Detecting gene and gene-environment effects of common and uncommon variants on quantitative traits: A marker-set approach using gene-trait similarity regression

Jung-Ying Tzeng[1,2], Daowen Zhang[1], Monnat Pongpanich[2], Chris Smith[2], Mark I. McCarthy[3], Michèle M. Sale[4], Bradford B. Worrall[5], Fang-Chi Hsu[6], Duncan C. Thomas[7], Patrick F. Sullivan[8]

[1]: Department of Statistics, North Carolina State University, Raleigh NC, USA

[2]: Bioinformatics Research Center, North Carolina State University, Raleigh NC, USA

[3]:Wellcome Trust Center for Human Genetics, University of Oxford, Oxford, UK

[4]:Center for Public Health Genomics, Department of Medicine, Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville VA

[5]: Center for Public Health Genomics, Department of Public Health Sciences, Department of Neurology, University of Virginia, Charlottesville, VA

[6]: Department of Biostatistical Sciences, Wake Forest School of Medicine, Winston-Salem, NC

[7]: Department of Preventive Medicine, University of Southern California, Los Angeles CA, USA

[8]: Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill NC, USA

ADDRESS FOR CORRESPONDENCE:

Jung-Ying Tzeng, Department of Statistics, North Carolina State University, Campus Box 7566, Raleigh NC, 27695. Tel: 919-513-2723. Fax: 919-515-7315. E-mail: jytzeng@stat.ncsu.edu.

RUNNING TITLE: G & GxE test for common and rare variant

KEY WORDS: gene, pathway, exon level association test; gene-environment-wide interaction studies; uncommmon mutations

Summary

Genomic association analyses of complex traits demand statistical tools that are capable of detecting small effects of common and rare variants, modeling complex interaction effects, and yet be computationally feasible. In this work, we introduce a similarity-based regression method for assessing genetic main and interaction effects of a group of markers on quantitative traits. The method uses genetic similarity to aggregate information from multiple polymorphic sites, with adaptive weights dependent on allele frequencies for accommodating common and uncommon variants. Collapsing information at the similarity level instead of the genotype level avoids canceling signals with opposite etiological effects, and is applicable to any class of genetic variants without having to dichotomize the allele types. To assess gene-trait associations, we regress trait similarities for pairs of unrelated individuals on their genetic similarities, and assess association using a score test whose limiting distribution is derived. The proposed regression framework allows for covariates, has the capacity to model both main and interaction effects, can be applied to a mixture of different polymorphism types, and is computationally efficient. These features make it an ideal tool for evaluating associations between phenotype and marker sets defined by LD blocks, genes, or pathways in whole-genome analysis.

Introduction

Marker-set analysis refers to the joint evaluation of a group of markers for genetic association. These markers may be of various polymorphism types (e.g., a mixture of single nucleotide polymorphisms (SNP), insertion-deletion variants (INDEL), block substitutions, copy number variants, or inversion variants), but share certain common genomic features, such as being involved in the same pathway, in high linkage disequilibrium (LD), or located within the same gene or conserved functional region. Marker-set analysis has drawn great attention in recent genome-wide and sequence-based association studies. It assesses the joint association of potentially correlated and interacting loci. It amplifies the detectability of the causal signals by aggregating small effects from multiple individual loci. Furthermore, because sequences and functions of genes are highly consistent across populations and species, a marker-set analysis increases the interpretability and replicability of the association findings. For whole genome scans, it also offers a natural way to reduce the total number of tests, and hence improves power by reducing the multiple-testing burden. For sequence-based studies, it accumulates information across multiple rare mutations, and greatly enhances the power to detect rare variants that are hard to identify by traditional analysis methods.

A variety of methods are available for detecting marker-set association, ranging from minimum p-value or Fisher's combined methods[1,2] for single-marker tests to multi-marker tests using a genotype or haplotype based scoring. Many recent methods fall in between the two extremes. These methods collapse information from all markers in the set, and achieve a better balance between information and degrees of freedom. Depending on how the individual marker information is combined, we can roughly classify these approaches into four categories. The first type of methods uses the weighted sum of genotypes across markers, e.g., the LD-based weighting method[3], the weighted Fourier transform[4], and the PCA-based methods[5,6]. Recently, special versions of the weighted sum methods based on allele frequencies were proposed to target rare variants[7–10]. The second type of methods models genetic similarity of pairs of individuals, and are also referred to as U-statistics approaches[11–19]. The third type are variance-component

(VC) methods, which treat individual genetic effects as random effects, and test for the corresponding VC to detect the global effect of a gene. Methods of this type include the SNP random effects model[20,21], haplotype random effects model[22], and kernel-based methods[23–25]. The fourth category includes other approaches that do not fit into the above categories, such as the c-alpha test[26], group additive regression model[27], Tukey's model[28], and entropy-based methods[29,30].

While most marker-set methods have concentrated on detecting genetic main effects, here we focus on methods for studying gene-environment (G×E) interactions. Identifying genetic variants with heterogeneous effects under different environmental exposures is crucial for individualized medicine, for pharmacogenetics, for characterizing underlying biological mechanisms, and for uncovering unexplained heritability[31,32]. Marker-set analysis provides an ideal framework to study G×E interactions. The marker set, either defined by genes, pathways, or functions, provides a biologically sensible unit for the "G" component, and the loci in a set can be assessed jointly for whether their effects are modified under different environmental exposures. In addition, the potential power gain brought by the marker-set analysis—either through aggregating genetic signals or by reducing multiple testing penalty—can alleviate the "data-hungry" nature of detecting G×E interactions. Typically, a G×E test would require sample sizes at least four times larger than a main effect test for detecting an effect of comparable magnitude[31–34]. Furthermore, many G×E studies are based on conceptual models for candidate pathways, in which a set of genes are selected and studied together[32,35]. Marker-set analysis offers a suitable tool to evaluate the overall effect of the postulated pathways when assessing G×E interactions.

The marker-set G×E method we present focuses on quantitative traits and uses pairwise genetic similarity as a tool to aggregate marker information (i.e., category 2 in the above method categorization). Our approach differs from those in the literature on gene/pathway level analysis in the following aspects. First, we introduce a framework for incorporating interaction effects in similarity-based methods. To be useful for G×E studies with either confirmatory or exploratory aims, we develop a series of tests to suit

different purposes, including a test for detecting G×E interactions, a test for detecting marginal main effects, and a joint test for detecting the overall association induced either by genetic main effects or by G×E interactions. The joint test serves as a good tool when little is known *a priori* about the genetic heterogeneity across exposure strata, and provides power across a wide range of the unknown underlying true structures. Second, the proposed method can be used to collapse information from a mixture of different types of variants and is designed to detect common and uncommon variants. Both are desirable features when more classes of DNA variants are available. Finally, we illustrate how similarity-based collapsing methods can be equivalent to VC methods (i.e., category 3 in the method categorization), which are found to have better main-effect performance than several other marker-set approaches[24,37–39]. Through simulation, we show the validity of the test, and investigate the power of the proposed approach under a wide range of scenarios. We illustrate the utility of the proposed method using the samples from the Vitamin Intervention for Stroke Prevention (VISP) trial. In this study, candidate genes across the genome were selected to evaluate the gene and gene-age interaction effects on the change in fasting homocysteine (Hcy) level following a 2-hour methionine load test.

Materials and Methods

Gene-Trait Similarity Regression for G and G×E Effects

We use the following notations. For individual i $(i = 1, 2, \ldots, n)$, let Y_i be the continuous trait, X_i be the $K \times 1$ covariate vector excluding the intercept term and standardized to mean 0 and variance 1, and let $G_{m,i}$ be the allele-count vector of marker m for person i, with length equal to the number of distinct alleles at marker m (denoted by ℓ_m), $m = 1, 2, \cdots, M$. For example, $G_{m,i} = [2,\ 0]$ if person i has genotype '11' at SNP m. To fix idea, we consider $K = 1$, but the method described here also applies to $K > 1$.

For each pair of individuals i and j, we measure the trait similarity Z_{ij} and genetic similarity S_{ij} of the targeted marker set. We then regress the trait similarity on the genetic similarity, and detect gene-trait association by testing for the significance of relevant regression coefficients. The trait similarity Z_{ij} is quantified through trait covariance by taking the product of the trait residuals of subjects i and j. Define μ_i the subject-specific mean of trait value adjusted for the covariate information, we set $Z_{ij} = (Y_i - \mu_i)(Y_j - \mu_j)$, where $\mu_i = \gamma_0 + X_i\gamma$, and (γ_0, γ) is the covariate effects including the intercept. The genetic similarity S_{ij} is measured by the average of the weighted allele matching score (weighted matching score for short) between subjects i and j across the M markers. It takes the form of $S_{ij} = 1/M \times \sum_{m=1}^{M} G_{m,i}^T W_m G_{m,j}$, with W_m an $\ell_m \times \ell_m$ matrix that specifies the weighting scheme. To illustrate, consider a SNP and weight $W_m = I_{2\times 2}$. Then $S_{AA,AA} = 4$, $S_{AA,Aa} = 2$, and $S_{AA,aa} = 0$. When quantifying genetic similarity, one can use weights based on allele frequencies, the degree of evolutionary conservation, or the functionality of the variations to better target genetic variants of certain features (e.g., rare, functional)[15,25,41]. For example, to upweight similarities contributed by rare variants, we define the frequency of allele a at marker m as $q_{a,m}$, and set $W_m = diag\{1/q_{a,m}\}$ or $diag\{1/\sqrt{q_{a,m}}\}$ to upweight the similarity in rare alleles[23,24].

The proposed gene-trait similarity regression model has the following form:

$$E\left(Z_{ij} \mid X, H\right) = b \times S_{ij} + d \times S_{ij} \times X_i X_j,\ i \neq j. \tag{1}$$

Because baseline and covariate effects have been adjusted for in Z_{ij}, the regression has a zero intercept and does not have the covariate term $X_i X_j$. This contention will become more obvious from the viewpoint of variance components in the following paragraph. Model (1) incorporates information about genetic main effects and gene-environment interactions, and hence allows the possibility of a genetic effect to be modified by an environmental exposure. Under model (1), one can evaluate the overall genetic association by performing a joint test of genetic main effects and gene-environment interactions for $H_0 : b = d = 0$. To assess gene-environment interactions only, one can perform a G×E test by examining $H_0 : d = 0$. Finally, one can evaluate the marginal main effects by examining the main effect term and test for $H_0 : b = 0$, under the constraint of $d = 0$. We refer this test as the G test. The G test can be used as a subsequent test when a G×E test fails to reject H_0, or it can be used as an alternative way to detect the overall genetic association. Because interactive factors can often exhibit a marginal effect even when the interaction terms are not modeled[42,43], the G test is often used to perform genome screening in common practice. Compared with the joint test, the G test uses fewer degrees of freedom and hence is more powerful when there are no gene-environment interactions or when the interaction effects are big, but it may be less powerful when the genetic effect is restricted to exposure group[44].

The test statistics for G×E, G, and joint tests can be derived through the equivalence between the similarity regression models and the haplotype random effect model[17]. Consider a working haplotype random effect model:

$$Y_i = \gamma_0 + X_i\gamma + H_i^T\beta + X_i H_i^T\lambda + e_i, \tag{2}$$

where $e_i \sim N(0, \sigma)$, H_i is the $L \times 1$ haplotype vector with L being the number of distinct haplotypes observed in the population, $\beta_{L\times 1} \sim \mathrm{N}\left(0, \tau R\right)$, $\lambda_{L\times 1} \sim \mathrm{N}(0, \phi R)$, and $R_{L\times L}$ is an $L \times L$ matrix with the

(h,k)-th entry equal to the similarity between haplotypes h and k quantified by the weighted matching score divided the number of loci M. Under the working mixed model (2), the trait covariance between individuals i and j ($i \neq j$) is

$$\begin{aligned} \text{cov}\left(Y_i, Y_j \mid X, H\right) &= H_i^T \text{cov}\left(\beta\right) H_j + X_i H_i^T \text{cov}\left(\lambda\right) H_j X_j \\ &= \tau \times H_i^T R H_j + \phi \times X_i X_j \times H_i^T R H_j \\ &= \tau \times S_{ij} + \phi \times X_i X_j \times S_{ij}. \end{aligned} \tag{3}$$

The last line follows from the fact that $H_i^T R H_j = 1/M \times \sum_{m=1}^{M} G_{m,i}^T W_m G_{m,j}$ [17]. Comparing equations (1) and (3), we have $b = \tau$ and $d = \phi$. That is, the regression coefficients in the similarity regression are the variance components in the mixed model (2). Therefore, following similar derivations in [22] and [45], we obtain the score test statistics for G×E test, G test and the joint test as follows:

$$T_{G\times E} = Y^T P_1 D S D P_1 Y\big|_{\phi=0,\tau=\hat{\tau},\sigma=\hat{\sigma}},$$

$$T_{G} = Y^T P_0 S P_0 Y\big|_{\phi=0,\tau=0,\sigma=\tilde{\sigma}},$$

and

$$T_{\text{joint}} = Y^T P_0 \left(S + DSD\right) P_0 Y\big|_{\phi=0,\tau=0,\sigma=\tilde{\sigma}}.$$

In the above, $Y_{n\times 1}^T = (Y_1, \cdots, Y_n)$, $D_{n\times n} = diag\left\{X_i\right\}$, and $S = \{S_{ij}\}$ with $S_{ij} = H_i^T R H_j$; matrix $P_t = V_t^{-1} - V_t^{-1} X \left(X^T V_t^{-1} X\right)^{-1} X^T V_t^{-1}$, $t = 0, 1$, with $V_0 = \sigma I$, $V_1 = \tau S + \sigma I$. The quantities $(\hat{\tau}, \hat{\sigma})$ are the REML estimates for (τ, σ) obtained under $H_0 : \phi = 0$, and $\tilde{\sigma}$ is the REML estimate for σ under $H_0 : \phi = \tau = 0$. These estimates are given in Appendix A. As shown in Appendix B, these test statistics follow a weighted χ^2 distribution, and the p-values can be calculated using the three-moment approximation [46, 47].

There are a few remarks regarding the similarity-based marker set methods. The similarity regression aggregates marker information through a sum of genotype similarity across markers instead of a sum of

genotypes. Compared to genotype sums, aggregating information through similarity can prevent signals of opposite directions from being cancelled. In addition, because $G_{m,i}$ takes integer or dosage counts and can be of any length, this approach can work with typed and imputed genotype calls, and is applicable to a mixture of different types of variants without having to dichotomize the variants.

Simulation Studies

We performed simulations based on HapMap 3 data to assess the performance of the proposed tests. We obtained a haplotype population consisting of 234 phased haplotypes from chromosome 21 of the CEU samples in HapMap 3. To obtain a variety of risk allele frequencies and LD patterns of a marker set, we defined a marker set as a 10-SNP region, and used a non-overlapping sliding window on chromosome 21 to obtain 1,734 regions. Given a marker-set region, we generated haplotypes for 500 individuals by randomly sampling 500 pairs of haplotypes with replacement from the 234 haplotypes under a Hardy-Weinberg equilibrium assumption. Because the rarest allele frequency we can obtain is $1/234 \approx 0.004$, we used a relatively small sample size (n=500) to assure genetic heterogeneity attributable to rare mutations.

Given a 10-SNP region, the 5th and the 10th SNPs were set to be the risk loci, and their genotypes for individual i are denoted by $\mathcal{G}_{1i}$ and $\mathcal{G}_{2i} \in \{0, 1, 2\}$, respectively. We generated $X_i \sim N(0, 1)$. Then based on the genetic and covariate information of individual i, the trait value Y_i was sampled from a normal distribution with mean $\gamma_0 + \gamma_1 X_i + \gamma_{G_1}\mathcal{G}_{1i} + \gamma_{G_2}\mathcal{G}_{2i} + \gamma_{GE_1} X_i\mathcal{G}_{1i} + \gamma_{GE_2} X_i\mathcal{G}_{2i}$ and variance v^2, where γ_0 and γ_1 were set to be 1, and v^2 was determined so that the heritability was around 0.1 to 0.2. For type I error rate analysis, we set $(\gamma_{G_1}, \gamma_{G_2}, \gamma_{GE_1}, \gamma_{GE_2}) = (0, 0, 0, 0)$ for all 3 tests and also $(0.2, 0.2, 0, 0)$ for $G\times$ E test. For power analysis, we set $(\gamma_{G_1}, \gamma_{G_2}, \gamma_{GE_1}, \gamma_{GE_2}) = (0.25, 0.25, 0.3, 0.3)$. These values were chosen so that the power of the joint tests is not too close to 1, while the power of G×E and G tests is not too close to the nominal level of 0.0005.

Each region was analyzed using the proposed similarity regression with three weighting schemes

considered in the literature[23,24]: (1) $W_m = diag\{1/q_{a,m}\}$ (referred to as SIM1), (2) $W_m = diag\{1/\sqrt{q_{a,m}}\}$ (referred to as SIM2), and (3) $W_m = diag\{1\}$ (referred to as SIM0). The results were compared with two benchmark methods, the single-SNP minimum p-value method (referred to as SNP) and the multi-SNP haplotype based method (referred to as HAP). In all analyses, the two risk loci were excluded and the phase information was removed. For the minimum p-value G×E/G/joint method, we used the minimum of the p-values from the G×E, G, and joint tests for the eight SNPs, and the significance threshold was determined using the multiple testing correction method of Moskvina and Schmidt[48]. This method estimates the effective number of independent tests for correlated SNPs at a given overall type I error rate, and calculates the significance level for the individual tests accordingly. For the haplotype-based analysis, we used the widely used R package haplo.stats to carry out standard haplotype regression analysis. Specifically, we used haplo.glm[49] for the G×E test and haplo.score[50] for the G test. We did not perform the joint test at the haplotype level as it is not supported by this program. Haplotypes with frequencies less than the program default threshold (i.e., 0.01) were pooled into the baseline haplotype.

Results

Simulation Studies

To evaluate type I error rates, we randomly selected 6 of 1,734 regions on chromosome 21 to represent 6 different scenarios: two levels of disease allele frequencies ($q = 0.1$ and 0.3) combined with three levels of LD pattern (high, medium, and low). The LD pattern was summarized using the average of the 16 R^2 values, where each value is the LD between an observed marker (8 in total) and a risk locus (2 in total). A larger LD value reflects stronger correlation between the observed markers and the unobserved risk loci, hence the value reflects the informativeness of the observed markers for the risk loci. Each of the type I error rates was calculated based on 50,000 replications for $(\gamma_{G_1}, \gamma_{G_2}, \gamma_{GE_1}, \gamma_{GE_2}) = (0, 0, 0, 0)$ for all tests and 20,000 replications for $(0.2, 0.2, 0, 0)$ for G$\times$ E test. The results (Figure 1) indicate that the type I error rates were around the nominal levels considered (i.e., $\alpha = 0.05$, 0.005 and 0.0005) for all methods in most scenarios. The exceptions tend to occur in the haplotype G$\times$E tests, where the type I errors can be inflated due to the presence of rare haplotypes. Inflation at larger α levels can often be eliminated by using a slightly higher threshold (e.g., 0.02, as opposed to the default value of 0.01) that pools uncommon haplotypes into the baseline group. To avoid any potential impact that modifying the default threshold might induce, we still used the threshold value of 0.01 in our power analysis.

The power was evaluated for each of the 1,734 regions based on 100 replications at the nominal level of 0.0005. The results are shown in Figure 2 (G$\times$E test), Figure 3 (G test), and Figure 4 (Joint test). The 1,734 regions were grouped into 12 categories: combination of four scenarios of allele frequencies and three LD patterns. The risk allele frequencies from rare to common are categorized as follows: (a) both allele frequencies are less than 0.05, (b) sums of allele frequencies that are less than 0.3 but excluding those in (a), (c) sums of allele frequencies that are between 0.3 and 0.6, and (d) sums of allele frequencies that are greater than 0.6. The clustering of LD patterns is based on the following thresholds: average $R^2 > 0.6$

for High, average $R^2 \in (0.25, 0.6)$ for Medium, and average $R^2 < 0.25$ for Low.

A similar pattern was observed across Figures 2 to 4, hence we concentrate on explaining Figure 2. In regions that exhibit low LD (LD-L), all three methods lacked power and had roughly equal performance. The exception is in column (a), where the SIM1 method performed worse than the other two. The situation that all three methods had similarly low power is not surprising because LD-L represents regions that contained markers with little information about the two risk loci. The lone exception in LD-L (a) can be explained by the fact that the SIM1 method is best applied in scenarios where a large number of markers have at least medium-level LD with the risk loci, but in LD-L (a), such scenario only occurred in 13% of the regions. On the opposite, in 60% of the regions, the majority of the markers had no LD with the risk loci, but either one single marker was in perfect LD with one of the risk loci, or two markers were in very high LD with each of the rick loci. The former cases tend to favor the SNP methods, while the latter tend to favor the HAP methods. (And the remaining 27% were regions where all markers had extremely low LD with the risk loci.) In the scenarios of LD-L with (b), (c) and (d), we did not observe such a large proportion of extreme cases, resulting in a more comparable performance of the three methods. Finally, compared to LD-L, in the regions with medium LD (LD-M), we observed a uniform increase of power in all three methods, with SIM1 having a slightly greater power. The power gain was more pronounced for high LD regions (LD-H), where SIM1 showed more power than the other two methods.

To understand the impact of different weighting schemes in the similarity regression, we repeated the same analysis using SIM1, SIM2 and SIM0 (Figure 5). Because the overall patterns were similar across different tests, we present the results from the G×E and G tests. Figure 5 presents the box plots of power for the same regions as shown previously, except that categories (c) and (d) in Figures 2 to 4 were grouped together to represent common-variant scenarios. We also marked the corresponding average power of SNP (solid line) and HAP (dotted line) for comparison. We observed the following features: (1) SIM0 and SIM2 had very similar power in almost all situations. (2) When risk alleles are common (i.e., panels (c) and (f)),

SIM2 and SIM0 had similar or slightly better power than SIM1, although the difference was not very obvious. (3) When the risk alleles are uncommon or rare, SIM1 started to gain some traction in improving power. The power improvement became more substantial for rarer alleles. For example, in situations with a moderate LD level, SIM1 had higher power than SNP and HAP, while SIM2 and SIM0 did not.

Application to Real Data

We applied the similarity regression on samples collected from the VISP trial. VISP was a multi-center, double blind, randomized, controlled clinical trial that aimed to study the effect of vitamin on preventing recurrent stroke. The trial enrolled patients who were 35 or older with a non-disabling cerebral infarction [MIM 601367] within 120 days of randomization and Hcy levels in the top quartile for the U.S. population. Subjects were randomly assigned to receive daily doses of either a high-dose formulation (containing 25mg vitamin B_6, 0.4mg vitamin B_{12}, and 2.5 mg folic acid), or a low-dose formulation (containing 200μg vitamin B_6, 6μg vitamin B_{12}, and 20μg folic acid). The patients were followed up for a maximum of two years, and the average follow-up time was 1.7 years. About twenty-one hundreds of the VISP participants provided DNA samples, and genotype information was collected from candidate genes selected across the genome that are involved in homocysteine metabolism, stroke risk, and atherosclerosis [MIM 209010]. After quality control , the dataset consists of 1944 subjects, with genotypes of 1393 SNPs collected from 215 candidate genes. More details on the VISP trial and VISP genetic study can be found in Toole et al.[51] and Hsu et al.[52] respectively.

Our analysis here focused on the genetic influence on the Hcy level obtained from a 2-hour methionine load test measured at baseline. It has been suggested that Hcy level can be used to predict risk of recurrent stroke and symptomatic coronary heart disease, and genetic variations may attribute to mild to moderate hyperhomocystinemia [MIM 603174]. Given that the Hcy level tends to increase with age, we also investigated the potential gene-age interaction effects on Hcy. We conducted gene-based analyses—we

used the proposed SIM1 method to assess the significant level of each gene and compared it with available benchmark methods, SNP and/or HAP methods. As did in the original study[52], we adjusted for age, sex and race in each analysis. The Bonferroni threshold for p-value is $0.05/215 = 2.33 \times 10^{-4}$.

We first used the joint test to perform a gene-based scan to evaluate the gene and gene-age effects on the change in post-methionine load Hcy level (i.e., post-methionine load test Hcy - baseline fasting Hcy). If a gene is rejected by a joint test, the G×E and G tests can be used to further refine the sources of identified signals. The joint test is a suitable screening tool for scenarios where the underlying gene-age interaction mechanism is little known[33,44], as it assesses the genetic main effect and gene-age interactions simultaneously. The p-values of the testing results for each gene (sorted by gene names) are shown in Figure 6. For joint tests, there is one gene found significant (*CBS* [MIM 613381]), and both SIM1 and SNP tests yield significant p-values. The p-value of SIM1 joint is 2.46×10^{-5}, and the follow-up analysis reveals that the signal is caused by the genetic main effect instead of gene-age interactions. (The p-value of SIM GxE is 0.614 and the p-value of SIM G is 1.99×10^{-6}). The SNP joint test has the adjusted minimum p-value (adjusted for the 10 typed SNPs in gene CBS) as 2.06×10^{-5}. The adjusted minimum p-value is obtained by $1 - (1-\text{raw p-value})^{k_{eff}}$, where $k_{eff} = 7.59$ is the effective number of independent tests estimated using the method of Moskvian and Schmidt[48] after accounting the LD in gene CBS. The adjusted minimum p-value for SNP GxE test is 0.700, and for SNP G test is 9.42×10^{-6}. Finally, the HAP GxE test yielded a p-value of 0.362, and HAP G test yielded a significant p-value of 1.02×10^{-5}. Variants in the *CBS* gene have previously been associated with post-methionine load Hcy levels and change in Hcy levels[52–55]. A common 68 bp insertion at the intron 7-exon 8 boundary of the CBS gene (844ins68) and the 31 bp variable number of tandem repeats (VNTR) may be genetic determinants of post-methionine load Hcy levels. Because post-methionine load Hcy levels are found to have an increased risk for cardiovascular disease, the CBS gene may be also considered a risk factor for cardiovascular disease.

Discussion

Association analyses at the gene, pathway, and exon levels (here by marker-set analysis) hold great promise in evaluating modest etiological effects of genes using data from GWAS or next-generation sequencing. However, currently available methods tend to target either rare or common variants but not both, assume same-direction effects for loci within a marker set, use a testing framework that cannot accommodate covariates, or do not have the capacity to assess interaction effects. In this article, we propose a flexible, powerful and computationally efficient method to conduct marker-set analysis for assessing gene and gene-environment interactions on quantitative traits. The proposed method is constructed using a similarity regression framework under which we regress trait similarity on genetic similarity. The framework incorporates interaction effects, can adjust for covariates, and is applicable to both observed and imputed dosage genotypes. We develop a series of statistical tests that can be used for genetic marginal main effects, G$\times$E interactions, or the joint effect of the two. We demonstrated that a similarity regression is equivalent to a haplotype random effect model. The equivalence enabled us to analytically derive the asymptotic distributions of the test statistics and provide a permutation-free procedure to assess significance. The software implementing the proposed methods is available at the authors' website.

The proposed method uses genetic similarity to aggregate information across markers, and integrates adaptive weights dependent on allele frequencies to accommodate common and uncommon variants. Collapsing information at the similarity level instead of the genotype level avoids canceling signals with opposite etiological effects, and is applicable to any class of genetic variants without having to dichotomize the allele types. As demonstrated in the simulation, incorporating frequency weights gives the method satisfactory power for detecting both common and uncommon variants. The simulation results also reveal that its performance is sensitive to the "signal-to-noise" ratio (e.g., LD) among all loci included in the marker-set analysis. The higher the ratio is, the greater the power gain for the proposed methods. As discussed in the next paragraph, it is feasible to increase the signal-to-noise ratio to maximize the chance of

power gain, such as using functional, biological or LD information to down-weight the contribution from noise markers. In practice, the underlying LD levels are not known and will vary from regions to regions, it is less likely to choose one best performing method in advance. In addition, in GWAS, the low LD scenario would occur less frequently by design, and in sequencing studies, the number of risk loci in a set should be higher than what we considered in the simulation. Given these considerations, the proposed method can serve as a sensible and robust tool for evaluating association of complex traits in whole genome marker-set analyses.

The inclusion of nonfunctional loci (i.e., non-risk markers that are in not LD with the risk loci) is a major factor influencing the performance of all marker-set approaches. Intelligently incorporating LD information and biological knowledge into the collapsing process and down-weighting the contribution of nonfunctional markers will be a useful solution. In our framework, biological and functional information, as pioneered and comprehensively reviewed in [10,41], can be naturally incorporated through the weight matrix W_m. One unique feature of our weighting framework is that it allows functional weights at the allele-specific level (as opposed to locus-specific level), such as the impact of specific mutation sequence on protein functions, structures or stability. We are exploring mechanisms to include genomic knowledge based on functionality, biological pathways and system biological networks.

One key factor for the proposed method to have power for both common and uncommon variants is to weight the similarity level by allele frequency at order k (i.e., q^{-k}). Although the principle is to upweight similarities that are contributed by rare variants, there are no clear rules for what the specific form of the weights should be as a function of the allele frequencies. Kwee et al.[23] considered both $k = 1$ and $k = 1/2$ when calculating the IBS kernel, and concluded that the former may be too strong and the latter is more suitable in their setting. Wu et al.[24] therefore used $k = 1/2$ in their work. When aggregating information of multiple loci through weighted genotype sum, Madsen and Browning[8]) considered their weights in the order of $k = 1/2$ from the binomial standard deviation (SD) viewpoint. Here we evaluated these different

choices of k under our framework (i.e., SIM1 $(k = 1)$, SIM2 $(k = 1/2)$, and SIM0 $(k = 0)$. We found that SIM2 may be too mild and tends to yield similar results as the unweighted SIM0. One main difference between our weighting framework and others is that we assign weights for every allele, while others only assign weights for minor alleles. So instead of a $q_{minor}^{-1/2} : q_{major}^{-1/2}$ ratio between the minor allele and the major allele, those weights placed only on minor alleles yield a $q_{minor}^{-1/2} : 1$ ratio and give a bigger contrast.

Simulation results also suggest that larger values of k can greatly boost power for detecting rare variants, but it also risks losing power when the risk variant is common. We focused on SIM1 based on its superior power for rare variants and comparable power for common variants. It is possible that the optimal weights would lie somewhere between $k = 1$ and $k = 1/2$, and we are investigating further how to identify an optimal order. Alternatively, one can use centered genotype scoring to account for sharing of rarer alleles[57]. To center the allele count vector $G_{m,i}$, define $G^*_{m,i} = G_{m,i} - \bar{G}_m$, where $\bar{G}_m$ is the vector of population allele frequency for marker m. Then the similarity score S^*_{ij} is obtained by $1/M \times \sum_{m=1}^{M} G^{*T}_{m,i} G^*_{m,j}$. The centering strategy bypasses the need of allele-frequency-dependent weights and hence avoids the choice of an order k. Studies to understand the pros and cons of centering vs. weighting strategies are underway.

Acknowledgments

The authors thank the two anonymous reviewers for their constructive comments, and Drs. Alison Motsinger-Reif and Dmitri Zaykin for their helpful discussions. This work was supported by National Institutes of Health grants R01 MH074027 (to JYT, DZ, MP, CS, and PFS), P01 CA142538 (to JYT), R37 AI031789-20 (to DZ), and R01 CA85848 (to DZ).

Appendix A: EM Algorithm for the REML Estimates of τ and σ When Testing for G×E $H_0 : \phi = 0$

Let $u = K^T Y$ be a set of $n - d$ linearly independent contrasts of Y with $KK^T = I - X\left(X^T X\right)^{-1} X^T$ and $K^T K = I_{n\times n}$. Then the conditional distribution of u given β, denoted by $f\left(u|\beta\right)$, is Normal with mean $K^T H\beta$ and variance σI and does not depend on the fixed effect γ.Therefore, the REML estimations of τ and σ can be based on its marginal distribution $f\left(u\right) = \int f\left(u|\beta\right) f\left(\beta\right) d\beta$. This motivated an EM algorithm based on observed data u and missing data β. The complete-data log likelihood is given by

$$\begin{aligned}\log f\left(u,\beta;\tau,\sigma\right) &= \log f\left(u \mid \beta;\tau,\sigma\right) + \log f\left(\beta;\tau,\sigma\right)\\ &= -\frac{n-d}{2}\log\sigma - \frac{1}{2\sigma}\left(u - K^T H\beta\right)^T\left(u - K^T H\beta\right)\\ &\quad -\frac{L}{2}\log\tau - \frac{1}{2}\log|R| - \frac{1}{2\tau}\beta^T R^{-1}\beta.\end{aligned}$$

In the expectation step (E-step), we compute $Q\left(\tau,\sigma;\hat{\tau}^{(t)},\hat{\sigma}^{(t)}\right)$, the conditional expected value of $\log f\left(u,\beta;\tau,\sigma\right)$ given the observed data u assuming $(\tau,\sigma) = \left(\hat{\tau}^{(t)},\hat{\sigma}^{(t)}\right)$, where $\hat{\tau}^{(t)}$ and $\hat{\sigma}^{(t)}$ are the estimates at the t-th iteration.

$$\begin{aligned}Q\left(\tau,\sigma;\hat{\tau}^{(t)},\hat{\sigma}^{(t)}\right) &\equiv \mathrm{E}\left[\log f\left(u,\beta;\tau,\sigma\right) \mid u;\hat{\tau}^{(t)},\hat{\sigma}^{(t)}\right]\\ &= -\frac{n-d}{2}\log\sigma - \frac{1}{2\sigma}\mathrm{E}\left[\left(u - K^T H\beta\right)^T\left(u - K^T H\beta\right) \mid u;\hat{\tau}^{(t)},\hat{\sigma}^{(t)}\right]\\ &\quad -\frac{L}{2}\log\tau - \frac{1}{2}\log|R| - \frac{1}{2\tau}\mathrm{E}\left[\beta^T R^{-1}\beta \mid u;\hat{\tau}^{(t)},\hat{\sigma}^{(t)}\right].\end{aligned}$$

In the maximization step (M-step), we solve for $\partial Q/\partial\tau = 0$ and $\partial Q/\partial\sigma = 0$ and obtain

$$\hat{\tau}^{(t+1)} = \frac{1}{L}\mathrm{E}\left[\beta^T R^{-1}\beta \mid u;\hat{\tau}^{(t)},\hat{\sigma}^{(t)}\right] = \frac{1}{L}\tilde{\beta}R^{-1}\tilde{\beta}^{-1} + \mathrm{tr}\left(R^{-1}\tilde{W}\right),$$

and

$$\hat{\sigma}^{(t+1)} = \frac{1}{n-d}\mathrm{E}\left[\left(u - K^T H\beta\right)^T\left(u - K^T H\beta\right) \mid u;\hat{\tau}^{(t)},\hat{\sigma}^{(t)}\right] = \left(Y - H\tilde{\beta}\right)^T A\left(Y - H\tilde{\beta}\right) + \mathrm{tr}\left(H^T AH\,\tilde{W}\right).$$

In the above equations, $A = KK^T = I - X\left(X^TX\right)^{-1}X^T$, $\tilde{\beta} \equiv \mathrm{E}\big(\beta|u;;\hat{\tau}^{(t)},\hat{\sigma}^{(t)}\big) = \tau RH^TP_1Y$, and $\tilde{W} \equiv \mathrm{var}\big(\beta|u;\hat{\tau}^{(t)},\hat{\sigma}^{(t)}\big) = \tau R - \tau^2RH^TPHR$. The conditional moments of β given u are obtained directly from the normality of the joint distribution of (u,β) .The calculation of the project matrix P_1 requires inverting the $n \times n$ non-sparse matrix V_1, which can be computational burdensome. To speed up, we rewrite

$$V_1 = \tau S + \sigma I = \sigma\left\{I + \frac{\tau}{\sigma}S\right\} = \sigma\left\{I + \frac{\tau}{\sigma}E\Lambda E^T\right\}$$

where $S = E\Lambda E^T$ the eigenvalue decomposition of matrix S. Then by the fact that $(I + B_1B_2)^{-1} = I - B_1\left(I + B_2B_1\right)^{-1}B_2$, $V_1^{-1} = \frac{1}{\sigma}\left\{I - \frac{\tau}{\sigma}E\Lambda\left[I + E^T\frac{\tau}{\sigma}E\Lambda\right]^{-1}E^T\right\}$ $= \frac{1}{\sigma}\left\{I - \tau E\Lambda\left[\sigma I + \tau E^TE\Lambda\right]^{-1}E^T\right\}$, in which the calculation involves only an inversion of an $L \times L$ matrix.

Appendix B: Derivation of the Score Test Statistics and Their Asymptotic Distribution

For quantitative traits that follow a normal distribution directly or after appropriate transformations, model (2) reduces to the following linear mixed model (LMM) in matrix notation

$$\mathbf{Y} = \mathbf{1}\gamma_0 + X\gamma + H\beta + DH\lambda + \varepsilon, \text{ with } \beta \sim \mathrm{N}\left(0, \tau R\right), \ \lambda \sim \mathrm{N}\left(0, \phi R\right), \text{ and } \varepsilon \sim \mathrm{N}\left(0, \sigma I\right) \quad (4)$$

where $Y^T = [Y_1, \cdots, Y_n]$, $\mathbf{1}$ is an $n \times 1$ vector of 1's, $X^T = [X_1, \cdots, X_n]$, $D = diag\{X_i\}$, and $\varepsilon_{n\times 1}^T = [e_1, \cdots, e_n]$. Since our primary interest is to test the variance components ϕ and τ, we consider the restricted maximum likelihood (REML) log-likelihood function of variance components (τ, ϕ, σ) : $\ell_{REML}(\tau, \phi; Y) = -\left\{\log|V| + \log|X^T V^{-1} X| + Y^T P Y\right\}/2$, where V is the marginal variance of Y and $V = \phi\Sigma + \tau S + \sigma I$, with $S = HRH^T$ and $\Sigma = DSD$; $P = V^{-1} - V^{-1}X(X^T V^{-1} X)^{-1} X^T V^{-1}$ is the projection matrix for the LMM (4).

Let $U_\phi(\phi, \tau, \sigma)$ and $U_\tau(\phi, \tau, \sigma)$ denote the score functions based on the REML function for ϕ and τ, respectively. Simple algebra[56] shows that under $H_0 : \phi = 0$,

$$U_\phi(0, \hat{\tau}, \hat{\sigma}) = \left.\frac{\partial \ell_{REML}(\tau, \phi, \sigma)}{\partial \phi}\right|_{\phi=0, \tau=\hat{\tau}, \sigma=\hat{\sigma}} = \frac{1}{2}\left\{Y^T P_1 \Sigma P_1 Y - \mathrm{tr}(P_1\Sigma)\right\}, \quad (5)$$

and under $H_0 : \tau = 0$ (and with the constrain of $\phi = 0$),

$$U_\tau(0, 0, \hat{\sigma}) = \left.\frac{\partial \ell_{REML}(\tau, \phi, \sigma)}{\partial \tau}\right|_{\phi=0, \tau=0, \sigma=\tilde{\sigma}} = \frac{1}{2}\left\{Y^T P_0 S P_0 Y - \mathrm{tr}(P_0 S)\right\}. \quad (6)$$

In the above equations, $(\tilde{\tau}, \tilde{\sigma})$ are the REML estimates of (τ, σ) under $H_0 : \phi = 0$ as given in Appendix A, and $\tilde{\sigma}$ the REML estimate of σ when $\tau = \phi = 0$. Recall that $P_t = V_t^{-1} - V_t^{-1} X \left(X^T V_t^{-1} X\right)^{-1} X^T V_t^{-1}$ with $t \in \{0, 1\}$, and $V_1 = \tau S + \sigma I$ and $V_0 = \sigma I$.

Null Distribution of the score statistics for G×E test and G test. As shown in[22], the score statistics under H_0 are not asymptotically normal because the design matrix H for the random effects β is not block

diagonal and the dimension of β is fixed. We hence use the first terms of the score statistics as the testing statistics and obtain $T_{G\times E} = Y^T P_1 \Sigma P_1 Y/2$ and $T_G = Y^T P_0 S P_0 Y/2$. Below we derive the asymptotic null distribution of $T_{G\times E}$, and similar steps can be used to obtain the distribution for T_G. Defining the vector $\mathbf{Z} = V^{-\frac{1}{2}}(\mathbf{Y} - \mu)$ with $\mu = \mathbf{1}\gamma_0 - X\gamma$, then $\mathbf{Z}$ follows a standard multivariate normal distribution. Rewrite $T_{G\times E} = \mathbf{Z}^T \left(\frac{1}{2} V^{\frac{1}{2}} P_1 \Sigma P_1 V^{\frac{1}{2}}\right) \mathbf{Z} \equiv \mathbf{Z}^T C_{G\times E} \mathbf{Z}$, which is true because $\mu^T P_1 = 0$ by the fact of P_1 being a projection matrix. Define e_i and η_i the eigenvector and eigenvalue of matrix $C_{G\times E}$ respectively. Then $T_{G\times E} = \sum_{i=1}^{c} \eta_i \left(e_i^T Z\right)^2 \equiv \sum_{i=1}^{L} \eta_i \tilde{Z}_i^2$, with $\tilde{Z}_i^2$ follows a 1 degree-of-freedom chi-square distribution. In reality, (τ, σ) is evaluated at their rest ricted maximum likelihood estimates $(\hat{\tau}, \hat{\sigma})$. Following[22], the distribution of $T_{G\times E}$ can be approximated by the distribution of $\sum_{i=1}^{c} \hat{\eta}_i \chi_{i1}^2$, where $\hat{\eta}_i$'s are the non-zero eigenvalues of matrix $C_{G\times E}|_{\tau=\hat{\tau},\sigma=\hat{\sigma}}$. The distribution of $T_{G\times E}$ can be approximated by the three-moment approximation method of[46]. The level-α significance threshold is estimated by $\kappa_1 + (\chi_\alpha - h') \times \sqrt{\frac{\kappa_2}{h'}}$, where $\kappa_j = \sum_i \eta_i^j$, $h' = \kappa_2^3/\kappa_3^2$ 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 $T_{G\times E}$ by $P\left(\chi_{h'}^2 > \chi^*\right)$, where $\chi^* = (T_{G\times E} - \kappa_1) \times \sqrt{h'/\kappa_2} + h'$.

By the same manner, the distribution of T_G can also be approximated by the three-moment approximation as above, except that the eigenvalues η_is are obtained from matrix $C_G = \frac{1}{2} V_0^{\frac{1}{2}} P_0 S P_0 V_0^{\frac{1}{2}}\Big|_{\sigma=\tilde{\sigma}}$.

Null Distribution of the score statistics for joint test. The test statistic for the joint hypothesis $H_0 : \phi = \tau = 0$ is $T_{\text{joint}} = T_G + T_{G\times E}^{(0)}$, where T_G is defined as before and $T_{G\times E}^{(0)} = \frac{1}{2} Y^T P_0 \Sigma P_0 Y$, i.e., $T_{G\times E}$ evaluated at $\phi = \tau = 0$ and $\sigma = \tilde{\sigma}$. A direct (unweighted) sum is used here because X has been pre-standardized to mean 0 and variance 1 and hence T_G and $T_{G\times E}$ are on the same scale. We found that its performance is very similar to a weighted sum version, $T_{\text{joint}}^{wt} = w_G \times T_G + w_{G\times E} \times T_{G\times E}^{(0)}$, where the weights $w_i = E(T_i)/var(T_i)$. By the similar derivation as in the G$\times$ E test, it can be shown that the null distribution of T_{joint} also has a weighted chi-square distribution, and can be approximated by the

three-moment approximation. The procedure is the same as what mentioned for the G$\times$ E test, except that the eigenvalues should be obtained from the matrix $C_{joint} = \frac{1}{2} V_0^{\frac{1}{2}} P_0 (S + \Sigma) P_0 V_0^{\frac{1}{2}} \Big|_{\sigma=\tilde{\sigma}}$.

Web Resources

The URLs for data presented herein are as follows:

UCSC genome browser (NCBI36/hg18), http://genome.ucsc.edu/cgi-bin/hgTracks

Authors' web site, http://www4.stat.ncsu.edu/˜tzeng/software.php

Online Mendelian Inheritance in Man (OMIM), http://www.omim.org.

References

1. De la Cruz, O., Wen, X., Ke, B., Song, M., and Nicolae, D.L. (2010). Gene, region and pathway level analyses in whole-genome studies. Genet. Epidemiol. *34*, 222-231.

2. Fisher, R.A. (1932). Statistical methods for research workers (London: Oliver and Boyd).

3. Li, M., Wang, K., Grant, S.F., Hakonarson, H., and Li, C. (2008). ATOM: a powerful gene-based association test by combining optimally weighted markers. Bioinformatics *25*, 497-503.

4. Wang, T., and Elston, R.C. (2007). Improved power by use of a weighted score test for linkage disequilibrium mapping. Am. J. Hum. Genet. *80*, 353-60.

5. Gauderman, W.J., Murcray, C., Gilliland, F., and Conti, D.V. (2007). Testing association between disease and multiple SNPs in a candidate gene. Genet. Epidemiol. *31*, 383-395.

6. Wang, K., and Abbott, D. (2008). A principal components regression approach to multilocus genetic association studies. Genet. Epidemiol. *32*, 108-18.

7. Li, B., and Leal, S.M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am. J. Hum. Genet. *83*, 311-321.

8. Madsen, B.E., and Browning, S.R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. *5*, e1000384.

9. Morgenthaler, S., and Thilly, W.G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat. Res. *615*, 28-56.

10. Price, A.L., Kryukov, G.V., de Bakker, P.I., Purcell, S.M., Staples, J., Wei, L.J., and Sunyaev, S.R. (2010). Pooled association tests for rare variants in exon-resequencing studies. Am. J. Hum. Genet. *86*, 832-838.

11. Tzeng, J.Y., Byerley, W., Devlin, B., Roeder, K. and Wasserman, L. (2003a) Outlier detection and false discovery rates for whole-genome DNA matching. J Am Stat Assoc *98*, 236-246.

12. Tzeng, J.Y., Devlin, B., Wasserman, L., and Roeder, K. (2003b). On the identification of disease mutations by the analysis of haplotype similarity and goodness-of-fit. Am. J. Hum. Genet. *72*, 891-902.

13. Schaid, D.J., McDonnell, S.K., Hebbring, S.J., Cunningham, J.M., and Thibodeau, S.N. (2005). Nonparametric tests of association of multiple genes with human disease. Am. J. Hum. Genet. *76*, 780-93.

14. Beckmann, L., Thomas, D.C., Fischer, C., and Chang-Claude, J. (2005). Haplotype sharing analysis using mantel statistics. Hum. Hered. *59*, 67-78.

15. Wessel, J., and Schork, N.J. (2006). Generalized genomic distance-based regression methodology for multilocus association analysis. Am. J. Hum. Genet. *79*, 792-806.

16. Dempfle, A., Hein, R., Beckmann, L., Scherag, A., Nguyen, T.T., Schäfer, H., and Chang-Claude, J. (2005). Comparison of the power of haplotype-based versus single- and multilocus association methods for gene x environment (gene x sex) interactions and application to gene x smoking and gene x sex interactions in rheumatoid arthritis. BMC Proc. *1 Suppl 1*,S73.

17. Tzeng, J.Y., Zhang, D., Chang, S.M., Thomas, D.C., and Davidian, M. (2009). Gene-trait similarity regression for multimarker-based association analysis. Biometrics *65*, 822-32.

18. Mukhopadhyay, I., Feingold, E., Weeks, D.E., and Thalamuthu, A. (2010). Association tests using kernel-based measures of multi-locus genotype similarity between individuals. Genet. Epidemiol. *34*, 213-221.

19. Wei, Z., Li, M., Rebeck, T., and Li, H. (2008). U-statistics-based tests for multiple genes in genetic association studies. Ann. Hum. Genet. *72*, 821-833.

20. Goeman, J.J., van de Geer, S.A., de Kort, F., van Houwelingen, H.C. (2004). A global test for groups of genes: testing association with a clinical outcome. Bioinformatics *20*, 93-99.

21. Goeman, J.J., van de Geer, S.A., and van Houwelingen, H.C. (2005). Testing against a high dimensional alternative. J R Stat Soc Series B Stat Methodol *68*, 477-493.

22. Tzeng, J.Y., and Zhang, D. (2007). Haplotype-based association analysis via variance component score test. Am. J. Hum. Genet. *81*, 927-938.

23. Kwee, L.C., Liu, D., Lin, X., Ghosh, D., and Epstein, M.P. (2008). A powerful and flexible multilocus association test for quantitative traits. Am. J. Hum. Genet. *82*, 386-397.

24. Wu, M.C., Kraft, P., Epstein, M.P., Taylor, D.M., Chanock, S.J., Hunter, D.J., and Lin, X. (2010). Powerful SNP-set analysis for case-control genome-wide association studies. Am. J. Hum. Genet. *86*, 929-42.

25. Schaid, D.J. (2010a). Genomic similarity and kernel methods I: advancements by building on mathematical and statistical foundations. Hum. Hered. *70*, 109-131.

26. Neale, B.M., Rivas, M.A., Voight, B.F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S.,, Purcell, S.M., Roeder, K., and Daly, M.J. (2011). Testing for an Unusual Distribution of Rare Variants. PLoS Genet *7*:e1001322.

27. Luan, Y., and Li, H. (2008). Group additive regression models for genomic data analysis. Biostatistics *9*, 100-113.

28. Chatterjee, N., Kalaylioglu, Z., Moslehi, R., Peters, U., and Wacholder, S. (2006). Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. Am. J. Hum. Genet. *79*, 1002-1016.

29. Zhao, J., Boerwinkle, E., and Xiong, M. (2008). An entropy-based statistic for genomewide association studies. Am. J. Hum. Genet. *77*, 27-40.

30. Zhao, J., Boerwinkle, E., Xiong, M. (2005). An entropy-based statistic for genomewide association studies. Am. J. Hum. Genet. *77*, 27-40.

31. Dempfle, A., Scherag, A., Hein, R., Beckmann, L., Chang-Claude, J., and Schäfer, H. (2008). Gene-environment interactions for complex traits: definitions, methodological requirements and challenges. Eur. J. Hum. Genet. *16*, 1164-1172.

32. Thomas, D. (2010). Gene-environment-wide association studies: emerging approaches. Nat. Rev. Genet. *11*, 259-272.

33. Lindström, S., Yen, Y.C., Spiegelman, D., and Kraft, P. (2009). The impact of gene-environment dependence and misclassification in genetic association studies incorporating gene-environment interactions. Hum. Hered. *68*, 171-181.

34. Smith, P.G., and Day, N.E. (1984). The design of case-control studies: the influence of confounding and interaction effects. Int. J. Epidemiol. *13*, 356-65.

35. Thomas, D. (2010). Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. Annu. Rev. Public. Health. *31*, 21-36.

36. Schork, N.J., Wessel, J., and Malo, N. (2008). DNA sequence-based phenotypic association analysis. Adv. Genet. *60*, 195-217.

37. Ballard, D.H., Cho, J., and Zhao, H. (2010). Comparisons of multi-marker association methods to detect association between a candidate region and disease. Genet. Epidemiol. *34*, 201-12.

38. Chapman, J., and Whittaker, J. (2008). Analysis of multiple SNPs in a candidate gene or region. Genet. Epidemiol. *32*, 560-566.

39. Fridley, B.L., Jenkins, G.D., and Biernacka, J.M. (2010). Self-contained gene-set analysis of expression data: an evaluation of existing and novel methods. PLoS ONE *5*, e12693.

40. Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661-78.

41. Schaid, D.J. (2010b). Genomic similarity and kernel methods II: methods for genomic information. Hum. Hered. *70*, 132-140.

42. Cordell, H.J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum. Mol. Genet. *11*, 2463-2468.

43. Hirschhorn, J.N., and Daly, M.J. (2005). Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. *6*, 95-108.

44. Kraft, P., Yen, Y.C., Stram, D.O., Morrison, J., and Gauderman, W.J. (2007). Exploiting gene-environment interaction to detect genetic associations. Hum. Hered. *63*, 111-119.

45. Zhang, D., and Lin, X. (2003). Hypothesis testing in semiparametric additive mixed models. Biostatistics *4*, 57-74.

46. Pearson, E.S. (1959). Note on an approximation to the distribution of non-central χ^2. Biometrika*46*, 364.

47. Imhof, J.P. (1961). Computing the Distribution of Quadratic Forms in Normal Variables. Biometrika*48*, 419-426.

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

49. Lake, S.L., Lyon, H., Tantisira, K., Silverman, E.K., Weiss, S.T., Laird, N.M., and Schaid, D.J. (2003). Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. Hum. Hered. *55*, 56-65.

50. 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. Am. J. Hum. Genet. *70*, 425-34.

51. Toole, J.F., Malinow, M.R., Chambless, L.E., Spence, J.D., Pettigrew, L.C., Howard, V.J., Sides, E.G., Wang, C.H., and Stampfer, M. (2004) Lowering homocysteine in patients with ischemic stroke to prevent recurrent stroke, myocardial infarction, and death: the Vitamin Intervention for Stroke Prevention (VISP) randomized controlled trial. J. Am. Med. Assoc. *291*, 565-575.

52. Hsu, F.C., Sides, E.G., Mychaleckyj, J.C., Worrall, B.B., Elias, G.A., Liu, Y, M.D., Chen, W.M., Coull, B.M., Toole, J.F., Rich, S.S., et al. (2011) A Transcobalamin 2 gene variant associated with post-stroke homocysteine modifies recurrent stroke risk. Submitted.

53. Tsai, M.Y., Yang, F., Bignell, M., Aras, O., and Hanson, N.Q. (1999) Relation between plasma homocysteine concentration, the 844ins68 variant of the cystathionine beta-synthase gene, and pyridoxal-5'-phosphate concentration. Mol. Genet. Metab. *67*, 352-356.

54. Lievers, K.J., Kluijtmans, L.A., Heil, S.G., Boers, G.H., Verhoef, P., van Oppenraay-Emmerzaal, D., den Heijer, M., Trijbels, F.J., Blom, H.J. (2001) A 31 bp VNTR in the cystathionine beta-synthase (CBS) gene is associated with reduced CBS activity and elevated post-load homocysteine levels. Eur J Hum Genet. *9*, 583-589.

55. Lievers, K.J., Kluijtmans, L.A., Blom, H.J., Wilson, P.W., Selhub, J., and Ordovas, J.M. (2006)

Association of a 31 bp VNTR in the CBS gene with postload homocysteine concentrations in the Framingham Offspring Study. Eur. J. Hum. Genet. *14*, 1125-1129.

56. Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and related problems. J. Am. Stat. Assoc. *72*, 322-340.

57. Qian, D., and Thomas, D.C. (2001). Genome scan of complex traits by haplotype sharing correlation. Genet. Epidemiol. *21 Suppl 1*, S582-S587.

Figure Legend

Figure 1: The type I error rates shown on the scale of 10^2, 10^3 and 10^4 for nominal level $\alpha = 0.05$, 0.005, and 0.0005, respectively. The regions are randomly selected from chromosome 21 to represent 6 different scenarios listed on the X axis: two levels of disease allele frequencies ($q = 0.1$ and 0.3) combined with three levels of LD pattern (high, medium, and low). The LD pattern is summarized using the average of the 16 R^2 values, where each value was the LD between an observed marker (8 in total) and a risk locus (2 in total). A high LD value reflects stronger correlation between the observed markers and the two unobserved risk loci. The column titles indicate the value of $(\gamma_{G_1}, \gamma_{G_2}, \gamma_{GE_1}, \gamma_{GE_2})$, i.e., the effect sizes of the genetic main effects and gene-environment interactions at the two risk loci used in generating simulated data. Each of the type I error rates is calculated based on 50,000 replications for $(\gamma_{G_1}, \gamma_{G_2}, \gamma_{GE_1}, \gamma_{GE_2}) = (0, 0, 0, 0)$ and $20,000$ replications for $(0.2, 0.2, 0, 0)$. The type I error rates for HAP-G at $\alpha = 0.0005$ are given below as some are beyond the plotting range: (0.00454, 0.00266, 0.0023, 0.00158, 0.00794, 0.00072).

Figure 2: Boxplot of power of G×E test from the 1,734 regions on chromosome 21. The "×" sign indicates the average power. The power at a region is calculated based on 100 replications at a nominal level 0.0005. The results are grouped into 12 categories based on allele frequencies of the risk alleles and LD patterns. The risk allele frequencies from rare to common are categorized: (a) both allele frequencies < 0.05; (b) sums of allele frequencies < 0.3 but excluding (a); (c) sums of allele frequencies between 0.3 and 0.6; and (d) sums of allele frequencies > 0.6. The clustering of LD patterns is done according to the following thresholds: average $R^2 > 0.6$ for High (LD-H), average $R^2 \in (0.25, 0.6)$ for Medium (LD-M), and average $R^2 < 0.25$ for Low (LD-L).

Figure 3: Boxplot of power of G test from the 1,734 regions on chromosome 21. The "×" sign indicates the average power. The power at a region is calculated based on 100 replications at a nominal level 0.0005. The results are grouped into 12 categories based on allele frequencies of the risk alleles and LD patterns.

The risk allele frequencies from rare to common are categorized: (a) both allele frequencies < 0.05; (b) sums of allele frequencies < 0.3 but excluding (a); (c) sums of allele frequencies between 0.3 and 0.6; and (d) sums of allele frequencies > 0.6. The clustering of LD patterns is done according to the following thresholds: average $R^2 > 0.6$ for High (LD-H), average $R^2 \in (0.25, 0.6)$ for Medium (LD-M), and average $R^2 < 0.25$ for Low (LD-L).

Figure 4: Boxplot of power of joint test from the 1,734 regions on chromosome 21. The "$\times$" sign indicates the average power. The power at a region is calculated based on 100 replications at a nominal level 0.0005. The results are grouped into 12 categories based on allele frequencies of the risk alleles and LD patterns. The risk allele frequencies from rare to common are categorized: (a) both allele frequencies < 0.05; (b) sums of allele frequencies < 0.3 but excluding (a); (c) sums of allele frequencies between 0.3 and 0.6; and (d) sums of allele frequencies > 0.6. The clustering of LD patterns is done according to the following thresholds: average $R^2 > 0.6$ for High (LD-H), average $R^2 \in (0.25, 0.6)$ for Medium (LD-M), and average $R^2 < 0.25$ for Low (LD-L).

Figure 5: Boxplot of power of G$\times$E test and G test with different weights (SIM1, SIM2 and SIM0) from the 1,734 regions on chromosome 21. The "$\times$" sign indicates the average power of the method shown on the X axis. The solid and dotted lines indicate the average power of SNP test and HAP test, respectively. The power at a region is calculated based on 100 replications at a nominal level 0.0005. The results are grouped into 9 categories based on allele frequencies of the risk alleles and LD patterns. The risk allele frequencies from rare to common are categorized: (a) both allele frequencies < 0.05; (b) sums of allele frequencies < 0.3 but excluding (a); (c) sums of allele frequencies > 0.3. The clustering of LD patterns is done according to the following thresholds: average $R^2 > 0.6$ for High (LD-H), average $R^2 \in (0.25, 0.6)$ for Medium (LD-M), and average $R^2 < 0.25$ for Low (LD-L).

Figure 6: The negative log 10 transformation of the p-values for the VISP trial analysis. The X axis shows the gene IDs sorted by the alphabetic order of the gene names, and gene ID 39 is gene *CBS*. The red line indicates results for SIM1, "+" for SNP method, and "x" for HAP method. The results for the SNP methods are based on the adjusted minimum p-values that adjust for the multiple SNPs in a gene. The adjusted minimum p-value is obtained by $1-(1-\text{raw p-value})^{k_{eff}}$, where k_{eff} is the effective number of independent tests estimated using the method of Moskvian and Schmidt[48] after accounting the LD among SNPs in a gene. A few genes are not plotted on the graph for the HAP methods due to convergence failure at these locations. This failure is mostly attributed to excessive number of SNPs in the gene.

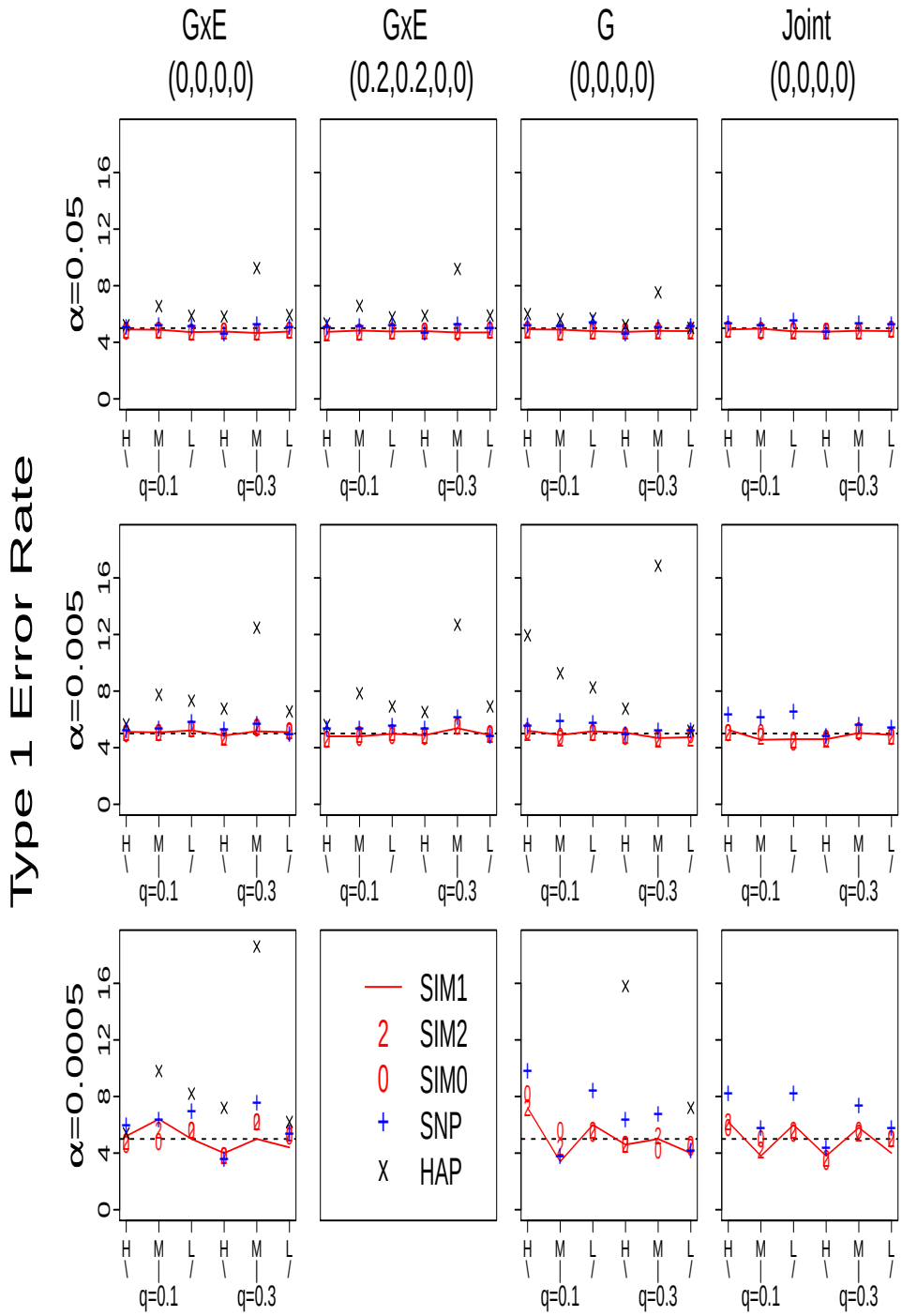

Figure 1: Type I error rates.

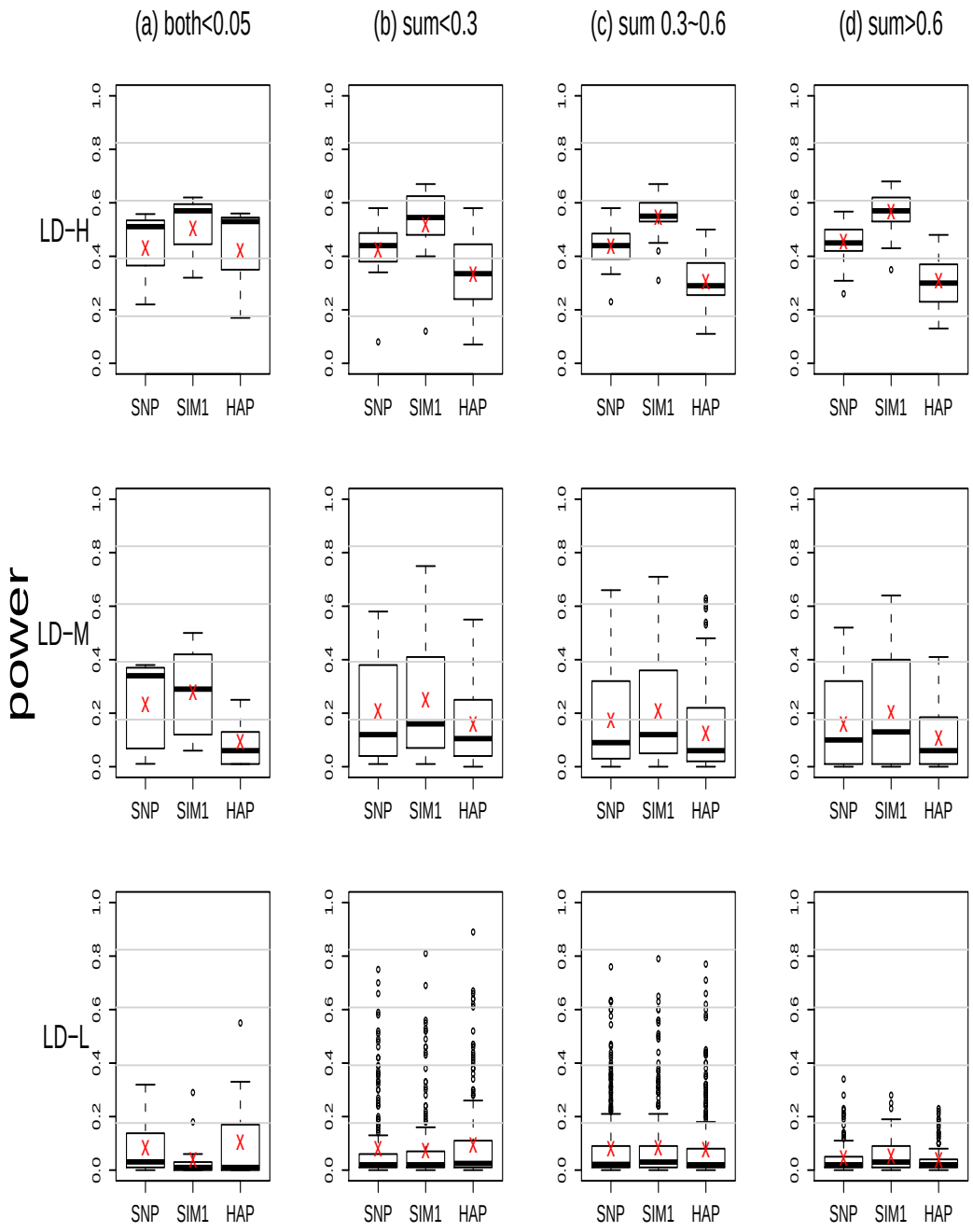

Figure 2: Power of GxE tests.

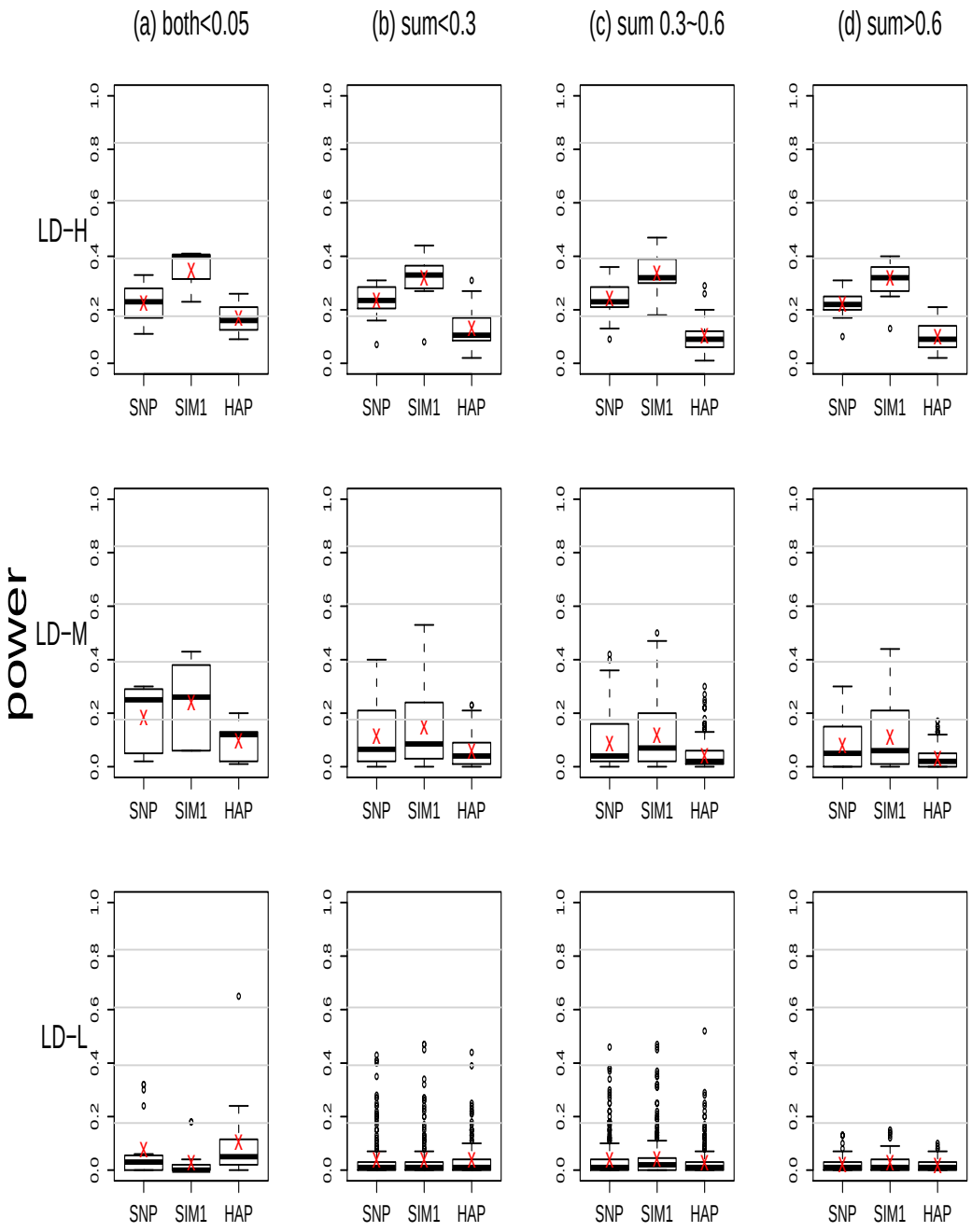

Figure 3: Power of G tests.

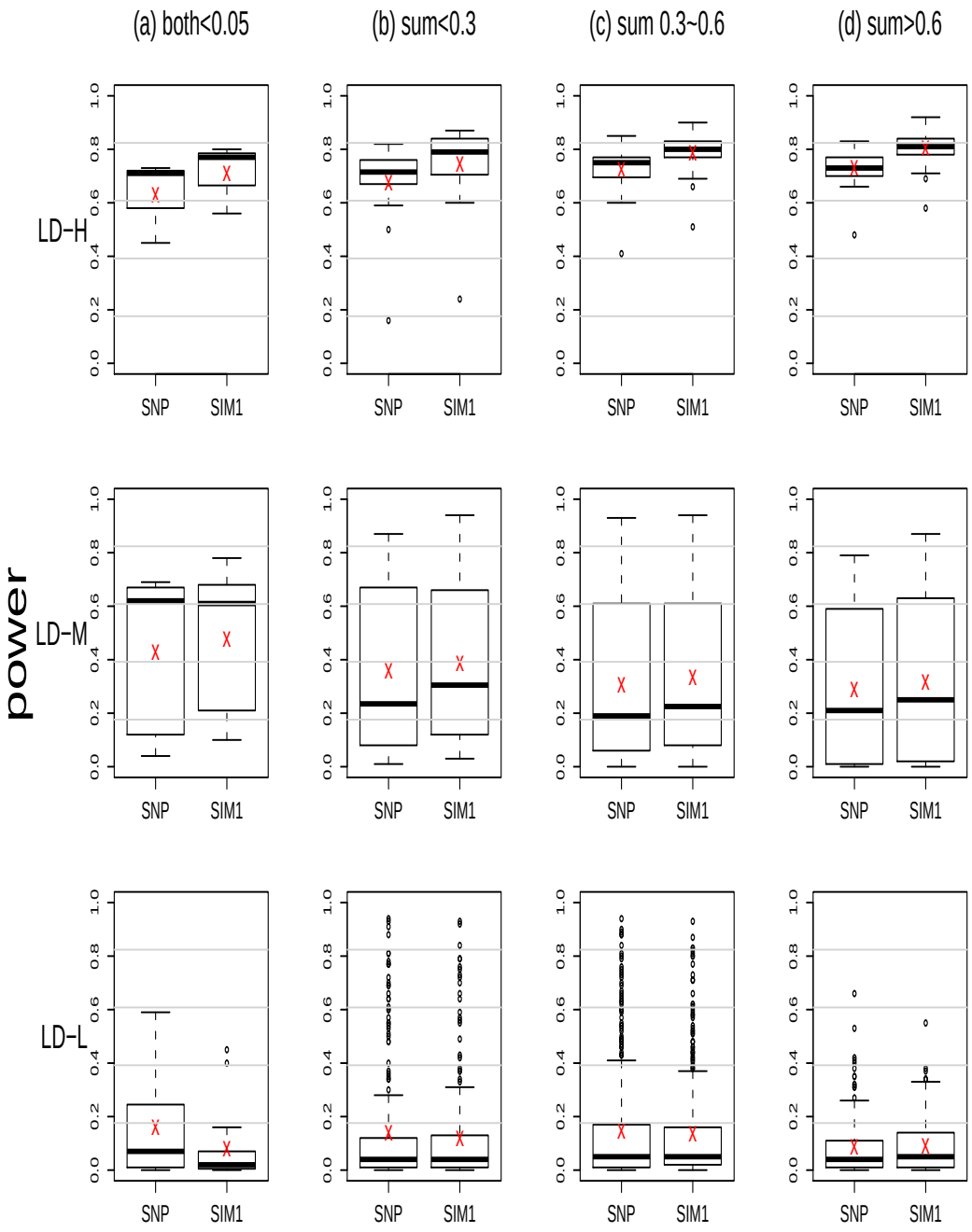

Figure 4: Power of joint tests.

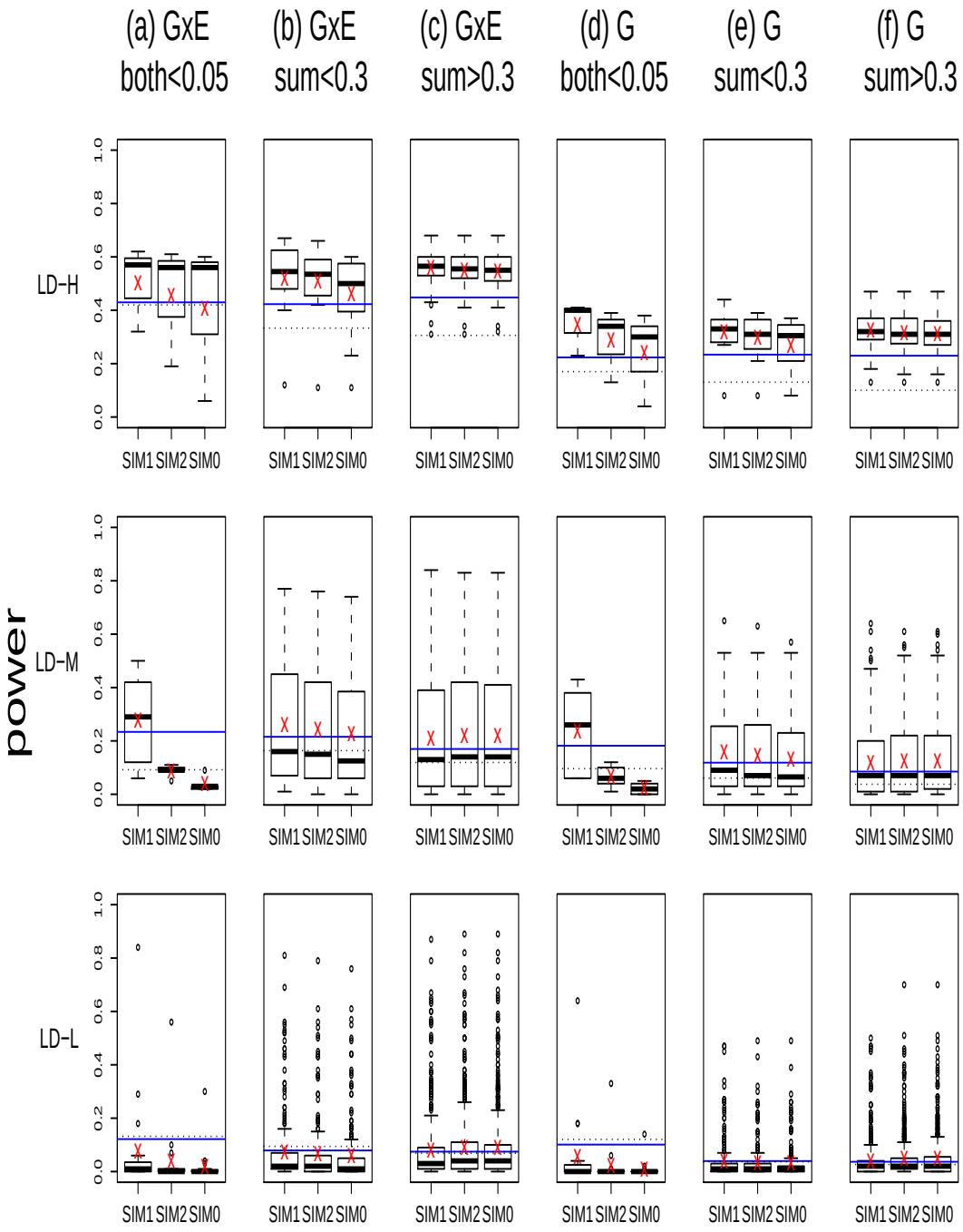

Figure 5: Power of SIM1, SIM2 and SIM0 tests.

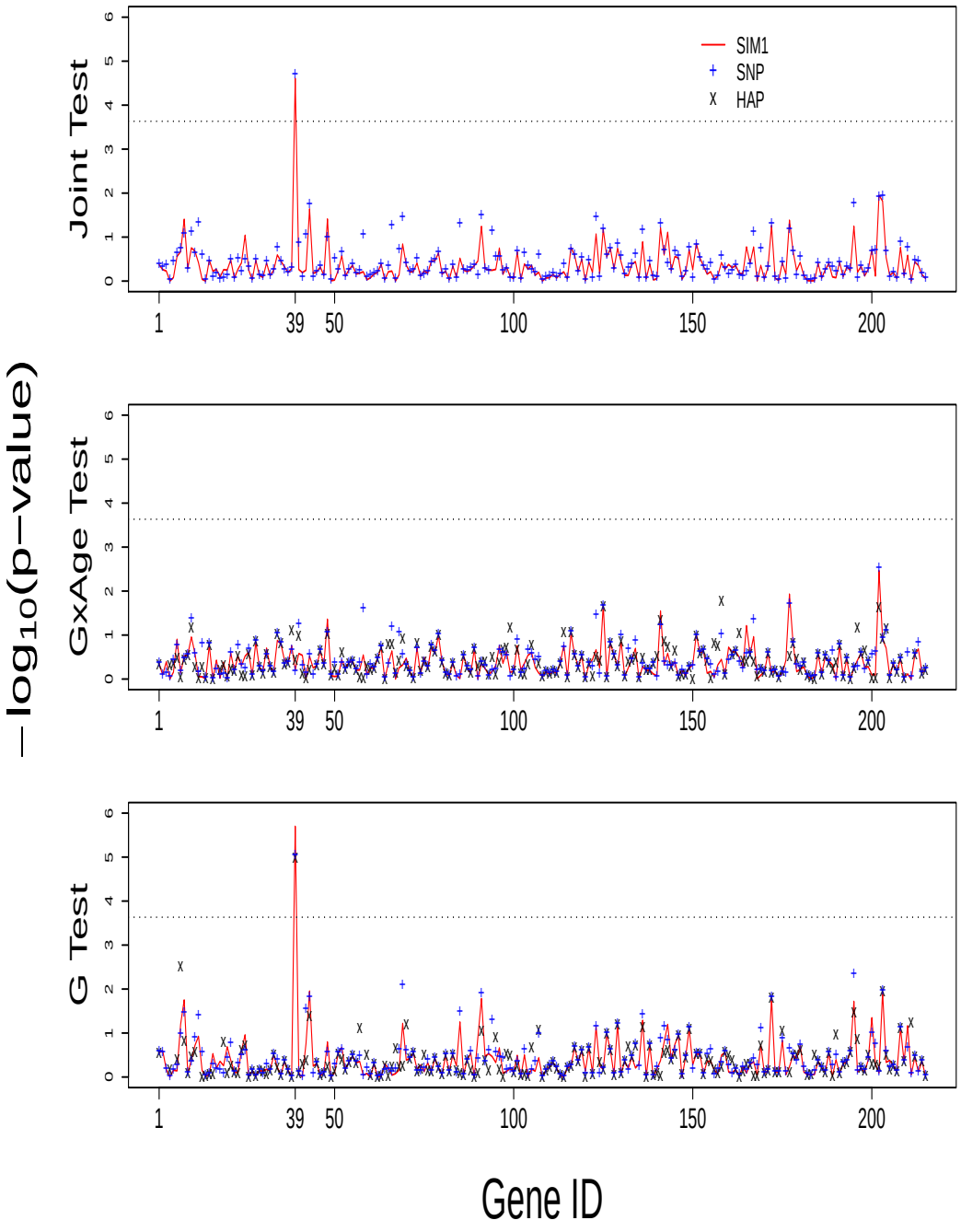

Figure 6: P-values of VISP analysis.