INSTITUTE OF STATISTICS MIMEO SERIES #2606

Regression-based Multi-marker Analysis for Genome-wide Association Studies Using Haplotype Similarity

Jung-Ying Tzeng* Shang-Mao Chang† Daowen Zhang*
Duncan Thomas‡ Marie Davidian*

Abstract

Although haplotype analyses have been prevalent in genetic association studies, the main statistical analyses in genome-wide association (GWA) studies still mostly focus on single-SNP analyses. Several practical issues hinder practitioners' enthusiasm to perform haplotype-based analysis under a GWA setting, including the need for large degrees of freedom and the treatment of missing phase information. To avoid these pitfalls, we propose a similarity-based regression model. It captures genetic variants via haplotype similarity to reduce the degrees of freedom, and uses phase-independent similarity measures to bypass the needs to impute phase information. We construct the score test to detect association between trait similarity and genetic similarity, and identify the limiting distribution of the score statistic. We show that the gene-trait similarity regression is closely connected with the random effects haplotype analysis although commonly they are considered as separate modeling tools in haplotype analysis. The proposed method is computationally efficient, allows for covariates and is applicable to both quantitative and binary traits. It serves as an effective tool for multi-marker analysis in genome-wide association studies.

Keywords: Haplotype-based association test, Haplotype analysis in GWAS, Haplotype similarity

*Department of Statistics, North Carolina State University, Raleigh NC, USA
†Department of Statistics, National Cheng Kung University, Tainan, Taiwan
‡Department of Preventive Medicine, University of Southern California, Los Angeles CA, USA

1 Introduction

With the rapid development of detailed genomic information and cost-effective genotyping technologies, genome-wide association studies (GWAS) have become a powerful new tool for scientists to identify genetic variants responsible for human complex diseases. In GWAS, a high-density set (550K or more) of single nucleotide polymorphisms (SNPs) across the genome are genotyped, and the association between these genetic contents and disease phenotypes is evaluated region by region to identify candidate loci for disease susceptibility genes.

It is generally appreciated that the genetic contents can be better captured by considering haplotypes (i.e., the ordered allele sequence of multi-SNPs) than by single SNPs. However, most statistical analyses in GWAS focus primarily on single-SNP analyses. Several practical issues complicate haplotype analyses under a GWA setting. First, it is usually not clear how a haplotype segment should be defined. Second, a regular haplotype analysis consumes more degrees of freedom per test than a single-SNP analysis. The large degrees of freedom can limit the power of a haplotype association test, and the overall performance of a haplotype association scan can be further diminished by the multiple testing adjustment. Finally, common solutions for dealing with missing phases may no longer be suitable in GWAS. For instance, one usually treats the unobserved phase information as missing data and intermediately impute them using the EM algorithm. However, this strategy relies on assumptions, such as common haplotype frequencies across different subpopulations and Hardy-Weinberg equilibrium, that may not hold for all regions across the genome. Consequently the phases cannot be inferred accurately and neither can be the subsequent association inference.

In this work we address the concerns listed above except the haplotype definition, and develop an approach for haplotype association analysis in GWAS by modeling haplotype similarity. Haplotype similarity approaches look for unusual sharing of chromosomal segments within homogeneous trait groups (Houwen et al. 1994; McPeek and Strahs 1999). This strategy has been used to reduce the degrees of freedom and to avoid problems caused by rare haplotypes. The underlying rationale is that cases tend to share 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 sharing common haplotypes in the region flanking a disease mutation (Feder et al. 1996; Puffenberger et al. 1994).

The excessive 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 et al. 2003ab; Yu et al. 2004). The former approaches were used in the successful positional cloning of Mendelian disorders since the 1980's. 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 et al. (2007) addressed the latter concern by contrasting similarity of concordant pairs (case-case and control-control) with similarity of discordant pairs (case-control).

Current developments have shifted the focus from two-sample tests to regression models that correlate trait similarity with genetic 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 to 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 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 pair-wise samples, as the observation unit in the regression is now pairs of individuals instead of single individuals. Consequently, test statistic distributions are hard to derive analytically, and permutation is needed to find p-values. This greatly complicates their usage in GWAS.

Our proposed similarity method also follows this model-based direction. For trait similarity, we measure the trait covariance of all 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 propose to use the similarity metrics that can measure haplotype similarity directly from genotypes. In section 3, we show the similarity regression is closely connected to an alternative haplotype analysis approach, the variance component 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 h th 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 measured by a certain similarity metric $s\left(h,k\right)$. From the definition of H, we have

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

Let $\mu_i^0 = \mathrm{E}(Y_i \mid X_i, H_i) = \delta\left(X_i^T\gamma\right)$ 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 $X_i^T\gamma$ such as that specified by a 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 subject pair in linkage studies (Elston et al. 2000; Thomas et al. 1999). The weight ω_i may be used to account for the fact that Y_i is not homogeneous. In principle, ω_i can be any pre-specified positive values 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 the gene-trait association:

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

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.

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 $\mathrm{v}\,_i^0 = \mathrm{var}(Y_i \mid X_i, H_i) = m_i^{-1}\phi v\left(\mu_i^0\right)$ under the condition of no haplotype effects, where m_i is a known prior weight, ϕ is the dispersion parameter, and $v\left(\mu_i\right)$ 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 to construct the score test:

$$U \;=\; \begin{bmatrix} U_b \\ U_\gamma \\ U_\phi \end{bmatrix} = -\,\mathrm{E}\begin{bmatrix} \frac{\partial\{Z-\mathrm{E}(Z|X,H)\}}{\partial(b,\gamma^T,\phi)} \\ \frac{\partial\{Y-\mathrm{E}(Y|X,H)\}}{\partial(b,\gamma^T,\phi)} \end{bmatrix}^T \begin{bmatrix} \mathrm{var}\left(Z \mid X,H\right) & \mathrm{cov}\left(Z,Y \mid X,H\right) \\ \mathrm{cov}\left(Y,Z \mid X,H\right) & \mathrm{var}\left(Y \mid X,H\right) \end{bmatrix}^{-1} \begin{bmatrix} Z-\mathrm{E}\left(Z \mid X,H\right) \\ Y-\mathrm{E}\left(Y \mid X,H\right) \end{bmatrix} = 0$$

Assume that Y_i's are independent under the null hypothesis of no haplotype effect (i.e., $b = 0$). This assumption will be justified after we postulate a model on haplotype effects (Section 3). Following the independence assumption and by the definition of variance, we have that under H_0, $\mathrm{cov}(Z_{ij}, Y_l \mid X, H) = 0\ \forall i \neq j$ and $\forall l$, $\mathrm{var}(Y \mid X, H) = V = diag\left\{\mathrm{v}_i^0\right\}$, and $\mathrm{var}(Z \mid X, H) = diag\left\{\omega_i^2\omega_j^2\mathrm{v}_i^0\mathrm{v}_j^0\right\}$. Thus U_b, the score statistic for b, is

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

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 $\widehat{\mu}^0 = \delta\left(X_i^T\widehat{\gamma}\right)$. To derive its distribution, we rewrite U_b in a quadratic form as

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

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

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

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

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

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 $P\left(\chi_{h'}^2 > \chi^*\right)$, 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 $s\left(h,k\right)$ is the ”counting measure” of Tzeng et al. (2003a), which calculates the proportion of alleles in common between haplotype h and haplotype k. Define $G_{m,i}^T$ 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, $m = 1, 2, \cdots, M$. It can be shown that

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

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 SNP.

Concerns may arise in terms of choosing between phase-dependent and phase-independent metrics for performing haplotype-similarity analysis. In theory, metrics that make use of the phase information should be more powerful as they may capture the identical-by-descent (IBD) 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 et al. 2003a; Sha et al. 2007).

3 Connection to the Variance-Component Score Test

Similar to the linkage analysis where the regression-based approaches have been shown equivalent to the variance-component methods (Sham and Purcell 2001), the proposed gene-trait similarity regression is also connected with the variance-component 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 the following generalized linear mixed model (GLMM):

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

where $\mu_i = \mathrm{E}(Y_i \mid \beta, X_i, H_i)$, $g\left(\cdot\right)$ is the link function, γ is the fixed effect of environmental covariates, and β is the random effect of the haplotypes. Also define that $\mathrm{var}(Y_i \mid \beta, X_i, H_i) = \mathrm{v}_i = m_i^{-1} \phi v\left(\mu_i\right)$. Under this model, $\mu_i = g^{-1}\left(X_i^T \gamma + H_i^T \beta\right)$ and $\mu_i^0 = g^{-1}\left(X_i^T \gamma\right)$, 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 $s\left(h, k\right)$ $(0 \leq s(h, k) \leq 1)$ being the similarity metric that measures the similarity level between haplotypes h and k (Schaid 2004; Tzeng and Zhang 2007).

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

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

Therefore

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

Consequently, the expected trait similarity is

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

The result indicates that trait similarity is (approximately) in a linear relationship with haplotype similarity if we choose $\omega_i = g'(\mu_i^0)$. (In that case, $\mathrm{E}(Z_{ij} \mid X, H) \approx \tau S_{ij}$.) For the canonical link, this choice is equivalent to $\omega_i = 1/v\left(\mu_i^0\right)$. Comparing (6) to (3), we see that testing $b = 0$ in the similarity regression is the same as testing for $\tau = 0$ in a variance component model. This implies that under the null hypothesis of $b = 0$, τ 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 variance component score test for testing $H_0 : \tau = 0$ based on the GLMM (5) as

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

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

$$U_b = \frac{1}{2}\left(Y - \widehat{\mu}^0\right) \Delta W \Delta \Omega^{-1} S_0 \Omega^{-1} \Delta W \Delta \left(Y - \widehat{\mu}^0\right), \tag{7}$$

as $V^{-1} = \Delta W \Delta$. Comparing T_τ to U_b, 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 $\omega_i = g'(\mu_i^0)$ is used, in which case $U_b = \frac{1}{2}(Y - \widehat{\mu}^0)^T \Delta W S_0 W \Delta (Y - \widehat{\mu}^0)$. 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 comparisons (i.e., $i \neq j$), while T_τ of variance component includes the comparisons 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-comparison and hence not informative. As a result, although more data information is utilized in the variance-component test, we expect a similar performance of the two approach in detecting haplotype-phenotype association.

4 Simulation Studies

We performed simulation studies to investigate the behaviors of the score test based on the 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 Prichard (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 three criteria: (a) the allele frequency, (b) the haplotype diversity, and (c) the correlation between the disease SNP and its 5 flanking SNPs. We set the allele frequency at 0.1 and 0.3, classified the haplotype diversity levels into *high* (11-16 distinct haplotypes), *moderate* (9-11) and *low* (6-9), and labeled a disease SNP as *tagged* if it is in high correlation with at least one neighboring SNPs (i.e., $R^2 > 0.7$) and *untagged* if not. We paired two random draws from the 100 haplotypes and formed an individual.

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 $\mu_i \equiv \mathrm{E}(Y_i \mid X_i, \mathrm{G}_i)$, X_i is a standard normal variable, and G_i is determined by the number of disease alleles that individual i has. We considered two types of traits: normal traits and binary traits. For normal traits, we use identity link and set G_i as the number of the disease allele minus 1. This set-up simplified the trait variance as $\mathrm{Var}(Y_i \mid X_i, \mathrm{G}_i) = 2q(1-q)(1-h^2)/h^2$ with h^2 the heritability and q the disease allele frequency. We set $h^2 = 0.1$, $\gamma_0 = \gamma_1 = 1$, and $\theta = 1$. For binary traits, we use the logit link and let G_i equal to the number of the disease allele. We set $\gamma_0 = -4.5$, $\gamma_1 = 0$, and $\theta = \log 2$, resulting a disease rate of 1. For both traits, we set $\theta = 0$ when evaluating the size of the proposed test.

Under each simulation scenario, we used random sampling to obtain 200 individuals with normal traits, and used balanced case-control sampling to obtain 100 cases and 100 controls. 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), the variance component method of Tzeng and Zhang (2007), and the gene-trait similarity regression method with $\omega_i = 1/v(\mu_i)$. The simulation results are obtained based on 10,000 replications under each scenario.

Table 1 and Table 2 show the type I error rates for normal traits and binary traits respectively at the nominal levels of 0.05, 0.01, 0.005 and 0.001. To evaluate the appropriateness of the score test based on the proposed gene-trait similarity regression and the three-moment approximation separately, we also calculated the empirical p-values by resampling 10^4 U_b's using the fact that $U_b = \sum_{i=1}^{n} \lambda_i \chi^2_{1,i}$. The type I error rates obtained by resampling are shown in the parentheses.

Overall speaking the ”resampling” type I error rates and the ”3-moment approximation” type I error rates do not differ much from each other. Though slightly conservative (mostly at the 0.05 nominal level), the type I error rates are around the desired size, and the general patterns of normal traits and binary traits are similar. We also note that the over-conservativeness faded away when we increased the sample size to 500 and 1000 (results not shown). These results suggest the proposed similarity-model based score test and the three-moment method used to approximate the null distribution performed reasonably for both trait types.

Table 3 and Table 4 show the power comparisons for normal traits and binary traits respectively. We presented the results for nominal levels 0.05 and 0.005, and note that the findings with nominal levels of 0.01 and 0.001 are similar (results not shown). When comparing the similarity regression (*Similarity*) with the variance-component method (*VC*), we see similar power performance from these two approaches. This is within our anticipation as the two tests have comparable forms of statistics and utilized genetic information through haplotype similarity. Next we compare these two methods (*Similarity* and *VC*) to the standard regression method (*Standard*). We found the factor ”tagged” (i.e., whether the unobserved disease SNP is highly correlated ($R^2 > 0.7$) with its neighboring SNPs) plays a key role in predicting the performance. When the disease SNP was tagged by at least one of surrounding SNPs, we observed that the methods of *Similarity* and *VC* attain the same or higher level of power than the standard regression. On the other hand, if none of the adjacent SNPs captured the information of the unobserved disease SNP, all methods suffer from power loss, and the standard method often gets affected less and retains higher power than *Similarity* and *VC*. These observations holds for both binary traits and normal traits. However, the power improvement may be less obvious with normal traits because the power of the ”tagged” scenario is set a bit too high, leaving little room for the potential power gain resulting from the reduction of degrees of freedom by the methods of *Similarity* and *VC*.

5 Application to the ALS study

We applied our method to a case-control study of amyotrophic lateral sclerosis (ALS) obtained from the National Institute of Neurological Disorders and Stroke (NINDS) neurogenetics Repository at the Coriell Institute website (http://www.coriell.org/index.php/). 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

Table 1: Type I error rates for normal traits using the 3-moment approximation to the limiting distribution of U_b. The type I error rates obtained from resampling are shown in parentheses.

Normal Traits	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.005$	$\alpha = 0.001$
Tagged				
High Diversity				
$q = 0.1$	0.039 (0.043)	0.011 (0.010)	0.0070 (0.0052)	0.0020 (0.0010)
$q = 0.3$	0.045 (0.050)	0.011 (0.011)	0.0068 (0.0055)	0.0025 (0.0018)
Moderate Diversity				
$q = 0.1$	0.039 (0.046)	0.010 (0.009)	0.0053 (0.0045)	0.0017 (0.0008)
$q = 0.3$	0.038 (0.042)	0.009 (0.008)	0.0042 (0.0033)	0.0011(0.0007)
Low Diversity				
$q = 0.1$	0.034 (0.038)	0.008 (0.007)	0.0048 (0.0036)	0.0019 (0.0010)
$q = 0.3$	0.047 (0.050)	0.009 (0.009)	0.0060 (0.0054)	0.0011 (0.0009)
Untagged				
High Diversity				
$q = 0.1$	0.043 (0.049)	0.011 (0.011)	0.0068 (0.0054)	0.0019 (0.0007)
$q = 0.3$	0.047 (0.051)	0.013 (0.012)	0.0064 (0.0051)	0.0014 (0.0010)
Moderate Diversity				
$q = 0.1$	0.040 (0.044)	0.009 (0.009)	0.0052 (0.0045)	0.0011 (0.0008)
$q = 0.3$	0.044 (0.050)	0.011 (0.010)	0.0067 (0.0054)	0.0015 (0.0009)
Low Diversity				
$q = 0.1$	0.044 (0.048)	0.010 (0.010)	0.0062 (0.0050)	0.0018 (0.0014)
$q = 0.3$	0.022 (0.024)	0.006 (0.006)	0.0040 (0.0026)	0.0011 (0.0007)

Table 2: Type I error rates for binary traits using the 3-moment approximation to the limiting distribution of U_b. The type I error rates obtained from resampling are shown in parentheses.

Binary Traits	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.005$	$\alpha = 0.001$
Tagged				
High Diversity				
$q = 0.1$	0.038 (0.043)	0.011 (0.011)	0.0077 (0.0063)	0.0018 (0.0008)
$q = 0.3$	0.043 (0.048)	0.011 (0.011)	0.0065 (0.0056)	0.0020 (0.0012)
Moderate Diversity				
$q = 0.1$	0.039 (0.043)	0.011 (0.010)	0.0054 (0.0043)	0.0013 (0.0007)
$q = 0.3$	0.044 (0.049)	0.011 (0.011)	0.0060 (0.0050)	0.0020 (0.0012)
Low Diversity				
$q = 0.1$	0.033 (0.037)	0.009 (0.008)	0.0055 (0.0039)	0.0025 (0.0012)
$q = 0.3$	0.044 (0.049)	0.011 (0.010)	0.0049 (0.0045)	0.0013 (0.0008)
Untagged				
High Diversity				
$q = 0.1$	0.043 (0.048)	0.012 (0.011)	0.0065 (0.0061)	0.0027 (0.0014)
$q = 0.3$	0.042 (0.047)	0.010 (0.009)	0.0055 (0.0046)	0.0018 (0.0011)
Moderate Diversity				
$q = 0.1$	0.041 (0.044)	0.010 (0.009)	0.0059 (0.0052)	0.0012 (0.0006)
$q = 0.3$	0.040 (0.044)	0.010 (0.010)	0.0063 (0.0054)	0.0020 (0.0017)
Low Diversity				
$q = 0.1$	0.041 (0.045)	0.010 (0.010)	0.0055 (0.0045)	0.0010 (0.0007)
$q = 0.3$	0.020 (0.022)	0.004 (0.004)	0.0028 (0.0022)	0.0009 (0.0003)

Table 3: Power comparison of different approaches for normal traits.

Normal Traits	$\alpha = 0.05$			$\alpha = 0.005$		
	Standard	VC	Similarity	Standard	VC	Similarity
Tagged						
High Diversity						
$q = 0.1$	0.86	0.87	0.85	0.60	0.54	0.58
$q = 0.3$	0.69	0.93	0.93	0.38	0.74	0.77
Moderate Diversity						
$q = 0.1$	0.88	0.92	0.91	0.63	0.65	0.68
$q = 0.3$	0.85	0.98	0.98	0.57	0.89	0.91
Low Diversity						
$q = 0.1$	0.91	0.95	0.93	0.71	0.76	0.79
$q = 0.3$	0.96	0.99	0.99	0.81	0.95	0.95
Untagged						
High Diversity						
$q = 0.1$	0.57	0.49	0.47	0.26	0.14	0.16
$q = 0.3$	0.31	0.13	0.13	0.08	0.02	0.02
Moderate Diversity						
$q = 0.1$	0.58	0.29	0.28	0.26	0.07	0.08
$q = 0.3$	0.62	0.79	0.78	0.29	0.47	0.49
Low Diversity						
$q = 0.1$	0.79	0.55	0.53	0.48	0.22	0.24
$q = 0.3$	0.45	0.56	0.41	0.18	0.20	0.17

Table 4: Power comparison of different approaches for binary traits.

Binary Traits	$\alpha = 0.05$			$\alpha = 0.005$		
	Standard	VC	Similarity	Standard	VC	Similarity
Tagged						
High Diversity						
$q = 0.1$	0.29	0.37	0.34	0.06	0.10	0.12
$q = 0.3$	0.29	0.55	0.54	0.06	0.24	0.27
Moderate Diversity						
$q = 0.1$	0.31	0.45	0.43	0.07	0.16	0.18
$q = 0.3$	0.39	0.68	0.66	0.11	0.36	0.39
Low Diversity						
$q = 0.1$	0.35	0.46	0.42	0.10	0.17	0.19
$q = 0.3$	0.55	0.77	0.76	0.21	0.45	0.47
Untagged						
High Diversity						
$q = 0.1$	0.21	0.19	0.17	0.04	0.03	0.04
$q = 0.3$	0.13	0.08	0.07	0.02	0.01	0.01
Moderate Diversity						
$q = 0.1$	0.18	0.12	0.12	0.03	0.02	0.02
$q = 0.3$	0.25	0.40	0.38	0.05	0.14	0.16
Low Diversity						
$q = 0.1$	0.26	0.21	0.20	0.06	0.04	0.05
$q = 0.3$	0.22	0.25	0.15	0.04	0.06	0.04

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 mot significant SNP (rs4363506) lives in the 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 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 of Schymick et al's, and followed their haplotype definition to perform 3-SNP sliding window haplotype association tests using the proposed similarity-regression based score test. Our method identified the most significant association SNP right around rs4363506 (p-value=1.2×10^{-7}), and the p-values for the region around rs4363506 are shown in Figure 1 on the scale of negative logarithm of base 10 (upper panel). We see that haplotype analyses exhibited a smoother association signals across SNPs than the single SNP tests. Although the less-noisy haplotypic signals came at the cost of modeling multi-marker variations, we see that our similarity method achieved a level of significance that is comparable to single SNPs analyses. To compare, we also implemented the standard haplotype analysis (Schaid et al. 2002) and the variance component haplotype analysis (Tzeng and Zhang 2007). The results are presented in the lower panel of Figure 1. We observed that the p-values of the standard method are slightly less significant around rs4363506, and as expected, the variance component tests yield very similar results as the similarity regression.

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 cross products, and haplotype similarity to be the sharing degrees of the haplotypes between two individuals. We constructed a score statistic to test the null hypothesis of zero association between trait similarity and haplotype similarity. The score statistic is shown to follow a weighted chi-squared distribution under the null hypothesis, and the distribution can be approximated by the three-moment approximation of Imhof (1961). The statistic is simple to calculate, eliminates the needs to perform permutation, and is computationally efficient to be applied directly for GWAS. The proposed method applies to both qualitative and quantitative traits, and can incorporate covariate effects such as population structure and other environmental confounders. Simulation results showed that the test has its size around the nominal level, and exhibits the same or higher power than the standard

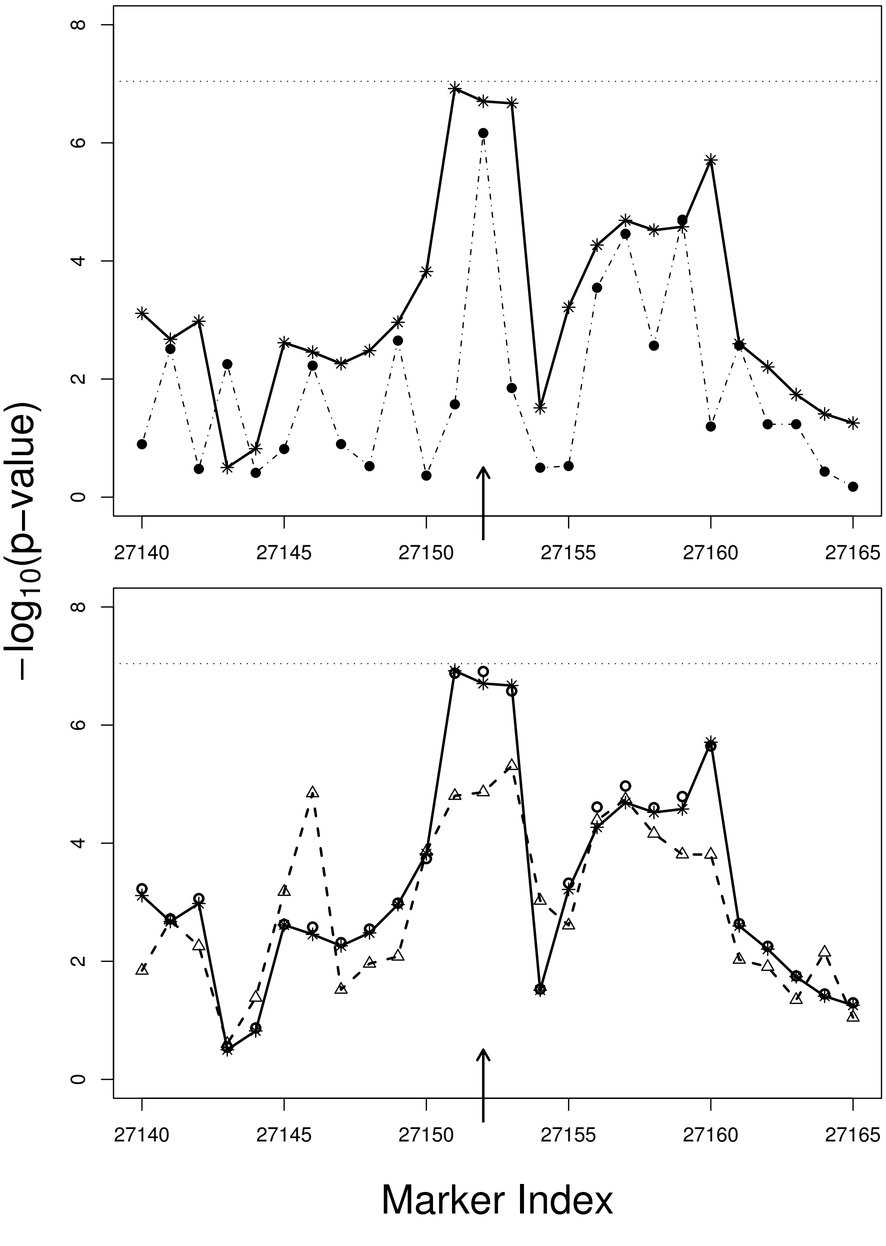

Figure 1: P-values of association analyses around SNP rs4363506

haplotype regression analysis in cases where the disease SNP is tagged by the flanking SNPs.

The idea of our haplotype-trait similarity regression can be traced back to the Haseman-Elston regression model for linkage analysis, where the trait covariance of sib pairs is regressed on their identical-by-decent sharing probability. In our case, we treat the entire population as a family, and regress trait similarity of pair-wise samples on their identical-by-state status. Just like the Haseman-Elston regression, which is shown to be equivalent to the variance-component methods in linkage analysis (Sham and Purcell 2001), our gene-trait similarity regression is also analytically united with the variance-component approaches. Specifically, we showed that testing the regression coefficient of the haplotype similarity is the same as testing the genetic variance component 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 variance-component 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 simulation results and data analysis.

There are limitations in our proposed method. Our method implicitly assumes that the adjacent SNPs have high probability to be passed down together with the disease locus from generation to generation. This is indeed a natural constrain to the sharing-based approaches. Low correlation among SNPs indicates that these SNPs are less likely to descend from some common ancestors and hence results in low degrees of sharing among the current population even in regions harboring a disease mutation. Under this scenario, we observed that our method, compared to the standard regression method, tend to have a larger drop in power. This suggests that the similarity-based methods would be most ideally to be coupled with a careful determination of haplotype regions based on the LD structure. On the other hand, given a predefined haplotype region, the LD pattern might be used as an indicator in practice to choose between similarity-based approaches and standard haplotype regression.

Due to computational limit, we examined the performance of the test with the three-moment approximation at the nominal levels from 0.05 to 0.001. More extreme tail behaviors, such as to the popular genome-wide significance level of 10^{-7}, have not been thoroughly studied here. While we do not suspect that the test statistics will behave drastically different, we describe a resampling algorithm to calculate p-values as an alternative if one feels uncomfortable with the performance of the approximation. The resampling can be implemented based on the fact that $U_b \overset{d}{\equiv} Z^T \Lambda Z$, where Z is a $n \times 1$ standard multivariate normal random vector, and $\Lambda = diag(\lambda_i)$ with λ_i the eigenvalues of matrix C.

Acknowledgments

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 Dr. Kathryn Roeder, Dr. Silviu-Alin Bacanu, and Dr. Andrew Allen for their helpful discussion and suggestions on this work. J.Y.T. was supported by NSF grant DMS-0504726 and National Institutes of Health grant R01 MH074027. S.M.C. was supported by NSC grant 96-2119-M-006-010. D.Z. was supported by NIH grant R01 CA085848.

References

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

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

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

4. 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.

5. 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.

6. 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.

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

8. 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.

9. 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.

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

11. McPeek, M. S. and Strahs, A. (1999). Assessment of linkage disequilibrium by the decay of haplotype sharing with application to fine-scale genetic mapping. *The American Journal of Human Genetics* **65,** 858–875.

12. 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.

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

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

15. 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. *he American Journal of Human Genetics* **70,** 425–434.

16. 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.

17. Service, S. K., Lang, D. W., Freimer, N. B., 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.

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

19. Sham, P. C., and Purcell, S. (2001). Equivalence between Haseman-Elston and Variance-Components Linkage Analyses for Sib Pairs. *The American Journal of Human Genetics* **68,** 1527–1532.

20. 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.

21. Thomas, D. C., Morrison, J. L., and Clayton, D. G. (2001). Bayes estimates of haplotype effects. *Genetic Epidemiology* **21,** Suppl 1:S712–S717.

22. 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.

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

24. 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. Genet Epidemiol 14:915

25. 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.

26. 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.

27. 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.

Appendix

.1 A. Derivation of the distribution of U_b

From equation (4), we have the score statistic expressed as $U_b = (Y - \widehat{\mu}^0)^T V^{-1} \Omega^{-1} S_0 \Omega^{-1} V^{-1} (Y - \widehat{\mu}^0)/2$. To derive its distribution, we first apply Taylor expansion on $\widehat{\mu}^0$ and get that under the null hypothesis,

$$\widehat{\mu}^0 = \delta(X\widehat{\gamma}) \;\approx\; \mu^0 + DX(\widehat{\gamma} - \gamma),$$

where $D = diag\left\{\partial \delta\left(\eta_i\right)/\partial \eta_i\right\}$ and $\eta_i = X_i^T \gamma$. By the score function $U_\gamma = X^T D V^{-1}\left(Y - \mu^0\right)$ and that $\mathrm{E}(\partial U_\gamma / \partial \gamma) = -X^T D V^{-1} D X$ under H_0,

$$(\widehat{\gamma} - \gamma) \;\approx\; (X^T D V^{-1} D X)^{-1} X^T D V^{-1} (Y - \mu^0).$$

Therefore

$$\begin{aligned} Y - \widehat{\mu}^0 \;&\approx\; Y - \left\{\mu^0 + DX\left(X^T D V^{-1} D X\right)^{-1} X^T D V^{-1}\left(Y - \mu^0\right)\right\} \\ &= \left\{I - DX\left(X^T D V^{-1} D X\right)^{-1} X^T D V^{-1}\right\}\left(Y - \mu^0\right). \end{aligned}$$

Multiplying both sides by matrix V^{-1}, we have $V^{-1}\left(Y - \widehat{\mu}\right) \approx Q\left(Y - \mu^0\right)$ with $Q = V^{-1} - V^{-1} DX\left(X^T D V^{-1} D X\right)^{-1} X^T D V^{-1}$. Therefore,

$$\begin{aligned} U_b \;&\approx\; \frac{1}{2}\left(Y - \mu^0\right)^T Q \Omega^{-1} S_0 \Omega^{-1} Q \left(Y - \mu^0\right) \\ &= \frac{1}{2}\widetilde{Y}^T C \widetilde{Y}, \end{aligned}$$

where $\widetilde{Y} = V^{-\frac{1}{2}}\left(Y - \mu^0\right)$ and is the standardized Y under H_0, i.e., its ith element is equal to $\left(Y_i - \mu_i^0\right)/\sqrt{\mathrm{v}_i^0}$. Matrix $C = V^{\frac{1}{2}} Q \Omega^{-1} S_0 \Omega^{-1} Q V^{\frac{1}{2}}/2$.

Let $\Lambda = diag\left(\lambda_i\right)$ with $\lambda_1 \geq \cdots \geq \lambda_n$ the ordered eigenvalues of matrix C, and let E be an $n \times n$ matrix consisting of vector e_i, the corresponding orthonormal eigenvectors of λ_i. We then have

$$U_b \approx \widetilde{Y}^T C \widetilde{Y} = \widetilde{Y}^T E \Lambda E^T \widetilde{Y}.$$

Provided that each e_i is not dominated by a few elements, $Z_i \equiv e_i^T \widetilde{Y}$ will be approximately independently standard normal random variables. Therefore under H_0, the distribution of U_b is the same as that of $\sum_{i=1}^n \lambda_i \chi_{1,i}^2$, the weighted chi-squared distribution.

.2 B. Variance of Y

$$\mathrm{cov}\left(Y_i, Y_j \mid X, H\right) = \mathrm{E}_\beta\left\{\mathrm{cov}\left(Y_i, Y_j \mid \beta, X, H\right) \mid X, H\right\} + \mathrm{cov}_\beta\left\{\mu_i\left(\beta\right), \mu_j\left(\beta\right) \mid X, H\right\}.$$

Here we tentatively rewrite μ_i as $\mu_i\left(\beta\right)$ to denote that μ_i is a function of β. By the Taylor expansion of $\mu_i\left(\beta\right)$ around E$\beta = 0$,

$$\begin{aligned}
\text{cov}_\beta\left(\mu_i, \mu_j \mid X, H\right) &\approx \text{cov}_\beta\left\{\mu_i\left(0\right) + \mu_i'\left(0\right)^T\beta,\ \mu_j\left(0\right) + \mu_i'\left(0\right)^T\beta \mid X, H\right\} \\
&= \mu_i'\left(0\right)^T \text{var}\left(\beta\right)\mu_j'\left(0\right) \\
&= \left\{g'\left(\mu_i^0\right)g'\left(\mu_j^0\right)\right\}^{-1} \times \tau H_i^T R_\beta H_j,
\end{aligned}$$

as $\mu_i'\left(\beta\right) = \partial\mu_i\left(\beta\right)/\partial\beta = \left\{g'\left(\mu_i\right)\right\}^{-1}H_i$.