Haplotype-Based Association Analysis via Variance-Components Score Test

Jung-Ying Tzeng and Daowen Zhang

Haplotypes provide a more informative format of polymorphisms for genetic association analysis than do individual single-nucleotide polymorphisms. However, the practical efficacy of haplotype-based association analysis is challenged by a trade-off between the benefits of modeling abundant variation and the cost of the extra degrees of freedom. To reduce the degrees of freedom, several strategies have been considered in the literature. They include (1) clustering evolutionarily close haplotypes, (2) modeling the level of haplotype sharing, and (3) smoothing haplotype effects by introducing a correlation structure for haplotype effects and studying the variance components (VC) for association. Although the first two strategies enjoy a fair extent of power gain, empirical evidence showed that VC methods may exhibit only similar or less power than the standard haplotype regression method, even in cases of many haplotypes. In this study, we report possible reasons that cause the underpowered phenomenon and show how the power of the VC strategy can be improved. We construct a score test based on the restricted maximum likelihood or the marginal likelihood function of the VC and identify its nontypical limiting distribution. Through simulation, we demonstrate the validity of the test and investigate the power performance of the VC approach and that of the standard haplotype regression approach. With suitable choices for the correlation structure, the proposed method can be directly applied to unphased genotypic data. Our method is applicable to a wide-ranging class of models and is computationally efficient and easy to implement. The broad coverage and the fast and easy implementation of this method make the VC strategy an effective tool for haplotype analysis, even in modern genomewide association studies.

Haplotypes of multiple SNPs are considered a more informative format of polymorphisms for genetic association analysis than single SNPs.[1] Haplotypes are more informative because they preserve the joint linkage disequilibrium (LD) structure among multiple adjacent markers.[2] Even when only tag SNPs are used, haplotypes serve as a proxy for unobserved SNPs and increase the predictive power for the genomic variation.[3,4] However, in terms of practical efficacy, the power of haplotype-based association analysis is challenged by a trade-off between the benefits of modeling abundant variation and the cost of the extra degrees of freedom for modeling the multi-marker variations. To avoid the curse of dimensionality encountered in haplotype association analysis, various strategies have been proposed in the literature. They include (1) clustering evolutionarily close haplotypes,[5–8] (2) modeling the level of haplotype sharing instead of the haplotypes themselves,[9–11] and (3) smoothing haplotype effects by introducing a correlation structure for the effects of similar haplotypes.[12–14] Although these strategies appear to be different, the fundamental principle is to use the evolutionary history of haplotypes to reduce the parameter space from individual haplotypes to haplotypes with similar ancestry. However, although the approaches of haplotype clustering and haplotype sharing enjoy a fair amount of power gain, empirical studies found that the smoothing approach may exhibit only similar or less

power than the standard methods that regress trait values on haplotypes and impose no assumptions on haplotypes, even when there are many haplotypes.[14]

In haplotype smoothing, a dependence structure is introduced to the effects of different haplotypes, according to the similarity between haplotypes, under a Bayesian hierarchical model or a mixed-model framework, and the overall gene-trait association can be studied via the variance components (VC).[12–14] The idea of correlating haplotype effect is based on the assumption that the present mutation-bearing haplotypes have descended from a small number of ancestral haplotypes, and, as a result, the disease haplotypes tend to be correlated because of this shared ancestry. Without losing generality, in this work, we refer to these methods as "VC" approaches and discuss them under a mixed-model framework. We also refer to the standard haplotype regression method as a "fixed-effect" approach. Schaid[14] first noted the underpowered phenomenon of the VC method, using the likelihood-ratio test (LRT), and explored potential reasons based on the noncentrality (NC) parameter of the distribution of the LRT statistics. The NC parameter reflects the distance between the alternative distribution and the null distribution of the test statistics, and the larger the null-to-alternative distance is, the higher the power a test possesses. By expressing the NC parameter as a function of heritability (h^2), it can be seen that, although the NC

parameter of a fixed-effect model is proportional to $h^2/(1 - h^2)$, the NC parameter of a VC model is much smaller (proportional to h^4). As a result, the power gain brought by the low degrees of freedom can be compromised with the small NC parameter in a VC-LRT approach.

Here, we report other key factors that contribute to this underpowered phenomenon. In brief, unlike the usual VC model in which the VC represents the potential variability from a source that is independently distributed in the population (e.g., the family effect in the study of linkage or familial aggregation), in the population-based haplotype analysis, the source of variability is not independent. That is, the design matrix of the random haplotype effect does not have a diagonal or block-diagonal structure. Furthermore, the dimension of the random haplotype effect is fixed. Therefore, the data under the alternative hypothesis cannot be represented as a collection of independent data vectors. As a result, the distribution of the LRT statistic does not converge to the conventional 50:50 mixture of χ_0^2 and χ_1^2 (i.e., the limiting distribution predicted by the usual asymptotic theory[15]). Instead, empirical evidence indicates that the distribution of VC-LRT statistics has higher weighting of χ_0^2. Hence, the threshold value obtained from the 50:50 χ^2 mixture is overstringent and causes a too-conservative testing result. Such overconservative findings of the LRT was obtained also by Crainiceanu and Ruppert[16] in certain linear mixed models.

To overcome the problem of a lack of independence and also to generalize the VC approach to all types of trait values, we propose a score test under the generalized linear mixed-model (GLMM) framework. Specifically, we construct a score statistic based on the restricted maximum likelihood (REML) or the marginal likelihood function of the VC and identify its nontypical asymptotic distribution. The proposed test is easy to implement and computationally efficient yet is general enough to accommodate a broad class of phenotypes and correlation structures. It allows for covariate information and can be used for phase-unknown genotypic data. Through simulation, we demonstrate the validity of the test and investigate the power performance of the VC approach and the fixed-effect approach under general scenarios. We also apply the proposed method to a case-control data set from a genomewide association study of amyotrophic lateral sclerosis (ALS) conducted by Schymick et al.[17] In the analysis, we test for gene-trait association on chromosome 10 with the 275 ALS cases and 271 controls and examine statistical significance at the genomewide level. We verify the findings from the proposed method by comparing them with the results reported by Schymick et al.[17]

Material and Methods
VC Method for Association Analysis

We denote the data with the following notations. For individual i ($i = 1,2,\ldots,n$), we have trait value Y_i, environmental covariates $\mathbf{X}_i$ (a $K \times 1$ vector including the intercept term), and haplotype $\mathbf{H}_i$ (an $L \times 1$ vector, where L is the number of distinct haplotypes

observed in the population). Vector $\mathbf{H}_i$ records individual i's haplotype pair via a certain scoring rule, such as by setting its hth element as the number of haplotype h that individual i carries. Throughout this article, we treat explanatory variables (e.g., $\mathbf{X}_i$ and $\mathbf{H}_i$) as constants and will omit them in the lists of the conditional variables. This means that, for example, we will use $\mathrm{Var}(Y_i)$ instead of $\mathrm{Var}(Y_i \mid \mathbf{X}_i, \mathbf{H}_i)$.

Assume that the trait value Y_i follows some distribution with conditional mean $\mathrm{E}(Y_i \mid \boldsymbol{\beta}) = \mu_i$ and conditional variance $\mathrm{Var}(Y_i \mid \boldsymbol{\beta}) = m_i^{-1}\phi v(\mu_i)$, where m_i is a known prior weight (e.g., binomial denominator), ϕ is the dispersion parameter (e.g., measurement-error variance for a normal quantitative trait), and $v(\mu_i)$ is the variance function. Then, the VC model can be expressed under the framework of GLMM as

$$g(\mu_i) = \mathbf{X}_i^T \boldsymbol{\gamma} + \mathbf{H}_i^T \boldsymbol{\beta}$$
$$\boldsymbol{\beta} \sim \mathrm{MN}(0, \tau R_\beta) \; , \qquad (1)$$

where $g(\cdot)$ is a link function that connects the conditional mean μ_i and the explanatory variables, $\boldsymbol{\gamma}_{K \times 1}$ represents the fixed effect of environmental covariates, and $\boldsymbol{\beta}_{L \times 1}$ is the random effect of haplotypes. The haplotype effect is assumed to have a multivariate normal (MN) prior. With model (1), the marginal phenotypic variance, $\mathrm{Var}(Y_i)$, can be partitioned into genetic components and environment components, and the association between haplotypes and traits can be detected by testing for zero genetic VC (i.e., $\tau = 0$). Intuitively, $\tau = 0$ implies that all β_h share the same value, and this is essentially the null hypothesis of the standard fixed-effect approaches.

The correlation structure of β_h is specified through the $L \times L$ matrix R_β. Here, we consider a general formulation for R_β by letting its (h,k) element, denoted by r_{hk}, depend on the similarity level between haplotypes h and k, which is quantified by a certain similarity metric, $s(h,k)$. One simple choice of the correlation structure is to let $R_\beta = I$, where I is the identity matrix. This independence structure imposes no correlation among distinct haplotypes and reflects the "unstructured" variation among haplotypes. The independence prior may be reasonable if haplotype variants were created mainly by recombinations instead of mutations. In contrast, one can introduce local-dependence structures to account for the role of mutation and to reflect the conjecture that evolutionarily close haplotypes tend to have similar effects on traits. One convenient choice of such R_β is the conditional autoregressive (CAR) structure. The CAR structure assumes that all β_h are correlated but that the correlation diminishes as the haplotype similarity decays. With our representation, a CAR structure is to let $R_\beta = C$, where C^{-1} has diagonal elements equal to 1 and off-diagonal elements equal to $-s(h,k)$.[18] Alternatively, to avoid choosing between an independence prior and a sole CAR prior, an intermediate option, in practice, is the convolution model that combines the two: $\tau R_\beta = \tau_1 I + \tau_2 C$.[12,13] In this work, we focus on the model that was considered by Schaid[14] and set $r_{hk} = s(h,k)$, with $0 \leqslant s(h,k) \leqslant 1$. This model uses the haplotype similarity to reflect the correlation directly. It is more extreme but uses a simpler concept than the convolution model, by compromising between the dependence and the independence priors. It allows for correlation induced from partially similar haplotypes but assumes independence among haplotypes that share zero similarity.
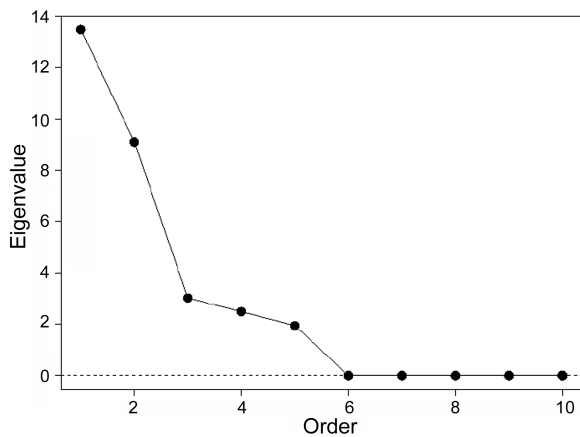
Figure 1. The 10 largest eigenvalues of matrix $W^{-1/2}P_0SP_0W^{-1/2}$. The eigenvalues are dominated by the first few and decrease rapidly to 0. The eigenvalues are averages across the 1,000 replications in the simulation scenario of high haplotype diversity and an allele frequency of 0.1 for binary traits.

VC-Score Test for Haplotype-Phenotype Association

To motivate our VC-score test for haplotype-phenotype association, we illustrate the method, assuming a normally distributed trait (perhaps after some transformation, such as the logarithm transformation) with a known dispersion parameter, ϕ. We then present the VC-score test for general scenarios of unknown ϕ and trait values with an arbitrary distribution. We provide the derivation of the generalization in appendixes A and B.

Quantitative traits with known dispersion parameter ϕ.—For quantitative traits that follow a normal distribution directly or after appropriate transformations, model (1) reduces to a linear mixed-model in matrix notation:

$$\mathbf{Y} = X\boldsymbol{\gamma} + H\boldsymbol{\beta} + \boldsymbol{\epsilon} , \qquad (2)$$

where X is the design matrix for $\boldsymbol{\gamma}$, whose ith row is $\mathbf{X}_i^T$; H is the design matrix for $\boldsymbol{\beta}$, whose ith row is $\mathbf{H}_i^T$; $\boldsymbol{\beta} \sim \mathrm{MN}(0, \tau R_\beta)$ is the same as described in model (1); and $\boldsymbol{\epsilon} \sim \mathrm{N}(0, \phi I)$ represents the uncertainty in measuring traits Y. Since our primary interest is to test $H_0:\tau = 0$, we consider the REML log-likelihood function of VC (τ, ϕ). It is well known that the REML estimating equation for (τ, ϕ) is unbiased and will produce less biased estimates compared with the maximum-likelihood approach.[19]

Denote by $\ell_{\mathrm{REML}}(\tau, \phi; \mathbf{Y})$ the REML log-likelihood function of τ and ϕ, which is given by

$$\ell_{\mathrm{REML}}(\tau, \phi; \mathbf{Y}) =$$
$$-\frac{1}{2}\log|V| - \frac{1}{2}\log|X^TV^{-1}X| - \frac{1}{2}\mathbf{Y}^TP\mathbf{Y} , \qquad (3)$$

where $V = \tau HR_\beta H^T + \phi I \equiv \tau S + \phi I$ is the marginal variance of $\mathbf{Y}$ and where $P = V^{-1} - V^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1}$ is the projection matrix for the linear mixed model (2). The REML log-likelihood function (3) can also be viewed as the marginal log-likelihood of (τ, ϕ) from the Bayesian perspective obtained by specifying a flat prior for γ and integrating out γ from $f(Y; \gamma, \tau, \phi)$.

Simple algebra[20] shows that the score statistic of τ evaluated under H_0 on the basis of the REML function (3) is equal to

$$U_\tau = \frac{\partial \ell_{\mathrm{REML}}(\tau, \phi)}{\partial \tau}\bigg|_{\tau=0} = \frac{1}{2}\{\mathbf{Y}^TP_0SP_0\mathbf{Y} - \mathrm{tr}(P_0S)\} , \qquad (4)$$

where $P_0 = \phi^{-1}\{I - X(X^TX)^{-1}X^T\} = \phi^{-1}Q$ is the projection matrix P evaluated under $H_0:\tau = 0$ and where $Q = I - X(X^TX)^{-1}X^T$. It is immediately seen from equation (4) that $\mathrm{E}(U_\tau) = 0$ under $H_0:\tau = 0$, and, when $\tau > 0$, $\mathrm{E}(U_\tau) = \tau \cdot \mathrm{tr}(QSQS)/(2\phi^2)$, which is a strictly increasing function of τ unless $QS = 0$. Therefore, larger values of U_τ provide stronger evidence against H_0. This suggests that the testing procedure for $H_0:\tau = 0$ using U_τ should be one sided.

In a situation where the VC τ represents the potential variability due to a source that is independently distributed in the population such as the subject-specific effects in a longitudinal study, the score statistic U_τ given in equation (4) under $H_0:\tau = 0$ has an asymptotic normal distribution with zero mean and some variance when the number of independent clusters goes to infinity.[21] However, this condition does not satisfy in our case. In model (1), the design matrix H for the random effects β is not block diagonal and the dimension of β is fixed. Hence, the Lin's[21] asymptotic result does not directly apply to U_τ.

Since ϕ is known, the second term in U_τ is a constant. Therefore using the score statistic U_τ is equivalent to using the first term of U_τ (denoted by T_τ):

$$T_\tau = \frac{1}{2}\mathbf{Y}^TP_0SP_0\mathbf{Y} = \frac{1}{2\phi^2}\mathbf{Y}^TQSQ\mathbf{Y} . \qquad (5)$$

We show in appendix A that T_τ has the same distribution as the weighted χ^2 random variables $\sum_{i=1}^c \lambda_i \chi_{1,i}^2$, where $\chi_{1,i}^2$'s are independent χ^2 random variables with 1 df, and λ_i is the ordered nonzero eigenvalues of the semipositive definite matrix $QSQ/(2\phi)$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_c > 0$ ($c \leq L$). If the $(1 - \alpha)$th quantile of this weighted χ^2 distribution is denoted by $T_{(\alpha)}$, then a level α score test will reject H_0 if $T_\tau \geq T_{(\alpha)}$.

General traits with unknown dispersion parameter ϕ.—Here, we present the VC-score test for the general case in which the traits may

Table 1. Type I Error Rates of the VC-Score Test for Quantitative Traits

Haplotype Diversity and Allele Frequency q	Disease SNP Tagged		Disease SNP Not Tagged	
	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$
High:				
q = .1	.056	.015	.049	.010
q = .3	.047	.011	.047	.012
q = .5	.050	.010	.054	.012
Moderate:				
q = .1	.050	.015	.056	.010
q = .3	.048	.012	.044	.008
q = .5	.052	.012	.050	.012
Low:				
q = .1	.046	.012	.046	.009
q = .3	.047	.011	.054	.012
q = .5	.046	.013	.052	.012

NOTE.—The results are based on 2,000 replications.

Table 2. Type I Error Rates of the VC-Score Test for Binary Traits

Haplotype Diversity and Allele Frequency q	Disease SNP Tagged		Disease SNP Not Tagged	
	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$
High:				
$q = .1$	.040	.007	.050	.014
$q = .3$	.050	.014	.043	.008
$q = .5$	.060	.014	.051	.010
Moderate:				
$q = .1$	.050	.014	.041	.009
$q = .3$	.046	.014	.056	.016
$q = .5$	.046	.012	.046	.008
Low:				
$q = .1$	.052	.010	.050	.013
$q = .3$	.052	.010	.050	.013
$q = .5$	.057	.014	.041	.009

NOTE.—The results are based on 2,000 replications.

not be normally distributed and the dispersion parameter ϕ may or may not be known. As indicated by the derivation given in appendix B, our test statistic can be defined as

$$T_\tau = \frac{1}{2}(\mathbf{Y} - \boldsymbol{\mu})^T \Delta W S W \Delta (\mathbf{Y} - \boldsymbol{\mu}) \bigg|_{\tau=0, \phi=\hat{\phi}, \gamma=\hat{\gamma}}, \qquad (6)$$

where $\boldsymbol{\mu} = g^{-1}(X\gamma)$, $\Delta = \mathrm{diag}\{g'(\mu_i)\}$, $\hat{\gamma}$ is the maximum-likelihood estimate of γ under H_0, and $\hat{\phi}$ is the REML type of estimate (such as the one that uses Pearson residuals) of ϕ under H_0. Matrix

$W = \mathrm{diag}\{w_i\}$, with $w_i = \{\phi m_i^{-1} v(\mu_i)[g'(\mu_i)]^2\}^{-1}$. These quantities are readily available by fitting a standard generalized linear model, $g(\mu) = X\gamma$. We derive in appendix B that T_τ also follows approximately the weighted χ^2 distribution $\sum_{i=1}^c \lambda_i \chi_{1,i}^2$, where $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_c > 0$ ($c \leqslant L$) is the nonzero eigenvalues of matrix $W^{-1/2} P_0 S P_0 W^{-1/2}/2$. We note that the conclusions given in the previous section are a special case of the results given here. For normally distributed traits, $\Delta = I$, and $W = V^{-1}$, which equals $\phi^{-1} I$ under H_0. Hence, equation (6) reduces to equation (5), and the matrix $W^{-1/2} P_0 S P_0 W^{-1/2}/2$ reduces to $QSQ/(2\phi)$.

Gamma approximation of the distribution of test statistic T_τ.—Given the fact that T_τ follows a weighted χ^2 distribution, one can obtain the significance threshold $T_{(\alpha)}$ at level α from simulation. However, such a task may not be trivial when α is small. As an alternative, we introduce a Gamma approximation of the distribution of T_τ. Empirical evidence indicates that the eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_c$ of the matrix $W^{-1/2} P_0 S P_0 W^{-1/2}/2$ are dominated by the first few ones and decay rapidly to 0 (fig. 1). Following the work of Zhang and Lin,[22] we use the Satterthwaite method to approximate the null distribution of T_τ by a Gamma distribution with parameters (a,b). Let $\mathcal{E}$ and $\mathcal{V}$ denote the mean and variance of T_τ, respectively. We match the mean and the variance of the Gamma distribution and those of the test statistic by setting $ab = \mathcal{E}$ and $ab^2 = \mathcal{V}$, and we get $a = \mathcal{E}^2/\mathcal{V}$ and $b = \mathcal{V}/\mathcal{E}$. We can then obtain $T_{(\alpha)}$ or calculate the P value of the test statistic from the distribution of Gamma (a,b). The mean, $\mathcal{E}$, and variance, $\mathcal{V}$, of T_τ can be calculated (appendixes A and B) by

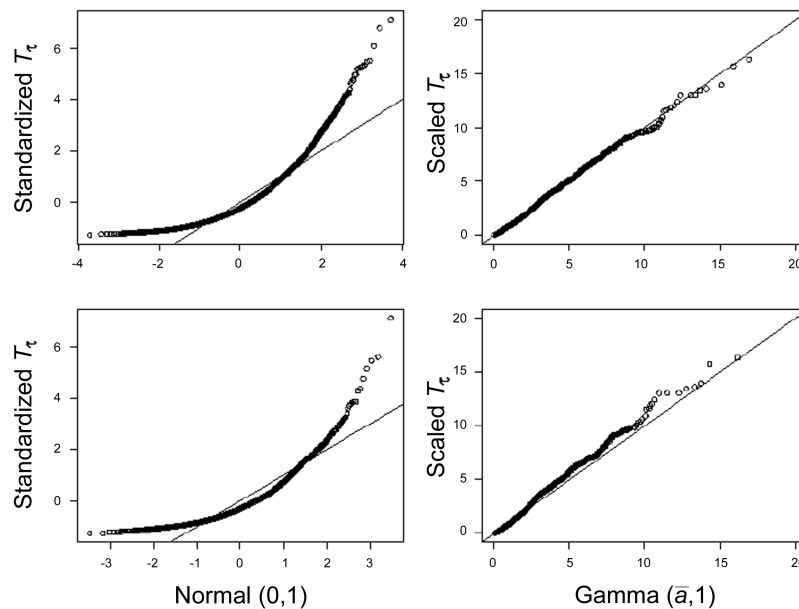$$\hat{\mathcal{E}} = \frac{1}{2}\mathrm{tr}(P_0 S)$$

Figure 2. QQ-plots of the test statistic T_τ. We defined the standardized T_τ as $(T_{\tau,i} - \mathcal{E}_i)/\sqrt{\mathcal{V}_i}$ and the scaled $T_{\tau,i}$ as $T_{\tau,i}/b_i$, where b_i is the scale parameter of the Gamma approximation. The left panels show the standardized T_τ against the expected quantiles from Normal(0,1). The right panels show the scaled T_τ against the expected quantiles from the Gamma($\bar{a}$,1). The upper panels are for quantitative traits, and the lower panels are for binary traits. These QQ-plots suggest that, for both trait types, T_τ values deviate from the normal distribution, as predicted by the usual asymptotic theory, and the empirical distribution of T_τ values can be approximated by a Gamma (a_i, b_i) distribution appropriately.

Table 3. Power for Quantitative Traits When the Disease SNP Is Tagged

Haplotype Diversity and Allele Frequency q	$\alpha = .05$		$\alpha = .01$	
	VC-Score Test	Fixed-Effect Test	VC-Score Test	Fixed-Effect Test
High:				
$q = .1$	.889[a]	.849	.701[a]	.658
$q = .3$	.952[a]	.714	.846[a]	.473
$q = .5$	.965[a]	.808	.884[a]	.587
Moderate:				
$q = .1$	.912[a]	.860	.759[a]	.704
$q = .3$	.987[a]	.853	.954[a]	.668
$q = .5$	.983[a]	.877	.951[a]	.689
Low:				
$q = .1$	.950[a]	.924	.845[a]	.790
$q = .3$	.994[a]	.957	.977[a]	.869
$q = .5$	.994[a]	.909	.979[a]	.766

NOTE.—The results are based on 1,000 replications.

[a] Cases for which the power of one method is significantly higher than that of the competing method at a nominal level of 0.05 by use of McNemar's test.

and

$$\hat{V} = \hat{I}_{\tau\tau} - \hat{I}_{\tau\phi}^2 / \hat{I}_{\phi\phi} ,$$

where

$$P_0 = W - WX(X^TWX)^{-1}X^TW\Big|_{\tau=0,\phi=\hat{\phi},\gamma=\hat{\gamma}} ,$$

$$\hat{I}_{\tau\tau} = \frac{1}{2}\text{tr}(P_0SP_0S) ,$$

$$\hat{I}_{\tau\phi} = \frac{1}{2\hat{\phi}}\text{tr}(P_0SP_0W^{-1}) = \frac{1}{2\hat{\phi}}\text{tr}(P_0S) ,$$

and

$$\hat{I}_{\phi\phi} = \frac{1}{2\hat{\phi}^2}\text{tr}(P_0W^{-1}) = \frac{n-K}{2\hat{\phi}^2} .$$

Phased haplotype data versus unphased genotype data.—Although we have described our test, assuming that the haplotype information H is observed, the phase information can be not crucial. From equations (5) and (6), we see that the haplotype information appears in T_τ only through $S = HR_\beta H^T$, whose (i,j) element, denoted by S_{ij}, can be rewritten as

$$S_{ij} = \mathbf{H}_i^T R_\beta \mathbf{H}_j = \sum_{h,k} H_{i,h}H_{j,k} \times s(h,k) .$$

The right-hand side of the equation states that S_{ij} is simply the similarity score between the haplotype pair of person i and that of person j measured by metric $s(h,k)$. As a result, by choosing those metrics that do not require phase information, we can calculate S without resorting to the phased data. One choice is to set $s(h,k)$ as the proportion of matching alleles between two haplotypes, h and k. As demonstrated by Tzeng et al.[11] and Schaid,[14] such S_{ij} is equivalent to the proportion of matching alleles between the genotypes of individual i and individual j and hence can be calculated directly from genotypes with unknown phase.

Simulation Studies

We conduct simulation studies to examine the performance of the proposed score test. In the simulation, we generated covariates X_i, haplotypes H_i, and trait values Y_i, given X_i and H_i, for each individual. The covariate X_i is drawn from a standard normal distribution, and the haplotype H_i is generated using a technique similar to those reported by Roeder et al.[23] and Tzeng et al.[8] Specifically, we simulated 100 haplotypes under the coalescent model,[24] with an effective population size of 10^4, a scaled mutation rate of 5.6×10^{-4} per bp, and a scaled recombination rate of $\sim 6 \times 10^{-3}$ per bp for the cold spots and a rate 45 times greater for the hotspots. These parameters are chosen to roughly match the genes observed in the SeattleSNP database. We discarded SNPs with minor-allele frequencies <0.05. The hypothetical disease locus is selected on the basis of a predetermined minor-allele frequency, q, and the diversity of haplotypes flanking the SNP. In the simulation, we considered $q = 0.1$, 0.3, and 0.5 and haplotype-diversity levels of high (11–16 distinct haplotypes), moderate (9–11 distinct haplotypes), and low (6–9 distinct haplotypes). We set a haplotype region to be a segment of five adjacent SNPs, including the two SNPs on the left and the three SNPs on the right of the disease locus. Given that the disease SNP is excluded, we also considered whether the disease SNP is "tagged" or "not tagged" by the surrounding five SNPs under each scenario. We defined that the disease SNP is "tagged" if there is at least one SNP whose R^2 with the disease SNP is >0.7, and it is "not tagged" otherwise. We then randomly sampled with replacement of 2 haplotypes from the 100 haplotypes to form an individual. The simulated haplotype data were then converted into unphased genotype data.

We next generated the trait values Y_i on the basis of X_i and the genotypes at the disease locus. We determined the trait value of individual i according to X_i and the number of disease alleles (G_i), using an additive-effect model. In the simulation study, we considered both quantitative traits and binary traits and adopted the same trait-generating scheme as did Lake et al.[25] and Tzeng et al.[8] For quantitative traits, we used a random-sampling scheme and generated 200 trait values from the normal conditional distribution of Y_i with mean $\gamma_0 + \gamma_1 \times X_i + (G_i - 1)$ and variance

Table 4. Power for Quantitative Traits When the Disease SNP Is Not Tagged

Haplotype Diversity and Allele Frequency q	$\alpha = .05$		$\alpha = .01$	
	VC-Score Test	Fixed-Effect Test	VC-Score Test	Fixed-Effect Test
High:				
$q = .1$	.470	.563[a]	.248	.347[a]
$q = .3$	.140	.282[a]	.039	.131[a]
$q = .5$	.595	.588	.341	.342
Moderate:				
$q = .1$	.309	.599[a]	.144	.364[a]
$q = .3$	.789[a]	.613	.581[a]	.389
$q = .5$	.959[a]	.874	.860[a]	.708
Low:				
$q = .1$	.558	.797[a]	.334	.588[a]
$q = .3$	.559[a]	.447	.294[a]	.253
$q = .5$	.484	.550[a]	.247	.293[a]

NOTE.—The results are based on 1,000 replications.

[a] Cases for which the power of one method is significantly higher than that of the competing method at a nominal level of 0.05 by use of McNemar's test.

Table 5. Power for Binary Traits When the Disease SNP Is Tagged

Haplotype Diversity and Allele Frequency q	$\alpha = .05$		$\alpha = .01$	
	VC-Score Test	Fixed-Effect Test	VC-Score Test	Fixed-Effect Test
High:				
$q = .1$	.372[a]	.274	.162[a]	.102
$q = .3$	.565[a]	.298	.359[a]	.102
$q = .5$	.573[a]	.327	.364[a]	.129
Moderate:				
$q = .1$	.462[a]	.279	.226[a]	.110
$q = .3$	.647[a]	.366	.473[a]	.167
$q = .5$	.584[a]	.298	.375[a]	.098
Low:				
$q = .1$	.457[a]	.348	.250[a]	.134
$q = .3$	.753[a]	.540	.542[a]	.282
$q = .5$	.717[a]	.377	.505[a]	.179

NOTE.—The results are based on 1,000 replications.

[a] Cases for which the power of one method is significantly higher than that of the competing method at a nominal level of 0.05 by use of McNemar's test.

$2q(1-q) \times (1-h^2)/h^2$. We set the heritability (h^2) at 0.1 and $\gamma_0 = \gamma_1 = 1$. For binary traits, we used a case-control sampling scheme and generated trait values of 0 or 1, using the penetrance function $\text{logit} P(Y = 1 \mid G_i, X_i) = \gamma_0 + \gamma_1 \times X_i + \theta \times G_i$. We set the odds ratio (OR) (e^θ) at 2.0 and set the disease prevalence at 0.01 by letting $\gamma_0 = -4.5$ and $\gamma_1 = 0$. We repeated the process until we collected 100 cases and 100 controls.

We analyzed these simulated data to evaluate the power performance of the VC-score method. To compare, we also conducted haplotype analyses, using the fixed-effect method and, in addition, the VC method via regular LRT (VC-LRT) under some scenarios. These analyses were performed assuming unknown phases. For fixed-effect analysis, we used the haplotype-score test of Schaid et al.,[26] as implemented in the R function "haplo.score," and determined the P values by using the asymptotic χ^2 distribution. The P values of our VC-score method were obtained from the approximated Gamma distribution, and the P values of the VC-LRT method were obtained from 50:50 mixtures of χ_0^2 and χ_1^2.

Data Application

We considered the data set from the genetic association study of ALS conducted by Schymick et al.[17] The ALS data set consists of 276 patients with sporadic ALS and 271 neurologically normal control subjects[27] and contains their genotypes at the 550K SNPs across the genome in the Illumina chip assays. The original genotyping was performed in the laboratory of Drs. Singleton and Hardy at National Institute of Aging. The genotype data have been made publicly available in the SNP Database at the National Institute of Neurological Disorders and Stroke (NINDS) Human Genetics DNA and Cell Line Repository. Schymick et al.[17] performed a genomewide association analysis and reported 34 SNPs that have P values <.0001 on the basis of the single-SNP genotypic test with 2 df. They used the Bonferroni correction to adjust for multiple testing, and the threshold of significance at the nominal level of 0.05 is 9.1×10^{-8}. Although none of the 34 SNPs was significant after the Bonferroni correction, the most significant SNP (*rs4363506*) lay in close proximity to one of the actin cy-

toskeleton genes (the dedicator of cytokinesis 1 gene [*DOCK1* {MIM 601403}]) that are increasingly recognized as playing an important role in motorneuron disease. To assess the performance of our method, we applied the proposed VC-score method to part of this data set. We focused our analysis on chromosome 10, where the most significant SNP is located. We used the results reported by Schymick et al.[17] as a benchmark to evaluate our findings.

Results

Simulation Studies

In the comparison of our VC-score method with the fixed-effect method, we reported the results for quantitative traits and binary traits under 18 scenarios (3 values of allele frequency × 3 levels of haplotype diversity × 2 different tagging statuses of the disease SNP). Type I error rates were calculated on the basis of 2,000 replications, and power was calculated on the basis of 1,000 replications.

We listed the results of type I error rates of the VC-score test in table 1 for quantitative traits and in table 2 for binary traits. The values are around the nominal levels of $\alpha = 0.05$ and $\alpha = 0.01$, indicating that the Gamma distribution approximates the null distribution of T_τ adequately. To ensure this conclusion, we also examined the null distribution of the test statistics; the results are displayed in figure 2. The left panels of figure 2 show the two quantile plots (hereafter, "QQ-plots") that compared the quantiles of the standardized T_τ from the null distribution with the quantiles of the standard normal distribution. The upper panels are for quantitative traits, and the lower panels are for binary traits. In both cases, it is apparent that standardized T_τ does not have a standard normal distribution, following Lin's[21] asymptotic result. We then draw the QQ-plots of T_τ against the Gamma distribution (fig. 2, right panels). In the simulation, each replication i

Table 6. Power for Binary Traits When the Disease SNP Is Not Tagged

Haplotype Diversity and Allele Frequency q	$\alpha = .05$		$\alpha = .01$	
	VC-Score Test	Fixed-Effect Test	VC-Score Test	Fixed-Effect Test
High:				
$q = .1$	.209	.218	.077	.075
$q = .3$	.094	.135[a]	.024	.035
$q = .5$	.187	.183	.052	.055
Moderate:				
$q = .1$	.128	.190[a]	.044	.052
$q = .3$	.389[a]	.255	.200[a]	.087
$q = .5$	.482[a]	.326	.266[a]	.134
Low:				
$q = .1$	.209	.249[a]	.095	.112
$q = .3$	.246	.228	.097[a]	.076
$q = .5$	.213	.238[a]	.080	.087

NOTE.—The results are based on 1,000 replications.

[a] Cases for which the power of one method is significantly higher than that of the competing method at a nominal level of 0.05 by use of McNemar's test.

Table 7.   Comparison of the Fixed-Effect Test, VC-Score Test, and VC-LRT

Trait Type, Effect Size, and Parameter	$\alpha = .05$			$\alpha = .01$		
	Fixed-Effect Test	VC-Score Test	VC-LRT	Fixed-Effect Test	VC-Score Test	VC-LRT
Quantitative:						
Power[a] at $h^2 = .00$	.047	.048	.001	.015	.011	.000
Type I error[b] at $h^2 = .05$	.696	.724	.255	.384	.488	.194
Type I error[b] at $h^2 = .10$	.964	.978	.722	.880	.914	.722
Binary:						
Power[a] at OR = 1.0	.075	.051	...	.028	.015	...
Type I error[b] at OR = 2.0	.226	.314	...	.112	.138	...
Type I error[b] at OR = 2.5	.418	.534	...	.218	.310	...

NOTE.—Simulations assume a high haplotype diversity, a disease-allele frequency of 0.1, and a sample of 200 individuals obtained from random sampling.

[a] Power calculated on the basis of 500 replications.

[b] Type I error rates calculated on the basis of 1,000 replications.

generated a $T_{\tau,i}$ that follows approximately a Gamma distribution with a unique shape parameter, a_i, and a unique scale parameter, b_i. To create a QQ-plot against these nonidentical Gamma variables, we first created a scaled $T_{\tau,i} = T_{\tau,i}/b_i$ that follows Gamma $(a_i, 1)$. Then, we used a single shape parameter of $\bar{a} = \text{mean}(a_i)$ to obtain the theoretical quantiles. Although use of a single shape parameter can cause some deviation in the QQ-plot (such as what can be observed in the right section of the graph), overall we see that the data points agree with the 45° line, indicating that the Gamma approximation works reasonably well.

The results of power comparison are displayed in tables 3 and 4 for quantitative traits and in tables 5 and 6 for binary traits. We highlighted those cases in which the power gain is significant at a 0.05 level by use of McNemar's test. We found that the correlation between the disease SNP and its nearby SNPs plays a key role in predicting the performance of the VC-score test compared with the fixed-effect test. When the disease SNP was tagged by at least one of surrounding SNPs, we observed a systemic power improvement of the VC-score method over the fixed-effect method. This is consistent for both trait types across all scenarios (tables 3 and 5). If none of the adjacent SNPs was highly correlated (i.e., $R^2 > 0.7$) with the unobserved disease SNP, we saw a power drop compared with that seen for the tagged SNPs. In these cases, the fixed-effect method tends to retain a higher power than that of the VC-score method, although the pattern is not universal (tables 4 and 6).

We also examined the performance of the VC-score test by varying the strength of genetic effects. We set the heritability h^2 at 0.00, 0.05, and 0.10 for quantitative traits and set the OR at 1.0, 2.0, and 2.5 for binary traits. This simulation considered the scenario of high haplotype diversity, a tagged disease SNP, and an allele frequency of 0.1 and used a sample size of 200 individuals generated from random sampling for both trait types. The power is calculated on the basis of 500 replications, and the type I error rate (i.e., for $h^2 = 0$ and OR = 1) is calculated on

the basis of 1,000 replications. As a quick verification, we see from table 7 that the power of the VC-score method increases as the genetic effect becomes stronger.

As a comparison, we also conducted the fixed-effect analysis and the VC analysis with the regular LRT. We were unable to obtain the result of the VC-LRT for binary traits because of computational limitations, since one has to calculate a c-dimensional (c = the number of the nonzero eigenvalues) numerical integration to obtain the LRT statistics. Table 7 shows that, as expected, the VC-LRT produces the lowest power among the three methods. The analysis of type I error rate helps to explain the low power of the VC-LRT; the size determined from the regular 50:50 mixture of χ_0^2 and χ_1^2 is extremely small, and the overconservative threshold obtained from the χ^2 mixture leads to a loss of power. When compared with the fixed-effect method, we noticed that the power loss of the VC-LRT method in our simulation is more substantial than seen in the results of Schaid.[14] We think this is probably because Schaid simulated data from a VC model, which would favor the performance of a VC approach.

Analysis of the ALS Data Set

Using unphased genotype data from the ALS study, we replicated the single-SNP genotypic test of Schymick et al.[17] and performed two haplotype association analyses, one with the fixed-effect method and the other with the proposed VC-score method. There are a total of 28,818 SNPs genotyped on chromosome 10, and we removed the 26,258th SNP because of its ambiguous marker information. Following the same haplotype definition as that in the work of Schymick et al.,[17] we also defined haplotypes by using a sliding window of three SNPs in the haplotype analyses.

The results from the VC-score method showed that the most significant association signal, with a P value of 1.2×10^{-7}, is near *rs4363506*. The P values are presented in figure 3, with the location of *rs4363506* indicated by the arrows. The two adjacent windows around the most
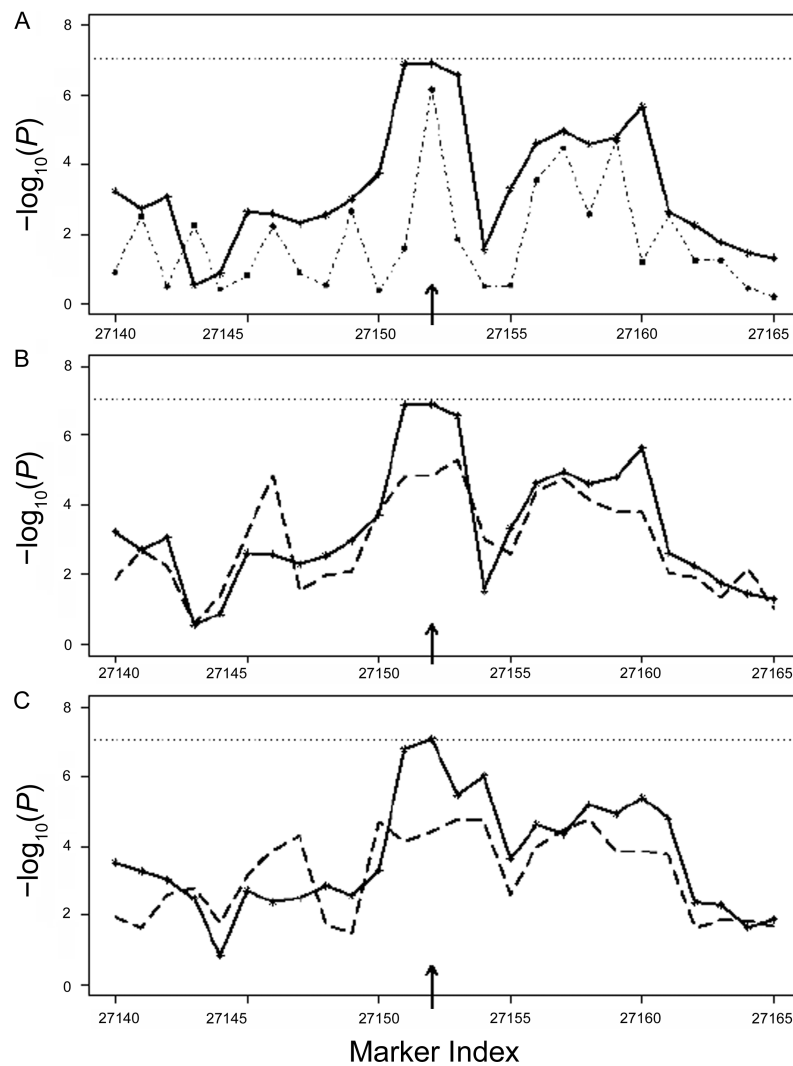
Figure 3. *P* values from the ALS data analysis around the most promising SNP reported by Schymick et al.[17] (i.e., SNP *rs4363506,* with location indicated by the arrows). The *P* values are presented on the scale of negative logarithm of base 10. *A, P* values of the three-SNP haplotype test by use of the VC-score method (*solid line*) and *P* values of a single-SNP genotypic test (*dotted dashed line*). *B, P* values of the three-SNP haplotype test by use of the VC-score method (*solid line*) and *P* values of the three-SNP haplotype test by use of the fixed-effect method (*dashed line*). *C, P* values of the four-SNP haplotype test by use of the VC-score method (*solid line*) and *P* values of the four-SNP haplotype test by use of the fixed-effect method (*dashed line*).

significant signal also have similar *P* values: 1.3×10^{-7} and 2.7×10^{-7}. These locations agree with the findings of Schymick et al.,[17] who reported that *rs4363506* has a *P* value of 6.8×10^{-7} for the genotypic test and a *P* value of 4.8×10^{-6} for the three-marker haplotype test. Although our *P* values appeared smaller, they were not significantly different from the results of the single-SNP analyses and were not significant after Bonferroni correction with the threshold of 9.1×10^{-8}. We also compared the VC-score results with those of the fixed-effect method (fig. 3*B*). The fixed-effect method also indicated a peak signal around *rs4363506,* with the peak *P* value (4.9×10^{-6}) slightly larger than those of the VC-score test and the single-SNP test. In general, we observed that analyses at the haplotype

level reduced the noisy association signals of single SNPs. Although the clearer association pattern of haplotypic analyses came with the cost of extra degrees of freedom used in multimarker variations, we see that the VC-score haplotype test achieved a level of significance that is comparable to that of single-SNP analyses.

To ensure that the results are not sensitive to the definition of the haplotypes, we repeated our analysis with window sizes of three, four, and five SNPs. The *P* values for a window size of four SNPs for fixed-effect and VC-score methods are shown in figure 3*C*, in which the *P* value curves retain a similar pattern to the *P* value curves in the three-SNP analysis (fig. 3*B*). Indeed, the *P* value curves of various window sizes for the same method are

similar (data not shown), except that larger window sizes led to a more smoothing effect on the P values across SNPs. With window sizes of four and five SNPs, the P values at the windows containing *rs4363506* are 8.3×10^{-8} and 7.7×10^{-8}, respectively; both are significant at the threshold determined by the Bonferroni correction. Judging by the criteria of locating the signal around *rs4363506* and producing reasonable P values, we think that the VC-score method performed competitively with the standard analyses of these data.

Discussion

The VC approaches have been considered as one common strategy to reduce the degrees of freedom required in haplotype association analyses. However, it is noticed that direct application of the VC-LRT to haplotype association often fails to increase power. In this article, we reported possible reasons that contribute to this phenomenon, which include that (a) the LRT statistic has a nontypical limiting distribution in a haplotype random-effect model and (b) none of the surrounding SNPs is highly correlated with the unobserved disease SNP. Although the latter reason naturally limits the performance of the type of VC approaches discussed in this article, the former can be overcome. In essence, our work tackles this limiting-distribution problem; we introduced a VC-score test based on the REML or the marginal likelihood function of the VC under the GLMM. We showed that the test statistic follows a weighted χ^2 distribution and provided a Gamma approximation. We demonstrated the validity of the proposed method through simulation. Constructed under the GLMM framework, our VC method can be applied to a broad class of data, allowing for traits of various types, different choices of correlation structure, and a flexible range of model assumptions. Finally, by choosing suitable similarity metrics, the proposed method can be directly applied to unphased genotypic data.

We note that the LRT statistic in the VC haplotype model does not converge to the distribution derived from the typical asymptotic theory that assumes independent clusters. As a result, use of the conventional limiting distribution could lead to an overconservative testing result. Crainiceanu and Ruppert[16] reported similar findings and provided a practical procedure to find the distribution of the LRT statistic for continuous Y variables. However, we still recommend the score test over the LRT, for several reasons. First, the correct LRT procedure of Crainiceanu and Ruppert[16] is applicable only to continuous traits. Second, the LRT is generally a more difficult test to implement under the GLMM framework. For example, with binary traits, it is almost impractical to obtain the exact maximum-likelihood estimates of the VC and, hence, the LRT statistic. On the contrary, our score test is applicable to a wide-ranging class of models, and it requires only the estimates under H_0 that can be easily obtained from standard statistical software. The broad coverage and the fast and easy implementation of the score test makes the VC strategy an effective tool for haplotype analysis, even in modern genomewide association studies. We have implemented the VC-score test in R and have distributed the R code at our Authors' Web site.

An additional factor that would influence the power performance is the LD pattern between the observed SNPs and the unobserved disease SNP. Our power analysis showed that, if at least one marker is in high LD with the disease SNP (i.e., $R^2 > 0.7$), the VC-score method performs better than the fixed-effect method. If all markers are in low LD with the disease SNP, both methods suffer from power loss. This is understandable because low correlation indicates lack of information from neighboring SNPs to make the correct inference. However, we noticed that the VC-score test appears to have a larger drop in power. The VC method is developed from an evolutionary point of view, which implicitly assumes high correlation among adjacent SNPs with the disease locus and uses the correlation to reduce the degrees of freedom. On the other hand, the fixed-effect approach does not rely on this assumption. We conjecture that this is the explanation for the observation of different degrees of power loss under the "untagged" scenarios, as well as for the observation of the power gain of the VC method under the "tagged" scenarios. We plan to continue exploring this aspect of the problem.

In this work, we suggested the use of similarity metrics that do not require phase information. For example, a candidate is the metric that counts the number of matching alleles between two haplotypes. Theoretically, metrics that make use of the phase information should be more powerful. A corresponding metric to our candidate that incorporates phase information is one that counts the number of longest consecutive matching alleles. Metrics of this type are more likely to capture the identity-by-descent sharing and to better reflect the results of the decay of haplotype sharing. However, these metrics are not robust to genotyping errors and recent market mutations, which often limit their power in practice. Our previous work found that phase-dependent and phase-independent metrics have similar performance.[11] In addition, the use of phase-independent metrics bypasses the need to infer the phase information, which is often achieved under unrealistic assumptions. For example, the expectation-maximization algorithm is typically used to impute haplotype phases, and it assumes that the population has common haplotype frequencies and is in Hardy-Weinberg equilibrium (HWE). These assumptions tend to not hold with the existence of population substructure; the haplotype frequencies could vary across subpopulations, and the HWE may hold only within the subpopulations but not for the entire population. Consequently, the phases are not inferred accurately—hence, the subsequent association inference. Use of phase-independent measures does not rely on imputing the haplotype information and, therefore, avoids these issues naturally.

Although we have introduced a Gamma approximation of the distribution of the score test for the sake of simplicity, we note that there exists alternative approaches to estimate the distribution. One approach is to use a three-moment approximation,[28,29] as opposed to the two-moment matching method described in this article. The three-moment method uses the information of the non-zero eigenvalues $\lambda_1, \ldots, \lambda_c$ of the matrix $W^{-1/2}P_0SP_0W^{-1/2}/2$ and approximates the α-level significance threshold $T_{(\alpha)}$ as $k_1 + (\chi_\alpha - h') \times \sqrt{k_2/h'}$, where $k_j = \Sigma_{i=1}^c \lambda_i^j$, $h' = k_2^3/k_3^2$, and χ_α is the αth quantile of $\chi_{h'}^2$ (i.e., a χ^2 distribution with h' df). The P value of the test statistic T_τ is then the right-tail portion of $(T_\tau - k_1) \times \sqrt{h'/k_2} + h'$ of the $\chi_{h'}^2$ distribution. Another approach is to obtain directly the empirical distribution of T_τ through simulation based on the fact that $T_\tau = \Sigma_{i=1}^c \lambda_i\chi_{1,i}^2$. To do so, find eigenvalues $\lambda_1, \ldots, \lambda_c$, and generate c sets of random values from the χ_1^2 distribution, each set with a certain sample size—say, 500,000. Then, the weighted sums of these random values form the empirical distribution of the VC-score statistics. The simplicity of the simulation carries over to genomewide association studies, in which case, the sample size of the simulated χ_1^2 values has to be reasonably large with respect to the stringent genomewide significance level. Fortunately, these simulated χ_1^2 random values can be used repeatedly in every haplotype region, and all that needs to be recalculated is the nonzero eigenvalues of matrix $W^{-1/2}P_0SP_0W^{-1/2}/2$ obtained in each region.

The VC-score method introduced here focuses on detecting the haplotype main effect, but the framework can be extended to consider interactions. For example, one can incorporate terms for gene-environment and gene-gene interactions in the model and can examine their significance by testing the corresponding VC. We are exploring these ideas in an ongoing work. We also note that the VC-score method has a direct connection to the strategy of reducing the degrees of freedom in haplotype association tests through haplotype sharing. The haplotype information appears in the formula of the score statistic T_τ as $S = HR_\beta H^T$, and the elements of S record the level of haplotype similarity between two subjects by a certain haplotype metric, $s(h,k)$. On the one hand, in the VC methods, the selection of R_β (and hence the similarity metric) largely emphasizes the evolutionary relationship of haplotypes in history. On the other hand, in the haplotype-sharing approaches, we choose the similarity metric to quantify the similarity level in the current population. We believe that taking advantage of the implicit relationship between the two methods can offer more insights into both strategies. We plan to further investigate this issue.

Appendix A

Distribution and Moments of the VC-Score Statistic for Normal Traits

From equation (5), we have $T_\tau = \mathbf{Y}^T QSQ\mathbf{Y}/(2\phi^2)$ for normal traits. Define vector $\mathbf{Z} = (\mathbf{Y} - \mu)/\sqrt{\phi}$, which follows a standard multivariate normal distribution. We then can rewrite T_τ as

$$T_\tau = \mathbf{Z}^T\frac{QSQ}{2\phi}\mathbf{Z} = \sum_{i=1}^c \lambda_iZ_i^2 \ , \qquad (A1)$$

where $Z_i^2 \sim \chi_1^2$ distribution. Equation (A1) holds as $\boldsymbol{\mu}^TQ = 0$ by the fact of Q, a projection matrix.

When ϕ is unknown, we replace ϕ in our test statistic T_τ with its REML estimate under H_0, which yields the regular mean squared error for the linear regression model $\mathbf{Y} = X\boldsymbol{\gamma} + \boldsymbol{\epsilon}$:

$$\hat{\phi} = \frac{\mathbf{Y}^TQ\mathbf{Y}}{n - K} \ .$$

Since, under $H_0: \tau = 0$, traits $\mathbf{Y}$ reduce to independent data, the variance of $\hat{\phi}$ will be negligible when the total sample size, n, is large. Therefore, the exact distribution of T_τ with ϕ replaced by $\hat{\phi}$ can be approximated by the distribution of $\Sigma_{i=1}^c \hat{\lambda}_i\chi_{i1}^2$, where $\hat{\lambda}_i$ are the eigenvalues of $QSQ/(2\hat{\phi})$, if such "exact" distribution is critically needed.

We can approximate the distribution of T_τ by a Gamma distribution by matching the first two moments. Following the work of Harville,[20] the mean, $\mathcal{E}$, and variance, $\mathcal{V}$, of T_τ are

$$\mathcal{E} = \frac{1}{2\phi}\text{tr}(QS)$$

and

$$\mathcal{V} \equiv I_{\tau\tau} = \frac{1}{2\phi^2}\text{tr}(QSQS) \ .$$

To take into account the fact that ϕ is estimated, we obtained the following result by Taylor expansion:

$$T_\tau(\hat{\phi}) \approx T_\tau(\phi) - I_{\tau\phi}I_{\phi\phi}^{-1}U_\phi \ ,$$

where $I_{\tau\phi} = E(\partial T_\tau/\partial\phi)$, $I_{\phi\phi} = E[\partial^2\ell_{\text{REML}}(\tau,\phi)/\partial\phi^2]$, and $U_\phi = \partial\ell_{\text{REML}}(\tau,\phi)/\partial\phi$. Hence, we can estimate the mean of T_τ by

$$\mathcal{E} \equiv E[T_\tau(\hat{\phi})] = E[T_\tau(\phi)] \approx \text{tr}(QS)/(2\hat{\phi})$$

and can estimate the variance of T_τ by

$$\mathcal{V} = \text{Var}\,[T_\tau(\hat{\phi})] \approx \text{Var}\,[T_\tau(\phi)] - I_{\tau\phi} I_{\phi\phi}^{-1} \text{Var}\,(U_\phi) I_{\phi\phi}^{-1} I_{\phi\tau}$$

$$\approx \hat{I}_{\tau\tau} - \hat{I}_{\tau\phi}^2 / \hat{I}_{\phi\phi} \,,$$

where

$$\hat{I}_{\tau\tau} = \frac{1}{2\hat{\phi}^2} \text{tr}(QSQS) \,,$$

$$\hat{I}_{\tau\phi} = \frac{1}{2\hat{\phi}^2} \text{tr}(QS) \,,$$

and

$$\hat{I}_{\phi\phi} = \frac{1}{2\hat{\phi}^2} \text{tr}(Q) = \frac{n-K}{2\hat{\phi}^2} \,.$$

Appendix B

Derivation of the VC-Score Test for General Traits

Motivated by the nice properties of the REML or the marginal log-likelihood function of the VC for a normal trait, presented in equation (3), and the success of the score test for testing a polynomial covariate effect in semiparametric additive mixed models by use of mixed-model representation,[22] we take a similar approach for a nonnormal trait, Y, such as a binary disease trait. Under model specification (1), the marginal log-likelihood function $\ell_M(\tau,\phi;Y)$ of τ and the possible dispersion parameter ϕ are given by

$$e^{\ell_M(\tau,\phi;Y)} = \int f(\mathbf{Y};\boldsymbol{\gamma},\tau,\phi)d\boldsymbol{\gamma}$$

$$= \int \int f(\mathbf{Y}|\boldsymbol{\beta};\boldsymbol{\gamma},\phi)f(\boldsymbol{\beta}|\tau)d\boldsymbol{\beta}d\boldsymbol{\gamma} \,. \qquad \text{(B1)}$$

The marginal log-likelihood in (B1) usually involves a high-dimensional and mathematically intractable integration. To overcome this problem, Zhang and Lin[22] used Laplace approximation to derive an approximate score statistic for testing $H_0:\tau = 0$ on the basis of a similar marginal log-likelihood function. For our model, their score statistic (eq. [16] in the work of Zhang and Lin[22]) reduces to

$$U_\tau \approx \frac{1}{2}\Big\{(\mathbf{Y} - \boldsymbol{\mu})^T \Delta WSW \Delta (\mathbf{Y} - \boldsymbol{\mu})$$

$$- \text{tr}(P_0 S)\Big\}\Big|_{\tau=0,\phi=\hat{\phi},\boldsymbol{\gamma}=\hat{\boldsymbol{\gamma}}} \,,$$

where $\boldsymbol{\mu} = g^{-1}(X\boldsymbol{\gamma})$ is the mean of $\mathbf{Y}$ under $H_0:\tau = 0$; $\Delta = \text{diag}\{g'(\mu_i)\}$; $W = \text{diag}\{w_i\}$, with $w_i^{-1} = \phi m_i^{-1} v(\mu_i)\{g'(\mu_i)\}^2$; $P_0 = W - WX(X^TWX)^{-1}X^TW$; $\hat{\boldsymbol{\gamma}}$ is the maximum-likelihood estimate of $\boldsymbol{\gamma}$ under H_0; and $\hat{\phi}$ is the REML type of estimate of ϕ under H_0.

Under $H_0:\tau = 0$, $\hat{\boldsymbol{\gamma}}$ and $\hat{\phi}$ are consistent estimates of $\boldsymbol{\gamma}$ and ϕ, respectively. Hence, there is not much variability in the second term (relative to the first term) of U_τ when sample size n is large. Therefore, we can again use the first term for testing $H_0:\tau = 0$:

$$T_\tau \equiv \frac{1}{2}(\mathbf{Y} - \boldsymbol{\mu})^T \Delta WSW \Delta (\mathbf{Y} - \boldsymbol{\mu})\Big|_{\tau=0,\phi=\hat{\phi},\boldsymbol{\gamma}=\hat{\boldsymbol{\gamma}}} \,.$$

To derive the distribution of T_τ, first we show, through Taylor expansion, that, under $H_0:\tau = 0$,

$$\hat{\boldsymbol{\mu}} = g^{-1}(X\hat{\boldsymbol{\gamma}}) \approx \boldsymbol{\mu} + \Delta^{-1}X(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})$$

and

$$(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \approx (X^TWX)^{-1}X^TW\Delta(\mathbf{Y} - \boldsymbol{\mu}) \,.$$

Then, we have

$$W\Delta(\mathbf{Y} - \hat{\boldsymbol{\mu}}) = [W\Delta - W(X^TWX)^{-1}X^TW\Delta](\mathbf{Y} - \boldsymbol{\mu})$$

$$= P_0\Delta(\mathbf{Y} - \boldsymbol{\mu}) \,.$$

As a result,

$$T_\tau = \frac{1}{2}(\mathbf{Y} - \hat{\boldsymbol{\mu}})^T\Delta WSW\Delta(\mathbf{Y} - \hat{\boldsymbol{\mu}}) \approx \frac{1}{2}(\mathbf{Y} - \boldsymbol{\mu})^T\Delta P_0 SP_0\Delta(\mathbf{Y} - \boldsymbol{\mu})$$

$$= \frac{1}{2}\tilde{\mathbf{Y}}^TW^{-\frac{1}{2}}P_0SP_0W^{-\frac{1}{2}}\tilde{\mathbf{Y}} \,,$$

where the ith element of $\tilde{\mathbf{Y}}$ is defined by $(Y_i - \mu_i)/\sqrt{\text{Var}(Y_i)}$ and where μ_i and $\text{Var}(Y_i)$ are the true mean and variance, respectively, of Y_i under H_0. The result indicates that T_τ has approximately the same distribution as

$$\frac{1}{2}\tilde{\mathbf{Y}}^TW^{-\frac{1}{2}}P_0SP_0W^{-\frac{1}{2}}\tilde{\mathbf{Y}} \,.$$

Again, denote by $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_c > 0$ $(c \leqslant L)$ the eigenvalues of $W^{-1/2}P_0SP_0W^{-1/2}/2$, and denote by $u_1, u_2, \ldots, u_c$ the corresponding orthonormal eigenvectors. Then, under H_0, T_τ will have approximately the same distribution as $\sum_{j=1}^c \lambda_j Z_j^2$, where $Z_j = u_j^T\tilde{\mathbf{Y}}$. Under the condition that each u_j is not dominated by a few elements, $Z_1, Z_2, \cdots, Z_c$ will be approximately independent standard normal random variables. So, T_τ has approximately the same distribution as that of $\sum_{j=1}^c \lambda_j \chi_{1(j)}^2$, which is similar to the case of normal traits.

We can approximate the distribution of T_τ by a Gamma distribution by matching the first two moments, as was done in appendix A. The approximate mean of T_τ is given by $\mathcal{E} = \text{tr}(P_0S)/2$, and the approximate variance by $\mathcal{V} = \hat{I}_{\tau\tau} - \hat{I}_{\tau\phi}^2/\hat{I}_{\phi\phi}$, where $\hat{I}_{\tau\tau} = \text{tr}(P_0SP_0S)/2$, $\hat{I}_{\tau\phi} = \text{tr}(P_0SP_0W^{-1})/(2\hat{\phi}) = \text{tr}(P_0S)/(2\hat{\phi})$, and $\hat{I}_{\phi\phi} = \text{tr}(P_0W^{-1})/(2\hat{\phi}^2) = (n-K)/(2\hat{\phi}^2)$.

Web Resources

The URLs for data presented herein are as follows:

Authors' Web site, http://www4.stat.ncsu.edu/~tzeng/Softwares/Hap-VC/R/ (for R code for implementing the VC-score test)
NINDS Human Genetics DNA and Cell Line Repository, http://ccr.coriell.org/ninds/ (for the ALS study by Schymick et al.)
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for *DOCK1*)

References

1. The International HapMap Consortium (2003) The International HapMap Project. Nature 426:789–796
2. Akey J, Jin L, Xiong M (2001) Haplotypes vs single marker linkage disequilibrium tests: what do we gain? Eur J Hum Genet 9:291–300
3. Pe'er I, de Bakker PI, Maller J, Yelensky R, Altshuler D, Daly MJ (2006) Evaluating and improving power in whole-genome association studies using fixed marker sets. Nat Genet 38:663–667
4. Zaitlen N, Kang HM, Eskin E, Halperin E (2007) Leveraging the HapMap correlation structure in association studies. Am J Hum Genet 80:683–691
5. Seltman H, Roeder K, Devlin B (2003) Evolutionary-based association analysis using haplotype data. Genet Epidemiol 25:48–58
6. Durrant C, Zondervan KT, Cardon LR, Hunt S, Deloukas P, Morris AP (2004) Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. Am J Hum Genet 75:35–43
7. Tzeng JY (2005) Evolutionary-based grouping of haplotypes in association analysis. Genet Epidemiol 28:220–231
8. Tzeng JY, Wang CH, Kao JT, Hsiao CK (2006) Regression-based association analysis with clustered haplotypes through use of genotypes. Am J Hum Genet 78:231–242
9. McPeek MS, Strahs A (1999) Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. Am J Hum Genet 65:858–875
10. der Meulen MAV, te Meerman GJ (1997) Haplotype sharing analysis in affected individuals from nuclear families with at least one affected offspring. Genet Epidemiol 14:915–920
11. Tzeng JY, Devlin B, Wasserman L, Roeder K (2003) On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. Am J Hum Genet 72:891–902
12. Thomas DC, Morrison JL, Clayton DG (2001) Bayes estimates of haplotype effects. Genet Epidemiol Suppl 1 21:S712–S717
13. Molitor J, Marjoram P, Thomas D (2003) Application of Bayesian spatial statistical methods to analysis of haplotypes effects and gene mapping. Genet Epidemiol 25:95–105
14. Schaid DJ (2004) Evaluating associations of haplotypes with traits. Genet Epidemiol 27:348–364
15. Self SG, Liang KY (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J Am Stat Assoc 82:605–610
16. Crainiceanu CM, Ruppert D (2004) Likelihood ratio tests in linear mixed models with one variance component. J R Statist Soc B 66:165–185
17. Schymick JC, Scholz SW, Fung HC, Britton A, Arepalli S, Gibbs JR, Lombardo F, Matarin M, Kasperaviciute D, Hernandez DG, et al (2007) Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: first stage analysis and public release of data. Lancet Neurol 6:322–328
18. Carlin B, Louis T (2000) Bayes and empirical Bayes methods for data analysis. 2nd ed. Chapman & Hall, New York, p 262
19. Searle S, Casella G, McCulloch C (1992) Variance components. Wiley, New York, pp 232–257
20. Harville D (1977) Maximum likelihood approaches to variance component estimation and related problems. J Am Stat Assoc 72:322–340
21. Lin X (1997) Variance component testing in generalized linear models with random effects. Biometrika 84:309–326
22. Zhang D, Lin X (2003) Hypothesis testing in semiparametric additive mixed models. Biostatistics 4:57–74
23. Roeder K, Bacanu SA, Sonpar V, Zhang X, Devlin B (2005) Analysis of single-locus tests to detect gene-disease associations. Genet Epidemiol 28:207–219
24. Wall JD, Pritchard JK (2003) Assessing the performance of the haplotype block model of linkage disequilibrium. Am J Hum Genet 73:502–515
25. Lake SL, Lyon H, Tantisira K, Silverman EK, Weiss ST, Laird NM, Schaid DJ (2003) Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. Hum Hered 55:56–65
26. Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 70:425–434
27. Simon-Sanchez J, Scholz S, Fung H, Matarin M, Hernandez D, Gibbs J, Britton A, Wavrant de Vrieze F, Peckham E, Gwinn-Hardy K, et al (2007) Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. Hum Mol Genet 16:1–14
28. Imhof J (1961) Computing the distribution of quadratic forms in normal variables. Biometrika 48:419–426
29. Allen A, Satten G (2007) Statistical models for haplotype sharing in case-parent trio data. Hum Hered 64:35–44