Published in final edited form as:

Ann Hum Genet. 2012 May ; 76(3): 246–260. doi:10.1111/j.1469-1809.2012.00706.x.

Similarity-based multi-marker association tests for continuous traits

Wan-Yu Lin¹, Hemant K. Tiwari¹, Guimin Gao², Kui Zhang¹, John J Arcaroli³, Edward Abraham⁴, and Nianjun Liu¹

¹Section on Statistical Genetics, Department of Biostatistics, University of Alabama at Birmingham, Birmingham, Alabama

²Department of Biostatistics, Virginia Commonwealth University, Richmond, Virginia

³Division of Medical Oncology, University of Colorado Denver, Anschutz Medical Campus, Aurora, Colorado

⁴Department of Medicine, University of Alabama at Birmingham, 1530 3rd Avenue South, BDB 420, 35294-0012 Birmingham, Alabama, USA

Summary

Testing multiple markers simultaneously not only can capture the linkage disequilibrium patterns but also can decrease the number of tests and thus alleviate the multiple-testing penalty. If a gene is associated with a phenotype, subjects with similar genotypes in this gene should also have similar phenotypes. Based on this concept, we have developed a general framework that is applicable to continuous traits. Two similarity-based tests (namely, SIMc and SIMp tests) were derived as special cases of the general framework. In our simulation study, we compared the power of the two tests with that of the single-marker analysis, a standard haplotype regression, and a popular and powerful kernel machine regression. Our SIMc test outperforms other tests when the average r -square (a measure of linkage disequilibrium) between the causal variant and the surrounding markers is larger than 0.3 or when the causal allele is common (say, frequency = 0.3). Our SIMp test outperforms other tests when the causal variant was introduced at common haplotypes (the maximum frequency of risk haplotypes > 0.4). We also applied our two tests to an adiposity data set to show their utility.

Keywords

Haplotype; Similarity; Genomic distance; Linkage disequilibrium; Multi-marker test; Body-mass index; *CPE* gene

1. Introduction

Multi-marker association testing has become an important approach for investigating putative loci affecting common human disorders (de Bakker *et al.*, 2005). Testing multiple genetic markers simultaneously not only can capture the joint effects of these adjacent markers while considering the linkage disequilibrium (LD) patterns between them but also can decrease the number of tests. For detecting multi-marker associations, in addition to the traditional regression-based approach, testing associations by using the similarities

Correspondence to: Nianjun Liu.

The authors declare that they have no conflict of interest.

constructed by DNA sequences among pairs of subjects is another approach. If a gene is associated with a phenotype, subjects with similar genotypes in this gene should have similar phenotypes as well. A traditional regression analysis usually captures the main effects of genetic variants without considering higher-order interactions among them. However, similarity-based methods can capture some complex higher-order interactions by using genomic similarity metrics (Han & Pan, 2010).

A number of similarity methods have been proposed for detecting multi-marker associations. For binary traits, Allen and Satten proposed a general framework that can be reduced to a p test or a *cross* test, corresponding to two different choices of a parameter vector in their test statistic (Allen & Satten, 2009). Their p test compares within-case similarities with within-control similarities. It is based on the same concept as that in some previous works (Schaid *et al.*, 2005, Tzeng *et al.*, 2003). Their *cross* test contrasts within-group similarities (including within-case similarities and within-control similarities) with between-group similarities and is based on the same concept as that in other works (Lin & Lee, 2010, Nolte *et al.*, 2007, Sha *et al.*, 2007). Allen and Satten (2009) have established a general framework for similarity-based approaches. However, their method can only deal with binary traits. To analyze continuous traits such as body-mass index (BMI) and triglyceride, we here extend Allen and Satten's (2009) framework. Based on our framework for continuous traits, we propose two tests corresponding to Allen and Satten's (2009) p test and *cross* test (we call them SIMp and SIMc, respectively).

We then compare the performance of the two proposed tests with that of the single-marker analysis, a standard haplotype regression (Schaid *et al.*, 2002), and a popular and powerful kernel machine regression (referred to as KMR) (Kwee *et al.*, 2008, Wu *et al.*, 2010), through systematic simulations. In addition, we apply our tests to a data set on human adiposity study. Our work provides a useful general framework of similarity-based approaches for continuous traits.

2. Methods

Suppose that the chromosomal region under study contains L single nucleotide polymorphisms (SNPs). A measure of genomic similarity can be constructed with genotypes or haplotypes formed by these L SNPs.

2.1 Genotype-based similarity measure

Let S_{ij} be the similarity between the i^{th} and the j^{th} subjects for this multi-marker set.

Summing the similarities over the L loci, we have $S_{ij} = \sum_{l=1}^L s(g_i^l, g_j^l)$, where $s(g_i^l, g_j^l)$ is the similarity of the i^{th} and j^{th} subjects at the l^{th} locus, and g_i^l and g_j^l are the genotypes of the i^{th} and j^{th} subjects at the l^{th} locus, respectively. A commonly used measure of genotype similarity is the allele-match measure. This measure counts the number of alleles that match between a pair of subjects, i.e., $s(g_i^l = a_{i1}^l a_{i2}^l, g_j^l = a_{j1}^l a_{j2}^l) = \sum_{m=1}^2 \sum_{n=1}^2 I[a_{im}^l = a_{jn}^l]$, where a_{i1}^l and a_{i2}^l are the two alleles at the l^{th} locus for the i^{th} subject, and $I[\cdot]$ is the indicator function with a value of 1 or 0 according to whether the argument is true or false. Note that the similarity metric can be standardized to a range of from 0 to 1. We ignore this standardization part because it makes no essential change to the test results. This allele-match measure has been widely used in genetic linkage (Weeks & Lange, 1988) and association studies (Schaid *et al.*, 2005, Tzeng *et al.*, 2003, Tzeng *et al.*, 2009), and it is equivalent to the counting measure of haplotypes (see below).

2.2 Haplotype-based similarity measure

Genomic similarities can also be quantified with haplotype data. L SNPs can form at most $k \equiv 2^L$ “haplotype categories” (i.e., unique haplotypes in the sample, two haplotypes are classified into a same *category* if all observed alleles on the two haplotypes are the same), denoted as $\mathbf{H} = \{h^1, h^2, \dots, h^k\}$. Suppose that the haplotype phases can be directly observed, we denote $h_{i1}h_{i2}$ and $h_{j1}h_{j2}$ as the diplotypes (a diplotype is a pair of haplotypes a subject possesses) of the i^{th} and j^{th} subjects, respectively. Let S_{ij} be the similarity between the i^{th} and

the j^{th} subjects, and $S_{ij} = s(g_i = h_{i1}h_{i2}, g_j = h_{j1}h_{j2}) = \sum_{m=1}^2 \sum_{n=1}^2 s(h_{im}, h_{jn})$, where g_i and g_j are the multi-marker genotypes of the i^{th} and j^{th} subjects at the L SNPs, respectively. However, in most situations, haplotype phases cannot be directly observed and need to be inferred from genotypes. Let $h_{iu1}h_{iu2}$ be the u^{th} possible diplotype of the i^{th} subject. $P(h_{iu1}h_{iu2}|g_i)$ is the posterior probability that the i^{th} subject has the u^{th} possible diplotype, given the unphased multi-marker genotypes as g_i . Under the assumption of the Hardy-Weinberg equilibrium (HWE) (Excoffier & Slatkin, 1995, Hawley & Kidd, 1995, Long *et al.*, 1995, Schaid *et al.*, 2002), the posterior probability $P(h_{iu1}h_{iu2}|g_i)$ can be inferred by statistical methods such as the expectation-maximization (EM) algorithm (Dempster *et al.*, 1977). The expected haplotype similarity over the posterior distribution of haplotype pairs, given the observed unphased genotypes, is

$$S_{ij} = \sum_u \sum_v P(h_{iu1}h_{iu2}|g_i) P(h_{jv1}h_{jv2}|g_j) s(g_i = h_{iu1}h_{iu2}, g_j = h_{jv1}h_{jv2}) \\ = \sum_u \sum_v P(h_{iu1}h_{iu2}|g_i) P(h_{jv1}h_{jv2}|g_j) \sum_{m=1}^2 \sum_{n=1}^2 s(h_{ium}, h_{jvn}).$$

If the counting measure of haplotypes (Tzeng *et al.*, 2003) is used, i.e.,

$s(h_{im}, h_{jn}) = \sum_{l=1}^L s(a_{im}^l, a_{jn}^l) = \sum_{l=1}^L I[a_{im}^l = a_{jn}^l]$, where a_{im}^l is the allele on haplotype h_{im} at the l^{th} locus and $I[\cdot]$ is the indicator function, then

$$S_{ij} = \sum_u \sum_v P(h_{iu1}h_{iu2}|g_i) P(h_{jv1}h_{jv2}|g_j) \sum_{m=1}^2 \sum_{n=1}^2 s(h_{ium}, h_{jvn}) \\ = \sum_{l=1}^L \sum_u \sum_v P(h_{iu1}h_{iu2}|g_i) P(h_{jv1}h_{jv2}|g_j) \sum_{m=1}^2 \sum_{n=1}^2 I[a_{ium}^l = a_{jvn}^l] \\ = \sum_{l=1}^L \sum_{m=1}^2 \sum_{n=1}^2 I[a_{im}^l = a_{jn}^l].$$

This equation holds because all possible diplotypes of a subject should be consistent with his/her multi-marker genotypes. Therefore, the similarity metric based on the counting measure of haplotypes is equivalent to that based on the allele-match measure of genotypes (Section 2.1).

2.3 Test statistics

With certain haplotype phases—If the haplotype phases can be observed directly, the haplotype-frequency vector of the i^{th} subject is

$$\Gamma(h_i) = \frac{1}{2} \begin{pmatrix} I[h_{i1}=h^1] + I[h_{i2}=h^1] \\ I[h_{i1}=h^2] + I[h_{i2}=h^2] \\ \vdots \\ I[h_{i1}=h^k] + I[h_{i2}=h^k] \end{pmatrix},$$

where $I[\cdot]$ is the indicator function. In Sections 2.1 and 2.2, we have defined the similarities of pairs of subjects. The average of similarities between any two subjects is

$$\frac{\sum_{i < j} S_{ij}}{\binom{n}{2}} = \bar{\mathbf{p}}^T \mathbf{S} \bar{\mathbf{p}} + O\left(\frac{1}{n}\right) \quad (\text{Schaid } et al., 2005, \text{ Tzeng } et al., 2003),$$

where n is the number of subjects, $\bar{\mathbf{p}}$ is the vector of the average haplotype frequency of the n subjects, and $\mathbf{S}$ is a $k \times k$ matrix whose (v, v) element is the similarity between the v^{th} and v^{th} categories of haplotypes. Using similarities of pairs of ‘haplotype categories’ (i.e., unique haplotypes in the sample) instead of similarities of pairs of subjects in the following test statistic, we can reduce the dimension of the similarity matrix from $n \times n$ to $k \times k$ and decrease the computational burden (because the number of haplotype categories observed in a sample is usually smaller than the number of subjects under study). If we let a_{hm}^l be the allele on h^m at the l^{th} locus and $I[\cdot]$ be the indicator function, then the counting-measure similarity between the v^{th} and v^{th} categories of haplotypes is $s(h^v, h^v) = \sum_{l=1}^L I[a_{hv}^l = a_{hv}^l]$.

Let $\mathbf{Y}$ be the vector of continuous traits of n subjects, let $\mathbf{C}$ be an $n \times (m+1)$ matrix with the i^{th} row of $\mathbf{C}_i^T = [1 \quad c_{i,1} \quad c_{i,2} \quad \dots \quad c_{i,m}]$ coding 1 (for the intercept term) and m covariates (age, gender, ethnicity, etc.) of the i^{th} subject, and let $\mathbf{x}$ be an n -length vector with the i^{th} element $x_i = \gamma^T \cdot \mathbf{S} \cdot \Gamma(h_i)$ coding the genetic information (regarding the region under investigation) of the i^{th} subject. For $x_i = \gamma^T \cdot \mathbf{S} \cdot \Gamma(h_i)$, $\Gamma(h_i)$ is the haplotype-frequency vector of the i^{th} subject, γ is a specified vector accounting for the aggregate haplotype information of all the n subjects, and $\mathbf{S}$ is a $k \times k$ matrix whose (v, v) element is the similarity between the v^{th} and v^{th} categories of haplotypes. The scalar $x_i = \gamma^T \cdot \mathbf{S} \cdot \Gamma(h_i)$ is a quantity comparing the i^{th} subject's haplotypes against haplotypes of all the other subjects. We consider the model:

$$E(\mathbf{Y}) = \mathbf{C}\alpha + \mathbf{x}\beta, \quad (1)$$

where α is the $(m+1)$ -element vector of covariate effects including the intercept term, and β is the regression coefficient of the genetic information (regarding the region) represented by $\mathbf{x}$. Note that the regression model of Equation (1) is different from the standard haplotype regression model, in which trait values are regressed on haplotypes of subjects, i.e., $E(\mathbf{Y}) = \mathbf{C}\eta + \mathbf{Z}\psi$, where $\mathbf{Z}$ is an $n \times k$ matrix with the i^{th} row of $\Gamma(h_i)^T$ (the transpose of the haplotype-frequency vector of the i^{th} subject), η is the $(m+1)$ -element vector of covariate effects including the intercept term and ψ is the k -element vector of regression coefficients for the k categories of haplotypes in the region. We can see that the standard haplotype regression tests the association between phenotypes and haplotypes, while the regression model of Equation (1) tests the association between phenotypes and quantities of contrasting individual haplotypes with haplotypes of all the other individuals (eventually, this creates ‘similarity’).

Let $\widehat{\mu}_i = \mathbf{c}_i^T (\mathbf{C}^T \mathbf{C})^{-1} (\mathbf{C}^T \mathbf{Y})$ be the predicted mean of y_i under the null hypothesis of no association between the gene variants (in the region under investigation) and the traits, i.e., $H_0: \beta = 0$ in Equation (1). Based on the model in Equation (1) and under the assumption of gene-covariate independence, the score statistic is

$$U = \gamma^T \mathbb{S} \sum_{i=1}^n (y_i - \widehat{\mu}_i) \Gamma(h_i) \quad (2)$$

(see Appendix I for derivation). In Equation (2), $(y_i - \widehat{\mu}_i)$'s are the residuals obtained from regressing the traits on the covariates. We consider two choices for γ . One is

$\gamma = \frac{1}{n} \sum_{i=1}^n \Gamma(h_i) \equiv \widehat{\mathbf{p}}$, where $\widehat{\mathbf{p}}$ is the vector of the average haplotype frequency of all the n subjects. We call the resulting test the 'SIMp' test, which extends Allen and Satten's p test (Allen & Satten, 2009) to deal with continuous traits. Another choice for γ is

$\gamma = \sum_{i=1}^n (y_i - \widehat{\mu}_i) \Gamma(h_i)$, and the resulting test is called the 'SIMc' test, which extends Allen and Satten's *cross* test to be applicable to continuous traits. In the following, we consider the asymptotic properties of the two tests respectively.

(1) The SIMp test: The test statistic

$$T_{SIMp} = \frac{\left[\widehat{\mathbf{p}}^T \mathbb{S} \sum_{i=1}^n (y_i - \widehat{\mu}_i) \Gamma(h_i) \right]^2}{\widehat{\mathbf{p}}^T \mathbb{S} \widehat{\Omega} \mathbb{S} \widehat{\mathbf{p}}} \quad (3)$$

has an asymptotic chi-square distribution with one degree of freedom (because it is the square of a standard normal variable), where Ω is the variance-covariance matrix of

$\sum_{i=1}^n (y_i - \widehat{\mu}_i) \Gamma(h_i)$. The matrix Ω is estimated by

$$\widehat{\Omega} = \sum_{i=1}^n [(y_i - \widehat{\mu}_i) \Gamma(h_i)] [(y_i - \widehat{\mu}_i) \Gamma(h_i)]^T - \frac{1}{n} \left[\sum_{i=1}^n (y_i - \widehat{\mu}_i) \Gamma(h_i) \right] \left[\sum_{i=1}^n (y_i - \widehat{\mu}_i) \Gamma(h_i) \right]^T.$$

(2) The SIMc test: The test statistic

$$T_{SIMc} = \left[\sum_{i=1}^n (y_i - \widehat{\mu}_i) \Gamma(h_i) \right]^T \mathbb{S} \sum_{i=1}^n (y_i - \widehat{\mu}_i) \Gamma(h_i) \quad (4)$$

is asymptotically distributed as $\sum_{i=1}^{\varpi} \lambda_i \chi_{1,i}^2$, where $\chi_{1,i}^2$'s are independent chi-square variables with one degree of freedom, an $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{\varpi}$ are the ordered eigenvalues of the matrix $\widehat{\Omega} \mathbb{S}$ (by the theory of quadratic forms of normal variables (Scheffe, 1959)). The distribution of T_{SIMc} can be approximated by the three-moment approximation method (Allen & Satten, 2007, Imhof, 1961, Tzeng *et al.*, 2009, Allen & Satten, 2009). The P value of the observed SIMc test statistic is given by

$$P\left(\chi_b^2 > (T_{SIMc} - c_1) \times \sqrt{\frac{b}{c_2}} + b\right),$$

where $c_j = \sum_{i=1}^w \lambda_i^j$, $b = c_2^2 / c_3^2$, and χ_b^2 is the chi-square distribution with b degrees of freedom.

To perform the SIMp and the SIMc tests, the trait values $\mathbf{Y}$ are first regressed on the covariates ($\mathbf{C}$) alone, and the subject-specific predicted means are obtained by

$\hat{\mu}_i = \mathbf{C}_i^T (\mathbf{C}^T \mathbf{C})^{-1} (\mathbf{C}^T \mathbf{Y})$, where $i = 1, 2, \dots, n$. Then the test statistics and the P values can be computed through the formulas of T_{SIMp} and T_{SIMc} . An R code to perform these two tests is available upon request.

With uncertain haplotype phases—If the haplotype phases are uncertain, the haplotype-frequency vector of the i^{th} subject needs to be inferred by his/her multi-marker genotypes:

$$\Gamma(h_i | g_i) = \frac{1}{2} \begin{pmatrix} \sum_u \left[P(h_{iu1}=h^1 | g_i) + P(h_{iu2}=h^1 | g_i) \right] \\ \sum_u \left[P(h_{iu1}=h^2 | g_i) + P(h_{iu2}=h^2 | g_i) \right] \\ \vdots \\ \sum_u \left[P(h_{iu1}=h^k | g_i) + P(h_{iu2}=h^k | g_i) \right] \end{pmatrix},$$

under the assumption of the Hardy-Weinberg equilibrium (HWE). The HWE is a typical assumption when inferring haplotypes from multi-marker genotypes (Excoffier & Slatkin, 1995, Hawley & Kidd, 1995, Long *et al.*, 1995, Liu *et al.*, 2008, Liu *et al.*, 2006), but how to make this assumption is an issue for case-control studies (e.g., assuming the HWE in a pooled sample or only in a control sample) (Allen & Satten, 2008). However, the issues related to phase ambiguity are not crucial here because we use a phase-independent similarity measure (i.e., the counting measure of haplotypes, or equivalently, the allele-match measure of genotypes). Although the test statistic (Equation (2)) is presented with a haplotype-frequency vector, using the inferred haplotypes in the downstream test statistic will be equivalent to using the observed multi-marker genotypes in the test statistic (In Section 2.2, we have shown that the expected haplotype similarity over the posterior distribution of haplotype pairs is equivalent to the similarity based on the allele-match measure of multi-marker genotypes.) We will elaborate on the pros and cons of using phase-independent and phase-dependent similarity measures in our discussion.

3. Simulation Study

With simulations, we evaluate the performances of our methods. We first generated 200 data sets, each containing 10,000 chromosomes of 1 Mb regions with the Cosei program (Schaffner *et al.*, 2005). The chromosomes were generated in consistency with the HapMap CEU (CEPH people from Utah, U.S.A.) samples. We considered five levels of causal allele frequency: $q = 0.01, 0.05, 0.1, 0.3$, and 0.5 . For each data set, we chose SNPs with MAFs within the region of $[0.01 \pm 0.01/100]$, $[0.05 \pm 0.05/100]$, $[0.1 \pm 0.1/100]$, $[0.3 \pm 0.3/100]$, or $[0.5 \pm 0.5/100]$ as the causal SNPs, respectively. In each data set, we randomly chose 120 from the 10,000 chromosomes to mimic the Phase II HapMap CEU data, and we randomly paired them to form 60 subjects. Based on the LD patterns of the 60 subjects, we selected

tag SNPs according to the conventional cut-off $r^2 = 0.8$ and $MAF > 0.05$, with the *H-clust* method (Rinaldo *et al.*, 2005, Roeder *et al.*, 2005). These tag SNPs were served as markers used in our simulations. To generate chromosomes of one individual, we randomly selected two chromosomes from the remaining 9,880 ($= 10,000 - 120$) chromosomes. Following Wu *et al.* (2011), we created two covariates when generating trait values. Trait values were generated by $Y = 0.5C_1 + 0.5C_2 + g + e$, where C_1 was a continuous covariate generated from a standard normal distribution, C_2 was a dichotomous covariate taking values 0 and 1 with a probability of 0.5, g was the genetic effect (2, 1, or 0, depending on the genotype of the causal SNP), and e was the random error. The random error, e , was assumed to have a normal distribution with a mean of zero and a variance of V_e , where V_e was determined so that the heritability was 2% under the alternative hypothesis. After generating the trait values, the genotypes of the causal SNP were removed from our data sets. We randomly selected a ten-SNP region to encompass the causal SNP. We then evaluated the performances of the proposed SIMp and SIMc tests. We compared our similarity tests with the standard haplotype regression (referred to as HAP) (Schaid *et al.*, 2002), the single-marker analysis with P values adjusted by the Sidak correction (Sidak., 1967) (referred to as Single-marker), the single-marker analysis with P values adjusted by 5,000 permutations (referred to as Single-marker-perm), and a popular kernel machine regression (referred to as KMR) (Kwee *et al.*, 2008, Wu *et al.*, 2010).

When analyzing the simulated data, we first fitted a model under the null hypothesis of no association between the gene variants (in the region under investigation) and the traits. That is, we first regressed the trait values on the covariates C_1 and C_2 to obtain the predicted means of trait values under the null hypothesis ($\hat{\mu}_i$'s). Then the residuals ($y_i - \hat{\mu}_i$)'s were used in the test statistics of SIMp and SIMc, in the single-marker analysis, and in the package 'haplo.stats' (Schaid *et al.*, 2002).

The single-marker analysis was performed with the '--assoc' command implemented by PLINK-1.07 (Purcell *et al.*, 2007). In our simulations, each analysis region was formed by ten SNPs. We adjusted the minimum P value ($P_{\min}$) of the ten SNPs with the Sidak correction (Sidak., 1967), $P_{\min,adjusted} = 1 - (1 - P_{\min})^{M_{eff}}$, where M_{eff} was the effective number of independent tests and was estimated by the *Keffective* program (Moskvina & Schmidt, 2008) (referred to as Single-marker). This single-marker analysis is used to be compared with some popular multi-marker association tests (Tzeng *et al.*, 2009, Wu *et al.*, 2010, Tzeng *et al.*, 2011), and is commonly used in real data analyses (Harold *et al.*, 2009, Kirov *et al.*, 2009, Jones *et al.*, 2010). Therefore, we also include it into our comparisons.

We also performed the single-marker analysis with P values adjusted by permutations (referred to as Single-marker-perm). For each replication, we permuted the phenotypes 5,000 times and obtained a minimum P value of the ten SNPs from each of the 5,000 permuted data sets (so there were 5,000 minimum P values). We then estimated the significance by comparing the minimum P value (of the ten SNPs) from the observed data set with the 5,000 minimum P values from the permuted data sets. This analysis was performed by using PLINK-1.07 (Purcell *et al.*, 2007) with the command '--assoc --mperm 5000'.

The HAP was performed by the package 'haplo.stats' (Schaid *et al.*, 2002). The haplotypes with frequencies less than a threshold α_0 were pooled into a single baseline group by the default of the package 'haplo.stats'. In our simulations and the following data analysis, the threshold α_0 was set at 0.01.

In addition, we compared our tests with a popular multi-marker association test, the kernel machine regression (referred to as KMR) (Kwee *et al.*, 2008, Wu *et al.*, 2010). We

performed this test with the R code of KMR (<http://genetics.emory.edu/labs/epstein/software/lskm/index.html>). The trait value Y was treated as phenotype and the two covariates C_1 and C_2 were treated as environmental covariates when the KMR package was performed. The similarity measure used in the KMR package is the MAF-weighted identical-by-state (IBS) sharing (see Equation (5) of (Kwee *et al.*, 2008)). This measure downweights the similarities contributed by common SNPs and upweights the similarities contributed by rare SNPs. The implication is that subjects sharing rare alleles may have more similar genomes than do subjects sharing common alleles (Lin & Lee, 2010, Wessel & Schork, 2006).

4. Simulation Results

Type-I Error Rates

We created scenarios under the null hypothesis of no association by fixing the genetic effect g at zero. For each of the 200 data sets, we performed 1,000 replications. In each replication, the sample size was set at 1,000 subjects. Figure 1 presents the type-I error rates averaged over all of the 200,000 replications. The results show that the type-I error rates of five tests (SIMp, SIMc, HAP, Single-marker, and Single-marker-perm) are under control. KMR is slightly anticonservative (will be discussed later).

Statistical Power for Low-Frequency Variant Detection

For each of the 200 data sets, we performed 1,000 replications given each q (we have five levels of causal allele frequency, $q = 0.01, 0.05, 0.1, 0.3, \text{ or } 0.5$), and so the total number of replications was 200,000 for each q . The heritability (H^2) was set at 2%, and the sample size was set at 1,000. Figure 2 presents the power performances of the six tests (SIMp, SIMc, HAP, Single-marker, Single-marker-perm, and KMR) for low-frequency variant detection ($q = 0.01 \text{ or } 0.05$). We stratified the results based on the average R^2 between the causal SNP and the surrounding markers (the average R^2 was obtained by averaging the ten R^2 's of the causal SNP and each of the ten surrounding markers, where R was the correlation coefficient between two SNPs and was obtained by using the package 'genetics' (Warnes *et al.*, 2011)). We considered three ranges of the average R^2 : smaller than 0.2, between 0.2 and 0.3, and larger than 0.3. For $q = 0.01$, the average R^2 's of all the 200,000 replications were smaller than 0.2 (the left panel of Figure 2). For $q = 0.05$, the average R^2 's of 198,896 replications were smaller than 0.2 (the middle panel of Figure 2) and the remaining 1,104 replications were between 0.2 and 0.3 (the right panel of Figure 2). Overall, HAP is the most powerful test among the six tests, when the frequency of the causal variant is smaller than or equal to 0.05.

The test statistics of similarity-based methods are averages of genomic similarities weighted by some forms of covariate-adjusted phenotypes (see Equations (3) and (4)), depending on different choice of γ . When the causal variant is uncommon, the similarity-based method is an inferior choice because few subjects have this causal variant and most subjects are similar by having no causal variants. However, basically, HAP let each haplotype category (common or uncommon, as long as the frequency is larger than the pooling threshold α_0) account for one degree of freedom in the test statistic (see Equation (3) of (Schaid *et al.*, 2002)). Therefore, the association contributed by uncommon haplotypes is more likely to be preserved by HAP, compared with the similarity-based tests.

Statistical Power for Common Variant Detection

Figure 3 presents the power performances of the six tests, according to various combinations of causal allele frequency q and the average R^2 (between the causal SNP and the surrounding markers). With a causal allele frequency of 0.1, HAP is the most powerful test

when the average R^2 is smaller than 0.2; KMR and Single-marker-perm are the two most powerful tests when the average R^2 is between 0.2 and 0.3; SIMc and KMR are the most powerful tests when the average R^2 is larger than 0.3. With a causal allele frequency of 0.3 or 0.5, SIMc and KMR are the most powerful tests. Note that KMR is generally more powerful than SIMc when the causal allele frequency $q \leq 0.1$. However SIMc is more powerful than KMR when the causal allele frequency $q \geq 0.3$. This result is reasonable because the similarity measure used in the R code of KMR is the MAF-weighted IBS sharing that downweights the similarities contributed by common SNPs and upweights the similarities contributed by rare SNPs (Kwee *et al.*, 2008). This similarity measure favors the detection of uncommon variants, but suffers from power loss when the causal variant is common.

For single-marker analyses, adjusting for multiple testing with permutations (Single-marker-perm) is much more powerful than adjusting with the Sidak correction (Single-marker) (Sidak., 1967). Overall, Single-marker-perm has a good power performance, although SIMc, KMR, or HAP can outperform it in almost all combinations of causal allele frequency and average R^2 . From Figures 2 and 3, SIMp is generally not powerful. However, SIMp has a good power performance when the causal SNP was introduced at common haplotypes. Figure 4 presents the results (only for $q = 0.5$) stratified by the *Maximum Frequency of Risk Haplotypes* (MFRH) (where a risk haplotype is defined if the causal variant was introduced at that haplotype). We can see that SIMp outperforms other tests when the MFRH is larger than 0.4.

SIMp is a generalization of the p test (Tzeng *et al.*, 2003, Allen & Satten, 2009) that compares within-case similarities with within-control similarities. The p test is more powerful than the *cross* test when most cases share a common haplotype and controls have many different categories of haplotypes. Likewise, SIMp is more powerful than SIMc (a generalization of the *cross* test) when most subjects with undesirable traits (say, large BMIs that lead to complex disorders) share a common haplotype and subjects with normal traits have many different categories of haplotypes. When the MFRH is large, many subjects with undesirable traits share the risk haplotype with the highest frequency, and SIMp can become more powerful than SIMc. Note that even for phase-independent similarity metrics, as long as no severe errors in the haplotype inference, the MFRH is still a good index of the genomic similarity (in the region under study) between any two subjects with undesirable traits.

5. Application to Adiposity Data

We then applied our method to a data set on human adiposity (Chung *et al.*, 2009). The data set contains 1,982 unrelated European Americans living in the New York City metropolitan area. We investigated the association of 23 SNPs in the *carboxypeptidase E (CPE)* gene (located on chromosome 4q32.3) with BMI. To replicate the previous single-marker analysis results (Chung *et al.*, 2009), we first followed their analysis. BMI was log-transformed and then was adjusted for the covariates sex, age, age², and their respective interactions. Associations between BMI and the joint additive and dominance effects of each of the 23 SNPs were tested using the ordinary-least-squares (OLS) regression method.

To perform multi-marker analyses, we first need to specify the number of SNPs used in an analysis unit. Partitioning the gene into segments according to the LD patterns is a commonly used practice (Gabriel *et al.*, 2002, Twells *et al.*, 2003, Zhang *et al.*, 2002). Therefore, we first used the Haploview (Barrett *et al.*, 2005) and learned that the 22 SNPs with MAFs larger than 0.01 constitute three haplotype blocks, as shown in Figure 5. Then we applied the six tests (SIMp, SIMc, HAP, Single-marker, Single-marker-perm, and KMR)

to each of the three haplotype blocks. Table 1 presents the P values based on the six tests, for the three blocks, respectively. Given a significance level of 0.05, only SIMp suggests that the *CPE* gene, especially from 166,668,633 bp to 166,727,266 bp (Block 1), may potentially be interesting to investigators ($0.05 > 0.012 \times 3$ blocks/tests, if the Bonferroni correction is used to adjust for the three blocks/tests in the *CPE* gene). In Table 2, frequencies and P values of haplotype categories (with frequencies ≥ 0.01) in Block 1 are listed. We see that the second significant haplotype (Haplotype M) has a high frequency of 0.37. Although the P value of the score statistic of Haplotype M, 0.075, is not smaller than the commonly used significance level of 0.05, aggregating the information of these individual haplotypes in a SIMp test yields a significant result (P value = 0.012, from Table 1). Based on our simulation results, SIMp has a good power performance when the causal variant was introduced at a common haplotype. The situation of Block 1 may correspond to this scenario, and therefore SIMp yields the most significant result among all the six tests we evaluated. Note that studies have shown that a single point mutation in the *CPE* gene is sufficient to produce an animal with multiple disorders including obesity (Fricker *et al.*, 1996, Naggert *et al.*, 1995). Moreover, an independent study (Jeffrey *et al.*, 2008) also pointed out the association of the *CPE* gene with human obesity and diabetes. The role of the *CPE* gene in human adiposity may deserve further investigation.

6. Discussion

Many previous similarity-based methods can only deal with binary traits (Allen & Satten, 2009, Lin & Lee, 2010, Lin & Schaid, 2009, Nolte *et al.*, 2007, Schaid *et al.*, 2005, Tzeng *et al.*, 2003, Yuan *et al.*, 2006, Sha *et al.*, 2007). In this work, we developed a general framework for similarity-based methods that can handle continuous traits. To evaluate the robustness of our tests to the underlying distributions, we deliberately set the error term to be very far away from the normal distribution, and we performed more simulations. We find that our tests are very robust to the assumption of normality (this assumption is made in Appendix I).

Using multiple genetic markers in association analyses has been shown to be superior to the single-marker analysis in some situations (Schaid, 2004b, Schaid, 2004a). The method proposed here is applicable to chromosomal regions or genes, as well as genome-wide search for disease susceptibility loci. An issue is how to choose an 'appropriate' size of a multi-marker set. Multi-marker analyses with larger window sizes may allow for measuring sharing over longer genomic sequences and lead to more power gains, though this also increases the computational burdens (Allen & Satten, 2009). How to choose the size of a multi-marker set is still an open question (Schaid, 2004b). Analyses can be chosen based on sliding windows or haplotype blocks (Guo *et al.*, 2009). However, similar to single-marker analysis, a harsh multiple-testing penalty will be encountered by using the sliding-window strategy. Partitioning the chromosomal region into segments according to the LD patterns is a commonly used practice (Gabriel *et al.*, 2002, Twells *et al.*, 2003, Zhang *et al.*, 2002), as we have shown in the analysis of the adiposity data.

The computational burden of performing the SIMp and SIMc tests is reasonable, although it increases with the number of 'haplotype categories' (i.e., unique haplotypes in the sample). For data sets simulated based on the LD patterns of the HapMap CEU samples, the burden of analyzing 20 SNPs in a multi-marker test is still reasonable. It takes 1.5 hours to run 1,000 replications (each with 1,000 subjects), given an Intel Xeon workstation with 3.0 GHz CPU and 2.0 GB of memory. When using 20 SNPs to form a multi-marker set for simulations, the relative power performances of the four tests were very similar to our Figures 2 and 3. However, each method was slightly more powerful than that by using ten

SNPs in a multi-marker set. This is because larger multi-marker sets allow for measuring sharing over longer genomic sequences and lead to more power gains.

The genomic distance-based regression (GDBR) (Wessel & Schork, 2006) and the gene-trait similarity regression (GTSR) (Tzeng *et al.*, 2009, Tzeng *et al.*, 2011) can also deal with continuous traits. These two approaches, our SIMc test, the sum-of-squared-score (SSU) test (Pan, 2009), the variance-component score test (Tzeng & Zhang, 2007), and the KMR method (Kwee *et al.*, 2008, Wu *et al.*, 2010) are closely related to one another. Motivated differently, these test statistics end up with very similar forms. We show the theoretical connection between these methods in Appendix II. Note that we observed a slight inflation of type-I error rates for KMR (Figure 1). The KMR uses a Satterthwaite procedure to approximate the mixture chi-square distribution, while SIMc and other related tests use the three-moment approximation (Allen & Satten, 2007, Imhof, 1961, Tzeng *et al.*, 2009, Allen & Satten, 2009). This may be a possible reason for the slight inflation of type-I error rates of KMR.

More than just SIMc, we would like to emphasize that we have proposed a general framework of similarity-based tests for continuous traits. By employing different γ in Equation (2), our test statistic results in various tests. SIMc (and thus GDBR and GTSR) is a special case of our general framework. The SIMp test is another special case (which was shown to identify the signal of the *CPE* gene that was pointed out to be associated with obesity by several independent studies (Jeffrey *et al.*, 2008, Fricker *et al.*, 1996, Naggert *et al.*, 1995)). In addition, we have performed systematic simulations to evaluate the performance of several popular methods for detecting causal variants, common or uncommon. Through these simulations, we further clarify the merits and limitations of similarity-based methods.

Our proposed framework results in various test given different choice of γ , and no one test performs well universally. There is no uniformly most-powerful test in the situation of multi-marker association testing (Cox & Hinkley, 1974, Han & Pan, 2010). These different tests capture various aspects of information from the data. One test can outperform other tests in certain scenarios, as we have shown in Figures 2-4. Note that the three features (1) causal allele frequency, (2) the average R^2 between the causal SNP and the surrounding markers, and (3) the maximum frequency of risk haplotypes (MFRH) are unknown in real-data analysis (because we do not know where the causal variant is). Therefore, we cannot choose the most powerful test in advance. Our aim is to provide a general framework harboring various tests that are complementary to one another. A research direction is to come up with a third test based on our general framework, by specifying a different γ . More interestingly, developing a strategy that can determine an optimal γ based on data, as illustrated by the test selection of a previous study (Pan *et al.*, 2010).

The similarity measure used in this work is a phase-independent metric (i.e., the counting measure of haplotypes, or equivalently, the allele-match measure of genotypes), although we infer haplotype phases with the EM algorithm to facilitate the computation of the test statistic (Equation (2)). In Section 2.2, we have shown that the expected haplotype similarity over the posterior distribution of haplotype pairs is equivalent to the similarity based on the allele-match measure of multi-marker genotypes. There are other similarity measures, including phase-independent and phase-dependent metrics (Durrant *et al.*, 2004, Lin & Schaid, 2009, Tzeng *et al.*, 2003, Van der Meulen & te Meerman, 1997, Wessel & Schork, 2006). These similarity measures can also be plugged into our framework (Equation (2)). The merit of a phase-independent metric is its robustness to genotyping errors and the HWE assumption, although it might be less powerful than a phase-dependent metric that can capture the identical-by-descent sharing more precisely (Tzeng *et al.*, 2009). Although a

phase-dependent metric (e.g., a matching measure of haplotypes (Tzeng *et al.*, 2003)) may be more powerful, the mis-specification of haplotypes given multi-marker genotypes can cause adverse results such as inflated type-I error rates (Allen & Satten, 2008).

Finally, low-frequency ($0.01 \leq \text{MAF} < 0.05$) and rare ($\text{MAF} < 0.01$) variants are thought to be responsible for a non-negligible portion of heritability that cannot be explained by common variants (Manolio *et al.*, 2009, Zeggini, 2011). Recently, Wu *et al.* (2011) proposed a sequence kernel association test (SKAT) to analyze rare variants. SKAT is closely related to KMR (Kwee *et al.*, 2008, Wu *et al.*, 2010) and our SIMc test. However, there are two fundamental differences between Wu *et al.*'s (2011) work and our work. First, SKAT was developed for sequencing data, and most variants were uncommon in their simulations (Wu *et al.*, 2011). Our work is mainly for association analyses using tag SNPs as markers, and the MAF of each marker was larger than 5% (see our Simulation Study). Second, Wu *et al.* (2011) simulated multiple uncommon causal variants for each simulated region, while we simulated only one uncommon or common causal variant in this work. In addition, Tzeng *et al.* (2011) proposed a GTSR for uncommon and common variants. The GTSR is also closely related to the SIMc test here. Tzeng *et al.* (2011) found a weighting scheme on GTSR that can improve the power for detecting rare variants and have comparable power for detecting common variants (referred to as 'SIM1' by Tzeng *et al.* (2011)). However, the standard haplotype regression (HAP) (Schaid *et al.*, 2002) is still more powerful than that 'SIM1' test, when the causal variant is rare and when the average R^2 between the causal variant and the surrounding markers is low (this is also a feature of rare causal variants) (Tzeng *et al.*, 2011). Unifying the standard haplotype regression (Schaid *et al.*, 2002) with the similarity-based tests may further increase the power to detect rare causal variants. This topic may deserve further research.

Acknowledgments

We thank the anonymous reviewers and Dr. David B. Allison for their constructive comments; Drs. Wendy K. Chung and Rudolph L. Leibel for providing the adiposity data (sponsored by NIH grant DK52431 and the New York Health Project); and Dr. Valentina Moskvina for providing the *Keffective* program. This work was supported in part by NIH grants GM081488 (NL), GM073766 (GG), GM074913 (KZ), GM087748 (EA), and HL076206 (EA) from the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Appendix

Appendix I: Derivation of the score statistic, given phase-certain data

Consider the model in Equation (1) : $E(Y) = C\alpha + x\beta$, where the i^{th} element of the n -length vector x is $x_i = \gamma^T \cdot \mathbb{S} \cdot \Gamma(h_i)$, β is the regression coefficient of the genetic information (in the region) represented by x , C is an $n \times (m+1)$ matrix with the i^{th} row $c_i^T = [1 \ c_{i,1} \ c_{i,2} \ \dots \ c_{i,m}]$ coding 1 (for the intercept term) and the m covariates of the i^{th} subject, and α is the $(m+1)$ - element vector of covariate effects including the intercept term. Assume that the error term $\varepsilon_i = y_i - c_i^T \alpha - \beta x_i$ is normally distributed with a mean of zero and a variance of σ^2 , where $i = 1, \dots, n$. The likelihood is

$$L = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{[y_i - c_i^T \alpha - \beta x_i]^2}{2\sigma^2} \right\}.$$

The log-likelihood is

$$l \propto \sum_{i=1}^n [y_i - \mathbf{c}_i^T \alpha - \beta x_i]^2.$$

The score function for β is

$$\frac{\partial l}{\partial \beta} = -2 \sum_{i=1}^n x_i [y_i - \mathbf{c}_i^T \alpha - \beta x_i].$$

The score function for α is

$$\frac{\partial l}{\partial \alpha} = -2 \sum_{i=1}^n \mathbf{c}_i [y_i - \mathbf{c}_i^T \alpha - \beta x_i].$$

When $\beta = 0$, solving α from the equation

$$0 \equiv \frac{\partial l}{\partial \alpha} = -2 \sum_{i=1}^n \mathbf{c}_i [y_i - \mathbf{c}_i^T \alpha - \beta x_i] = -2 \sum_{i=1}^n \mathbf{c}_i [y_i - \mathbf{c}_i^T \alpha],$$

and we obtain $\hat{\alpha} = \left(\sum_{i=1}^n \mathbf{c}_i \mathbf{c}_i^T \right)^{-1} \left(\sum_{i=1}^n \mathbf{c}_i y_i \right) = (\mathbf{C}^T \mathbf{C})^{-1} (\mathbf{C}^T \mathbf{Y})$. Then we have

$$\frac{\partial l}{\partial \beta} = -2 \sum_{i=1}^n x_i [y_i - \mathbf{c}_i^T \alpha - \beta x_i] = -2 \sum_{i=1}^n x_i [y_i - \mathbf{c}_i^T (\mathbf{C}^T \mathbf{C})^{-1} (\mathbf{C}^T \mathbf{Y}) - \beta x_i],$$

and

$$\left. \frac{\partial l}{\partial \beta} \right|_{\beta=0} = -2 \sum_{i=1}^n x_i [y_i - \mathbf{c}_i^T (\mathbf{C}^T \mathbf{C})^{-1} (\mathbf{C}^T \mathbf{Y})].$$

We thus have the score statistic,

$$U = \sum_{i=1}^n x_i [y_i - \mathbf{c}_i^T (\mathbf{C}^T \mathbf{C})^{-1} (\mathbf{C}^T \mathbf{Y})].$$

Because $x_i = \gamma^T \cdot \mathbf{S} \cdot \Gamma(h_i)$, the score statistic is

$$U = \sum_{i=1}^n \gamma^T \mathbb{S} \Gamma(h_i) \left[y_i - \mathbf{c}_i^T (\mathbf{C}^T \mathbf{C})^{-1} (\mathbf{C}^T \mathbf{Y}) \right] = \gamma^T \mathbb{S} \sum_{i=1}^n \left[y_i - \mathbf{c}_i^T (\mathbf{C}^T \mathbf{C})^{-1} (\mathbf{C}^T \mathbf{Y}) \right] \Gamma(h_i).$$

Let $\widehat{\mu}_i = \mathbf{c}_i^T (\mathbf{C}^T \mathbf{C})^{-1} (\mathbf{C}^T \mathbf{Y})$ be the predicted mean of y_i under the null hypothesis of no association between the gene variants (in the region under study) and the traits, i.e., $H_0: \beta = 0$.

The score statistic is

$$U = \gamma^T \mathbb{S} \sum_{i=1}^n [y_i - \widehat{\mu}_i] \Gamma(h_i).$$

Appendix II: Relationship between GDBR, GTSR, SSU, KMR, and our SIMc test

(I) GDBR

The relationship between the pseudo- F statistic in the GDBR and the SSU test in logistic regression has been elucidated by Han and Pan (2010). This relationship also holds for continuous traits. Let $\mathbf{G}$ be the Gower matrix (Gower, 1966) harboring the information of genomic similarities among subjects. If $\mathbf{G}$ is an outer product matrix, we can find an $n \times p$ matrix $\mathbf{Z}$ such that $\mathbf{G} = \mathbf{Z}\mathbf{Z}^T$. If $\mathbf{G}$ is not an outer product matrix, we can still approximate $\mathbf{G} \approx \mathbf{Z}\mathbf{Z}^T$ by using the principal coordinates analysis (Gower, 1966). Suppose we have an $n \times p$ matrix $\mathbf{Z}$ such that $\mathbf{G} = \mathbf{Z}\mathbf{Z}^T$. Then the GDBR can be reformulated as a multivariate linear model $\mathbf{Z} = \mathbf{y}_c \mathbf{B} + \boldsymbol{\varepsilon}$, where $\mathbf{y}_c$ is the $n \times 1$ vector of covariate-adjusted trait values, $\mathbf{B}$ is a $1 \times p$ vector of unknown regression coefficients, and $\boldsymbol{\varepsilon}$ is an $n \times p$ matrix of random errors (Han & Pan, 2010) (in contrast to the standard regression approaches, the GDBR treats the genetic part as the response variable and treats the trait value as the explanatory variable).

Therefore, we have the projection matrix $\mathbf{H} = \mathbf{y}_c (\mathbf{y}_c^T \mathbf{y}_c)^{-1} \mathbf{y}_c^T$ and the fitted matrix $\hat{\mathbf{Z}} = \mathbf{H}\mathbf{Z}$. The association between $\mathbf{y}_c$ and $\mathbf{Z}$ can be assessed with a simple linear regression: $\mathbf{y}_c = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\xi}$, where $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression coefficients and $\boldsymbol{\xi}$ is an $n \times 1$ vector of random errors. With permutations, the kernel of the pseudo- F statistic in the GDBR is

$$\text{tr}(\widehat{\mathbf{Z}}^T \widehat{\mathbf{Z}}) = \text{tr}(\mathbf{Z}^T \mathbf{H}^T \mathbf{H} \mathbf{Z}) = \text{tr}(\mathbf{Z}^T \mathbf{y}_c (\mathbf{y}_c^T \mathbf{y}_c)^{-1} \mathbf{y}_c^T \mathbf{Z}) = (\mathbf{y}_c^T \mathbf{y}_c)^{-1} \mathbf{y}_c^T \mathbf{Z} \mathbf{Z}^T \mathbf{y}_c.$$

The scalar $(\mathbf{y}_c^T \mathbf{y}_c)^{-1}$ does not vary with permutations. Therefore, the pseudo- F statistic in the GDBR is proportional to $\mathbf{y}_c^T \mathbf{Z} \mathbf{Z}^T \mathbf{y}_c$, which is equivalent to the SSU test (Pan, 2009).

(II) GTSR

The kernel of the test statistic of GTSR is $\mathbf{y}_c^T \mathbf{S}_0 \mathbf{y}_c$ (Equation (4) of Tzeng *et al.* (2009), where $\mathbf{S}_0$ is an $n \times n$ matrix with diagonal elements equal to 0 (not 1) and off-diagonal

elements equal to S_{ij} (the similarity between the i^{th} and the j^{th} subjects). If we specify $\mathbf{G} = \mathbf{Z}\mathbf{Z}^T = \mathbf{S}_\theta$ in the GDBR, the GDBR will be equivalent to the GTSR.

(III) Our SIMc test

The test statistic is

$$T_{\text{SIMc}} = \left[\sum_{i=1}^n (y_i - \hat{\mu}_i) \mathbf{\Gamma}(h_i) \right]^T \mathbb{S} \sum_{i=1}^n (y_i - \hat{\mu}_i) \mathbf{\Gamma}(h_i),$$

where $\mathbb{S}$ is the $k \times k$ matrix whose (v, v) element is the similarity between the v^{th} and v^{th} categories of haplotypes; y_i and $\mathbf{\Gamma}(h_i)$ are the trait value and haplotype-frequency vector of the i^{th} subject, respectively; $\hat{\mu}_i$ is the predicted mean of y_i under the null hypothesis of no association between the gene variants (in the region under investigation) and the traits. Let $\mathbf{S}$ be an $n \times n$ matrix with diagonal elements equal to 1 and off-diagonal elements equal to S_{ij} (the similarity between the i^{th} and the j^{th} subjects). The test statistic of SIMc can be re-expressed as

$$\begin{aligned} T_{\text{SIMc}} &= \left[\sum_{i=1}^n (y_i - \hat{\mu}_i) \mathbf{\Gamma}(h_i) \right]^T \mathbb{S} \sum_{i=1}^n (y_i - \hat{\mu}_i) \mathbf{\Gamma}(h_i) \\ &= \sum_{i=1}^n \sum_{j=1}^n [(y_i - \hat{\mu}_i)(y_j - \hat{\mu}_j) \mathbf{\Gamma}(h_i)^T \mathbb{S} \mathbf{\Gamma}(h_j)] = \sum_{i=1}^n \sum_{j=1}^n [(y_i - \hat{\mu}_i)(y_j - \hat{\mu}_j) S_{ij}] = \mathbf{y}_c^T \mathbf{S} \mathbf{y}_c. \end{aligned}$$

If we specify $\mathbf{G} = \mathbf{Z}\mathbf{Z}^T = \mathbf{S}$ in the GDBR, the GDBR will be equivalent to our SIMc test. There is a subtle difference between the GTSR (with $\mathbf{S}_\theta$ in the test statistic) and the SIMc test (with $\mathbf{S}$ in the test statistic). To be more specific, the SIMc test is closer to the variance-component score test (VC-score) (Tzeng & Zhang, 2007, Tzeng *et al.*, 2009), which also includes $\mathbf{S}$ (instead of $\mathbf{S}_\theta$) in the test statistic. However, the amount of information contributed by the comparison of two haplotypes within a subject is relatively small compared with the between-subjects comparisons (Tzeng *et al.*, 2009). Therefore, the performances of the SIMc test, VC-score, and GTSR are very similar.

In addition, from $T_{\text{SIMc}} = \left[\sum_{i=1}^n (y_i - \hat{\mu}_i) \mathbf{\Gamma}(h_i) \right]^T \mathbb{S} \sum_{i=1}^n (y_i - \hat{\mu}_i) \mathbf{\Gamma}(h_i) = \mathbf{y}_c^T \mathbf{S} \mathbf{y}_c$, we can see that our SIMc test has a similar form as KMR (see Equation (9) of (Kwee *et al.*, 2008)). A major difference between these two tests is that the similarity measure (or ‘kernel’ called by the KMR paper) used in the R code of the KMR method is the MAF-weighted IBS sharing (see Equation (5) of (Kwee *et al.*, 2008)). This measure downweights the similarities contributed by common SNPs and upweights the similarities contributed by rare SNPs. Consequently, KMR is more powerful than SIMc when the causal variant is uncommon (MAF ≤ 0.1 , based on our simulation results), while SIMc is more powerful than KMR when the causal variant is common, as we can see from our simulation results.

In summary, the GDBR, GTSR, VC-score, SSU, KMR, and our SIMc test are all closely related. They detect the same aspect of information with a difference caused by various choices of similarity measures.

References

- Allen AS, Satten GA. Statistical models for haplotype sharing in case-parent trio data. *Hum Hered.* 2007; 64:35–44. [PubMed: 17483595]
- Allen AS, Satten GA. Robust estimation and testing of haplotype effects in case-control studies. *Genet Epidemiol.* 2008; 32:29–40. [PubMed: 17948229]
- Allen AS, Satten GA. A novel haplotype-sharing approach for genome-wide case-control association studies implicates the calpastatin gene in Parkinson's disease. *Genet Epidemiol.* 2009; 33:657–67. [PubMed: 19365859]
- Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics.* 2005; 21:263–5. [PubMed: 15297300]
- Chung WK, Patki A, Matsuoka N, Boyer BB, Liu N, Musani SK, Goropashnaya AV, Tan PL, Katsanis N, Johnson SB, Gregersen PK, Allison DB, Leibel RL, Tiwari HK. Analysis of 30 genes (355 SNPs) related to energy homeostasis for association with adiposity in European-American and Yup'ik Eskimo populations. *Hum Hered.* 2009; 67:193–205. [PubMed: 19077438]
- Cox, DR.; Hinkley, DV. *Theoretical Statistics.* London: Chapman and Hall; 1974.
- De Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. Efficiency and power in genetic association studies. *Nat Genet.* 2005; 37:1217–23. [PubMed: 16244653]
- Dempster AP, Laird NM, Rubin DB. Maximum likelihood estimation from incomplete data via the EM algorithm. *J R Stat Soc.* 1977; 39:1–38.
- Durrant C, Zondervan KT, Cardon LR, Hunt S, Deloukas P, Morris AP. Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. *Am J Hum Genet.* 2004; 75:35–43. [PubMed: 15148658]
- Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol.* 1995; 12:921–7. [PubMed: 7476138]
- Fricker LD, Berman YL, Leiter EH, Devi LA. Carboxypeptidase E activity is deficient in mice with the fat mutation. Effect on peptide processing. *J Biol Chem.* 1996; 271:30619–24. [PubMed: 8940036]
- Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, Defelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. The structure of haplotype blocks in the human genome. *Science.* 2002; 296:2225–9. [PubMed: 12029063]
- Gower JC. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika.* 1966; 53:325–338.
- Guo Y, Li J, Bonham AJ, Wang Y, Deng H. Gains in power for exhaustive analyses of haplotypes using variable-sized sliding window strategy: a comparison of association-mapping strategies. *Eur J Hum Genet.* 2009; 17:785–92. [PubMed: 19092774]
- Han F, Pan W. Powerful multi-marker association tests: unifying genomic distance-based regression and logistic regression. *Genet Epidemiol.* 2010; 34:680–8. [PubMed: 20976795]
- Harold D, Abraham R, Hollingworth P, Sims R, Gerrish A, Hamshere ML, Pahwa JS, Moskvina V, Dowzell K, Williams A, Jones N, Thomas C, Stretton A, Morgan AR, Lovestone S, Powell J, Proitsi P, Lupton MK, Brayne C, Rubinsztein DC, Gill M, Lawlor B, Lynch A, Morgan K, Brown KS, Passmore PA, Craig D, McGuinness B, Todd S, Holmes C, Mann D, Smith AD, Love S, Kehoe PG, Hardy J, Mead S, Fox N, Rossor M, Collinge J, Maier W, Jessen F, Schurmann B, Van Den Bussche H, Heuser I, Kornhuber J, Wiltfang J, Dichgans M, Frolich L, Hampel H, Hull M, Rujescu D, Goate AM, Kauwe JS, Cruchaga C, Nowotny P, Morris JC, Mayo K, Sleegers K, Bettens K, Engelborghs S, De Deyn PP, Van Broeckhoven C, Livingston G, Bass NJ, Gurling H, McQuillin A, Gwilliam R, Deloukas P, Al-Chalabi A, Shaw CE, Tsolaki M, Singleton AB, Guerreiro R, Muhleisen TW, Nothen MM, Moebus S, Jockel KH, Klopp N, Wichmann HE, Carrasquillo MM, Pankratz VS, Younkin SG, Holmans PA, O'donovan M, Owen MJ, Williams J. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nat Genet.* 2009; 41:1088–93. [PubMed: 19734902]
- Hawley ME, Kidd KK. HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *J Hered.* 1995; 86:409–11. [PubMed: 7560877]

- Imhof JP. Computing the distribution of quadratic forms in normal variables. *Biometrika*. 1961; 48:419–426.
- Jeffrey KD, Alejandro EU, Luciani DS, Kalynyak TB, Hu X, Li H, Lin Y, Townsend RR, Polonsky KS, Johnson JD. Carboxypeptidase E mediates palmitate-induced beta-cell ER stress and apoptosis. *Proc Natl Acad Sci U S A*. 2008; 105:8452–7. [PubMed: 18550819]
- Jones L, Holmans PA, Hamshe ML, Harold D, Moskvina V, Ivanov D, Pocklington A, Abraham R, Hollingworth P, Sims R, Gerrish A, Pahwa JS, Jones N, Stretton A, Morgan AR, Lovestone S, Powell J, Proitsi P, Lupton MK, Brayne C, Rubinsztein DC, Gill M, Lawlor B, Lynch A, Morgan K, Brown KS, Passmore PA, Craig D, McGuinness B, Todd S, Holmes C, Mann D, Smith AD, Love S, Kehoe PG, Mead S, Fox N, Rossor M, Collinge J, Maier W, Jessen F, Schurmann B, Heun R, Kolsch H, Van Den Bussche H, Heuser I, Peters O, Kornhuber J, Wiltfang J, Dichgans M, Frolich L, Hampel H, Hull M, Rujescu D, Goate AM, Kauwe JS, Cruchaga C, Nowotny P, Morris JC, Mayo K, Livingston G, Bass NJ, Gurling H, Mcquillin A, Gwilliam R, Deloukas P, Al-Chalabi A, Shaw CE, Singleton AB, Guerreiro R, Muhleisen TW, Nothen MM, Moebus S, Jockel KH, Klopp N, Wichmann HE, Ruther E, Carrasquillo MM, Pankratz VS, Younkin SG, Hardy J, O'donovan MC, Owen MJ, Williams J. Genetic evidence implicates the immune system and cholesterol metabolism in the aetiology of Alzheimer's disease. *PLoS One*. 2010; 5:e13950. [PubMed: 21085570]
- Kirov G, Zaharieva I, Georgieva L, Moskvina V, Nikolov I, Cichon S, Hillmer A, Toncheva D, Owen MJ, O'donovan MC. A genome-wide association study in 574 schizophrenia trios using DNA pooling. *Mol Psychiatry*. 2009; 14:796–803. [PubMed: 18332876]
- Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP. A powerful and flexible multilocus association test for quantitative traits. *Am J Hum Genet*. 2008; 82:386–97. [PubMed: 18252219]
- Lin WY, Lee WC. Discovering joint associations between disease and gene pairs with a novel similarity test. *BMC Genet*. 2010; 11:86. [PubMed: 20920333]
- Lin WY, Schaid DJ. Power comparisons between similarity-based multilocus association methods, logistic regression, and score tests for haplotypes. *Genet Epidemiol*. 2009; 33:183–97. [PubMed: 18814307]
- Liu N, Beerman I, Lifton R, Zhao H. Haplotype analysis in the presence of informatively missing genotype data. *Genet Epidemiol*. 2006; 30:290–300. [PubMed: 16528706]
- Liu N, Zhang K, Zhao H. Haplotype-association analysis. *Adv Genet*. 2008; 60:335–405. [PubMed: 18358327]
- Long JC, Williams RC, Urbanek M. An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am J Hum Genet*. 1995; 56:799–810. [PubMed: 7887436]
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. *Nature*. 2009; 461:747–53. [PubMed: 19812666]
- Moskvina V, Schmidt KM. On multiple-testing correction in genome-wide association studies. *Genet Epidemiol*. 2008; 32:567–73. [PubMed: 18425821]
- Naggert JK, Fricker LD, Varlamov O, Nishina PM, Rouille Y, Steiner DF, Carroll RJ, Paigen BJ, Leiter EH. Hyperproinsulinaemia in obese fat/fat mice associated with a carboxypeptidase E mutation which reduces enzyme activity. *Nat Genet*. 1995; 10:135–42. [PubMed: 7663508]
- Nolte IM, De Vries AR, Spijker GT, Jansen RC, Brinza D, Zelikovsky A, Te Meerman GJ. Association testing by haplotype-sharing methods applicable to whole-genome analysis. *BMC Proc*. 2007; 1(1):S129. [PubMed: 18466471]
- Pan W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genet Epidemiol*. 2009; 33:497–507. [PubMed: 19170135]
- Pan W, Han F, Shen X. Test selection with application to detecting disease association with multiple SNPs. *Hum Hered*. 2010; 69:120–30. [PubMed: 19996609]
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007; 81:559–75. [PubMed: 17701901]

- Rinaldo A, Bacanu SA, Devlin B, Sonpar V, Wasserman L, Roeder K. Characterization of multilocus linkage disequilibrium. *Genet Epidemiol.* 2005; 28:193–206. [PubMed: 15637716]
- Roeder K, Bacanu SA, Sonpar V, Zhang X, Devlin B. Analysis of single-locus tests to detect gene/disease associations. *Genet Epidemiol.* 2005; 28:207–19. [PubMed: 15637715]
- Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* 2005; 15:1576–83. [PubMed: 16251467]
- Schaid DJ. The complex genetic epidemiology of prostate cancer. *Hum Mol Genet.* 2004a; 13:R103–21. Spec No 1. [PubMed: 14749351]
- Schaid DJ. Evaluating associations of haplotypes with traits. *Genet Epidemiol.* 2004b; 27:348–64. [PubMed: 15543638]
- Schaid DJ, McDonnell SK, Hebring SJ, Cunningham JM, Thibodeau SN. Nonparametric tests of association of multiple genes with human disease. *Am J Hum Genet.* 2005; 76:780–93. [PubMed: 15786018]
- 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; 70:425–34. [PubMed: 11791212]
- Scheffe, H. *The Analysis of Variance.* New York: Wiley; 1959.
- Sha Q, Chen HS, Zhang S. A new association test using haplotype similarity. *Genet Epidemiol.* 2007; 31:577–93. [PubMed: 17443704]
- Sidak Z. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association.* 1967; 62:626–633.
- Twells RC, Mein CA, Phillips MS, Hess JF, Veijola R, Gilbey M, Bright M, Metzker M, Lie BA, Kingsnorth A, Gregory E, Nakagawa Y, Snook H, Wang WY, Masters J, Johnson G, Eaves I, Howson JM, Clayton D, Cordell HJ, Nutland S, Rance H, Carr P, Todd JA. Haplotype structure, LD blocks, and uneven recombination within the LRP5 gene. *Genome Res.* 2003; 13:845–55. [PubMed: 12727905]
- Tzeng JY, Devlin B, Wasserman L, Roeder K. On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *Am J Hum Genet.* 2003; 72:891–902. [PubMed: 12610778]
- Tzeng JY, Zhang D. Haplotype-based association analysis via variance-components score test. *Am J Hum Genet.* 2007; 81:927–38. [PubMed: 17924336]
- Tzeng JY, Zhang D, Chang SM, Thomas DC, Davidian M. Gene-trait similarity regression for multimarker-based association analysis. *Biometrics.* 2009; 65:822–32. [PubMed: 19210740]
- Tzeng JY, Zhang D, Pongpanich M, Smith C, McCarthy MI, Sale MM, Worrall BB, Hsu FC, Thomas DC, Sullivan PF. Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. *Am J Hum Genet.* 2011; 89:277–88. [PubMed: 21835306]
- Van Der Meulen MA, Te Meerman GJ. Haplotype sharing analysis in affected individuals from nuclear families with at least one affected offspring. *Genet Epidemiol.* 1997; 14:915–20. [PubMed: 9433600]
- Warnes, G.; Gorjanc, G.; Leisch, F.; Man, M. *genetics: Population Genetics.* R package version 1.3.6. 2011. Retrieved, from the World Wide Web: <http://cran.r-project.org/web/packages/genetics/>
- Weeks DE, Lange K. The affected-pedigree-member method of linkage analysis. *Am J Hum Genet.* 1988; 42:315–26. [PubMed: 3422543]
- Wessel J, Schork NJ. Generalized genomic distance-based regression methodology for multilocus association analysis. *Am J Hum Genet.* 2006; 79:792–806. [PubMed: 17033957]
- Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet.* 2010; 86:929–42. [PubMed: 20560208]
- Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet.* 2011; 89:82–93. [PubMed: 21737059]
- Yuan A, Yue Q, Apprey V, Bonney G. Detecting disease gene in DNA haplotype sequences by nonparametric dissimilarity test. *Hum Genet.* 2006; 120:253–61. [PubMed: 16807758]

- Zeggini E. Next-generation association studies for complex traits. *Nat Genet.* 2011; 43:287–8. [PubMed: 21445070]
- Zhang K, Calabrese P, Nordborg M, Sun F. Haplotype block structure and its applications to association studies: power and study designs. *Am J Hum Genet.* 2002; 71:1386–94. [PubMed: 12439824]

Figure 1. Type-I error rates

The x -axis is nominal significance level, and the y -axis is type-I error rate. The gray line represents $y = x$.

Figure 2. Power curves of the six tests for low-frequency variant detection

The x -axis is significance level, and the y -axis is power. The left panel shows the results when the causal allele frequency q is set at 0.01. The middle and the right panels present the results when the causal allele frequency q is 0.05, where the average R^2 (between the causal SNP and the surrounding markers) is smaller than 0.2, and between 0.2 and 0.3, respectively. In the bottom of each panel, we list the number of replications meeting the conditions of the causal allele frequency and the average R^2 .

Figure 3. Power curves of the six tests for common variant detection

The x -axis is significance level, and the y -axis is power. The top, middle, and bottom rows are the results when the causal allele frequency q is set at 0.1, 0.3, and 0.5, respectively. The left, middle, and right columns present the results based on the replications with average R^2 (between the causal SNP and the surrounding markers) smaller than 0.2, between 0.2 and 0.3, and larger than 0.3, respectively. In the bottom of each panel, we list the number of replications meeting the conditions of the causal allele frequency and the average R^2 .

Figure 4. Power curves of the six tests when $q = 0.5$, stratified by the Maximum Frequency of Risk Haplotypes (MFRH)

The x -axis is significance level, and the y -axis is power. The left and the right panels present the results based on the replications with MFRH smaller than 0.4 and larger than 0.4, respectively. In the bottom of each panel, we list the number of replications meeting the condition of MFRH.

Figure 5. The haplotype structure of the *carboxypeptidase E* (*CPE*) gene
 The number in each cell (the quantity between each pair of SNPs) is $100 \times D'$, a measure of linkage disequilibrium.

Table 1
Real data analysis: *P* values for the three haplotype blocks

	Block 1	Block 2	Block 3
SIMp	0.012	0.745	0.857
SIMc	0.049	0.262	0.964
HAP	0.571	0.490	0.844
Single-marker¹	0.178	0.517	0.800
Single-marker-perm²	0.139	0.454	0.712
KMR	0.092	0.382	0.883

¹ Single-marker: The *P* values have adjusted with the Sidak correction (Sidak., 1967), where the effective numbers of tests were estimated as 7.92, 5.80, and 2.99 for the three blocks, respectively. The effective numbers of tests were estimated by the *Keffective* program (Moskvina & Schmidt, 2008).

² Single-marker-perm: The *P* values for the three blocks have adjusted with 50,000 permutations.

Table 2
Real data analysis: Frequencies and P values of haplotype categories in Block 1

Haplotype	loc-1	loc-2	loc-3	loc-4	loc-5	loc-6	loc-7	loc-8	loc-9	loc-10	loc-11	loc-12	Hap-Freq ¹	p-val ²
A	2	1	2	1	1	2	2	1	2	1	2	2	0.01678	0.04491
B	2	1	2	2	2	1	1	2	1	1	2	2	0.01093	0.20489
C	2	1	2	1	1	2	2	1	2	2	2	2	0.24596	0.22458
D	1	1	1	1	1	1	1	2	1	1	1	1	0.10447	0.34698
E	1	2	2	2	2	1	2	1	2	2	2	2	0.02043	0.64974
F	1	2	2	2	1	2	2	1	2	2	2	2	0.02225	0.69037
G	1	1	1	1	1	1	1	2	1	1	2	2	0.01599	0.85261
H	1	2	2	1	1	1	1	2	1	1	1	1	0.01621	0.87198
I	1	1	2	1	1	1	1	1	1	1	1	2	0.02046	0.92624
J	1	1	1	1	1	2	1	2	1	1	2	2	0.01229	0.92330
K	2	1	2	1	1	2	1	2	1	1	2	2	0.01630	0.65679
L	1	2	2	2	2	1	1	2	2	2	2	2	0.05067	0.24820
M	1	2	2	2	2	1	1	2	1	1	2	2	0.37006	0.07503

This table came from the result of the function 'haplo.score' in the R package 'haplo.stats' (Schaid *et al.*, 2002), where all the 1,982 unrelated European Americans were used for analyses. The haplotype categories with frequencies less than 0.01 were not listed here.

¹ Hap-Freq: The frequencies of haplotypes.

² p-val: The P values of the score statistics of individual haplotypes, based on the chi-square distribution with one degree of freedom.