	0.989	0.984	0.983	0.760	0.989	0.960	0.990	**0.991**	0.987	0.988	0.939
0.25	54	0.2	0.817	0.897	0.880	0.650	0.877	0.691	0.915	**0.919**	0.885	0.874	0.711
0.25	54	0.3	0.973	0.979	0.983	0.732	0.974	0.940	0.987	**0.988**	0.981	0.972	0.898
0	72	0.2	0.894	0.938	0.940	0.684	0.940	0.816	0.948	**0.959**	0.933	0.937	0.800
0	72	0.3	0.993	0.990	0.991	0.745	0.987	0.985	0.992	**0.997**	0.990	0.988	0.940

NZ = Non-zero.

tiple test adjustment). Here, for illustration, we considered two of the 34 top SNPs, rs6013382 and rs2782931. We extracted 10 neighboring SNPs upstream and another 10 downstream of either of the two SNPs of interest, then used Haploview (v4.1) [Barrett et al., 2005] with its default setting to identify LD blocks around the two SNPs for the control group, leading to block sizes of 5 and 19 respectively. The MAF of one SNP in the second LD block was only 3% while all others were larger than 10%. Considering the relatively small sample size and large number of parameters, we removed the SNP with MAF = 3%, though the results remained almost the same in either way. Hence, we had 5 and 18 SNPs in the two blocks. Given that there was some marginal association between ALS and each of the two LD blocks, we would like to test whether there is any two-way interaction between the two SNP blocks. We applied the tests to the above data. We used $B = 10{,}000$ simulations for the adaptive tests.

Table 9. p values for testing the interaction between two LD blocks of sizes 5 and 18, surrounding rs6013382 and rs2782931, respectively, for the ALS data

	Score	SSU	SSUw	Sum	UminP	aScore	aSSU	aSSUw	aSum2	aUminP
p	0.3090	0.0020	0.0014	0.0007	0.0067	0.5268	0.0027	0.0014	0.0038	0.0066

The p values are shown in table 9. There seemed to be some evidence of interaction according to the more powerful SSU, SSUw, Sum, UminP tests and their adaptive versions, though none could reach the genome-wide significance level after adjusting for multiple testing. On the other hand, both the score and adaptive score tests gave much larger p values, possibly due to their low power for high-dimensional parameters (i.e. 90 interaction parameters tested here).

Discussion

In summary, based on our simulation studies, we have found that the adaptive SSU and SSUw tests consistently performed well, either with or nearly with the highest power, across a wide range of scenarios. In particular, if the true model is sparse with only few non-zero interaction effects, the adaptive SSU and SSUw could substantially improve the power over their original nonadaptive versions. Even if the true model is only moderately sparse, the adaptive tests could still result in some improvement. We also note that, analogous to the most popular single SNP-based univariate (or marginal) test for the main effects of SNPs in GWASs, the univariate test on each SNP-SNP pair for interaction is not powerful unless the true model is extremely sparse.

Our proposed tests are ideal for applications to candidate genes or regions. To detect gene-gene interactions in GWASs, it may be computationally too demanding to apply our tests with standard personal computers or workstations, though in principle, we can conduct a two-dimensional scan of a whole genome. For example, we may first apply an LD blocking algorithm to partition the whole genome into many haplotype or LD blocks [Cardon and Abecasis, 2003; Zhang et al., 2004], then consider all pair-wise combinations of these LD blocks to detect their interactions. Alternatively, we can use two fixed- or varying-width sliding windows across the genome [Durrant et al., 2004], and then test interactions between the SNPs inside the two sliding windows. These SNP blocks or sliding windows can also be used to detect gene-environment interactions. We may embed our powerful tests inside a computationally more efficient genome-wide algorithm, such as BOOST [Wan et al., 2010]. We may also use other computationally more efficient algorithms, e.g. single SNP-based analysis, BOOST or BEAM [Zhang and Liu, 2007], to detect some genetic regions of interest, then apply our tests to these candidate regions. This is an important topic to be addressed in the future.

It is noted that the adaptive Neyman test can be regarded as a simple approach to dimension reduction via variable (i.e. SNP) selection. It selects which of the first few components of a high-dimensional summary statistic (e.g. the score vector) are to be used. One may wonder how this restricted linear search compares with a more exhaustive search, such as a stepwise forward search or best subset selection. While it is possible that a restricted search may miss the best model (or best combination of the SNPs), it does have a reduced search space and thus lower DF and lower cost for multiple test adjustment than an exhaustive search, in addition to much lower computational cost. For example, with k SNP-SNP pairs, our restricted search compares k candidate models, compared to the much bigger $k(k + 1)/2$ and 2^k for the stepwise forward and best subset, respectively. There are some other forms of restricted search as a simple and possibly effective way for dimension reduction. For example, the two-stage approach of finding significant main effects and then considering the interactions between only significant main effects [Marchini et al., 2005] is also a simple restricted search. Some previous studies [e.g. Marchini et al., 2005; Evans et al., 2006; Storey et al., 2005; Basu et al., 2011] have demonstrated possible performance advantages of restricted searches over more exhaustive searches, though more studies are needed. Another future topic is how to order the components of the score vector (or other summary statistic) since the performance of the corresponding adaptive test depends on the chosen order. We have proposed some *simple and heuristic* methods for the adaptive SSU,

SSUw, Sum and UminP tests, but there is no guarantee that they are optimal; it is possible that other better methods exist.

R code will be posted on our web site at http://www.biostat.umn.edu~weip/prog.html.

Acknowledgments

This research was supported by NIH grants R01GM081535, R01HL65462, R01HL105397 and R21DK089351.

References

Barrett JC, Fry B, Maller J, Daly MJ: Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2005;21:263–265.

Basu S, Pan W, Oetting WS: A dimension reduction approach for modeling multi-locus interaction in case-control studies. Hum Hered 2011;71:234–245.

Buzkova P, Lumley T, Rice K: Permutation and parametric bootstrap tests for gene-gene and gene-environment interactions. Ann Hum Genet 2011;75:36–45.

Cardon LR, Abecasis GR: Using haplotype blocks to map human complex trait loci. Trends Genet 2003;19:135–140.

Chapman JM, Whittaker J: Analysis of multiple SNPs in a candidate gene or region. Genet Epidemiol 2008;32:560–566.

Chatterjee N, Carroll RJ: Semiparametric maximum-likelihood estimation exploiting gene-environment independence in case-control studies. Biometrika 2005;92:399–418.

Chatterjee N, Kalaliouglu Z, Moslehi R, Peters U, Wacholder S: Powerful multi-locus tests of genetic association in the presence of gene-gene and gene-environment interactions. Am J Hum Genet 2006;79:1002–1016.

Chen M, Cho J, Zhao H: Detecting epistatic SNPs associated with complex diseases via a Bayesian classification tree search method. Ann Hum Genet 2011;75:112–121.

Chen SX, Qin YL: A two-sample test for high-dimensional data with applications to gene-set testing. Ann Statist 2010;38:808–835.

Conneely KN, Boehnke M: So many correlated tests, so little time! Rapid adjustment of p values for multiple correlated tests. Am J Hum Genet 2007;81:1158–1168.

Cordell HJ: Detecting gene-gene interactions that underlie human diseases. Nature Reviews Genetics 2009;10:392–404.

Cox DR, Hinkley DV: Theoretical Statistics. London, Chapman and Hall, 1974.

Durrant C, Zongdervan KT, Cardon LR, et al: Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. Am J Hum Genet 2004; 75:35–43.

Evans DM, Marchini J, Morris AP, Cardon LR: Two-stage two-locus models in genome-wide association. PLoS Genet 2006;2:e157.

Fan J: Test of significance based on wavelet thresholding and Neyman's truncation. JASA 1996;91:674–688.

Goeman JJ, van de Geer S, van Houwelingen HC: Testing against a high dimensional alternative. J R Stat Soc B 2006;68:477–493.

Han F, Pan W: A data-adaptive sum test for disease association with multiple common or rare variants. Hum Hered 2010a;70:42–54.

Han F, Pan W: Powerful multi-marker association tests: unifying genomic distance-based regression and logistic regression. Genet Epidemiol 2010b;34:680-688.

He H, Oetting WS, Brott MJ, Basu S: Pair-wise multifactor dimensionality reduction method to detect gene-gene interactions in a case-control study. Hum Hered 2010;69:60–70.

Hindorff LA, Junkins HA, Hall PN, Mehta JP, Manolio TA: A catalog of published genome-wide association studies. Available at: http://www.genome.gov/gwastudies (accessed October 31, 2010).

Kooperberg C, LeBlanc ML: Increasing the power of identifying gene × gene inter-actions in genome-wide association studies. Genet Epi 2008;32:255–263.

Kooperberg C, LeBlanc ML, Dai JY, Rajapakse I: Structures and assumptions: strategies to harness gene × gene and gene × environment interactions in GWAS. Statistical Science 2009;24:472–488.

Lin DY: An efficient Monte Carlo approach to assessing statistical significance in genomic studies. Bioinformatics 2005;21:781–787.

Lou X, Chen G, Yan L, Ma JZ, Zhu J, Elston RC, Li MD: A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. Am J Hum Genet 2007;80:1125–1137.

Marchini J, Donnelly P, Cardon LR: Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 2005;37:413–417.

Maher B: Personal genomes: the case of the missing heritability. Nature 2008;456:18–21.

Millstein J, Conti DV, Gilliland FD, Gauderman WJ: A testing framework for identifying susceptibility genes in the presence of epistasis. Am J Hum Genet 2006;78:15–27.

Moore JH: The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum Hered 2003;56:73–82.

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

Musani SK, Shriner D, Liu N, Feng R, Coffey CS, Yi N, Tiwari HK, Allison DB: Detection of gene × gene interactions in genome-wide association studies of human population data. Hum Hered 2007;63:67–84.

Neyman J: Smooth test for goodness of fit. Skandinavisk Aktuarietidskrift 1937;20:149-199.

Pan W: Asymptotic tests of association with multiple SNPs in linkage disequilibrium. Genet Epidemiol 2009;33:497–507.

Pan W: Statistical tests of genetic association in the presence of gene-gene and gene-environment interactions. Hum Hered 2010;69:131–142.

Pan W, Shen X: Adaptive tests for association analysis of rare variants. Genet Epidemiol 2011;35:381–388.

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Plummer WD, Parl FF, Moore JH: Multifactor dimensionality reduction reveals high-order interactions among estrogen metabolism genes in sporadic breast cancer. Am J Hum Genet 2001;69:138–147.

Schymick JC, Scholz SW, Fung HC, et al: Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: first stage analysis and public release of data. Lancet Neurol 2007;6:322–328.

Seaman SR, Muller-Myhsok B: Rapid simulation of p values for product methods and multiple testing adjustment in association studies. Am J Hum Genet 2005;76:399–408.

Song M, Nicolae DL: Restricted parameter space models for testing gene-gene interaction. Genet Epidemiol 2009;33:386–393.

Storey JD, Akey JM, and Kruglyak L: Multiple locus linkage analysis of genomewide expression in yeast. PLoS Biology 2005;3:1380–1390.

Thomas D: Gene-environment-wide association studies: emerging approaches. Nature Reviews Genetics 2010;11:259–272.

Yu K, Li Q, Bergen AW, Pfeiffer RM, Rosenberg PS, Caporaso N, Kraft P, Chatterjee N: Pathway analysis by adaptive combination of P-values. Genet Epidemiol 2009;33:700–709.

VanderWeele TJ, Laird NM: Tests for compositional epistasis under single interaction-parameter models. Ann Hum Genet 2011;75:146–156.
Wan X, Yang C, Yang Q, Xue H, Fan X, Tang N, Yu W: BOOST: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. Am J Hum Genet 2010;87:325–340.
Wang K: Genetic association tests in the presence of epistasis or gene-environment interaction. Genet Epidemiol 2008;32:606–614.
Wang T, Elston RC: Improved power by use of a weighted score test for linkage disequilibrium mapping. Am J Hum Genet 2007;80:353–360.
Wang T, Ho G, Ye K, Strickler H, Elston RC: A partial least-square approach for modeling gene-gene and gene-environment interactions when multiple markers are genotyped. Genet Epidemiol 2009;33:6–15.
Wessel J, Schork NJ: Generalized genomic distance-based regression methodology for multilocus association analysis. Am J Hum Genet 2006;79:792–806.
Wu J, Devlin B, Ringquist S, Trucco M, Roeder K: Screen and clean: a tool for identifying interactions in genome-wide association studies. Genet Epidemiol 2010;34:275–285.
Zapala MA, Schork NJ: Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. Proc Natl Acad Sci USA 2006;103:19430–19435.
Zhang K, Qin ZS, Liu JS, Chen T, Waterman MS, Sun F: Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies. Genome Res 2004;14:908–916.
Zhang Y, Liu JS: Bayesian inference of epistatic interactions in case-control studies. Nat Genet 2007;39:1167–1173.
Zhang Y, Zhang J, Liu JS: Block-based Bayesian epistasis association mapping with application to WTCCC type 1 diabetes data. Annals of Applied Statistics, in press.
Zhao JY, Jin L, Xiong MM: Test for interaction between two unlinked loci. Am J Hum Genet 2006;79:831–845.
Zheng T, Wang H, Lo SH: Backward genotype-trait association (BGTA) – based dissection of complex traits in case-control design. Hum Hered 2006;62:196–212.