Biometrics 65, 822–832
September 2009

DOI: 10.1111/j.1541-0420.2008.01176.x

Gene-Trait Similarity Regression for Multimarker-Based Association Analysis

Jung-Ying Tzeng,[1,*] Daowen Zhang,[1] Sheng-Mao Chang,[2] Duncan C. Thomas,[3] and Marie Davidian[1]

[1]Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, U.S.A.
[2]Department of Statistics, National Cheng Kung University, Tainan, Taiwan
[3] Department of Preventive Medicine, University of Southern California, Los Angeles, California 90089, U.S.A.
[*]*email:* jytzeng@stat.ncsu.edu

Summary. We propose a similarity-based regression method to detect associations between traits and multimarker genotypes. The model regresses similarity in traits for pairs of "unrelated" individuals on their haplotype similarities, and detects the significance by a score test for which the limiting distribution is derived. The proposed method allows for covariates, uses phase-independent similarity measures to bypass the needs to impute phase information, and is applicable to traits of general types (e.g., quantitative and qualitative traits). We also show that the gene-trait similarity regression is closely connected with random effects haplotype analysis, although commonly they are considered as separate modeling tools. This connection unites the classic haplotype sharing methods with the variance-component approaches, which enables direct derivation of analytical properties of the sharing statistics even when the similarity regression model becomes analytically challenging.

Key words: Haplotype-based association test; Haplotype sharing; Haplotype similarity.

1. Introduction

A haplotype is an ordered allele sequence of multiple markers on the same chromosome. Compared to a single marker, a haplotype may better capture genetic variations because it incorporates information from multiple markers simultaneously and conserves the joint linkage disequilibrium (LD) structure among them (The International HapMap Consortium, 2003; de Bakker et al., 2005). However, modeling haplotype variations tends to require a large number of parameters or run into sparsity problems, as the majority of the polymorphism is concentrated on a relatively small number of haplotypes whereas the rest is sparsely spread over a fair number of categories. The existence of the many categories and sparse categories often leads to either power loss or unstable statistical inference, and consequently, limits the efficiency of haplotype-based association analysis in practice.

Modeling haplotype similarity instead of haplotype variants have been frequently considered as a tool to bypass the problems caused by many haplotypes and rare haplotypes. To detect the potential regions that may contain functional variants, approaches based on haplotype similarity look for unusual sharing of chromosomal segments within homogeneous trait groups (Houwen et al., 1994; McPeek and Strahs, 1999). The underlying rationale is that affected individuals tend to share similar genetic materials in close vicinity to the disease mutation. Even with complex diseases that exhibit greater etiologic heterogeneity, one can still expect to find disproportionately large clusters of cases sharingcommon haplotypes in the region flanking a disease mutation (Puffenberger et al., 1994; Feder et al., 1996).

The excess similarity among cases can be identified by either comparing to the similarity level expected from genealogical process (Durham and Feingold, 1997; Service et al., 1999), or by contrasting with control haplotypes (Van der Meulen and Te Meerman, 1997; Bourugain et al., 2000, 2001, 2002; Tzeng, Byerley, et al., 2003; Tzeng, Devlin, et al., 2003; Yu et al., 2004). The former approaches were used in the successful positional cloning of Mendelian disorders since the 1980s. However, extension of these methods to complex diseases makes comparisons to the genealogical expectations less practical. The latter approaches bypass the need to model the evolutionary process by which the observed haplotypes were produced. However, these methods tend to be feasible only with binary traits and do not incorporate covariate information. Moreover, these methods limit similarity calculations to the concordant samples only (i.e., case–case similarity and control–control similarity) and do not use information obtained from case–control similarity. Recently, Sha, Chen, and Zhang (2007) addressed the latter concern by contrasting similarity of concordant pairs (case–case and control–control) with similarity of discordant pairs (case–control), and showed that such contrast can improve power.

Current developments have shifted the focus from two-sample tests to regression models that correlate trait similarity withgenetic similarity (Beckmann et al., 2005; Wessel and Schork, 2006). This new direction incorporates similarity comparison between discordant pairs, and establishes a model-based framework that is ready for accommodating covariates and various trait types. The idea of gene-trait similarity was pioneered by Qian and Thomas (2001) with pedigree data.

Qian and Thomas (2001) quantified the similarities of phenotypes and of haplotypes within each family, and correlated these family statistics using the Mantel statistics. Beckmann et al. (2005) extended the framework for population-based samples. Although not accounting for covariate information yet, their methods can work on qualitative and quantitative traits. Wessel and Schork (2006) took one step further and developed a general regression framework for dissimilarity analysis between phenotypes and genotypes. Their model treats genetic similarity as the response variable, and treats trait similarity and environmental covariates as explanatory variables. However, because covariates tend to affect the disease risk rather than the genetic variants, it would be more desirable to switch the roles of genetic similarity and trait similarity. Finally, one major challenge of the gene-trait similarity approaches is the complex correlation structure introduced by the pairwise samples, as the observation unit in the regression is now pairs of individuals instead of single individuals. Consequently, the distributions of the test statistics are hard to derive analytically, and permutation is needed to find p-values.

Our proposed similarity method also follows this direction. For trait similarity, we measure the trait covariance of all distinct individual pairs conditional on covariates. For haplotype similarity, we measure the sharing level of haplotype pairs of two individuals. We then propose to regress trait similarity on haplotype similarity, and study gene-trait association by testing for zero regression coefficient of haplotype similarity. In Section 2, we formulate the gene-trait similarity regression model, construct a score test for association, and derive its limiting distribution to facilitate hypothesis testing in large scale. To tackle the issue of missing phases, we use the similarity metrics that can measure haplotype similarity directly from genotypes. In Section 3, we show that the similarity regression is closely connected to an alternative haplotype analysis approach, the variance-component (VC) method. In Section 4, we investigate the performance of the proposed method using simulations. In Section 5, we apply it to the case–control data obtained from the amyotrophic lateral sclerosis study of Schymick et al. (2007). Finally we conclude with discussions and remarks in Section 6.

2. Methods

2.1 *The Gene-Trait Similarity Model*

Let Y_i denote the trait value, X_i denote the $K \times 1$ covariate vector including the intercept term, and H_i denote the $L \times 1$ haplotype vector of the ith individual in a sample of n subjects. The hth element of H_i, denoted by $H_{i,h}$, records the number of copies of haplotype h that subject i carries.

For genetic similarity, define S_{ij} to be the haplotype similarity between subjects i and j $(i \neq j)$. From the definition of H, we have

$$S_{ij} = \sum_{h,k} H_{i,h} H_{j,k} \times s(h,k), \tag{1}$$

where $s(h, k)$ is a certain similarity metric that is used to measure the similarity level between haplotypes h and k. Let $\mu_i^0 = \mathrm{E}(Y_i \mid X_i, H_i) = \delta(X_i^T\gamma)$ under the condition of no haplotype effects, in which γ is the covariate effects including the intercept. We assume that the conditional mean can be modeled by some specific function of $X_i^T\gamma$ denoted by δ, such as that specified by the generalized linear model. Then for trait similarity, which is denoted by Z_{ij}, we define

$$Z_{ij} = \left\{\omega_i \left(Y_i - \mu_i^0\right)\right\}\left\{\omega_j \left(Y_j - \mu_j^0\right)\right\}, \tag{2}$$

which is the weighted cross product of the trait residuals with some weight ω_i. The cross product of residuals has been used to describe the level of trait similarity between a pair of subjects in linkage studies (Thomas et al., 1999; Elston et al., 2000). Here the residual is defined with respect to the covariate-adjusted mean for each subject. The weight ω_i may be used to account for the fact that Y_i is not necessarily homogeneous. In principle, ω_i can be any prespecified positive value such as 1, or some function of the trait variance. As we will illustrate later, optimal ω_i can be identified if a model is imposed on trait values and haplotype effects.

We propose a similarity regression model of the following form to study and test the gene-trait association:

$$Z_{ij} = b \times S_{ij} + e_{ij}, \quad \forall i < j, \tag{3}$$

where e_{ij}'s are some mean-zero error terms. By the definition of Z_{ij} (which has been adjusted for the effects of baseline and other covariates), the proposed regression has a zero intercept. Intuitively, in a chromosomal region that contains disease genes, one would expect that $b > 0$ as higher genetic similarity would lead to higher trait similarity. In "null" regions, $b \approx 0$ as genetic similarity and trait similarity would have little correlation.

Similar to Elston et al. (2000), one may estimate b in equation (3) using the least-square method and conduct inference accordingly. However, this approach is complicated by the need to invert a large variance–covariance matrix of e_{ij}'s. To bypass this issue, in this work we focus on the score test of $b = 0$, where the inverse of the covariance matrix takes place only under the null hypothesis and the covariance matrix becomes diagonal. In general, the estimation of b can be accomplished by using the VC estimation in a generalized linear mixed model (GLMM) if such model is posited on trait values.

2.2 *The Score Test of* $H_0 : b = 0$

To construct the score test for testing $H_0 : b = 0$, we further assume that $\mathrm{v}_i^0 = \mathrm{var}(Y_i \mid X_i, H_i) = m_i^{-1}\phi v(\mu_i^0)$ under the condition of no haplotype effects, where m_i is a known prior weight, such as the binomial denominator, ϕ is the dispersion parameter, and $v(\mu_i)$ is the variance function. With the two moment restrictions on μ_i^0 and v_i^0 in our model, we use the following estimating equations for (b, γ, ϕ) to construct the score test for testing $H_0 : b = 0$:

$$U = \begin{bmatrix} U_b \\ U_\gamma \\ U_\phi \end{bmatrix} = -\mathrm{E}\left[\begin{array}{c} \dfrac{\partial \{Z - \mathrm{E}(Z \mid X, H)\}}{\partial (b, \gamma^T, \phi)} \\ \dfrac{\partial \{Y - \mathrm{E}(Y \mid X, H)\}}{\partial (b, \gamma^T, \phi)} \end{array} \right]^T \times \begin{bmatrix} \mathrm{var}(Z \mid X, H) & \mathrm{cov}(Z, Y \mid X, H) \\ \mathrm{cov}(Y, Z \mid X, H) & \mathrm{var}(Y \mid X, H) \end{bmatrix}^{-1} \times \begin{bmatrix} Z - \mathrm{E}(Z \mid X, H) \\ Y - \mathrm{E}(Y \mid X, H) \end{bmatrix} = 0.$$

Assume that Y_i's are independent under the null hypothesis of no haplotype effect (i.e., $b = 0$). An intuitive explanation of this independence assumption is that, under H_0, the trait dependence should reflect the baseline relationship between the pair of the subjects. Here such relationship is zero because the study subjects are unrelated. This assumption will be more rigorously justified after we postulate a model on haplotype effects (Section 3).

Following the above null independence assumption of Y_i's and by the definition of variance, we have that under H_0, cov $(Z_{ij}, Y_l \mid X, H) = 0\ \forall i \neq j$ and $\forall l$, var $(Y \mid X, H) = V = \mathrm{diag}\{\mathrm{v}_i^0\}$, and var $(Z \mid X, H) = \mathrm{diag}\{\omega_i^2 \omega_j^2 \mathrm{v}_i^0 \mathrm{v}_j^0\}$. Thus U_b, the score statistic for b, is

$$\begin{aligned} U_b &= \sum_{i<j} S_{ij} \omega_i^{-2} \omega_j^{-2} \left(\mathrm{v}_i^0\right)^{-1} \left(\mathrm{v}_j^0\right)^{-1} Z_{ij} \Big|_{b=0, \gamma=\hat{\gamma}, \phi=\hat{\phi}} \\ &= \sum_{i<j} S_{ij} \omega_i^{-1} \omega_j^{-1} \left(\mathrm{v}_i^0\right)^{-1} \left(\mathrm{v}_j^0\right)^{-1} \left(Y_i - \widehat{\mu}_i^0\right) \left(Y_j - \widehat{\mu}_j^0\right), \end{aligned}$$

where $\widehat{\gamma}$ is the maximum likelihood estimate of γ under H_0, $\widehat{\phi}$ is the restricted maximum likelihood type of estimate of ϕ under H_0, and $\widehat{\mu}^0 = \delta(X_i^T \widehat{\gamma})$. To derive its distribution, we rewrite U_b in a quadratic form as

$$U_b = \frac{1}{2}(Y - \widehat{\mu}^0)^T V^{-1} \Omega^{-1} S_0 \Omega^{-1} V^{-1} (Y - \widehat{\mu}^0). \tag{4}$$

In the above equation, S_0 is a matrix with diagonal elements equal to 0 and off-diagonal elements equal to S_{ij}, and $\Omega = \mathrm{diag}\{\omega_i\}$. We show in Web Appendix A that equation (4) has approximately the same distribution as the weighted chi-squared random variable $\sum_{i=1}^n \lambda_i \chi_{1,i}^2$, where $\chi_{1,i}^2$'s are independent chi-squared variables with 1 degree of freedom, and $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$ are the ordered eigenvalues of the matrix C defined as

$$C = \frac{1}{2} V^{\frac{1}{2}} Q \Omega^{-1} S_0 \Omega^{-1} Q V^{\frac{1}{2}},$$

where Q is the projection matrix $V^{-1} - V^{-1}\ DX(X^T DV^{-1}\ DX)^{-1}\ X^T DV^{-1}$ under H_0 with matrix $D = \mathrm{diag}\{\partial\delta\ (\eta_i)/\partial\eta_i\}$ and $\eta_i = X_i^T \gamma$. With this result, we approximate the distribution of U_b by the three-moment approximation method of Imhof (1961) as did Allen and Satten (2007). The level-α significance threshold is estimated by

$$\kappa_1 + (\chi_\alpha - h') \times \sqrt{\frac{\kappa_2}{h'}},$$

where $\kappa_j = \sum_i \lambda_i^j$, $h' = \kappa_2^3/\kappa_3^2$ and χ_α is the αth quantile of $\chi_{h'}^2$ (i.e., chi-squared distribution with h' degrees of freedom). Alternatively, one can report the p-value of the observed statistic U_b by $P(\chi_{h'}^2 > \chi^*)$, where $\chi^* = (U_b - \kappa_1) \times \sqrt{h'/\kappa_2} + h'$.

2.3 *Analysis with Unphased Genotype Data*

To bypass the issues involved in phase missing and phase imputation, we use measures that do not require phase information to quantify haplotype similarity. One possible choice of such $s(h, k)$ is the "counting measure," $s_{count}(h, k)$, of Tzeng, Devlin, et al. (2003), which calculates the proportion of alleles in common between haplotype h and haplotype k. Define $G_{m,i}$ to be a 2×1 vector whose elements record the number of major alleles and the number of minor alleles for subject i at marker $m, m = 1, 2, \ldots, M$. It can be shown that

$$S_{ij} = \sum_{h,k} H_{i,h} H_{j,k} \times s_{count}(h, k) = \frac{1}{M} \sum_{m=1}^{M} G_{m,i}^T G_{m,j}.$$

That is, haplotype similarity score between subjects i and j is equal to the average allelic sharing across markers. With this transformation, it is possible to bypass the need of imputing haplotype phase, because all we need to measure haplotype similarity is the allele counts at each single nucleotide polymorphis (SNP).

Concerns may arise in terms of choosing between phase-dependent and phase-independent metrics for performing haplotype-similarity analysis. In theory, metrics using the phase information should be more powerful as they may capture the identical-by-descent sharing more precisely. However, these metrics are not robust to practical complications such as genotyping errors and recent marker mutations that often limit their performance in reality. Indeed, previous and recent works have found that phase-dependent and phase-independent metrics have very similar performance (Tzeng, Devlin, et al., 2003; Sha et al., 2007).

3. Connection with the Variance-Component Score Test

Similar to the linkage analysis where the regression-based approaches have been shown equivalent to the VC methods (Sham and Purcell, 2001), the proposed gene-trait similarity regression is also connected with the VC approaches of haplotype analysis. The connection can be obtained by the fact that $\mathrm{E}(Z_{ij} \mid X, H) \approx \omega_i \omega_j \times \mathrm{cov}(Y_i, Y_j \mid X, H)$. To see this, consider a GLMM:

$$\begin{aligned} g(\mu_i) &= X_i^T \gamma + H_i^T \beta \\ \beta &\sim MVN(0, \tau R_\beta), \end{aligned} \tag{5}$$

where $\mu_i = \mathrm{E}(Y_i \mid \beta, X_i, H_i)$, $g(\cdot)$ is the link function, γ is the fixed effect of environmental covariates, and β is the random effect of the haplotypes. Conditional on X_i, H_i, and β, Y_i's are assumed to be independent with conditional mean μ_i and conditional variance $\mathrm{var}(Y_i \mid \beta, X_i, H_i) = \mathrm{v}_i = m_i^{-1}\ \phi\ v(\mu_i)$. Under this model, $\mu_i = g^{-1}(X_i^T\ \gamma + H_i^T\ \beta)$ and $\mu_i^0 = g^{-1}(X_i^T\ \gamma)$, and under $H_0, \mu_i = \mu_i^0$ and $\mathrm{v}_i = \mathrm{v}_i^0$. Matrix $R_\beta = \{r_{hk}\}$ describes the correlation between effects of haplotypes h and k. One common correlation structure imposed on β is to allow evolutionarily close haplotypes to be more correlated, such as to set $r_{hk} = s(h, k)$ with r_{hk} being the (h, k) element of matrix R_β, and $s(h, k)$ $(0 \leqslant s(h, k) \leqslant 1)$ being the similarity

metric that measures the similarity level between haplotypes h and k (Schaid, 2004; Tzeng and Zhang, 2007).

Let $g'(\mu) = \partial g(\mu)/\partial\mu$. Then by Taylor expansion on the mean function $\mu_i = g^{-1}(X_i^T\gamma + H_i^T\beta)$ with respect to β around $\mathrm{E}(\beta) = 0$, we have

$$\begin{aligned}\mathrm{E}(Y_i \mid X_i, H_i) &= \mathrm{E}_\beta(\mu_i \mid X_i, H_i)\\ &\approx \mathrm{E}_\beta\left\{\mu_i^0 + \left[g'\left(\mu_i^0\right)\right]^{-1} H_i^T\beta \mid X_i, H_i\right\} = \mu_i^0.\end{aligned}$$

Therefore

$$\begin{aligned}\mathrm{cov}(Y_i, Y_j \mid X, H) &= \mathrm{E}[\{Y_i - \mathrm{E}(Y_i \mid X_i, H_i)\}\\ &\quad \times \{Y_j - \mathrm{E}(Y_j \mid X_j, H_j)\} \mid X, H]\\ &\approx \mathrm{E}\left\{\left(Y_i - \mu_i^0\right)\left(Y_j - \mu_j^0\right) \mid X, H\right\}.\end{aligned}$$

Consequently, the expected trait similarity is

$$\begin{aligned}\mathrm{E}(Z_{ij} \mid X, H) &= \mathrm{E}\left[\left\{\omega_i\left(Y_i - \mu_i^0\right)\right\}\left\{\omega_j\left(Y_j - \mu_j^0\right)\right\} \middle| X, H\right]\\ &\approx \omega_i\omega_j \times \mathrm{cov}\left(Y_i, Y_j \mid X, H\right)\\ &\approx \omega_i\omega_j \times \left\{g'\left(\mu_i^0\right)g'\left(\mu_j^0\right)\right\}^{-1}\\ &\quad \times \tau\sum_{h,k} H_{i,h}H_{j,k}r_{hk} \quad \text{(Web Appendix B)} \qquad (6)\\ &= \omega_i\omega_j \times \left\{g'\left(\mu_i^0\right)g'\left(\mu_j^0\right)\right\}^{-1}\\ &\quad \times \tau S_{ij}, \quad \text{if } r_{hk} = s\left(h,k\right). \qquad (7)\end{aligned}$$

The result indicates that trait similarity is (approximately) in a linear relationship with haplotype similarity if we choose $\omega_i = g'\left(\mu_i^0\right)$. (In that case, $\mathrm{E}(Z_{ij} \mid X, H) \approx \tau\, S_{ij}$.) For the canonical link $g(\cdot)$, this choice is equivalent to $\omega_i = 1/v(\mu_i^0)$. Comparing equation (7) to equation (3), we see that testing $b = 0$ in the similarity regression is the same as testing for $\tau = 0$ in a VC model. This implies that under the null hypothesis of $b = 0, \tau$ is zero and hence Y_i's are independent. The connection also suggests that the test of $H_0 : b = 0$ should be one-sided, and that the gene-trait regression model (3) can use a zero intercept.

Recently Tzeng and Zhang (2007) constructed the VC score test for testing $H_0 : \tau = 0$ based on the GLMM (5) as

$$T_\tau = \frac{1}{2}(Y - \widehat{\mu}^0)^T\,\Delta\, WSW\Delta(Y - \widehat{\mu}^0),$$

where matrix $S = \{S_{ij}\}, W = \mathrm{diag}\ \{w_i\}$ with $w_i = [m_i^{-1}\ \phi v(\mu_i)\ \{g'(\mu_i)\}^2]^{-1}$, and $\Delta = \mathrm{diag}\ \{g'(\mu_i)\}$. Notice that under GLMM (5), U_b of equation (4) becomes

$$U_b = \frac{1}{2}(Y - \widehat{\mu}^0)\Delta W\Delta\Omega^{-1}S_0\Omega^{-1}\Delta W\Delta(Y - \widehat{\mu}^0), \qquad (8)$$

as $V^{-1} = \Delta\, W\Delta$. Comparing T_τ to U_b in equation (8), we notice that both statistics incorporate genetic information solely through the form of haplotype similarity (i.e., S or S_0). We also note that both statistics share analogous quadratic forms; the forms are almost identical if $\omega_i = g'(\mu_i)$ is used, in which case $U_b = \frac{1}{2}(Y - \widehat{\mu}^0)^T\Delta WS_0W\Delta(Y - \widehat{\mu}^0)$. Thus while haplotype sharing (i.e., to detect unusual sharing of haplotypes among homogeneous trait groups) and haplotype smoothing (i.e., to smooth the haplotype effects by introducing correlation structure on similar haplotypes) are commonly considered as separate modeling strategies in haplotype analysis, they are unified through the framework of similarity regression and random effects haplotype analysis. With this comparison, we also learn the subtle difference between the two is that U_b of similarity regression uses information from between-individual comparison (i.e., $i \neq j$), whereas T_τ of VC includes the comparison of between and within individuals. Relatively speaking, the amount of information contributed by the comparison of the two haplotypes within a person is small (relative to the between-individual comparison) because two out of the four comparisons are self-comparisons and hence not informative. As a result, although more data information is utilized in the VC test, we expect a similar performance of the two approaches in detecting haplotype–phenotype association.

4. Simulation Studies

We performed simulation studies to investigate the behaviors of the proposed gene-trait similarity regression. We follow the same simulation scheme as Tzeng and Zhang (2007), where a coalescent process is first used to generate the SNP sequences, and then a causal SNP (rather than causal haplotypes or haplotype-similarity levels) is used to determine the trait values. Specifically, we implemented the coalescent program of Wall and Pritchard (2003) to generate SNP sequences using the following parameters: an effective population size of 10^4, a scaled mutation rate of 5.6×10^{-4} (per bp), and a scaled recombination rate around 6×10^{-3} (per bp) for the cold spots and 45 times greater for the hot spots. These parameters are chosen to produce a similar number of common SNPs to the European American sample in the SeattleSNP database and to mimic the linkage disequilibrium pattern of the SELP gene observed in it. A total number of 100 sequences were generated from this model. We selected certain SNPs as the disease loci and form a haplotype region by including the two SNPs to its left and the three SNPs to its right. The disease SNP is selected based on the disease allele frequency (0.1 and 0.3) and the pairwise LD pattern between the disease SNP and its flanking SNPs (R^2). We focus mostly on regions with max $R^2 > 0.7$, i.e., at least one of the neighboring SNPs is highly correlated with the unobserved disease SNP. This scenario reflects the common study designs where the disease SNP variation is captured by at least one tagSNPs. In the simulation studies, we further classify the haplotype regions by the level of average R^2, after seeing exploratory results that suggest the relevancy of the average LD level across markers. The average R^2 is obtained by averaging the 5 R^2's of the disease SNP and one of the nearby SNPs.

We next determined the trait value of an individual using the regression model $g(\mu_i) = \gamma_0 + \gamma_1 X_i + \theta \mathrm{G}_i$, where $\mu_i \equiv \mathrm{E}(Y_i \mid X_i, \mathrm{G_i})$ and X_i is a standard normal variable. The genetic value G_i is determined by the genotype of the disease locus (AA, Aa, or aa) and the disease effect (additive, dominant, and recessive). For genotypes AA, Aa, and aa, G_i = (2, 1, 0) for additive effect, (1, 1, 0) for dominant effect, and (1, 0, 0) for recessive effect. We considered two types of traits: binary traits and normal traits. For binary traits, we use the logit link and set $\gamma_0 = -4.5, \gamma_1 = 0$, and $\theta = \log 2$, resulting a disease rate of 1%. For normal traits, we use the identity link and set $\gamma_0 = 0, \gamma_1 = 1$, and $\theta = 1$. We set the trait variance $\mathrm{Var}\ (\mathrm{Y_i} \mid X_i, \mathrm{G_i})$ to be 2 and 4 for allele frequency $q = 0.1$ and $q = 0.3$, respectively. The resulting heritability is about

0.1 under the additive model, 0.06 ∼ 0.08 under the dominant model, and 0.01 ∼ 0.02 under the recessive model. For both traits, we set $\theta = 0$ when evaluating the sizes of the proposed test.

Under each simulation scenario, we used balanced case–control sampling to obtain 100 cases and 100 controls, and used random sampling to obtain 200 individuals with normal traits. We then removed the disease SNP information and converted the remaining haplotype data to unphased genotypes. We performed haplotype association analysis using three approaches: the standard haplotype regression of Schaid et al. (2002) (referred to as HAP method), the gene-trait similarity regression method with $\omega_i = 1/v(\mu_i)$ (referred to as HS method), and the single SNP analysis using 1 degree of freedom (referred to as SNP-1 method) and 2 degrees of freedom (referred to as SNP-2 method). The results of the SNP methods were obtained using the minimum p-value among the five SNPs, whose significant thresholds were determined using the multiple-testing correction method of Moskvina and Schmidt (2008). This method estimates the effect number of independent tests for correlated SNPs at a given overall type I error rate, and estimates the significance level for each individual test accordingly. In the HAP method, haplotypes with frequencies less than the program default threshold (i.e., 5/[2 × no. of individuals]) will be pooled into the category of the reference haplotype.

Type I error rates. Table 1 shows the type I error rates for the binary traits and normal traits at the nominal levels of 0.05 and 0.01. The results are obtained via 10,000 replications. For both trait types, the type I error rates are somehow conservative, especially at the 0.05 nominal level. There could be two possible causes for the differences: (a) the use of the quadratic form $Z^T \Lambda Z$ (see Web Appendix A) to approximate the distribution of U_b, and (b) the use of the three-moment approximation to approximate the distribution of the quadratic form. We ran a simulation to investigate and found that (a) was the more plausible cause. To reach this conclusion, we obtained the empirical p-values, for each replication, by resampling 10^4 U_b's using the fact that $U_b = \sum_{i=1}^{n} \lambda_i \chi^2_{1,i}$. The resulting type I error rates do not differ substantially from the three-moment approximation, which indicates that the three-moment method approximates the distribution of the quadratic form reasonably well. We also examined the type I error rates with an increased sample size of 1000. The resulting type I error rates based on 2000 replications are shown in Table 1. When the sample size increases, the type I error rates approach to the nominal levels. This suggests that the numerical differences we observed are likely due to the loss of accuracy in approximating U_b with $Z^T \Lambda Z$ when using moderate sample sizes. This is possibly caused by the fact that matrix C (and hence Λ) is not positive definite, and the use of large samples can improve the accuracy and alleviate the overconservative situation.

Power. Results of the power comparisons are shown in Figure 1 (for binary traits) and Figure 2 (for normal trait) at the nominal levels 0.05 and 0.01. We considered the scenarios that the true genetic effects are either additive, dominant, or recessive. The data were analyzed with an additive model, with the exception to the SNP-2 method. The power calculations are based on 5000 replications. The regions studied here have max $R^2 > 0.9$.

We first compare the HS method versus the other methods. The general power patterns are similar for both trait types so we can concentrate the comparison on either type. With allele

Table 1
Type I error rates obtained by the three-moment approximation (3-*M*) *with sample size* $n = 200$, *obtained by the resampling methods* (*Resamp*) *with* $n = 200$, *and obtained by the three-moment approximation with* $n = 1000$

		$\alpha = 0.05$			$\alpha = 0.01$		
q	AvgR^2	$n = 200$ 3-M	$n = 200$ Resamp	$n = 1000$ 3-M	$n = 200$ 3-M	$n = 200$ Resamp	$n = 1000$ 3-M
Binary traits							
0.1	0.2	0.042	0.046	0.046	0.012	0.010	0.009
	0.3	0.035	0.039	0.041	0.010	0.009	0.011
	0.4	0.022	0.024	0.050	0.006	0.006	0.014
	0.6	0.040	0.046	0.043	0.011	0.010	0.010
0.3	0.2	0.028	0.032	0.052	0.008	0.008	0.011
	0.3	0.032	0.038	0.043	0.009	0.008	0.013
	0.5	0.043	0.047	0.051	0.008	0.008	0.007
	0.7	0.047	0.050	0.049	0.010	0.010	0.008
Normal traits							
0.1	0.2	0.042	0.046	0.053	0.011	0.009	0.009
	0.3	0.033	0.037	0.044	0.011	0.010	0.009
	0.4	0.021	0.024	0.049	0.007	0.007	0.013
	0.6	0.037	0.041	0.049	0.010	0.009	0.009
0.3	0.2	0.027	0.029	0.044	0.008	0.008	0.008
	0.3	0.036	0.042	0.043	0.009	0.008	0.008
	0.5	0.042	0.047	0.051	0.012	0.012	0.009
	0.7	0.045	0.049	0.045	0.009	0.008	0.010

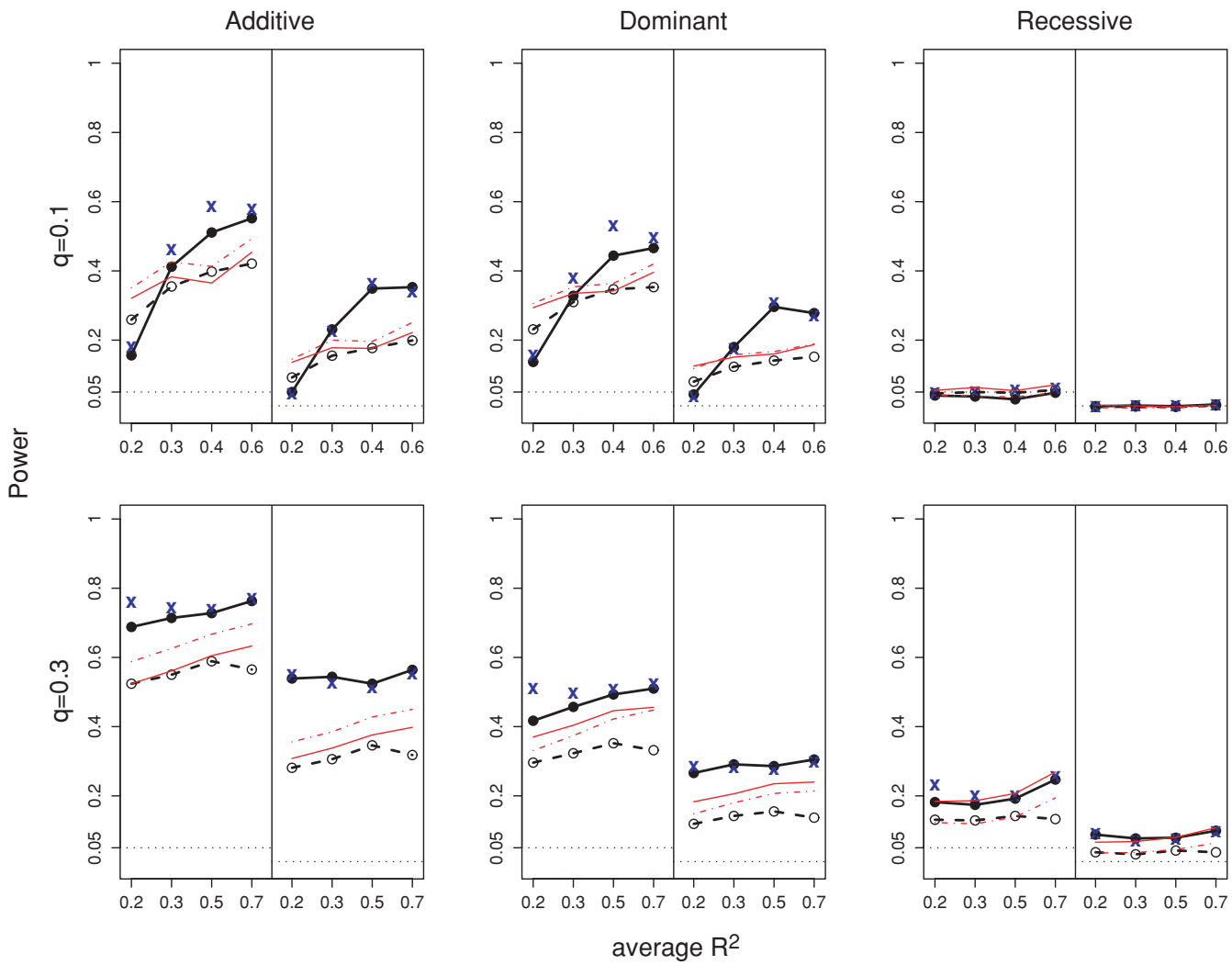

Figure 1. Power comparisons for binary traits. The power is calculated based on 5000 replications. In each plot, the left (right) panels are for nominal level 0.05 (0.01), and the horizontal dotted lines in the bottom indicates the corresponding nominal level. The upper (lower) panel is the results for allele frequency $q = 0.1(0.3)$. Lines with circles are results for haplotype-based analysis, and lines without circles are results for SNP-based analysis. Specifically, the solid lines with filled circles indicate the HS method, the dashed lines with open circles indicate the HAP method, the dotted-dashed lines indicate the SNP-1 method, and the solid lines indicate the SNP-2 method. The "x" signs indicate the VC method.

frequency of 0.1 (the upper panel) and under either additive or dominant effect, the HS method has higher power when the average R^2 is high, but exhibits a power drop when average R^2 is low (i.e., the case where only one SNP is correlated with the disease SNP). Under recessive effect, most of the methods have no power to detect the effect and hence all power is around the nominal level.

When the allele frequency is 0.3 (the lower panel), the HS method exhibits a clear power gain over the other methods (even in cases where the average R^2 is low) for the additive and dominant effect scenarios. Under the recessive effect scenario, the SNP-2 methods have comparable or better power than the HS methods.

In the cases presented here, the SNP method tends to have higher power than the HAP method. This is likely due to the fact that these selected regions contain disease SNPs that are in extremely high correlation with at least one neighboring SNP (i.e., max $R^2 > 0.9$). The SNP method is expected to perform well when one SNP is a perfect surrogate to the disease SNP. However, the standard haplotype analysis, which regresses the trait values on multiple haplotype polymorphisms, must spend extra degrees of freedom on the less or not significant variants and hence becomes less efficient than the SNP method in this case.

We also compared the HS method with the VC method of Tzeng and Zhang (2007), and observed similar power performance for these two approaches. This reconfirms the findings reported in Section 3.

To make sure that the power pattern observed above was not subject to the regions chosen, we ran additional

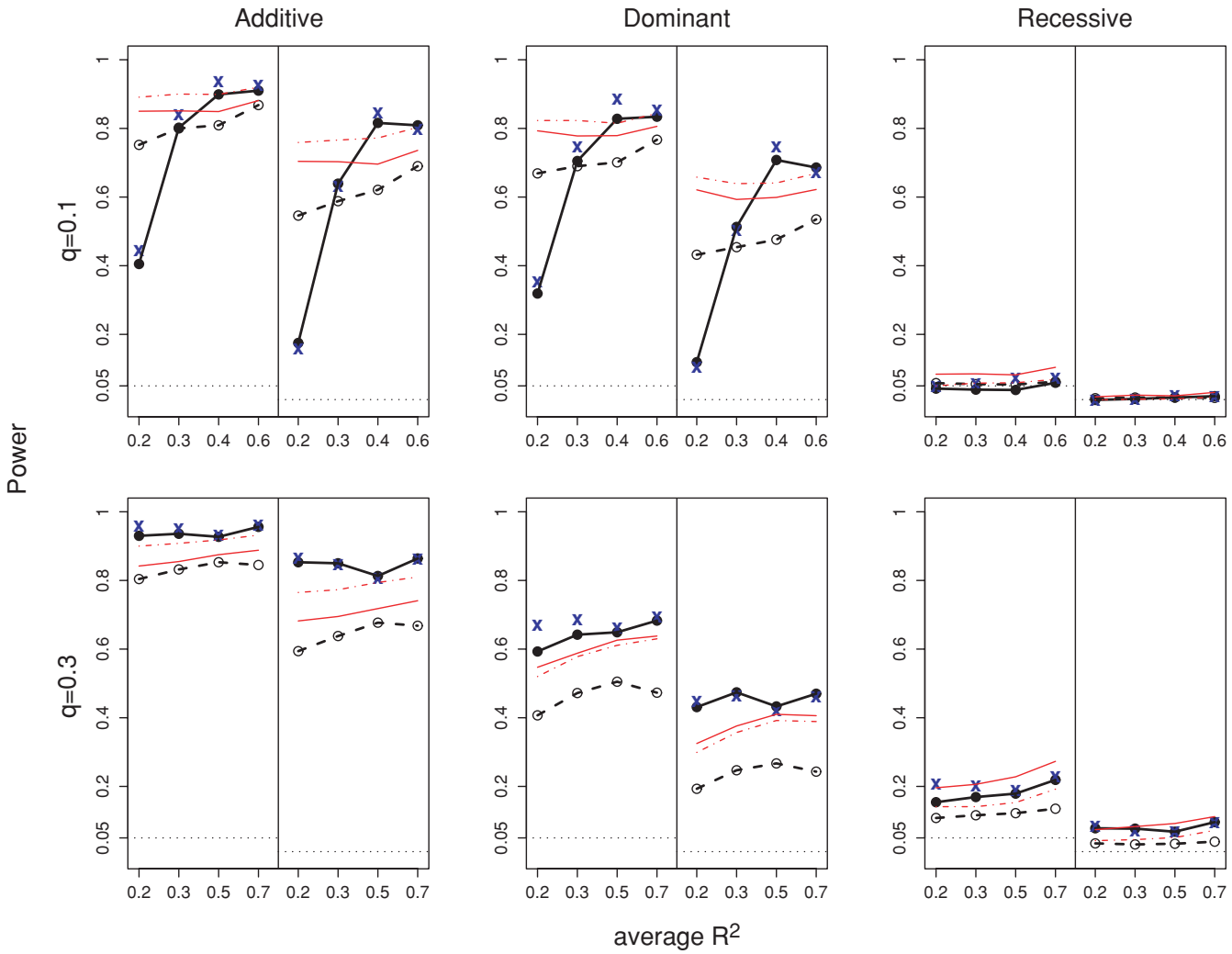

Figure 2. Power comparisons for normal traits. The powers are calculated based on 5000 replications. In each plot, the left (right) panels are for nominal level 0.05 (0.01), and the horizontal dotted lines in the bottom indicates the corresponding nominal level. The upper (lower) panel is the results for allele frequency $q = 0.1(0.3)$. Lines with circles are results for haplotype-based analysis, and lines without circles are results for SNP-based analysis. Specifically, the solid lines with filled circles indicate the HS method, the dashed lines with open circles indicate the HAP method, the dotted-dashed lines indicate the SNP-1 method, and the solid lines indicate the SNP-2 method. The "x" signs indicate the VC method.

simulations to systematically examine the power performance of each method. In this set of simulations, only the binary trait under the additive effect was considered, due to the similarity in results between different trait types and effects. Regions with max $R^2 > 0.7$ were examined and the disease allele frequency was set to be less than 0.4. There are a total of 207 regions with $q \in 0.1 \pm 0.045$, 208 regions with $q \in 0.2 \pm 0.045$, and 121 regions with $q \in 0.3 \pm 0.045$. Figure 3 shows boxplots of power across all regions for each method. The power was obtained based on 1000 replications at the nominal level of 0.01.

We present the power by different ranges of the average R^2: (a) >0.8, (b) (0.6, 0.8], (c) (0.4, 0.6], (d) (0.2, 0.4], and (e) $\leqslant$ 0.2. The scenarios (a) and (e) are considered to be cases at the end of the spectrum whereas the rest are intermediate cases. In scenario (a), all neighboring SNPs are in very high LD with the disease SNP (all $R^2 > 0.8$), and in (e), the disease SNP has high correlation with only 1 nearby SNP and has 0 correlation with the other four SNPs. Intermediate levels of multilocus LD are represented in cases (b) to (d). Given a certain disease allele frequency, we see in Figure 3 that the median power of the HAP method (left boxplots) and the HS method (middle boxplots) increase with average R^2. This trend is expected: when more SNPs start to capture a greater amount of information about the disease locus, the power of those methods that utilize multimarker information should improve as well. We also note that the HS power increases more sharply than the HAP method, suggesting that the HS method could be more sensitive to the changes in the joint LD structure of multiloci. In contrast to the increase pattern

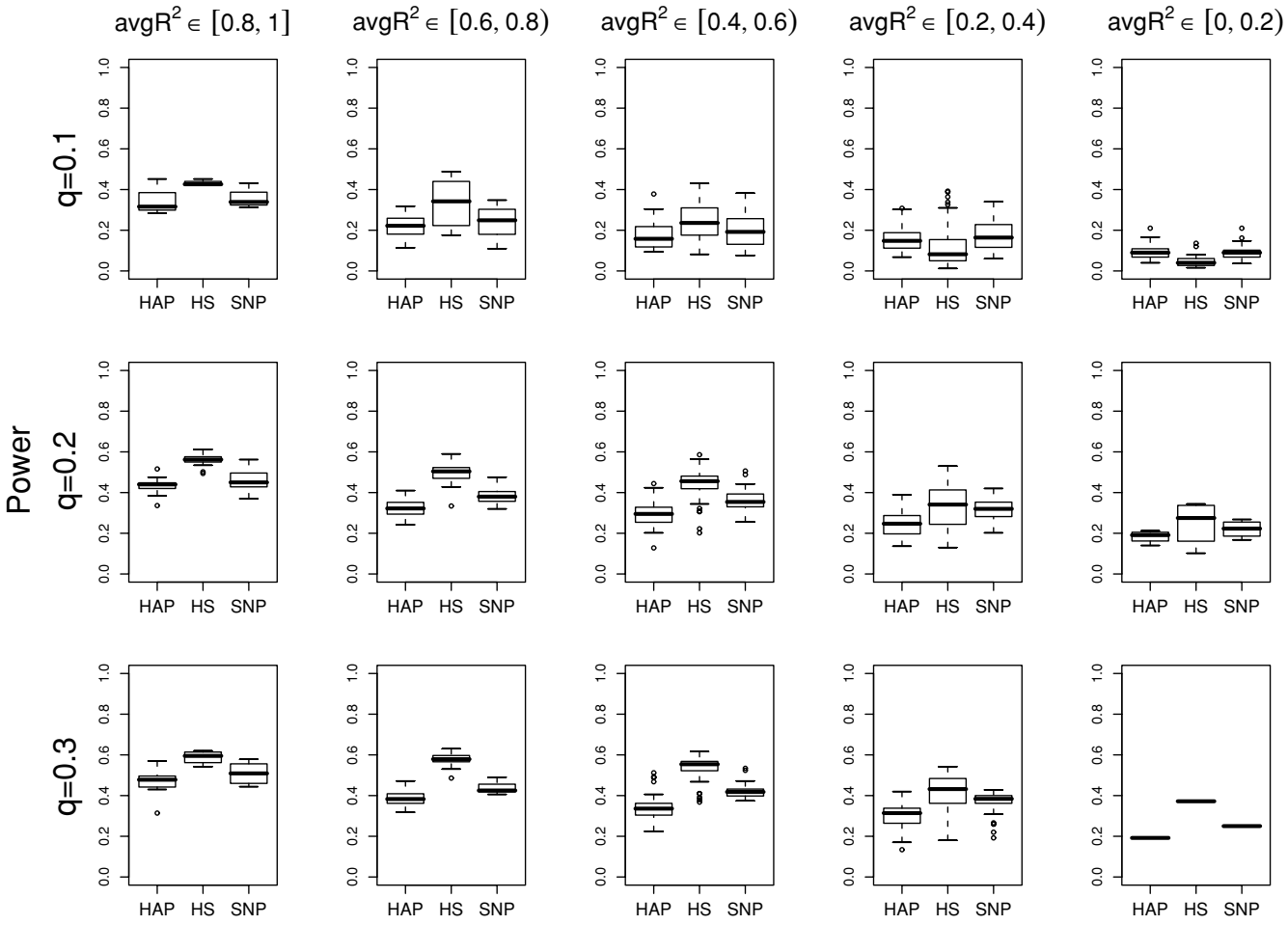

Figure 3. Boxplots of power of the HAP/HS/SNP-1 methods for all simulated genome regions with max $R^2 > 0.7$. The data consist of 100 cases and 100 controls that were generated under an additive model. The power is evaluated based on 1000 replications at nominal level 0.01. The upper/middle/lower panels are the results for allele frequency $q \approx 0.1/0.2/0.3$, respectively.

of the HS and HAP methods, the median power of the SNP method (right boxplots) stays relatively constant.

The boxplots support the general pattern identified in Figures 1 and 2: the HS method mostly has higher power than the HAP method and the SNP-1 method, and the exception is the cases with low average R^2 and low allele frequency. This low power scenario of the HS method can be seen in the top right two panels of Figure 3. We suspect that this is related to the HS method being more sensitive to multilocus LD structure, and are actively investigating the underlying mechanisms.

5. Analysis of the Amyotrophic Lateral Sclerosis (ALS) Study of Schymick et al.

We applied our method to a case–control study of ALS obtained from the National Institute of Neurological Disorders and Stroke (NINDS) Neurogenetics Repository at the Coriell Institute. The ALS study was conducted by Schymick et al. (2007) and one main objective of this study is to identify genetic factors that could contribute in the pathogenesis of sporadic ALS. The study recruited 276 patients with sporadic ALS and 271 neurologically normal controls, and genotyped 555,352 SNPs across the genome. Schymick et al. (2007) performed a genome-wide association analysis and reported the 34 most significant SNPs with p-values less than 0.0001 based on the single SNP tests. Although none of the 34 SNPs was significant after the Bonferroni correction for multiple testing, the most significant SNP (rs4363506) lives in close proximity to the dedicator of cytokinesis 1 gene (DOCK1), which is recognized to play an important role in motorneuron disease.

To illustrate the proposed HS method, we analyzed a portion of the ALS data. We concentrate on chromosome 10, where the most significant SNP is located. Due to ambiguous marker information, we exclude the 26,258th SNP of chromosome 10 and worked with the remaining 28,817 SNPs. We replicated the single SNP test (SNP-2 method) of Schymick et al., and followed their haplotype definition to perform 3-SNP sliding window haplotype association tests using the HS method. The HS method identified the most significant association SNP right around rs4363506 (p-value $= 1.2 \times 10^{-7}$), and the p-values for the region around rs4363506 are shown in

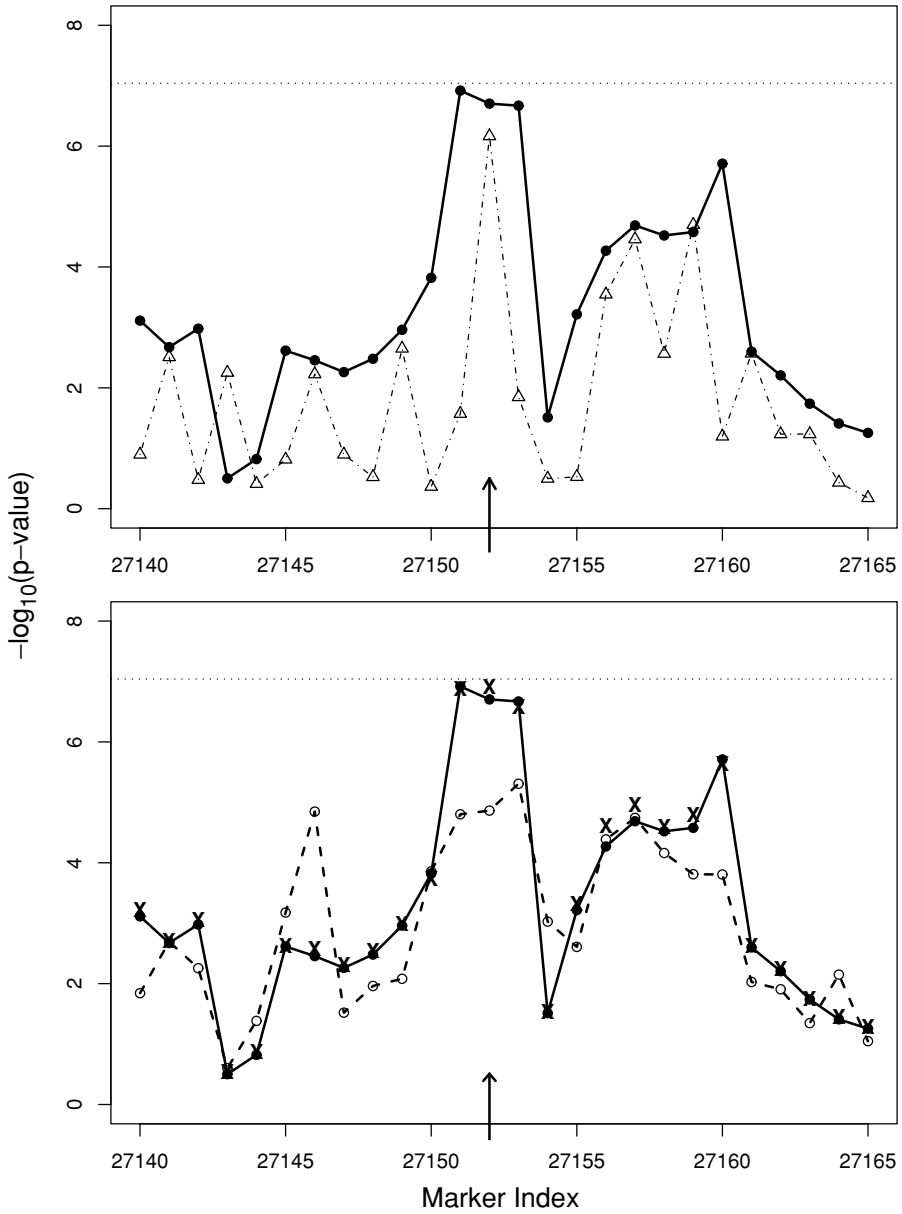

Figure 4. P-values from the ALS data analysis around the most promising SNP reported in Schymick et al. (i.e., SNP rs4363506 with location indicated by the arrow). The p-values are presented on the scale of negative logarithm of base 10, and the horizontal lines indicate the Bonferroni genome-wide threshold. Upper panel: p-values of the HS method based on 3-SNP haplotypes (solid line with filled circle) versus p-values of the SNP-2 method (dotted-dashed line with open triangle). Lower panel: p-values of the HS method (solid line with filled circle) versus p-values of the HAP method (dashed line with open circle). The "x" signs indicate the p-values of the VC method.

Figure 4 on the scale of negative logarithm of base 10 (upper panel). We see that haplotype analyses exhibited smoother association signals across SNPs than the single SNP tests. Although the less-noisy haplotypic signals came at the cost of modeling multimarker variations, we see that the HS method achieved a level of significance that is comparable to single SNP analyses. To compare, we also implemented the HAP method and the VC method. The results are presented in the lower panel of Figure 4. We observed that the p-values of the HAP method are slightly less significant around rs4363506, and as expected, the VC method yields very similar results as the HS method.

6. Discussion

In this article, we introduced a regression model of trait similarity and haplotype similarity to study haplotype association. We set trait similarity to be the weighted mean-corrected trait across products, and haplotype similarity to be the sharing

degrees of the haplotypes between two subjects. We constructed a score statistic based on the similarity regression to test the null hypothesis of zero association between genotypes and traits. The score statistic is shown to follow a weighted chi-squared distribution under the null hypothesis, which can be approximated by the three-moment approximation method of Imhof (1961). The proposed method uses a simple statistic, eliminates the need to perform permutation, and can be easily scaled up to the whole genome scale. In addition, the proposed method applies to both qualitative and quantitative traits, and can incorporate covariate effects such as population structure and other environmental confounders. The test could be conservative with moderate sample size, but the conservativeness is alleviated with an increased sample size. Through simulation studies, we explored its performance and observed that the proposed HS methods result in power increase in many cases.

The HS method assumes that the adjacent SNPs have a high probability to be passed down together from some common ancestors in a region containing the disease SNP, and hence models the sharing level of the genetic materials instead of the variants themselves. On one hand, this rationale indicates that the HS method is more effective/sensitive in taking advantage of the joint correlation of the SNPs. On the other hand, this also means that the performance of the HS method can be more vulnerable to a low multilocus LD. For example, Figures 1 to 3 show that a combination of low average LD and low allele frequency of the disease SNP could diminish the performance of the HS method. Although Figures 1 to 3 are based on regions with max $R^2 > 0.7$ (i.e., the disease SNP is "tagged" by at least one of the neighboring SNPs), we found that similar high/low power patterns by the allele frequency and the average LD are observed for regions with max $R^2 \in [0.5, 0.7)$ (i.e., moderately tagged) and for regions with max $R^2 < 0.5$ (i.e., not tagged) (results not shown). We are examining whether this feature of low power implies any of the following potential causes: (i) there is a joint acting effect between the allele frequency and the multilocus LD, (ii) the conditions required in the HS method are jeopardized with low average LD level and low allele frequency, or (iii) the performance of the HS method can be better predicted by other high-order LD structure than average R^2, such as the multiple-order Markov chains of Kim, Feng, and Zeng (2008). Nevertheless, the findings suggest that the HS methods would be most ideal if coupled with a careful determination of haplotype regions based on the LD structure of multimarkers to obtain the power gain.

The idea of our haplotype-trait similarity regression can be traced back to the Haseman–Elston regression model for linkage analysis (Haseman and Elston, 1972; Elston et al. 2000), where the trait similarity of sib pairs is regressed on their identical-by-descent sharing probability. In our case, we treat the entire population as a family, and regress trait similarity of pairwise samples on their identical-by-state status. Just like the Haseman–Elston regression, which is shown to be equivalent to the VC methods in linkage analysis (Sham and Purcell, 2001), our gene-trait similarity regression is also analytically united with the VC approaches. Specifically, we showed that testing the regression coefficient of the haplotype similarity is the same as testing the genetic VC in a mixed model. In addition, from the score statistics of the two approaches, we see both methods utilize haplotype similarity, with a difference that the within-individual haplotype comparison is used in the VC model but not in the gene-trait similarity regression. These connections suggest that the two approaches should have similar performance; this conjecture is supported by the simulations and data analysis. This equivalence offers a new way of looking at haplotype-sharing methods: it enables direct derivation of analytical solutions of the sharing statistics under the framework of VC methods. In the cases studied here, both the similarity regression and the VC model yield closed-form solutions. But for more complicated models (i.e., incorporating interaction effects), the similarity regression model becomes much more analytically challenging yet the VC method offers an efficient and reliable technique to draw inference.

7. Supplementary Materials

Web Appendices A and B, referenced in Sections 2, 3, and 4, are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

The authors are grateful to the NINDS Human Genetics Resource Center DNA and Cell Line Repository at Coriell and the laboratory for supplying the data of the ALS study used in this work. They also thank the associate editor and two reviewers for their helpful and constructive comments, and Drs Kathryn Roeder, Silviu-Alin Bacanu, Andrew Allen, and Dahlia Nielsen for their careful discussions and suggestions on this work. JYT was supported by NSF grant DMS-0504726 and NIH grant R01 MH074027. DZ and MD were supported by NIH grant R01 CA085848. SMC was supported by NSC grant 96-2119-M-006-010. DCT was supported by NIH grants R01 CA52862 and U01 ES015090.

References

Allen, A. and Satten, G. (2007). Statistical models for haplotype sharing in case-parent trio data. *Human Heredity* **64,** 35–44.

Beckmann, L., Thomas, D. C., Fischer, C., and Chang-Claude, J. (2005). Haplotype sharing analysis using Mantel statistics. *Human Heredity* **59,** 67–78.

Bourgain, C., Genin, E., Quesnevile, H., and Clerget-Darpoux, F. (2000). Search for multifactorial disease susceptibility genes in founder populations. *Annals of Human Genetics* **64,** 255–265.

Bourgain, C., Genin, E., Holopainen, P., Mustalahti, K., Maki, M., and Partanen, J. (2001). Use of closely related affected individuals for the genetic study of complex diseases in founder populations. *American Journal of Human Genetics* **68,** 154–159.

Bourgain C., Genin, E., Ober, C., and Clerget-Darpoux, F. (2002). Missing data in haplotype analysis: A study on the MILC method. *Annals of Human Genetics* **66,** 99–108.

de Bakker, P. I., Yelensky, R., Pe'er, I., Gabriel, S. B., Daly, M. J., and Altshuler, D. (2005). Efficiency and power in genetic association studies. *Nature Genetics* **37,** 1217–1223.

Durham, L. K. and Feingold, E. (1997). Genome scanning for segments shared identical by descent among distant relatives in isolated populations. *The American Journal of Human Genetics* **61,** 830–842.

Elston, R. C., Buxbaum, S., Jacobs, K. B., and Olson, J. M. (2000). Haseman and Elston revisited. *Genetic Epidemiology* **19,** 1–17.

Feder, J. N., Gnirke, A., Thomas, W., Tsuchihashi, Z., Ruddy, D. A., Basava, A., Dormishian, F., Domingo, R., Ellis, M. C., Fullan, A., Hinton, L. M., Jones, N. L., Kimmel, B. E., Kronmal, G. S., Lauer, P., Lee, V. K., Loeb, D. B., Mapa, F. A., McClelland, E., Meyer, N. C., Mintier, G. A., Moeller, N., Moore, T., Morikang, E., Prass, C. E., Quintana, L., Starnes, S. M., Schatzman, R. C., Brunke, K. J., Drayna, D. T., Risch, N. J., Bacon, B. R., and Wolff, R. K. (1996). A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nature Genetics* **13,** 399–408.

Haseman, J. K. and Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics* **2,** 3–19.

Houwen, R. H. J., Baharloo, S., Blankenship, K., Raeymaekers, P., Juyn, J., Sandkuijl, L. A., and Freimer, N. B. (1994). Genome screening by searching for shared segments: Mapping a gene for benign recurrent intrahepatic cholestasis. *Nature Genetics* **8,** 380–386.

Imhof, J. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika* **48,** 419–426.

Kim, Y., Feng, S., Zeng, Z. B. (2008). Measuring and partitioning the high-order linkage disequilibrium by multiple order Markov chains. *Genetic Epidemiology* **32,** 301–312.

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

Puffenberger, E. G., Kauffman, E. R., Bolk, S., Matise, T. C., Washington, S. S., Angrist, M., Weissenbach, J., Garver, K. L., Mascari, M., Ladda, R., Siaugenhaupt, S. A., and Chakravarti, A. (1994). Identity-by-descent and association mapping of a recessive gene for Hirschsprung disease on human chromosome 13q22. *Human Molecular Genetics* **3,** 1217–1225.

Qian, D. and Thomas, D. C. (2001). Genome scan of complex traits by haplotype sharing correlation. *Genetic Epidemiology* **21**(Suppl 1), S582–S587.

Schaid, D. J. (2004). Evaluating associations of haplotypes with traits. *Genetic Epidemiology* **27,** 348–364.

Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M., and Poland, G. A. (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. *The American Journal of Human Genetics* **70,** 425–434.

Schymick, J. C., Scholz, S. W., Fung, H. C., Britton, A., Arepalli, S., Gibbs, J. R., Lombardo, F., Matarin, M., Kasperaviciute, D., and Hernandez, D. G. (2007). Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: First stage analysis and public release of data. *Lancet Neurology* **6,** 322–328.

Service, S. K., Lang, D. W., Freimer, N. B., and Sandkuijl, L. A. (1999). Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. *The American Journal of Human Genetics* **64,** 1728–1738.

Sha, Q., Chen, H. S., and Zhang, S. (2007). A new association test using haplotype similarity. *Genetic Epidemiology* **31,** 577–593.

Sham, P. C. and Purcell, S. (2001). Equivalence between Haseman-Elston and variance-components linkage analyses for sib pairs. *American Journal of Human Genetics* **68,** 1527–1532.

The International HapMap Consortium. (2003). The International HapMap Project. *Nature* **426,** 789–796.

Thomas, D. C., Qian, D., Gauderman, W. J., Siegmund, K., and Morrison, J. L. (1999). A generalized estimating equations approach to linkage analysis in sibships in relation to multiple markers and exposure factors. *Genetic Epidemiology* **17**(Suppl 1), S737–S742.

Tzeng, J. Y. and Zhang, D. (2007). Haplotype-based association analysis via variance component score test. *The American Journal of Human Genetics* **81,** 927–938.

Tzeng, J. Y., Byerley, W., Devlin, B., Roeder, K., and Wasserman, L. (2003). Outlier detection and false discovery rates for whole-genome DNA matching. *Journal of the American Statistical Association* **98,** 236–246.

Tzeng, J. Y., Devlin, B., Wasserman, L., and Roeder, K. (2003). On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *The American Journal of Human Genetics* **72,** 891–902.

Van der Meulen, M. A. and Te Meerman, G. J. (1997) Haplotype sharing analysis in affected individuals from nuclear families with at least one affected offspring. *Genetic Epidemiology* **14,** 915–920.

Wall, J. D. and Pritchard, J. K. (2003). Assessing the performance of the haplotype block model of linkage disequilibrium. *The American Journal of Human Genetics* **73,** 502–515.

Wessel, J. and Schork, N. J. (2006). Generalized genomic distance-based regression methodology for multilocus association analysis. *The American Journal of Human Genetics* **79,** 792–806.

Yu, K., Gu, C., Province, M., Xiong, C., and Rao, D. (2004). Genetic association mapping under founder heterogeneity via weighted haplotype similarity analysis in candidate genes. *Genetic Epidemiology* **27,** 182–191.

Received November 2007. *Revised August* 2008.
Accepted August 2008.