Original Paper

Human Heredity

Hum Hered 2010;69:120–130
DOI: 10.1159/000264449

Received: April 27, 2009
Accepted after revision: July 8, 2009
Published online: December 4, 2009

Test Selection with Application to Detecting Disease Association with Multiple SNPs

Wei Pan[a] Fang Han[a] Xiaotong Shen[b]

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

Key Words
Fisher's method · Genome-wide association study · Logistic regression · Monte Carlo simulation · Power · Score test · Truncated product method

Abstract
We consider the motivating problem of testing for association between a phenotype and multiple single nucleotide polymorphisms (SNPs) within a candidate gene or region. Various statistical approaches have been proposed, including those based on either (combining univariate) single-locus analyses or (multivariate) multilocus analyses. However, it is known in theory that there is no single uniformly most powerful test to detect association with multiple SNPs. On the other hand, several tests have been shown to be among frequent winners across a range of practical situations, but the identity of the most powerful one changes with the situation in an unknown way. Here we propose a novel test selection procedure to select from five such tests: a so-called UminP test that combines multiple univariate/single-locus score tests by taking the minimum of their p values as its test statistic, a multivariate score test and its two modifications, and a so-called sum test. We also illustrate its application to selecting genotype codings for the sum test since the performance of the sum test depends on its genotype coding in an unknown way. Our major contributions include the methodology of estimating the power of a given test with a given dataset and the idea of using the estimated power as the criterion for test selection. We also propose a fast simulation-based method to calculate p values for the test selection procedure and for any method of combining p values. Our numerical results indicated that the proposed test selection procedure always yielded power close to the most powerful test among the candidate tests at any given situation, and in particular, our proposed test selection performed either better than or as well as the popular combining method of taking the minimum p value of the candidate tests.

1. Introduction

We consider the problem of testing for association between a phenotype and multiple single nucleotide polymorphisms (SNPs) within a gene or region, though our focus is on a binary phenotype, e.g. disease status, as arising from a genome-wide association study (GWAS) with the case-control design. Because the association between the phenotype and causal SNPs is often quite weak, it is compelling to find and use a statistical test with high power. Although there does not appear to exist any single uniformly most powerful test for multiple SNPs, some may tend to be more powerful than others in some situations. For example, Chapman and Whittaker [2008] found that either a Bayesian test of Goeman et al. [2005]

KARGER

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

© 2009 S. Karger AG, Basel
0001–5652/10/0692–0120$26.00/0

Accessible online at:
www.karger.com/hhe

Wei Pan
Division of Biostatistics, MMC 303, School of Public Health
University of Minnesota
Minneapolis, MN 55455-0392 (USA)
Tel. +1 612 626 2705, Fax +1 612 626 0660, E-Mail weip@biostat.umn.edu

or the UminP test that takes the minimum p value of univariate single-locus tests is often most powerful among some popular methods under some practical situations. Pan [2009] proposed two tests that are asymptotically equivalent to Goeman's test when permutation is used to calculate the p value for the latter; unlike Goeman's test, the asymptotic distributions of the former can be derived and used. On the other hand, a popular and standard genotype-based multivariate test is the score test based on joint logistic regression, and it is closely related to the generalized Hotelling's T^2 test [Fan and Knapp 2003; Xiong et al., 2002]. Zhong and Pan [2008] used the GAW16 Rheumatoid Arthritis data [Plenge et al., 2007] to empirically show that, depending on chromosome regions under study, each of the Goeman's test (or its variations as to be considered here), UminP test and the multivariate score test could be more powerful than the others, as to be shown subsequently.

Because there is no uniformly most powerful (unbiased) test for multiple parameters [Cox and Hinkley, 1974], as expected, several competitive tests (as discussed above) exist, and the answer to the question of which one is most powerful and thus to be used depends on the situation and is in general unknown. Hence, a practical approach is to apply multiple tests, and assess whether the smallest p value is significant after multiple-test adjustment. This strategy of choosing the minimum p value from multiple (dependent) tests, called the MinP method, is only one of several existing methods to combine p values from multiple tests. In this article, we will consider two other popular ones, Fisher's [1932] method and the truncated product method (TPM) [Zaykin et al., 2002]. Note that, although combining p values from multiple independent tests has been well studied [e.g. Loughin, 2004], this is not the case with multiple dependent tests as in the current context.

As an alternative to combining multiple tests, we propose selecting a most powerful test from a group of the candidate tests by estimating their power based on any given data. Although there is a huge literature on model selection, to our knowledge, there is no previous work on test selection, which we propose studying here. The two topics of test selection and test combination are closely related to each other. First, a test selection criterion as developed here may be incorporated into a more powerful methodology for combining tests, though we do not pursue it here. Second, in some situations, e.g. when it is known most of the tests may have low power, it may be a better idea to conduct test selection than test combination. For example, the performance of sum test is known to depend on the genotype coding in an unknown way, while most of the coding schemes may have low power [Chapman and Whittaker, 2008; Pan, 2009], it makes more sense to select a coding scheme giving high power than to combine many highly correlated and low powered sum tests. We will illustrate an application of our proposed methodology to select genotype coding schemes for the sum test, in addition to selecting from different types of tests. In general, it may be interesting to establish some analogy to model selection versus model combination [Shen and Huang, 2006; Yang, 2003] so that some insights into test selection and test combination can be provided.

An interesting observation as the premise for the application of our methodology is that quite a few tests, including all the five candidate tests to be considered here, are based on a component of or the whole score vector from a logistic regression model. Based on the asymptotic distribution of the score vector, we propose a fast simulation-based method to compute a p value for our test selection procedure and for any method to combine the p values from multiple tests. As an application, we also demonstrate the usefulness of the proposed method to select genotype coding schemes for the sum test, which has been shown to perform well in some situations but not in others due to its dependence on the genotype coding in an unknown way [Pan, 2009].

2. Methods

Although the methods may be extended to other study designs, we focus on the case-control study design with a binary phenotype, such as a disease indicator. We have m independent observations (Y_i, X_i), in which subject i has a binary phenotype (e.g. disease status) Y_i and genotype $X_i = (X_{i1}, ..., X_{ik})$. In this article, we consider only the dosage coding of X_{ij} for the additive mode of inheritance: X_{ij} = 0.1 or 2, representing the copy number of one of the two alleles present in SNP j of subject i, though other coding schemes, including a binary coding of X_{ij} = 0 or 1 for a dominant or recessive genetic model, can be equally used. Given the data, we would like to test whether there is any association between the phenotype and genotype.

2.1 Power Estimation

We assume that all the tests to be selected are based on a common multivariate statistic, such as the multivariate score statistic from logistic regression, which is true for the five candidate tests studied here (see Appendix A.1). In this way, we can take advantage of the nice property of the score statistic, whose asymptotic null distribution is multivariate normal with mean 0 and covariance matrix that can be consistently estimated. Specifically, we assume that based on a given dataset and a logistic regression model, we have a multivariate score statistic U, whose asymptotic null distribution is $N(0, V)$; see Appendix A.1 for closed forms of U and V. Now our goal is to estimate the power of any test with test statistic that is a function of U, say $T(U)$.

Since the null distribution of U is $N(0, V)$, in theory, the null distribution of $T(U)$, a function of U, is also known, though it may not have a simple or closed form. Similar to Conneely and Boehnke [2007], numerical method scan be used to calculate the rejection region $R(T, \alpha)$ for $T(U)$ at significance level α by solving the equation:

$$\text{Size}(T) = \int_{u \in R(T,\alpha)} \phi(u; 0, V)\, du = \alpha,$$

where $\Phi(.; u_1, V)$ is the density function of a multivariate normal distribution $N(u_1, V)$. If we know the true mean of U under the alternative hypothesis, say $E(U \mid H_1) = u_1$, then we can calculate the power of $T(U)$ simply as

$$\text{Power}(T, \alpha, u_1) = \int_{u \in R(T,\alpha)} \phi(u; u_1, V)\, du.$$

However, of course, we do not know u_1, which is the target of our hypothesis testing. An unbiased (and consistent) estimate of u_1 is U itself! Hence, we propose an estimator

$$\widehat{\text{Power}}(T, \alpha, U) = \int_{u \in R(T,\alpha)} \phi(u; U, V)\, du,$$

which, unfortunately, is biased. To see why, consider the case when H_0 holds. Then we have $u_1 = 0$ and thus $\text{Power}(T, \alpha) = \alpha$. On the other hand, for any finite sample, $U \neq 0$ with probability 1, implying that

$$\widehat{\text{Power}}(T, \alpha) > \alpha.$$

The bias and its corresponding estimate based on Monte Carlo simulations are respectively

$$\text{Bias}(T) = E\left[\widehat{\text{Power}}(T, \alpha, U^*) \mid U^* \sim N(0, V)\right] - \alpha,$$

and

$$\widehat{\text{Bias}}(T) = \sum_{b=1}^{B_0} \widehat{\text{Power}}\left(T, \alpha, U^{(b)}\right) / B_0 - \alpha,$$

where $U^{(1)}, \ldots, U^{(B_0)}$ are iid samples drawn from $N(0, V)$, the null distribution of the score statistics. The larger the replication number B_0, the more precise the resulting estimate, but also more time-consuming. We used $B_0 = 10$ throughout.

Finally, the estimated power of the test T is

$$\widetilde{\text{Power}}(T, \alpha, U) = \widehat{\text{Power}}(T, \alpha, U) - \max\left\{\widehat{\text{Bias}}(T), 0\right\}.$$

Note that the bias correction step is technically necessary: the bias of

$$\widehat{\text{Power}}(T, \alpha, U)$$

may depend on the test T; that is, without a bias correction, it may always favor one of the candidate tests. For example, among the tests based on the multivariate score statistic, say the score test versus one of its modifications called SSU test, if no bias correction is conducted, then we will always choose the score test, because it is known that the score test is the (estimated) most powerful test under the assumption of $u_1 = U$ as being used in our proposed test selection procedure [Cox and Hinkley, 1974; Pan 2009], though of course the score test may be less powerful than other tests.

2.2 Test Selection

2.2.1 Selecting from Candidate Tests

For a given dataset and a given list of candidate tests, say $T_1, \ldots, T_t$, we would like to select the test with the highest power. To do so, we only need to estimate the power of each test by calculating

$$\widetilde{\text{Power}}(T_i, \alpha, U)$$

for each $i = 1, \ldots, t$, and choose the test T_{i_0} with

$$i_0 = \text{argmax}_{i=1}^{t} \widetilde{\text{Power}}(T_i, \alpha, U),$$

such that T_{i_0} has the maximum estimated power among the candidate tests with the given data.

We use the p value of the test T_{i_0}, say p_{i_0}, as the test statistic for the test selection procedure. Such a chosen p value, as in choosing the minimum p value from multiple tests, is no longer a valid p value; a multiple test adjustment has to be made. We propose using a simulation based approach [Seaman and Muller-Myhsok, 2005; Chapman and Whittaker, 2008] to make an adjustment, which is faster than a permutation-based approach. Specifically, (i) we simulate some null score statistics $U_0^{(b)}$ iid from $N(0, V)$ for $b = 1, \ldots, B$; (ii) treating each $U_0^{(b)}$ as the observed score statistic, we apply the above test selection procedure to obtain its selected statistic $p_0^{(b)}$ (i.e. the p value from the estimated most powerful test among the candidates as applied to $U_0^{(b)}$); (iii) the final p value for the test selection procedure is the sample proportion of $p_{i_0} < p_0^{(b)}$. We used $B = 100$ throughout.

As the permutation-based method, the above simulation-based method takes account of the dependency among the candidate tests' being applied to the same data. Although the simulation-based method is in general faster than the permutation-based method, both methods are computationally demanding: their permutation or simulation number B, and thus their computational time, is inversely proportional to the largest p value to be significant after some multiple test adjustment. For example, if we would like to test disease association in 1,000 unlinked candidate regions, then, to achieve the statistical significance at 0.05 level, we need a p value for a significant association in a region to be no greater than 0.05/1,000, and hence at least $B = 10^5$ is needed. To save computing time, we may also use the Bonferroni adjustment, which, however, is known to be conservative. Specifically, with the estimated most powerful test T_{i_0} with p value p_{i_0}, the p value for the test selection procedure is min (1, $t * p_{i_0}$).

2.2.2 Selecting Genotype Coding for the Sum Test

As an illustration to a wide range of possible applications of the test selection procedure, we apply it to select the coding scheme for the sum test. A perceived drawback of the sum test is the dependence of its power on the genotype coding as clearly shown by its score statistic ([Chapman and Whittaker, 2008]; see Appendix A.4). For any given genotype coding X, we consider a different coding X^*: suppose that we partition the loci into two subsets, S_1 and its complement; for any i, $X_{ij}^* = X_{ij}$ if $j \in S_1$, and $X_{ij}^* = 2 - X_{ij}$ otherwise. If we replace X_{ij} by new X_{ij}^* in the logistic regression model (3), the corresponding score statistic for testing H_0: $\beta_c = 0$ is

$$U_c^* = \sum_{i=1}^{m} \left(Y_i - \bar{Y}\right) \sum_{j=1}^{k} X_{ij}^* = \sum_{j=1}^{k} 1_j^* U_j = 1^{*\prime} U, \qquad (1)$$

where

$$U = \left(U_1, \ldots, U_j\right)' = \sum_{i=1}^{m} \left(Y_i - \bar{Y}\right) \sum_{j=1}^{k} X_i$$

is the score vector based on the original genotype coding X, and $1^* = (1^*_1, \ldots, 1^*_k)'$ with $1^*_j = 1$ if $j \in S_1$ and $1^*_j = -1$ otherwise. In summary, the new score statistic U^*_c is just a known function of the score vector U for the original genotype coding. Furthermore, we have

$$Var(U^*_c) = 1^{*\prime}\ Cov(U)1^*,$$

where $Cov(U) = V$ is given in Appendix A.1. Hence, we can simulate U's from $N(0, V)$ to obtain the null statistics $U^*_c/\sqrt{Var(U^*)}$ for any new genotype coding X^*, from which its p value can be estimated. In this way, our proposed methods can be all equally applied to multiple sum tests with various genotype coding schemes for test selection and test combination (see Appendix B).

3. Results

3.1 Simulated Linkage Disequilibrium (LD) Patterns

We performed a simulation study following the set-ups given in Wang and Elston [2007] with $k = 10$ marker SNPs and sample size $n = 500$ or $n = 1{,}000$. The disease-causing SNP was assumed to be in the center of the marker SNPs, but was removed from the data. First, we generated a latent vector from a multivariate normal distribution with one of three covariance structures: a compound symmetry (CS) with an equal pairwise correlation $\rho = 0.4$, an AR-1 with the correlation $\rho_{ij} = 0.8^{|i-j|}$ between components i and j, and a correlation matrix with elements ρ_{ij} randomly between 0.3 and 0.7. Second, the latent vector was dichotomized to yield a haplotype with allele frequencies randomly chosen between 0.2 and 0.8 while the minor allele frequency (MAF) for the disease-causing SNP was fixed at 0.2, 0.3 or 0.4. Third, we combined two haplotypes and obtained marker genotype data $X_i = (X_{i1}, \ldots, X_{ik})'$ (and X_{0i} for disease-causing SNP) for subject i. Fourth, the disease status Y_i of subject i was generated from a logistic regression model:

$$\text{Logit Pr}(Y_i = 1) = \beta_0 + \log(OR)X_{i0}, \qquad (2)$$

where we chose $\beta_0 = -\log 4$ to give a background (i.e. not caused by the casual SNP) disease probability of 0.2, and the odds ratio (OR) ranged from 1 (i.e. no association) to 2. Finally, following the case-control design, we sampled n cases (with $Y_i = 1$) and n controls (with $Y_i = 0$). We excluded the disease-causing SNP, supplying $\{(Y_i, X_i): i = 1, 2, \ldots, 2n\}$ as a dataset to various statistical tests. For each set-up, we simulated 1,000 datasets, from which we obtained an empirical size or power for each test as its proportion of correctly or incorrectly rejecting its H_0; in particular, the Monte Carlo standard error of an empirical size/power $\hat{p}$ is $\sqrt{\hat{p}\,(1 - \hat{p})/1{,}000} \leq 0.016$.

Because a matrix with non-diagonal elements randomly between 0.3 and 0.7 may not be a proper correlation matrix that is positive definite, we generated the latent variables with a random correlation matrix in the following way. First, we generated iid random variates b_i, e_0 and e_{ij} for $i = 1, \ldots, 10$ and $j = 1, \ldots, 10$ from $N(0, 1/10)$. Second, for each $j = 1, \ldots, 10$, we generated a random integer j_0 uniformly from $U(3, 7)$. Suppose that the latent variables for the causal SNP and marker SNPs are Z_0 and $Z_1, \ldots, Z_{10}$. Then $Z_0 = \Sigma^{9}_{i=1}\ b_i + e_0$ and $Z_j = \Sigma^{j_0}_{i=1}\ b_i + \Sigma^{10}_{i=j_0+1}\ e_{ij}$ for $j = 1, \ldots, 10$. It is easy to verify that the correlation between any two latent variables was between 0.3 and 0.7, inclusively.

The simulation results are summarized in tables 1 and 2. Across all three correlation structures, the following conclusions can be drawn. First, among the five individual tests, SSU and SSUw had similar performance and were winners along with the sum test, followed by the UminP and then the (multivariate) score test. Second, among the three combining methods, the MinP method was the least powerful, while TPM seemed to have a slight edge over Fisher's method. Note that the power difference between the MinP and TPM could be often as large as 10% in absolute power. Third, in any case, the proposed test selection procedure always yielded power close to the highest one.

Note that, though some Type I error rates of the test selection procedure were slightly larger than the nominal level of 0.05, 0.05 fell within the 95% confidence intervals of the Type I error rates. Because the method depends on the asymptotic normality of the score statistic, the method could perform better with a larger sample size: the Type I error rates for $n = 1{,}000$ seemed to be closer to the nominal level than were those for $n = 500$. Other sources of approximation errors could be due to relatively small values of B_0 and B in the simulation-based method to calculate the p values. When we increased B_0 and B to 1,000 and 200 respectively, the Type I error rate estimates for the three correlation structures in table 1 decreased from 0.059, 0.063 and 0.065 to 0.049, 0.053 and 0.056, respectively. Furthermore, if we increase the number of simulations, we expect to have more accurate Type I error estimates.

3.2 HapMap Data for Gene IL21R

As in Chapman and Whittaker [2008], we also considered the region of gene IL21R, in which LD was low. We

Table 1. Empirical sizes and powers of various tests with nominal significance level $\alpha = 0.05$ for simulated data with three correlation structures (compound symmetry (CS), AR-1 and random (Rand)) and 10 SNPs; n = 500

Corr	OR	Sco	SSU	SSUw	UminP	Sum	Test combination			Test selection
							MinP	Fisher	TPM	
CS	1.0	0.045	0.043	0.044	0.046	0.051	0.043	0.045	0.045	0.059
	1.2	0.057	0.080	0.077	0.077	0.098	0.079	0.084	0.080	0.100
	1.4	0.085	0.193	0.199	0.149	0.235	0.153	0.193	0.198	0.201
	1.6	0.139	0.356	0.358	0.237	0.393	0.255	0.353	0.357	0.372
	1.8	0.245	0.519	0.518	0.365	0.577	0.405	0.504	0.520	0.540
	2.0	0.346	0.662	0.661	0.489	0.711	0.546	0.644	0.663	0.665
AR-1	1.0	0.051	0.048	0.048	0.040	0.055	0.043	0.046	0.050	0.063
	1.2	0.077	0.125	0.124	0.109	0.132	0.111	0.129	0.131	0.145
	1.4	0.182	0.352	0.354	0.296	0.350	0.298	0.356	0.354	0.361
	1.6	0.356	0.577	0.585	0.512	0.599	0.516	0.582	0.585	0.584
	1.8	0.543	0.785	0.783	0.712	0.797	0.720	0.784	0.781	0.774
	2.0	0.719	0.901	0.896	0.849	0.894	0.848	0.898	0.901	0.899
Rand	1.0	0.047	0.046	0.044	0.052	0.044	0.052	0.053	0.049	0.065
	1.2	0.076	0.114	0.116	0.086	0.134	0.086	0.116	0.121	0.137
	1.4	0.141	0.284	0.281	0.207	0.319	0.216	0.273	0.284	0.312
	1.6	0.238	0.500	0.505	0.357	0.546	0.392	0.495	0.507	0.508
	1.8	0.378	0.721	0.718	0.531	0.753	0.585	0.708	0.720	0.717
	2.0	0.517	0.836	0.837	0.687	0.863	0.742	0.830	0.836	0.837

Table 2. Empirical sizes and powers of various tests with nominal significance level $\alpha = 0.05$ for simulated data with three correlation structures and 10 SNPs; n = 1,000

Corr	OR	Sco	SSU	SSUw	UminP	Sum	Test combination			Test selection
							MinP	Fisher	TPM	
CS	1.0	0.033	0.033	0.032	0.032	0.040	0.029	0.033	0.035	0.043
	1.2	0.053	0.119	0.116	0.082	0.139	0.090	0.113	0.119	0.133
	1.4	0.145	0.349	0.344	0.237	0.386	0.253	0.341	0.345	0.361
	1.6	0.310	0.641	0.641	0.450	0.673	0.506	0.610	0.639	0.633
	1.8	0.499	0.836	0.835	0.655	0.872	0.716	0.821	0.841	0.841
	2.0	0.691	0.935	0.933	0.820	0.965	0.870	0.930	0.933	0.943
AR-1	1.0	0.053	0.029	0.030	0.045	0.036	0.043	0.036	0.034	0.053
	1.2	0.102	0.210	0.206	0.176	0.209	0.181	0.210	0.207	0.205
	1.4	0.370	0.593	0.604	0.511	0.606	0.535	0.606	0.601	0.606
	1.6	0.660	0.877	0.871	0.802	0.877	0.815	0.872	0.877	0.866
	1.8	0.869	0.978	0.978	0.959	0.976	0.962	0.981	0.980	0.978
	2.0	0.961	0.997	0.996	0.989	0.996	0.991	0.996	0.996	0.996
Rand	1.0	0.056	0.047	0.046	0.053	0.052	0.056	0.053	0.053	0.065
	1.2	0.106	0.194	0.196	0.144	0.200	0.150	0.186	0.196	0.205
	1.4	0.270	0.517	0.521	0.356	0.568	0.391	0.495	0.515	0.522
	1.6	0.495	0.801	0.804	0.636	0.843	0.701	0.785	0.799	0.806
	1.8	0.739	0.944	0.942	0.836	0.953	0.883	0.932	0.944	0.938
	2.0	0.895	0.988	0.988	0.942	0.991	0.968	0.983	0.987	0.986

Table 3. Empirical sizes and powers of various tests with nominal significance level $\alpha = 0.05$ for simulated data based on the LD patterns in gene IL21R with 27 SNPs

n	OR	Sco	SSU	SSUw	UminP	Sum	Test combination			Test selection
							MinP	Fisher	TPM	
500	1.0	0.039	0.042	0.044	0.055	0.050	0.053	0.040	0.041	0.068
	1.2	0.123	0.202	0.208	0.182	0.164	0.180	0.197	0.208	0.233
	1.3	0.205	0.402	0.402	0.417	0.304	0.413	0.432	0.419	0.453
	1.4	0.369	0.594	0.589	0.652	0.432	0.650	0.644	0.618	0.682
	1.5	0.526	0.737	0.740	0.829	0.527	0.829	0.800	0.779	0.831
	1.6	0.691	0.828	0.836	0.909	0.607	0.911	0.904	0.875	0.918
1,000	1.0	0.042	0.044	0.043	0.049	0.034	0.046	0.047	0.049	0.066
	1.2	0.214	0.390	0.391	0.411	0.286	0.404	0.403	0.397	0.445
	1.3	0.441	0.679	0.670	0.730	0.499	0.729	0.734	0.708	0.760
	1.4	0.726	0.826	0.826	0.917	0.604	0.919	0.892	0.873	0.922
	1.5	0.905	0.912	0.920	0.974	0.699	0.976	0.970	0.971	0.975
	1.6	0.973	0.960	0.966	0.991	0.767	0.993	0.993	0.991	0.989

followed exactly the same steps except that as in Chapman and Whittaker [2008], the disease-causing SNP was selected *randomly* and then excluded from the data in each simulation run. At the end, we had 28 SNPs. As shown in table 3, first, among the five individual tests, the UminP test was the winner, followed by SSU and SSUw, then by the sum test and finally by the score test. Second, among the three combining methods, the MinP and Fisher's methods appeared to be the winners, closely followed by TPM. Third, the test selection procedure had power close to the UminP, the winner among the individual candidate tests, and higher than any of the three combining methods for small ORs.

3.3 HapMap Data for a Region on Chromosome 9

Zhong and Pan [2008] found a region on chromosome 9 with the GAW16 RA data for which the power of the SSU or SSUw was low. The region of about 15 Kb contained eight SNPs: rs10985997, rs1891641, rs2491352, rs872863, rs4838057, rs12348586, rs4466467 and rs7865976. For the HapMap CEU samples, after randomly imputing for missing genotypes, we found that the genotype codings of rs1891641, rs12348586 and rs7865976 were perfectly correlated (with pairwise correlation coefficient 1); so were another three SNPs: rs872863, rs4838057 and rs4466467. For each simulated dataset, we randomly selected one of the eight SNPs as disease-causing and removed it from the subsequent analysis. First, among the five individual tests, the score test was the most powerful, followed by the UminP, SSUw and SSU tests sequentially, while the sum test had by far the lowest power (table 4). Second, among the three combining methods, the MinP method was the winner, closely followed by the TPM and Fisher's method. Third, the test selection procedure had either slightly higher or comparable power than or as the MinP method.

3.4 HapMap Data for Gene CHI3L2

We conducted a simulation study based on real LD patterns within the CHI3L2 gene for the 90 CEU samples. As in Wang and Elston [2007], first, we excluded SNPs with MAF ≤ 0.2, leaving 23 SNPs. Second, we did a single imputation for each of the missing genotypes by randomly drawing an observed genotype of the same SNP. Third, we used the dosage coding for the SNPs and tried to minimize the number of negative correlations among them. Fourth, we deleted redundant SNPs that were perfectly correlated with other SNPs, leading to 17 SNPs remaining. Fifth, we repeatedly sampled (with replacement) subjects from the 90 CEU individuals. Finally, as Wang and Elston [2007], we chose the SNP rs2182114 as disease-causing. As shown in table 5, first, among the five individual tests, the SSU and SSUw tests were the winners, followed by the sum and UminP tests, and finally by the score test. Second, among the three combining methods, the TPM and Fisher's methods appeared to be the winners, followed by MinP. Third, the test selection procedure had power that was comparable to that of the most powerful individual test, but slightly higher than any combining method for small ORs.

Table 4. Empirical sizes and powers of various tests with nominal significance level $\alpha = 0.05$ for simulated data based on the LD patterns in a chromosome 9 region with 7 SNPs

n	OR	Sco	SSU	SSUw	UminP	Sum	Test combination			Test selection
							MinP	Fisher	TPM	
500	1.0	0.055	0.051	0.047	0.052	0.046	0.055	0.056	0.059	0.069
	1.2	0.205	0.163	0.156	0.177	0.102	0.182	0.168	0.171	0.208
	1.4	0.615	0.443	0.459	0.522	0.261	0.569	0.514	0.535	0.602
	1.6	0.869	0.703	0.747	0.805	0.424	0.837	0.791	0.820	0.855
	1.7	0.933	0.784	0.810	0.882	0.495	0.915	0.879	0.898	0.913
	1.8	0.965	0.856	0.861	0.926	0.539	0.955	0.938	0.942	0.958
	2.0	0.992	0.955	0.924	0.978	0.626	0.988	0.981	0.984	0.988
1,000	1.0	0.061	0.053	0.049	0.053	0.065	0.059	0.060	0.056	0.068
	1.1	0.121	0.098	0.093	0.106	0.083	0.098	0.102	0.102	0.133
	1.2	0.394	0.277	0.286	0.318	0.179	0.343	0.302	0.324	0.378
	1.3	0.672	0.520	0.539	0.599	0.314	0.632	0.603	0.599	0.662
	1.4	0.862	0.711	0.742	0.794	0.440	0.839	0.798	0.805	0.860
	1.5	0.952	0.836	0.857	0.907	0.532	0.941	0.925	0.929	0.948
	1.6	0.984	0.925	0.914	0.963	0.606	0.977	0.973	0.978	0.981
	1.7	0.993	0.965	0.952	0.982	0.647	0.990	0.989	0.991	0.991

Table 5. Empirical sizes and powers of various tests with nominal significance level $\alpha = 0.05$ for simulated data based on the LD patterns in gene CHI3L2 with 17 SNPs

n	OR	Sco	SSU	SSUw	UminP	Sum	Test combination			Test selection
							MinP	Fisher	TPM	
500	1.0	0.050	0.050	0.049	0.050	0.050	0.049	0.056	0.056	0.056
	1.1	0.068	0.125	0.123	0.107	0.129	0.102	0.119	0.121	0.133
	1.2	0.144	0.363	0.358	0.288	0.346	0.291	0.334	0.343	0.373
	1.3	0.293	0.671	0.672	0.589	0.633	0.577	0.654	0.641	0.645
	1.4	0.470	0.891	0.894	0.852	0.839	0.827	0.882	0.872	0.878
	1.5	0.718	0.974	0.972	0.950	0.952	0.949	0.967	0.968	0.965
	1.6	0.883	0.995	0.996	0.990	0.989	0.987	0.992	0.993	0.992
1,000	1.0	0.049	0.042	0.045	0.052	0.045	0.048	0.046	0.053	0.056
	1.1	0.079	0.204	0.199	0.170	0.190	0.153	0.197	0.199	0.212
	1.2	0.271	0.644	0.636	0.569	0.608	0.545	0.613	0.618	0.642
	1.3	0.571	0.928	0.930	0.896	0.898	0.886	0.920	0.915	0.921
	1.4	0.871	0.993	0.992	0.989	0.985	0.986	0.992	0.991	0.992

3.5 Selecting Genotype Coding for the Sum Test

We applied the sum tests with various genotype coding schemes to the HapMap data for gene IL21R as before. For each locus, there are two possible coding schemes by counting the number of either alleles. Thus, there are 2^{27} possible genotype coding schemes for gene IL21R with 27 SNPs. Because it was not feasible to consider all possible coding schemes, we randomly chose 256 possible coding schemes, estimated their powers by simulations, and thus chose the ones with power at the maximum, first quartile (Q1), median (Q2), third quartile (Q3) and minimum, respectively, plus the original coding scheme (Ori) obtained from the heuristic algorithm of Pan [2009] that aims to minimize the number of SNP pairs with negative Pearson correlations. We then treated these six sum tests (with the six coding schemes) as the candidate tests and did a simu-

Table 6. Empirical sizes and powers of the sum tests with various genotype codings with nominal significance level $\alpha = 0.05$ for simulated data based on the LD patterns in gene IL21R with 27 SNPs

n	OR	Ori	Max	Q3	Q2	Q1	Min	Test combination			Test selection
								MinP	Fisher	TPM	
500	1.0	0.050	0.051	0.051	0.044	0.058	0.051	0.051	0.028	0.033	0.051
	1.2	0.164	0.159	0.173	0.167	0.112	0.062	0.136	0.111	0.121	0.141
	1.3	0.304	0.295	0.286	0.307	0.176	0.079	0.280	0.234	0.242	0.280
	1.4	0.432	0.438	0.424	0.417	0.246	0.097	0.427	0.358	0.378	0.422
	1.5	0.527	0.547	0.504	0.501	0.317	0.117	0.547	0.467	0.480	0.549
	1.6	0.607	0.633	0.596	0.564	0.378	0.139	0.643	0.553	0.572	0.645
1,000	1.0	0.034	0.035	0.042	0.040	0.068	0.058	0.058	0.026	0.028	0.055
	1.1	0.112	0.115	0.110	0.116	0.093	0.050	0.097	0.069	0.080	0.094
	1.2	0.286	0.285	0.286	0.287	0.177	0.077	0.256	0.211	0.229	0.258
	1.3	0.499	0.503	0.474	0.496	0.284	0.115	0.497	0.413	0.437	0.491
	1.4	0.604	0.633	0.608	0.572	0.388	0.152	0.672	0.584	0.594	0.670
	1.5	0.699	0.725	0.679	0.635	0.472	0.177	0.772	0.671	0.683	0.771
	1.6	0.767	0.788	0.749	0.686	0.531	0.214	0.834	0.756	0.752	0.834

lation study as before. The simulation results are shown in table 6, from which we observe the following. First, among the six candidate sum tests, the power difference could be large and, as explained by Pan [2009], the coding given by the heuristic algorithm yielded good power, but not necessarily the highest. Second, among the three combining methods, the MinP method was the consistent winner with substantial power improvement over the other two methods. Third, the test selection procedure was always as powerful as the most powerful sum test, and in this case also as powerful as the MinP method.

4. Discussion

As discussed in Pan [2009], among the five candidate association tests considered here, the SSU and SSUw perform well across a wide range of scenarios, but under some situations the UminP test (e.g. for gene IL21R as shown in table 3) and the multivariate score test (e.g. for a region in chromosome 9 as shown in table 4) can be more powerful. The sum test could be powerful if a suitable but unknown coding scheme is used. Hence, in general, the identity of the most powerful test for any given data is unknown, depending on the given situation, e.g. disease-genotype association patterns and LD patterns among the SNPs. The main motivation of this work is to develop a data-adaptive procedure to select the most powerful test from a set of candidate tests for any given data. The concept of the data-adaptivity adopted here is different from a criterion called 'efficiency robustness' (e.g. Freidlin et al., 2002; Zheng and Ng, 2008, and references therein): among a given set of candidate tests, the latter would select a test that has a high *minimum* power across various situations, whereas ours selects a test with the highest (estimated) power for a given situation.

For test combination, either the MinP or TPM method, but not both, seemed to work well with the power close to the most powerful candidate test across many situations considered here. However, when one of the two combining methods worked well, there might be a substantial power loss associated with the other. Because there is no uniformly most powerful test for multiple SNPs, as expected, no single combining method is uniformly most powerful either. The situation is similar to that for the individual association tests: which combining method or individual test is most powerful depends on the *unknown* data distribution. For example, the commonly used MinP method that selects the smallest p value, albeit powerful under many situations, might be less powerful than Fisher's method, and vice versa. Our proposed test selection procedure not only provides a means to estimate the power of a given test, but also to select the most powerful one and yield a valid p value. As shown in our numerical examples, even though the identity of the most powerful individual test or combining method changes with the situation in an unknown way, our proposed test selection procedure performed consistently well with its power always close to the most powerful individual test or combining method.

We have proposed and studied a simulation-based method to calculate the p value for the proposed test selection procedure, as for the UminP test and various test combination procedures. The goal is to account for multiple tests after test selection. Although the proposed simulation-based method, without refitting a model and recalculating test statistics, is much faster than a permutation-based method, both methods are computationally intensive: the required simulation or permutation number B has to be large if a highly significant p value is aimed, e.g. after a multiple test adjustment for a GWAS. For our simulations reported in tables 1 and 2, with B = 100, it took about half a minute for a dataset with either sample size n = 500 or n = 1,000 as implemented in R on a Linux PC. However, to achieve a p value in the order of 10^{-6}, we need $B = 10^6$, requiring several days for each dataset. Hence, our current implementation can only be applied to candidate genes or regions, not GWASs. If a more efficient implementation (e.g. in C or Fortran) is available with parallel computers, or the Bonferroni adjustment is adopted, it may be applied to GWASs.

We comment on some differences between test selection and model selection. Although there may be a correspondence between a candidate test and a candidate model, it is possible to have multiple tests based on the same model. For example, the multivariate score test, SSU and SSUw tests are all based on the same multiple logistic regression model (model (3) in Appendix A.1), but differing in how their test statistics are constructed. We have also showed how to select various versions of the sum test with different coding schemes on genotypes, all based on the same single logistic regression model (model (5) in Appendix A.4).

Although our proposed methodology is applicable to a large family of score-based tests, independent of the type of the phenotype or study design, extensions to other scenarios are warranted. First, in general, as long as the candidate tests are based on some statistics with known asymptotic distributions and the dependency of these statistics can be modeled or simulated, our methodology may be applicable. Second, even some existing tests are not score-based, a unified formulation of various tests in a general frame work, e.g. as some score-based tests in an expanded logistic regression model [Pan, 2009b], will largely facilitate the application of our proposed methodology. Further studies to extend the proposed test selection procedure to haplotype-based tests [e.g. Chapman et al., 2003; Schaid et al., 2002; Stephens et al., 2001; Zhao et al., 2003a, b, and references therein] and other tests, are needed.

R code will be posted on http://www.biostat.umn.edu/~weip, and available upon request.

Appendix A: Candidate Association Tests

A.1 A Global Test: The Multivariate Score Test
To test any possible association between the binary phenotype and any of the SNPs, we use a joint logistic regression model

$$\text{Logit Pr}\left(Y_i = 1\right) = \beta_0 + \sum_{j=1}^{k} X_{ij}\beta_j, \tag{3}$$

where Y_i = 0 or 1 indicates whether subject i is a control (i.e. without disease) or a case (i.e. with disease). A global test of any possible association between the phenotype and SNPs can be formulated as jointly testing on the multiple parameters β_j's with the null hypothesis H_0: $\beta = (\beta_1, ..., \beta_k) = 0$ by the likelihood ratio test (LRT), Wald test or score test in the context of logistic regression (or more generally, of generalized linear models for other types of responses); the three tests are asymptotically equivalent. In this paper, we will focus on the score test. As shown by Chapman et al. [2003], the score test statistics

$$T_{\text{score}} = U'V^{-1}U,$$

with the score vector as

$$U = \sum_{i=1}^{m}\left(Y_i - \bar{Y}\right)X_i,$$

and its covariance matrix as

$$V = \bar{Y}\left(1 - \bar{Y}\right)\sum_{i=1}^{m}\left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right)',$$

where

$$\bar{Y} = \sum_{i=1}^{m} Y_i \,/\, m \text{ and } \bar{X} = \sum_{i=1}^{m} X_i/m.$$

Under H_0, the score test statistic has an asymptotic chi-squared distribution χ_r^2 with degrees of freedom DF = rank (V). A potential problem with the test is that, for a large k, the test can be low-powered because of the large DF.

A.2 UminP Test: Combining Marginal or Univariate Tests
In contrast to the global or joint test, an other extreme is to conduct SNP-by-SNP single-locus tests: rather than including all the k SNPs in a joint model, we include only one SNP in a logistic model

$$\text{Logit Pr}(Y_i = 1) = \beta_{M,0j} + X_{ij}\beta_{M,j}, \tag{4}$$

where we explicitly distinguish $\beta_M = (\beta_{M,1}, ..., \beta_{M,k})'$ in marginal models (4) from β in joint model (3). Then we test $H_{0,j}$: $\beta_{M,j} = 0$ for each j = 1, ..., k sequentially. It turns out that the univariate score test statistic for $H_{0,j}$ is

$$T_{\text{score},j} = U_j^2 \,/\, v_j,$$

where U_j is the j-th component of the score vector U, and v_j the j-th diagonal element of V. Under $H_{0,j}$, $T_{\text{score}\,j}$ has an asymptotic distribution χ_1^2, from which we obtain a p value, say $p_{M,j}$. As usual, to test H_0, we can combine the k individual p values by taking $p_U = \min \{p_{M,1}, ..., p_{M,k}\}$, resulting in the so-called UminP test. Although other combining methods may be equally applied [Roeder et al., 2005], due to its simplicity and often good performance, we only consider the UminP test here.

Although each individual univariate test has DF = 1, a multiple test adjustment for p_U has to be made, often based on either permutation or Bonferroni adjustment. Because the Bonferroni adjustment is known to be conservative, permutation is more widely used, though it is computationally demanding. The multiple test adjustment may reduce the power of the test, as shown by Wang and Elston [2007]. Here we use simulations to approximate the asymptotic distribution of p_U, as to be discussed in Appendix B.

A.3 Modified Score Tests: SSU and SSUw

In contrast to the usual multivariate score test with statistic $T_{\text{score}} = U'V^{-1}U$, an alternative is to simply use

$$SSU = U'U,$$

ignoring the covariance matrix of the score vector U. This test is related to the test of Goeman et al. [2005]; in fact, the above SSU is equivalent to the permutation-based version of Goeman's test. Although Goeman's test was derived under an empirical Bayes framework to test on a large number of parameters, as arising in microarray gene expression data, Chapman and Whittaker [2008] found that Goeman's test worked impressively well across a wide range of scenarios, as also confirmed by Pan [2009].

A weighted form of the above test is the SSUw

$$SSUw = U'V_d^{-1}U,$$

with $V_d = \text{Diag}(V)$ as a diagonal matrix. SSU can be interpreted as an *estimated* most powerful test [Pan, 2009], which also partially explains the good performance of SSU. Often, SSU and SSUw perform similarly, but for some situations, SSUw can be more powerful [Zhong and Pan, 2008].

Asymptotically, each of the two test statistics is a quadratic form of normal variates, $Q = U'W^{-1}U$, with $W = I$ or $W = \text{Diag}(V)$ respectively. It is well known [e.g. Johnson and Kotz, 1970, p. 150] that the distribution of Q is a weighted sum of k independent chi-squared variates with DF = 1, $\Sigma_{j=1}^{k} c_j \chi^2{}_1$, where c_j's are the eigenvalues of VW^{-1}. Furthermore, by the results of Zhang [2005], $\Sigma_{j=1}^{k} c_j \chi^2{}_1$ can be well approximated by $a\chi_d^2 + b$ with

$$a = \frac{\sum_{j=1}^{k} c_j^3}{\sum_{j=1}^{k} c_j^2}, b = \sum_{j=1}^{k} c_j - \frac{\left(\sum_{j=1}^{k} c_j^2\right)^2}{\sum_{j=1}^{k} c_j^3}, d = \frac{\left(\sum_{j=1}^{k} c_j^2\right)^3}{\left(\sum_{j=1}^{k} c_j^3\right)^2}.$$

Note that the above approximation is independent of the sample size. To calculate a p value, for example for the SSU test,

$$\Pr(SSU > s \mid H_0) \approx \Pr(\chi_d^2 > (s - b) / a).$$

A.4 The Sum Test

To reduce the DF of the global test in model (3), Pan [2009] proposed a so-called sum test as a compromise. Under the possibly mis-specified working assumption of a common association strength with $\beta_1 = ... = \beta_k = \beta_c$, model (3) reduces to

$$\text{Logit}\Pr\left(Y_i = 1\right) = \beta_0 + \sum_{j=1}^{k} X_{ij}\beta_c, \tag{5}$$

then we can test $H_{0,c}$: $\beta_c = 0$ with a minimum DF = 1. Similar to the weighted score test (WST) proposed by Wang and Elston [2007], the sum test can have high power if $\beta_{M,j}$'s close; on the other hand, if $\beta_{M,j}$'s have different signs, the power can be low. Exactly for the latter reason, Chapman and Whittaker [2008] did not recommend the use of the sum test (or WST). The sum test can be also expressed in terms of the score statistic U, thus facilitating combining its use with other tests. It is easy to see that the univariate score statistic for (5) is

$$U_c = \sum_{i=1}^{m}\left(Y_i - \bar{Y}\right)\sum_{j=1}^{k} X_{ij} = \sum_{j=1}^{k} U_j = 1'U, \tag{6}$$

where 1 is a vector with all elements equal to 1. By the asymptotic normality of U, the asymptotic null distribution of U_c is normal $N(0, 1'V1)$.

Appendix B: Combining Multiple Tests

Given L p values, $p_1, ..., p_L$, obtained from L (possibly dependent) tests on H_0, we can combine the p values in several ways:

- The MinP method: $T_{\text{Min}} = \min(p_1, ..., p_L)$.
- The Fisher method: $T_F = \Pi_{j=1}^{L} p_{j.}$
- The truncated product method (TPM): $T_{\text{TPM}} = \Pi_{j=1}^{L} p_j I(p_j < \tau)$, where τ is some cutoff; as in Zaykin et al. [2002], we used $\tau = \alpha = 0.05$ throughout.

To obtain a p value for each combining function, say $C(p_1, ..., p_k)$, we can use a permutation method by permuting Y, which however is computationally demanding for its requirement of fitting models many times. Here we propose using a simulation-based approach [Seaman and Muller-Myhsok, 2005; Chapman and Whittaker, 2008]. First, we note that each individual test is based on a component of or the whole score vector U. Second, because of the asymptotic null distribution of U is known as $U \sim N(0, V)$, we can simulate B iid copies of U^b's from $N(0, V)$ with $b = 1, 2..., B$. Based on each U^b, we can calculate individual p values as $p_1^b, ..., p_k^b$, and thus $C(p_1^b, ..., p_k^b)$.

Third, the p value for $C(p_1, ..., p_k)$ is simply $\Sigma_{b=1}^{B} I\,[C(p_1, ..., p_k) < C(p_1^b, ..., p_k^b)] / B$. We used $B = 1{,}000$ throughout. An alternative method of Lin [2005] can be equally easily implemented.

We consider combining the p values from five tests: the UminP test, the multivariate score test, the SSU and SSUw tests, and the sum test. A combining function C combines the p values from the above five tests. We also consider combining multiple versions of the sum test with various genotype coding schemes.

Acknowledgments

This research was partially supported by NIH grants GM081535 and HL65462. We thank the two reviewers for helpful and constructive comments.

References

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

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

Clayton D, Chapman J, Cooper J: Use of unphased multilocus genotype data in indirect association studies. Genet Epidemiol 2004; 27:415–428.

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.

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

Fan R, Knapp M: Genome association studies of complex diseases by case-control designs. Am J Hum Genet 2003;72:850–868.

Fisher RA: Statistical Methods for Research Workers, ed 4. London, Oliver & Boyd, 1932.

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

Goeman JJ, van de Geer S, van Houwelingen HC: Testing against a high dimensional alternative. JR Stat Soc B 2005;68:477–493.

Johnson NL, Kotz S: Distributions in Statistics, Continuous Univariate Distributions, vol 2. Boston, Houghton-Mifflin, 1970.

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

Loughin TM: A systematic comparison of methods for combining p values from independent tests. Computational Statistics and Data Analysis 2004;47:467–485.

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

Pan W: A unified framework for detecting genetic association with multiple SNPs in a candidate gene or region: contrasting genotype scores and LD patterns between cases and controls. Hum Hered 2009b; to appear.

Plenge RM, et al: TRAF1-C5 as a risk locus for rheumatoid arthritis – a genome wide study. N Engl J Med 2007;357:1199–1209.

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

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA: Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 2002; 70:425–434.

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.

Shen X, Huang H: Optimal model assessment, selection and combination. J Am Stat Assoc 2006;101:554–568.

Stephens M, Smith NJ, Donnelly P: A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 2001;68:978–989.

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.

Xiong M, Zhao J, Boerwinkle E: Generalized T^2 test for genome association studies. Am J Hum Genet 2002;70:1257–1268.

Yang Y: Regression with multiple candidate models: selecting or mixing? Statistica Sinica 2003;13:783–809.

Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS: Truncated product method for combing p values. Genet Epidemiol 2002;22:170–185.

Zhang J-T: Approximate and asymptotic distributions of Chi-squared-type mixtures with applications. J Am Stat Assoc 2005;100:273–285.

Zhao H, Pfiffer R, Gail MH: Haplotype analysis in population genetics and association studies. Pharmacogenomics 2003a;4:171–178.

Zhao LP, Li S, Khalid N: Assessing haplotype-based association with multiple SNPs in case-control studies. Am J Hum Genet 2003b;72:1231–1250.

Zheng G, Ng HKT: Genetic model selection in two-phase analysis for case-control association studies. Biostatistics 2008;9:391–399.

Zhong W, Pan W: Power comparison of statistical tests of association with multiple SNPs with the GAW16 rheumatoid arthritis data. Available at www.biostat.umn.edu/rrs.php as Research Report rr2008–015, Division of Biostatistics, University of Minnesota, 2008.