A Generalized Similarity U Test

for Multivariate Analysis of Sequencing Data

Changshuai Wei[1,2] and Qing Lu[1]

[1]Department of Epidemiology and Biostatistics,

College of Human Medicine, Michigan State University,

East Lansing, MI, USA

[2]Department of Biostatistics and Epidemiology,

School of Public Health, University of North Texas Health Science Center,

Fort Worth, TX, USA

October 2, 2014

Author's Footnote:

Changshuai Wei, Assistant Professor of Biostatistics, Department of Biostatistics and Epidemiology, University of North Texas Health Science Center. (Email: changshuai.wei@unthsc.edu).

Qing Lu is the corresponding author, Assosicate Professor of Biostatistics, Department of Epidemiology and Biostatistics, Michigan State University.(Email: qlu@epi.msu.edu).

Abstract

Sequencing-based study is emerging as an essential tool for genetic research of complex diseases. Despite the promise of uncovering new disease-associated genetic variants, the high-dimensionality of data and the low frequency of genetic variants pose great challenges to statistical analysis of sequencing data. In addition, there is a great interest in biomedical research to identify genetic loci contributed to multiple disease phenotypes. The multiple phenotypes can often follow different distributions, which bring additional challenges. In this paper, we propose a generalized similarity U test, referred to as GSU, for multivariate analysis of sequencing data. GSU is a similarity-based test and can handle high-dimensional genetic and phenotypic data. We studied the theoretical properties of GSU, providing efficient p-value calculation for hypothesis testing and sample size calculation for sequencing study design. Through simulation, we found that GSU had advantages over existing methods in terms of power and robustness. Moreover, we showed that GSU was computationally more efficient than existing methods. Finally, we used GSU to perform a multivariate analysis of sequencing data in the Dallas Heart Study and identified a joint association of 4 genes with 5 metabolic related phenotypes.

KEYWORDS: Weighted U Statistic, Sequencing Study, Non-parametric Method

1. INTRODUCTION

In the last decades, genome-wide association studies (GWAS) have made substantial progress in discovering common genetic variants associated with complex diseases. Despite such success, a large proportion of heritability of complex diseases remains unexplained. Converging evidence has suggested that rare variants with minor allele frequency (MAF) smaller than 5% or 1%, hold promise in accounting for a significant proportion of the missing heritability (Fay et al., 2001; Pritchard, 2001; Kryukov et al., 2007; Boyko et al., 2008). With the advance of next-generation sequencing (NGS) technology, we are able to investigate the role of a wider scope of genetic variants, primarily rare variants, in human diseases. Evidence in early sequencing studies have already shown that rare variants played an important role in complex diseases (Cohen et al., 2004; Ahituv et al., 2007; Ji et al., 2008; Romeo et al., 2009). Although promising, the massive data generated from sequencing studies poses great challenges on the statistical methods. Rare mutations are recent mutations, and can be only found in a small fraction of individuals in the entire population. Even with a large effect size, a rare variant is hard to be detected because of its low MAF. Moreover, the massive number of rare variants raises the computational burden and the multiple comparison issue. Therefore, traditional single-locus analysis has low power to detect rare variants.

Many new statistical methods have been developed for the sequencing data in the last few years. Different from traditional methods for single-locus analysis, most of new methods perform a joint association test, namely, testing the joint effect of a set of single nucleotide variants (SNV) on a genomic region, a functional unit (e.g., a gene) or a functional pathway. The advantage of the joint association test over the single-locus analysis lies in the fact that, by combining multiple SNVs, not only the association information is aggregated but also the number of test is greatly reduced. The joint association tests can be briefly classified into two categories: burden tests and non-burden tests. Burden tests first summarize multiple rare variants into a univariate genetic score, and then test the association of the summary score

4

with the disease phenotype.(Morgenthaler and Thilly, 2007; Li and Leal, 2008; Madsen and Browning, 2009; Lin and Tang, 2011). Burden tests often assume the effect of the multiple variants have similar magnitude and direction. Non-burden tests, on the other hand, can take into account of the effect heterogeneity within the SNV-set, by considering the effects of the multiple variants as random effects or a function of genomic position.(Neale et al., 2011; Wu et al., 2011; Luo et al., 2012). Among non-burden methods, sequence kernel association test (SKAT) shares many nice computational and asymptotic properties with the variance component score test, and has been very popularly used(Wu et al., 2011).

Most of existing joint association tests are parametric-based methods, which often rely on certain assumptions (e.g., a normal distribution assumption). When the assumptions are violated, it may lead to power loss or an inflated type I error. Moreover, there lacks new methods for the multivariate analysis of sequencing data, especially with different types of phenotypes (e.g., some phenotypes are binary while the others are continuous). The study of multiple phenotypes is important in biomedical research. Many studies collect multiple bio-chemical measurements related to a disease status. These multiple measurements evaluate different aspects of the disease, and thus better reflect the underlying biological mechanism of the disease. It also becomes popular for human genetic studies to collect and study multiple disease phenotypes. For instance, studies have identified common gene variants contributed to co-morbidity of substance dependence. Finally, few statistical methods for sequencing study have targeted the computational issue. In the era of high-dimensional data analysis, we need more efficient software (e.g. C or C++ based) to handle the massive amount of sequencing data.

To achieve these goals, we propose a Generalized Similarity U test, referred to as GSU. We use two different similarity measurements to summarize genetic information and phenotypic information, and then form a test under the weighted U framework. We derive the asymptotic distribution of the test statistic and implement the new method in both R

5

and C++ programs. The proposed method has several remarkable features: 1) it is non-parametric and is thus robust to phenotype distributions; 2) it can handle multiple different phenotypes (e.g., a combination of binary and continuous phenotypes); 3)it has a nice statistical property and performs well under small sample sizes; and 4) it is computationally efficient for high-dimensional data. Simulation studies are conducted to evaluate the type 1 error and the power of our method with those of several commonly used methods. Finally, we applied the new method to the Dallas Heart Study to test the association of 4 candidate genes with 5 metabolic-related phenotypes.

2.   GENERALIZED SIMILARITY U

2.1   Weighted U Statistic

Suppose that n subjects are sequenced in the study, where we are interested in the association of L phenotypes ($y_{i,l}$, $1 \leq i \leq n$, $1 \leq l \leq L$) with M genetic variants ($g_{i,m}$, $1 \leq i \leq n$, $1 \leq m \leq M$). For each subject i, we observe a vector of phenotypes y_i ($y_i = (y_{i,1}, y_{i,2}, \cdots, y_{i,L})$) and a vector of genotypes g_i ($g_i = (g_{i,1}, g_{i,2}, \cdots, g_{i,M})$). In the special case when $L = 1$ (or $M = 1$), it is simplified to a univariate analysis (or a single-locus analysis). When $L > 1$, it can be used to perform a multivariate analysis. In GSU, we allow multiple phenotypes to be of different types (e.g., continuous or categorical), and do not assume any distribution of phenotypes. The number of genetic variants M and the number of phenotypes L can be larger than the sample size, for example, in the case of sequencing studies (high demensional genotypes) and imaging data (high demensional phenotypes).

Given the phenotypes and the genotypes for the subjects i and j ,we first define their phenotype similarity $S_{i,j}$ by,

$$S_{i,j} = h(y_i, y_j),$$

and define their genetic similarity $K_{i,j}$ by,

$$K_{i,j} = f(g_i, g_j).$$

The similarity measurements $h(\cdot, \cdot)$ and $f(\cdot, \cdot)$ defined above can be of a general form as long as they satisfy the finite second moment condition (i.e., $E(h^2(Y_1, Y_2)) < \infty$ and $E(f^2(G_1, G_2)) < \infty$). We further center the phenotype similarity by,

$$
\begin{aligned}
\tilde{S}_{i,j} &= \tilde{h}(y_i, y_j) \\
&= h(y_i, y_j) - E(h(y_i, Y_j)) - E(h(Y_i, y_j)) + E(h(Y_i, Y_j)), \quad (1)
\end{aligned}
$$

and center the genetic similarity ($\tilde{K}_{i,j} = \tilde{f}(g_i, g_j)$) in a similar manner. The generalized similarity U (GSU) is then defined as the summation of the centered phenotype similarities weighted by the centered genetic similarities,

$$U = \frac{1}{n(n-1)} \sum_{i \neq j} \tilde{K}_{i,j} \tilde{S}_{i,j}, \quad (2)$$

where $\tilde{K}_{i,j}$ is considered as the weight function and $\tilde{S}_{i,j}$ is considered as the U kernel. In our definition of GSU, the role of genetic similarity and phenotype similarity are interchangeable. In other words, we can also treat $\tilde{S}_{i,j}$ as the weight function and $\tilde{K}_{i,j}$ as the U kernel.

2.2 Similarity Measurement

The choices for the phenotype similarity $h(\cdot, \cdot)$ and the genetic similarity $f(\cdot, \cdot)$ are flexible. According to different types of genetic variants and the purpose of the analysis, we can choose different types of phenotype similarities and genetic similarities.

For the categorical SNVs data, we can use either the IBS function or the weighted IBS function to measure the genetic similarity(Lynch and Ritland, 1999). Assuming the genetic variants ($g_{i,m}$, $1 \leq i \leq n$, $1 \leq m \leq M$) are coded as 0, 1 and 2, respectively for AA, Aa and aa, we can define the IBS-based genetic similarity between subjects i and j as,

$$K_{i,j}^{IBS} = \frac{1}{2M} \sum_{m=1}^{M} 2 - |g_{i,m} - g_{j,m}|.$$

Alternatively, we can use a weighted-IBS (wIBS) genetic similarity to emphasize the effects of rare variants,

$$K_{i,j}^{wIBS} = \sum_{m=1}^{M} \frac{w_m(2 - |g_{i,m} - g_{j,m}|)}{\Upsilon},$$

where w_m represents the weight for the m-th SNV in the SNV-set, and Υ is a scaling constant, defined as $\Upsilon = 2\sum_{m=1}^{M} w_m$. w_m is usually defined as a function of MAF, γ_m. For example, we can define the weight w_m as $w_m = 1/\sqrt{\gamma_m(1 - \gamma_m)}$. When the genetic data are count data or continuous data, we can use other forms of $f(\cdot, \cdot)$ to measure genetic similarity (e.g., euclidian distance based similarity), which we will leave for further investigations.

For phenotype similarity, we define a unified measurement for both categorical and continuous phenotypes based on a normal quantile. For each phenotype, y_l $(1 \leq l \leq L)$, we define the corresponding normal quantile by,

$$q_{i,l} = \Phi^{-1}((rank(y_{i,l}) - 0.5)/n),$$

where $rank(\cdot)$ corresponds to the rank of the phenotype value $y_{i,l}$. When there are ties, we assign the averaged rank. $\Phi^{-1}(\cdot)$ is the inverse cumulative density function for a standard normal distribution. We can then calculate the Euclidian Distance (ED) based phenotype similarity between subjects i and j by,

$$S_{i,j}^{ED} = exp(-\sum_{l=1}^{L} \omega_l(q_{i,l} - q_{j,l})^2),$$

where ω_l represents the weight for the l-th phenotypes, which could be given based on prior knowledge. If there is no prior knowledge, we can use equal weights, $\omega_l = 1/L$. The ED-based phenotype similarity can be easily modified to take the correlations between the phenotypes into account,

$$S_{i,j}^{ED} = exp\left(-\frac{1}{L}(q_i - q_j)^T \Gamma(q_i - q_j)\right),$$

where $q_i = (q_{i1}, \cdots, q_{iL})^T$. Γ can be chosen to reflect the correlations among the phenotypes. For example, we can define Γ as,

$$\Gamma = (\frac{1}{n}\sum_{i=1}^{n} q_i q_i^T)^{-1}.$$

Other than the ED based phenotype similarity, we can define the phenotype similarity using cross product(Tzeng et al., 2009). The types of phenotype similarity or genetic similarity can be chosen based on different purposes, which may influence the power of the method. For simplicity, we used an ED-based phenotype similarity in this paper.

2.3   Hypothesis testing

Based on the definition of the centered similarity, we can show that $E(\tilde{f}(G_i, G_j)) = 0$ and $E(\tilde{h}(Y_i, Y_j)) = 0$ (Appendix A). Under the null hypothesis, when the genetic variants are not associated with multiple phenotypes, we have $E(U) = 0$ (Appendix A). Under the alternative hypothesis, when the genetic variants are associated with multiple phenotypes, we expect that the phenotype similarity is concordant with the genetic similarity. In other words, the positive phenotype similarities are weighted heavier and the negative phenotype similarities are weighted lighter, leading to a positive value of U statistic. A statistical test can be formed to test the association and its p-value can be calculated by $P(U > U_{obs})$, where U_{obs} is the observed value of U. Tests based on permutation or bootstrap would be a straight-forwad way of determining the statistical significance. Nevertheless, both methods are computationally expensive for high-dimensional data. Therefore, we derive the asymptotic distribution of GSU to assess the statistical significance of the association test.

By considering the genetic similarity as the weight function and the phenotype similarity as the U kernel, GSU is a weighted U statistic(Gregory, 1977; Shapiro and Hubert, 1979; O'Neil and Redner, 1993; Shieh, 1997; Lindsay et al., 2008). More specifically, because its kernel satisfied $Var(E(\tilde{h}(Y_1, Y_2)|Y_2)) = 0$ (Appendix A), GSU is a degenerated weighted U statistic. To derive the limiting distribution of GSU, we can decompose the centered phenotype similarity by,

$$\tilde{h}(y_1, y_2) = \sum_{s=1}^{\infty} \lambda_s \phi_s(y_1) \phi_s(y_2),$$

where $\{\lambda_s\}$ and $\{\phi_s(\cdot)\}$ are eigenvalues and eigenfunctions of the U kennel $\tilde{h}(\cdot, \cdot)$, and all

9

the eigenfunctions are orthogonal,

$$\int \phi_s(y_1)\phi_{s'}(y_1)dF(y_1) = \begin{cases} 0, & \text{if } s \neq s' \\ 1, & \text{if } s = s'. \end{cases}$$

Similarly, we can decompose the centered genetic similarity by,

$$\tilde{f}(G_i, G_j) = \sum_{t=1}^{\infty} \eta_t \varphi_t(g_1)\varphi_t(g_2).$$

We can then rewrite the GSU as (Appendix B),

$$U = \frac{1}{n-1} \sum_{t=1}^{\infty} \sum_{s=1}^{\infty} sign(\eta_t \lambda_s) \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \eta_t^\star(G_i)\phi_s^\star(Y_i) \right)^2$$
$$- \frac{1}{n-1} \sum_{t=1}^{\infty} \sum_{s=1}^{\infty} sign(\eta_t \lambda_s) \frac{1}{n} \sum_{i=1}^{n} (\eta_t^\star(G_i)\phi_s^\star(Y_i))^2,$$

where $\varphi_t^\star(G_i) = |\eta_t|^{0.5}\varphi_t(G_i)$ and $\phi_s^\star(Y_i) = |\lambda_s|^{0.5}\phi_s(Y_i)$.

Using the form above, we can show that the limiting distribution of GSU is a weighted sum of independent chi-square random variables. This is the result of theorem 1 below, which is proved in Appendix B.

Theorem 1. Suppose $E(h^2(Y_1, Y_2)) < \infty$, $E(f^2(G_1, G_2)) < \infty$, and $Y \perp G$. Let $\tilde{h}(Y_1, Y_2)$ and $\tilde{f}(G_1, G_2)$ be the centered similarity as defined in (1). Define U as $U = \frac{1}{n(n-1)} \sum_{i \neq j} \tilde{f}(G_i, G_j)\tilde{h}(Y_i, Y_j)$. Then, $nU \xrightarrow{D} \sum_{t=1}^{\infty} \eta_t \sum_{s=1}^{\infty} \lambda_s(\chi_{st}^2 - 1)$, where $\{\chi_{st}^2\}$ are independent chi-square random variables with 1 degree of freedom.

2.4 The power and sample size calculation

In this subsection, we study the properties of GSU under the alternative hypothesis. We derive the asymptotic distribution of GSU under the alternative hypothesis, and provide power and sample size calculations for sequencing association studies.

Assume $E(\tilde{f}(G_1, G_2)\tilde{h}(Y_1, Y_2)) = \mu > 0$ and $Var(\tilde{f}(G_1, G_2)\tilde{h}(Y_1, Y_2)|(G_2, Y_2)) = \zeta_1 > 0$ under the alternative hypothesis, where μ measures the strength of the association. It is

10

easy to show that GSU is the unbiased estimator of μ,

$$E(U) = \frac{1}{n(n-1)} \sum_{i \neq j} E(\tilde{f}(G_i, G_j)\tilde{h}(Y_i, Y_j)) = \mu.$$

Using the Hoeffding projection, we can show that GSU asymptotically follows a normal distribution, with mean μ and variance $4\zeta_1/n$. This is the result of Theorem 2, which is proved in Appendix C.

Theorem 2. Let $\tilde{h}(Y_1, Y_2)$ and $\tilde{f}(G_1, G_2)$ be the centered similarities as defined in (1). Suppose Y is associated with G, where the following conditions are satisfied: $E(\tilde{f}(G_1, G_2)\tilde{h}(Y_1, Y_2)) = \mu > 0$, $Var(\tilde{f}(G_1, G_2)\tilde{h}(Y_1, Y_2)) = \zeta_0 < \infty$, and $Var(\tilde{f}(G_1, G_2)\tilde{h}(Y_1, Y_2)|(G_2, Y_2)) = \zeta_1 > 0$. Let U defined as $U = \frac{1}{n(n-1)} \sum_{i \neq j} \tilde{f}(G_i, G_j)\tilde{h}(Y_i, Y_j)$. Then, $\sqrt{n}(U - \mu) \xrightarrow{D} N(0, 4\zeta_1)$.

The power of GSU at the significance level α can be calculated by,

$$P\{nU > q_{1-\alpha}\}$$
$$= P\left\{ \frac{\sqrt{n}(U - \mu)}{2\sqrt{\zeta_1}} > \frac{q_{1-\alpha} - n\mu}{2\sqrt{n\zeta_1}} \right\}$$
$$= \Phi\left(\frac{n\mu - q_{1-\alpha}}{2\sqrt{n\zeta_1}} \right),$$

where $q_{1-\alpha}$ is the $1-\alpha$ quantile for $\sum_{t=1}^{\infty} \eta_t \sum_{s=1}^{\infty} \lambda_s(\chi_{st}^2 - 1)$ and $\Phi(\cdot)$ is the CDF of a standard normal distribution. The sample size required to achieve power β can be calculated by solving $\Phi(\frac{n\mu - q_{1-\alpha}}{2\sqrt{n\zeta_1}}) \geq \beta$. By denoting Z_β as the β quantile for a standard normal distribution, the required sample size is given by,

$$n = \min_{n \in N} \left\{ n : n \geq \frac{\left(Z_\beta \sqrt{\zeta_1} + \sqrt{Z_\beta^2 \zeta_1 + \mu q_{1-\alpha}} \right)^2}{\mu^2} \right\}.$$

2.5 Computation and implementation

Let $S = \{S_{i,j}\}_{n \times n}$ and $K = \{K_{i,j}\}_{n \times n}$ be the matrix form of the phenotype similarity and genetic similarity, the centered similarity matrices $\tilde{S}$ and $\tilde{K}$ can be obtained as,

$$\tilde{S} = (I - J)S(I - J),$$

$$\tilde{K} = (I - J)K(I - J),$$

where I is an n-by-n identity matrix, and J is an n-by-n matrix with all elements being $1/n$ (Appendix D). The GSU can be expressed as,

$$U = \frac{1}{n(n-1)} \sum_{i \neq j} \tilde{K}_{i,j} \tilde{S}_{i,j}.$$

In this form, U can be viewed as a sum of the element-wise product of the two matrices, $\tilde{K}_0$ and $\tilde{S}_0$, which are obtained by assigning 0 to the diagonal elements of matrices $\tilde{K}$ and $\tilde{S}$. We then use matrix eigen-decomposition to approximate the eigen-values in function decomposition. Let $\{\tilde{\lambda}_s\}$ and $\{\tilde{\eta}_t\}$ respectively be the eigen-values for matrices $\tilde{K}_0$ and $\tilde{S}_0$ (Appendix E), the limiting distribution of U is given by,

$$nU \sim \sum_{t=1}^{n} \frac{\tilde{\eta}_t}{n} \sum_{s=1}^{n} \frac{\tilde{\lambda}_s}{n} (\chi_{st}^2 - 1),$$

where $\{\chi_{st}^2\}$ are independent chi-square random variables with 1 degree of freedom. The p-value can be calculated by using the Davies' method (Davies, 1980), the Liu's method(Liu et al., 2009a) or the Kuonen's method(Kuonen, 1999).

3. SIMULATION STUDY

3.1 Simulation method

To mimic real genetic structure, we used genetic data from the 1000 Genome Project(Abecasis et al., 2010). Based on the genetic data, we then simulated phenotype values. In particular, we used a 1Mb region of the genome (Chromosome 17: 7344328-8344327) from 1092 individuals in 1000 Genome Project. For each simulation replicate, we randomly chose a 30kb segment from the 1Mb region and formed a SNV-set for the analysis by using all SNVs in the segment. From the SNV-set, we set a proportion of the SNVs as causal. A number of individuals were randomly chosen as the simulation sample to study the performance

of the methods under different sample sizes. To investigate the robustness against different phenotype distributions, we simulated three types of phenotypes: a binary-distributed phenotype, a Gaussian-distributed phenotype and a Cauchy-distributed phenotype. The binary-distributed phenotype was simulated by using a logistic regression model,

$$logit(P(Y_i = 1)) = \mu + G_i^T \beta,$$

where Y_i and G_i were the phenotype value and the genotype vector (coded as 0, 1, and 2) for the i-th individual, respectively. β were the effects of the SNVs, which were sampled from a normal distribution with a mean of μ_β and a variance of σ_β^2. The Gaussian-distributed phenotype was simulated by using the conventional linear regression model,

$$Y_i = \mu + G_i^T \beta + \varepsilon_i, \ \varepsilon_i \sim N(0, \sigma^2),$$

The Cauchy-distributed phenotype represented a situation in which the continuous phenotype had more extreme values (i.e. heavy-tailed), and was simulated by using the following model,

$$Y_i \sim cauchy(a_i, b), \ a_i = \mu + G_i^T \beta,$$

where a_i and b are the location parameter and the scale parameter of the Cauchy distribution, respectively.

Two sets of simulation were performed. In simulation I, we considered a single phenotype, while in simulation II, we considered multiple-phenotypes.

Simulation I Setting

Under the null, the models were simulated by setting $\mu_\beta = 0$ and $\sigma_\beta^2 = 0$, and were evaluated by varying the sample size from 50 to 500 (i.e. 50, 100, 200 and 500). Under the alternatives, two sets of disease models were simulated (the details of the simulation setting can be found in Table S1 of Supplementary Materials):

1. Set $\mu_\beta = 0$ and $\sigma_\beta^2 > 0$ so that half of the causal SNVs were deleterious and the other half were protective.

2. Set $\mu_\beta > 0$ and $\sigma_\beta^2 > 0$ so that majority of the causal SNVs were deleterious.

Simulation II Setting

Two sets of models were simulated:

1. Assume the multiple phenotypes follow the same distribution. In particular, we simulated 3 binary-distributed phenotypes (BBB), 3 Gaussian-distributed phenotypes (GGG), and 3 Cauchy-distributed phenotypes (CCC).

2. Allow the multiple phenotypes to follow the different distributions. In particular, we simulated 3 phenotypes with 2 binary-distributed phenotypes and 1 Gaussian-distributed phenotype (BBG), 3 phenotypes with 1 binary-distributed phenotypes and 2 Gaussian-distributed phenotypes (BGG), and 3 phenotypes with 1 binary-distributed phenotype, 1 Gaussian-distributed phenotype and 1 Cauchy-distributed phenotype (BGC).

Similar as the simulation I, for the null model, we set $\mu_\beta = 0$ and $\sigma_\beta^2 = 0$. For the alternative models, we set $\mu_\beta = 0$ and $\sigma_\beta^2 > 0$, and allowed the multiple phenotypes to be influenced by different sets of causal SNVs. The details of the simulation setting were described in Table S2 of Supplementary Materials.

3.2 Simulation result

We evaluated the performance of GSU by comparing it with three existing methods, SKAT(Wu et al., 2011), AdjSKAT and SKATO(Lee et al., 2012). For each simulation, we created 1000 simulation replicates to evaluate type 1 error and power. For GSU, we used the wIBS-based genetic similarity and ED-based phenotype similarity to construct the U statistic. To be consistent, we used the same weighted IBS to construct the kernel for SKAT. Because neither AdjSKAT nor SKATO implement a weighted IBS kernel, we used the default kernel, the weighted linear kernel, when applying these two methods. SKAT, AdjSKAT and SKATO are designed for univariate analysis. To consider multiple phenotypes, we chose the most

significant p-value from the univariate analysis, and then adjust for multiple tests by using the Bonferroni correction.

Result for Simulation I

The type I error rates of the 4 methods were summarized in Table 1. GSU had a well-controlled type I error, regardless of phenotype distributions and sample sizes. Neverthless, SKAT, AdjSKAT and SKATO had inflated type I error rates (ranging from 0.101 to 0.19) for the Cauchy-distributed phenotype. When the sample size is small (e.g., 50 or 100), SKAT and SKATO also had conservative type I error (e.g., 0.001) for the binary-distributed phenotype.

The power comparison were summarized in Figures 1 and 2. For the disease model where half of the causal SNVs were deleterious (Figure 1), GSU had higher power than other methods for various sample sizes ranging from 50 to 500. For instance, in the simulation with binary phenotype and a sample size of 50, GSU (power=0.346) attained much higher power than AdjSKAT (power=0.114), SKATO (power=0.057), and SKAT (power=0.024). Similarly, for the Gaussian-distributed phenotype, GSU had the highest power among the four methods, and the three SKAT-based methods had similar performance. For the Cauchy-distributed phenotype, where the distribution assumption was violated, the power of the SKAT-based methods remained similar as the sample sizes increased, while GSU had increased power as the sample size increased. For the second disease model in which a majority of the SNVs were deleterious (Figure 2), we observed the same conclusion, i.e., GSU had highest power regardless of sample sizes and phenotype distributions.

Result for Simulation II

The type I error rates for the multivariate analysis were summarized in Table 2. Similar to

Table 1: type I error comparison for the univariate analysis

Sample size	Method	Distribution		
		Binary	Gaussian	Cauchy
50	SKAT	0.001	0.030	0.122
	SKATO	0.021	0.041	0.101
	AdjSKAT	0.054	0.028	0.123
	GSU	0.058	0.046	0.051
100	SKAT	0.014	0.028	0.149
	SKATO	0.035	0.038	0.120
	AdjSKAT	0.063	0.027	0.139
	GSU	0.046	0.050	0.053
200	SKAT	0.023	0.04	0.155
	SKATO	0.024	0.042	0.140
	AdjSKAT	0.043	0.042	0.156
	GSU	0.057	0.048	0.045
500	SKAT	0.039	0.055	0.190
	SKATO	0.052	0.056	0.158
	AdjSKAT	0.049	0.053	0.180
	GSU	0.038	0.046	0.043

Figure 1: Power comparison for univariate analysis when a majority of causal SNVs are deleterious
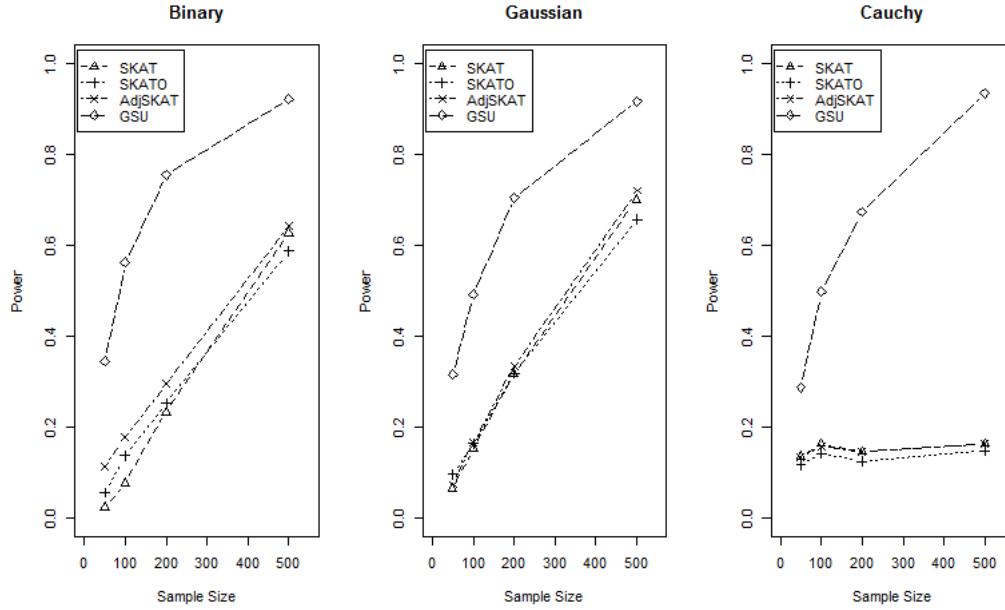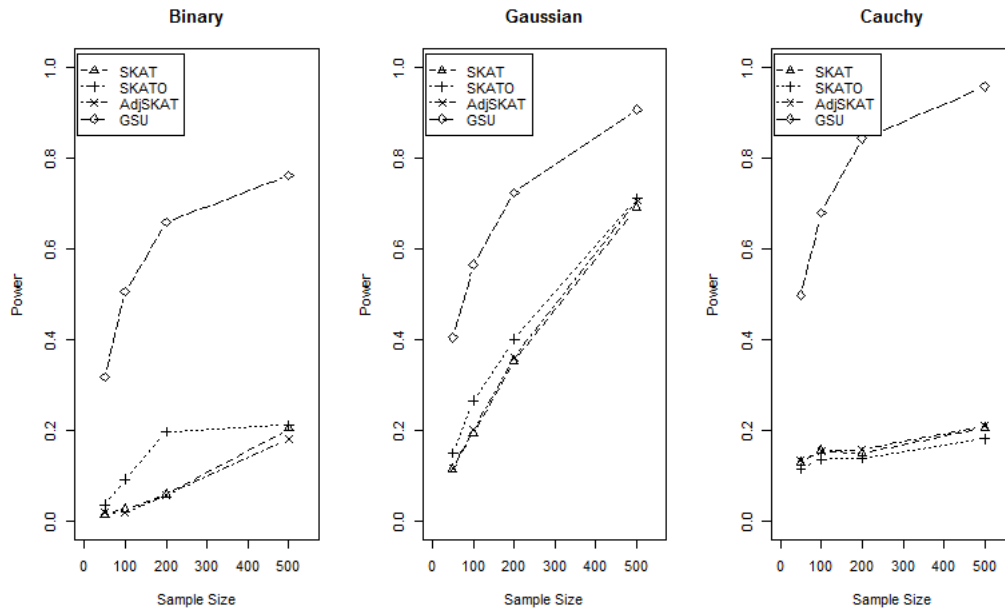

Figure 2: Power comparison for univariate analysis when half of causal SNVs are protective and the other half of causal SNVs are deleterious

the results of the univariate analysis, GSU can correctly control type I error at the level of 0.05 (Table 2), while the other three methods had inflated type I error when the distribution assumption was violated (e.g., CCC and BGC). When the distribution assumption was satisfied, AdjSKAT correctly controlled the type 1 error. SKAT and SKATO, however, had conservative type I error rates for small sample sizes (50 and 100), especially when the multiple phenotypes comprised of binary-distributed phenotypes (e.g., BBB, BBG and BGG).

The power comparison for the two disease models were summerized in Figures 3 and 4. For the disease model where the multiple phenotypes followed the same distribution, GSU had higher power than the three SKAT-based methods (Figure 3). For simulations with BBB phenotypes and GGG phenotypes, GSU obtained higher power than the other three methods when the sample size was 50, and obtained substantial power improvement than the other three methods as sample sizes increased. For the simulation with CCC phenotypes, we also observed substantial power improvement of GSU than the other three methods as sample size increased. Yet, the power of the three SKAT-based methods were higher than GSU when the sample size was 50. This can be explained by the inflated type 1 error rates of the SKAT-based methods for the Cauchy-distributed phenotypes (Table 2). Similarly, for disease models with a combination of different types of phenotype distribution (i.e., BBG, BGG, and BGC), GSU had higher power than SKAT, SKATO and AdjSKAT, regardless of sample size and phenotype distributions (Figure 4).

3.3   Computation and Software

We implemented GSU in both R and C++. Because SKAT was written in R, to be comparable, we compared the computational efficiency of R-based GSU with those of the SKAT, SKATO and AdjSKAT. We found GSU attained highest computational efficiency among four methods. For example, for the analysis of the BBG phenotype and a sample size of 50 for 1000 simulation replicates, SKAT, SKATO and AdjSKAT took 10 (95.606s), 217

Table 2: type I error comparison for multivariate analysis

Sample size	Method	Same Distr.[1]			Diff Distr.		
		BBB	GGG	CCC	BBG	BGG	BGC
50	SKAT	0.002	0.028	0.232	0.008	0.021	0.097
	SKATO	0.019	0.027	0.207	0.016	0.024	0.085
	AdjSKAT	0.046	0.028	0.237	0.049	0.040	0.113
	GSU	0.057	0.047	0.045	0.034	0.043	0.056
100	SKAT	0.001	0.023	0.295	0.011	0.013	0.122
	SKATO	0.025	0.039	0.260	0.025	0.036	0.117
	AdjSKAT	0.055	0.025	0.290	0.059	0.042	0.135
	GSU	0.048	0.047	0.044	0.051	0.059	0.058
200	SKAT	0.020	0.048	0.325	0.033	0.020	0.134
	SKATO	0.030	0.058	0.288	0.035	0.036	0.118
	AdjSKAT	0.059	0.046	0.319	0.064	0.038	0.136
	GSU	0.050	0.047	0.049	0.058	0.056	0.043
500	SKAT	0.044	0.035	0.364	0.035	0.04	0.166
	SKATO	0.054	0.036	0.315	0.038	0.047	0.156
	AdjSKAT	0.06	0.038	0.342	0.041	0.049	0.172
	GSU	0.045	0.054	0.044	0.046	0.039	0.062

[1] B, G, and C represent Binary-distributed, Gaussian-distributed, and Cauchy-distributed phenotypes, respectively.

Figure 3: Power comparison for the multivariate analysis when the multiple phenotypes follow the same type of distribution
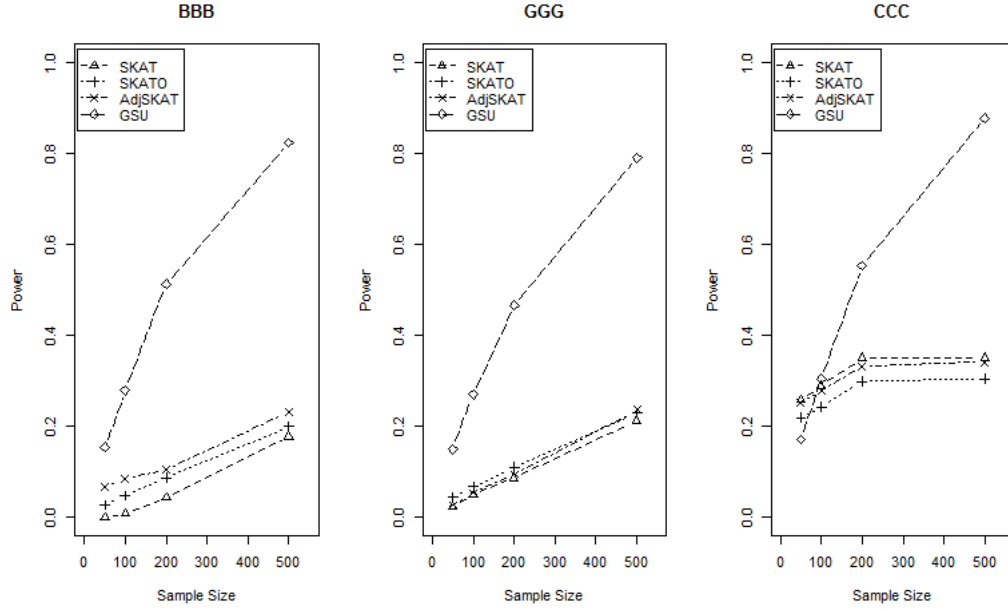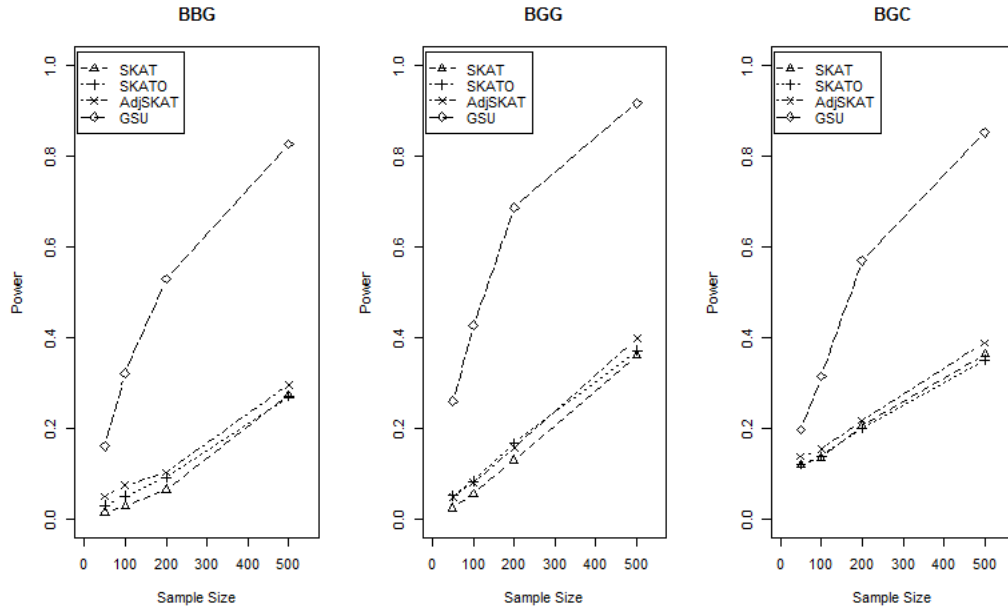

Figure 4: Power comparison for the multivariate analysis when the multiple phenotypes follow the different types of distributions

(2155.145s) and 532 (5286.115s) times longer than GSU (9.925s) on a personal computer with 2.3GHz CPU and 4G memory. The detailed computational efficiency comparison of four methods for analyzing data with various distributed phenotypes and different sample sizes was given in Table S3 of Supplementary Materials. To handle high-dimensional data and further improve computational efficiency, we also developed a C++ based package for GSU (https://www.msu.edu/ changs18/software.html#GTSU)for both windows and unix systems. The current version of the C++ based GSU use the default weighted IBS for the genetic similarity, and the ED-based similarity for the phenotype similarity. Additional features will be added to the package in the future.

4. APPLICATION TO THE DALLAS HEART STUDY

To evaluate the performance of GSU on real data, we applied it to the Dallas Heart Study (DHS) sequence data(Ahituv et al., 2007) and compared the result of GSU with those of three other SKAT-based methods. The DHS sequencing data is comprised of 4 genes, ANGPTL3, ANGPTL4, ANGPTL5 and ANGPTL6. After completing the quality control procedure (e.g., removing SNVs with high missing rate), 230 SNVs (54 SNVs, 63 SNVs, 61 SNVs, and 52 SNVs were from ANGPTL3, ANGPTL4, ANGPTL5 and ANGPTL6, respectively) and 2598 subjects remained for the analysis. In the real data analysis, we were interested in testing the association of the SNVs in these genes with multiple metabolic-related phenotypes, including obesity (dichotomized from BMI using a cut-off of 35), cholesterol, high-density lipoprotein cholesterol (HDL), low-density lipoprotein cholesterol (LDL) and very-low-density lipoprotein cholesterol (VLDL).

In order to consider the potential confounding effects of age, gender and race, we adjusted these covariates in the analysis and used the residuals to build phenotype similarity for GSU. Because the three SKAT-based methods cannot directly analyze multiple phenotypes, we obtained smallest p-values from the univariate analysis of all single phenotypes and

21

then used the Bonferroni correction to determine whether there was a significant association. The results were summarized in Table 3. Because all 4 genes were metablolic candidate genes, we first combined SNVs in the 4 genes into a single SNV-set. For the joint analysis of 4 genes, GSU could detect a significant association of 4 genes with 4 metabolic-related phenotypes (p-value=0.028), while SKAT, SKATO and AdjSKAT did not detect the association. For further exploration, we analyzed each gene separately and test its association with 4 metabolic-related phenotypes. By using GSU, we found a marginal association of ANGPTL4 with the multiple phenotypes(p-value=0.057).

Table 3: The multivariate analysis of 4 metabolic-related phenotypes in the DHS study

gene	rare variants	p-value[1]			
		GSU	SKAT	AdjSKAT	SKATO
ANGPTL3	54	0.300	0.083	0.071	0.112
ANGPTL4	63	0.057	0.079	0.086	0.155
ANGPTL5	61	0.075	0.046	0.073	0.107
ANGPTL6	52	0.297	0.059	0.065	0.101
ALL	230	0.028	0.250	0.174	0.245

[1] Multiple phenotypes considered in the analysis include BMI, cholesterol, HDL, LDL and VLDL. For SKAT-based methods, the p-values listed in the table are the smallest p-value attained from the multiple p-values, where their Bonferroni corrected significance level should be 0.05/5.

5. DISCUSSION

Driven by recent developments in sequencing technologies and the common-diseases-rare-variants hypothesis, sequencing studies are now emerging as a major study design for the genetic association of complex diseases. Yet, the characteristics of sequencing data, including low MAF and high dimensionality, pose daunting challenges to statistical analysis. The conventional methods, such as a single-locus analysis using logistic regression, had low power for sequencing data analysis. Instead, joint association analysis, or SNV-set analysis, is becoming popular owing to its increased power and reduced multiple testing issue. Although the existing joint association methods have nice features and are easy to use, they also have certain limitations. For example, most of burden tests assume homogeneous effects within the SNV-set, which may not reflect the true underlying disease scenarios. Besides burden tests, most of the existing methods are parametric or semi-perimetric, which often assume certain distribution of a phenotype and mode of inheritance. When the assumptions are violated, the results can be unreliable.

To overcome these limitations, we have proposed a generalized similarity U test for multivariate analysis of different types of phenotypes. We conducted extensive simulation studies using data from the 1000 Genome Project and compared the performance of GSU with three other popular methods: SKAT, AdjSKAT and SKATO. For all of the simulation scenarios, including single phenotype with various distributions, and multiple phenotypes with various combinations of phenotype distribution, GSU outperformed the other three methods in terms of robust type I error and higher power. Although the simulation results depend on the simulation settings, and should always be interpreted in the context of the simulation setting, we believe the results reflect the advantage of GSU in a broader sense, because 1) the genetic data used in the simulation comes from the 1000 Genome Project, which reflects the LD pattern and the allelic frequency distribution of in the general population; and 2) we simulated a wide range of disease models to mimic common disease scenarios.

In recent years, U-statistic-based methods have been popularly used in genetic data analysis, and have shown their robustness and flexibility in the analysis of genetic data(Schaid et al., 2005; Zhang et al., 2010; Wei et al., 2013). GSU is a general framework based on similarity measurements. Although in this paper we used the weighted-IBS to calculate genetic similarity based on the SNV-set, other forms of genetic similarity can also be used. For the weighted-IBS similarity, we can modify the weights of the weighted-IBS (i.e., the original weight $w_m = 1/\sqrt{\gamma_m(1 - \gamma_m)}$) to reflect the importance of each SNV. Besides IBS-based similarity, we can also use distance-transformed similarity to model the effect in a nonlinear way. For example, we can use the Euclidean-Distance-based similarity, $K_{i,j}^{ED} = exp(-\sum_{m=1}^{M} w_m(g_{i,m} - g_{j,m})^2)$. In this paper, we have focused on the analysis of categorical sequencing data (SNV data). By using appropriate genetic similarity measurements, GSU can easily be extended to analyze other types of genetic data, such as count data (CNV data) and continuous data (expression data).

The flexibility of GSU also comes from the construction of the phenotype similarity. By using phenotype similarity, GSU can analyze not only a single phenotype, but also multiple phenotypes following different distributions. Many genetic studies collect multiple secondary phenotypes, or use intermediate biomarkers, to study complex diseases. By considering multiple phenotypes that measure the different aspects of underlying diseases, the power of association analysis can be potentially improved (Zhang et al., 2010; Liu et al., 2009b; Maity et al., 2012). Nevertheless, few methods have been developed for multivariate analysis of sequencing data. Methods were recently developed for multivariate analysis of genetic data, but relies on the normal assumption for the trait distribution (Maity et al., 2012). GSU is developed for both univariate and multivariate analysis of sequencing data. It is distribution-free and can analyze multiple phenotypes with a combination of different types of phenotype. By using the appropriate similarity measurement, GSU can be potentially extended to analyze longitudinal data, survival data and even high-dimensional phenotypes, such as brain

imaging data.

GSU is not only statistically powerful, but also computationally efficient. We derived the asymptotic distribution of GSU so as to efficiently calculate the statistical significance for the association. As we showed in the results section, GSU is computationally efficient for sequencing data analysis. We have also developed a C++ version of GSU. With the current fast-moving sequencing technology, large-scale whole-genome sequencing will soon become popular, and GSU could be a computationally efficient and powerful tool to facilitate the sequencing data analysis.

APPENDICES: PROOF AND COMPUTATION DETAILS

Appendix A   Centered Similarity

By the following definition,

$$\tilde{h}(y_1, y_2) = h(y_1, y_2) - E(h(y_1, Y_2)) - E(h(Y_1, y_2)) + E(h(Y_1, Y_2))$$

we can obtain conditional expectation for the centered phenotype similarity,

$$E(\tilde{h}(Y_1, Y_2)|Y_1) = E(h(Y_1, Y_2|Y_1)) - E(h(Y_1, Y_2|Y_1)) - E(h(Y_1, Y_2)) + E(h(Y_1, Y_2)) = 0.$$

Therefore, we have $E(\tilde{h}(Y_1, Y_2)) = 0$ and $Var(E(\tilde{h}(Y_1, Y_2)|Y_1)) = 0$. Using the same argument, we can have the same result for the centered genetic similarity.

Under the null hypothesis, when the genetic similarities are independent of the phenotype similarities, we have,

$$
\begin{aligned}
E(U) &= \frac{1}{n(n-1)} E(\sum_{i \neq j} \tilde{f}(G_i, G_j) \tilde{h}(Y_i, Y_j)) \\
&= \frac{1}{n(n-1)} \sum_{i \neq j} E(\tilde{f}(G_i, G_j)) E(\tilde{h}(Y_i, Y_j)) \\
&= 0.
\end{aligned}
$$

Appendix B   Proof of Theorem 1

We can decompose the centered phenotype similarity by,

$$
\tilde{h}(y_1, y_2) = \sum_{s=1}^{\infty} \lambda_s \phi_s(y_1) \phi_s(y_2),
$$

where $\{\lambda_s\}$ and $\phi_s(\cdot)$ are eigenvalues and eigenfunctions of the kennel $\tilde{h}(\cdot, \cdot)$.

Because of the orthogonality of $\{\phi_s(\cdot)\}$, we have

$$
\begin{aligned}
& E(\tilde{h}(Y_1, Y_2) \phi_{s'}(Y_2) | Y_1) \\
&= \int \tilde{h}(Y_1, y_2) \phi_{s'}(y_2) dF(y_2) \\
&= \sum_{s=1}^{\infty} \lambda_s \phi_s(Y_1) \int \phi_s(y_2) \phi_{s'}(y_2) dF(y_2) \\
&= \lambda_{s'} \phi_{s'}(Y_1).
\end{aligned}
$$

We also showed $E(\tilde{h}(Y_1, Y_2) \times 1 | Y_1) = 0 \times 1$ in Appendix A, which forced $\phi_1(\cdot) = 1$ and $\lambda_1 = 0$ to be an eigenfunction-eigenvalue pair in the decomposition of $\tilde{h}(\cdot, \cdot)$. Again, because $\phi_1(\cdot)$ is orthogonal with $\{\phi_s(\cdot)\}_{s>1}$, for $s > 1$, we have

$$
\begin{aligned}
E\phi_s(Y_1) &= \int \phi_s(y_1) \times 1 dF(y_1) \\
&= \int \phi_s(y_1) \phi_1(y_1) dF(y_1) \\
&= 0.
\end{aligned}
$$

26

Using the same argument, we can have the corresponding results for the decomposition of the centered genetic similarity:

$$\tilde{f}(G_i, G_j) = \sum_{t=1}^{\infty} \eta_t \varphi_t(g_1) \varphi_t(g_2).$$

Then,

$$\begin{cases} E\phi_s(Y) = 0, & \forall\, s > 1 \\ E\varphi_t(G) = 0 & \forall\, t > 1. \end{cases} \tag{Ax.1}$$

Using the function decomposition, we can write GSU as,

$$\begin{aligned}
U &= \frac{1}{n(n-1)} \sum_{i \neq j} \tilde{f}(G_i, G_j) \tilde{h}(Y_i, Y_j) \\
&= \frac{1}{n(n-1)} \sum_{i \neq j} \sum_{t=1}^{\infty} \eta_t \varphi_t(G_i) \varphi_t(G_j) \sum_{s=1}^{\infty} \lambda_s \phi_s(Y_i) \phi_s(Y_j) \\
&= \frac{1}{n(n-1)} \sum_{t=1}^{\infty} \eta_t \sum_{s=1}^{\infty} \lambda_s \sum_{i \neq j} \varphi_t(G_i) \varphi_t(G_j) \phi_s(Y_i) \phi_s(Y_j) \\
&= \frac{1}{n-1} \sum_{t=2}^{\infty} \eta_t \sum_{s=2}^{\infty} \lambda_s \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_t(G_i) \phi_s(Y_i) \right)^2 \\
&\quad - \frac{1}{n-1} \sum_{t=2}^{\infty} \eta_t \sum_{s=2}^{\infty} \lambda_s \frac{1}{n} \sum_{i=1}^{n} (\varphi_t(G_i) \phi_s(Y_i))^2 \\
&= \frac{1}{n-1} \sum_{t=2}^{\infty} \sum_{s=2}^{\infty} sign(\eta_t \lambda_s) \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \eta_t^\star(G_i) \phi_s^\star(Y_i) \right)^2 \\
&\quad - \frac{1}{n-1} \sum_{t=2}^{\infty} \sum_{s=2}^{\infty} sign(\eta_t \lambda_s) \frac{1}{n} \sum_{i=1}^{n} (\eta_t^\star(G_i) \phi_s^\star(Y_i))^2 ,
\end{aligned}$$

where $\varphi_t^\star(G_i) = |\eta_t|^{0.5} \varphi_t(G_i)$ and $\phi_s^\star(Y_i) = |\lambda_s|^{0.5} \phi_s(Y_i)$.

Under the null hypothesis, genotype (G_i) is independent of phenotypes (Y_i). Therefore, for $s > 1$ and $t > 1$,

$$\begin{aligned}
&E(\eta_t^\star(G_1) \phi_s^\star(Y_1)) \\
&= |\eta_t|^{0.5} E\varphi_t(G_1) |\lambda_s|^{0.5} E\phi_s(Y_1) \\
&= 0, \tag{Ax.2}
\end{aligned}$$

and

$$E(\eta_t^\star(G_1)\phi_s^\star(Y_1)\eta_{t'}^\star(G_1)\phi_{s'}^\star(Y_1))$$

$$=|\eta_t\lambda_s\eta_{t'}\lambda_{s'}|^{0.5}E(\varphi_t(G_1)\varphi_{t'}(G_1))E(\phi_s(Y_1)\phi_{s'}(Y_1))$$

$$=\begin{cases} |\eta_t\lambda_s|, & \text{if } s = s' \text{and } t = t' \\ \\ 0, & \text{otherwise.} \end{cases} \tag{Ax.3}$$

Therefore, for any finite subset Δ of $\{(s,t)\}_{s>1,t>1}$, $\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^n \eta_t^\star(G_i)\phi_s^\star(Y_i)\right\}_{(s,t)\in\Delta}$ converges to a multivariate normal distribution by using results from equation Ax.2, equation Ax.3 and multivariate CLT.

Additionally, we can show that,

$$\sum_{s>1,t>1} E(\eta_t^\star(G_1)\phi_s^\star(Y_1))^2$$

$$=\sum_s |\lambda_s| \sum_t |\eta_t|$$

$$\leq \sum_s \lambda_s^2 \sum_t \eta_t^2$$

$$=E(\tilde{f}(G_1,G_2))^2 E(\tilde{h}(Y_1,Y_2))^2$$

$$<\infty$$

Under the condition $\sum_{s>1,t>1} E(\eta_t^\star(G_1)\phi_s^\star(Y_1))^2 < \infty$, the countable sequence of function $\{\eta_t^\star(\cdot)\phi_s^\star(\cdot)\}$ is a Donsker class(van der Vaart and Wellner, 2000). Therefore, the empirical process, $\frac{1}{\sqrt{n}}\sum_{i=1}^n \eta_t^\star(G_i)\phi_s^\star(Y_i)$, converges weakly to the Gaussian process $Z_{s,t}$ with mean zero and covariance function,

$$cov(Z_{s,t}, Z_{s',t'}) = E(\eta_t^\star(G_1)\phi_s^\star(Y_1)\eta_{t'}^\star(G_1)\phi_{s'}^\star(Y_1)) = \begin{cases} |\eta_t\lambda_s|, & \text{if } s = s' \text{and } t = t' \\ \\ 0, & \text{otherwise.} \end{cases}$$

With this uniform convergence (for all $s > 1$ and $t > 1$), we can show that,

$$nU \xrightarrow{D} \sum_{t=2}^{\infty} \sum_{s=2}^{\infty} sign(\eta_t \lambda_s)(Z_{s,t})^2$$

$$- \sum_{t=2}^{\infty} \sum_{s=2}^{\infty} sign(\eta_t \lambda_s)|\eta_t \lambda_s|$$

$$= \sum_{t=1}^{\infty} \eta_t \sum_{s=1}^{\infty} \lambda_s(\chi_{st}^2 - 1)$$

where χ_{st}^2 are i.i.d chi-squared random variables with a d.f. of 1.

Appendix C   Proof of Theorm 2

To simplify the notation, we denote $X = (Y, G)$ and $u(X_1, X_2) = \tilde{f}(G_1, G_2)\tilde{h}(Y_1, Y_2)$. GSU can then be rewritten as:

$$U = \frac{1}{n(n-1)} \sum_{i \neq j} u(X_i, X_j).$$

Define a centered kernel $\tilde{u}(x_1, x_2)$ by:

$$\tilde{u}(x_1, x_2) = u(x_1, x_2) - u_1(x_1) - u_1(x_2) - \mu,$$

where $u_1(x) = E(u(X_1, X_2)|X_1 = x)$.

We can decompose the GSU as follows:

$$
\begin{aligned}
U &= \frac{1}{n(n-1)} \sum_{i \neq j} u(X_i, X_j) \\
&= \frac{1}{n(n-1)} \sum_{i \neq j} (\tilde{u}(X_i, X_j) + u_1(X_i) + u_1(X_j) - \mu) \\
&= \frac{1}{n(n-1)} \sum_{i \neq j} \tilde{u}(X_i, X_j) + \frac{2}{n} \sum_{i=1}^{n} (u_1(X_i) - \mu) + \mu.
\end{aligned}
$$

Thus,

$$\sqrt{n}(U - \mu) = \frac{2}{\sqrt{n}} \sum_{i=1}^{n} (u_1(X_i) - \mu) + \frac{\sqrt{n}}{n(n-1)} \sum_{i \neq j} \tilde{u}(X_i, X_j).$$

Becuase $E(u_1(X)) = \mu$ and $Var(u_1(X)) = \zeta_1$, the first term converges to normal distribution by applying CLT:

$$\frac{2}{\sqrt{n}} \sum_{i=1}^{n} (u_1(X_i) - \mu) \xrightarrow{D} N(0, 4\zeta_1).$$

29

Then we need to show:

$$R = \frac{\sqrt{n}}{n(n-1)} \sum_{i \neq j} \tilde{u}(X_i, X_j) \xrightarrow{p} 0.$$

This can be done by proving $ER^2 \to 0$, using the fact that $E(\tilde{u}(X_1, X_2)) = 0$, $Var(\tilde{u}(X_1, X_2)) < \infty$ and $E(\tilde{u}(X_1, X_2)|X_1) = 0$. In fact, by using the similar technique in Appendix C, we can show that $\sqrt{n}R$ asymptotically follows the distribution of a weighted sum of independent chi-square random variables.

Appendix D   Matrix Similarity

In the study sample, we can get the centered phenotype similarity by:

$$\tilde{h}(y_i, y_j) = h(y_i, y_j) - \frac{1}{n} \sum_{j=1}^{n} h(y_i, y_j) - \frac{1}{n} \sum_{i=1}^{n} h(y_i, y_j) + \frac{1}{n^2} \sum_{i,j} h(y_i, y_j).$$

Denote $\tilde{S}_{i,j} = \tilde{h}(y_i, y_j)$ and $S_{i,j} = h(y_i, y_j)$, and the above equation becomes:

$$\tilde{S}_{i,j} = S_{i,j} - \frac{1}{n} \sum_{j=1}^{n} S_{i,j} - \frac{1}{n} \sum_{i=1}^{n} S_{i,j} + \frac{1}{n^2} \sum_{i,j} S_{i,j}.$$

The equations can be written in a matrix form,

$$\begin{aligned}
\tilde{S} &= S - JS - SJ + JSJ \\
&= (I - J)S(I - J),
\end{aligned}$$

where $\tilde{S} = \{\tilde{h}(y_i, y_j)\}_{n \times n}$, $S = \{h(y_i, y_j)\}_{n \times n}$, $I = \{1_{\{i=j\}}\}_{n \times n}$, and $J = \{\frac{1}{n}\}_{n \times n}$.

Similarly, the centered genetic similarity can also be written in the matrix form:

$$\tilde{K} = (I - J)K(I - J),$$

where $\tilde{K} = \{\tilde{f}(g_i, g_j)\}_{n \times n}$, and $K = \{f(g_i, g_j)\}_{n \times n}$.

Appendix E   Limiting Distribution

In the actual computation, we will use a matrix eigen-decomposition to obtain the eigen-value and eigenvectors as a finite-dimension approximation of the eigenfunction. For a matrix

eigen-decomposition, a computer algorithm usually gives the eigenvalue λ_s with the eigenvector ϕ_s, which satisfies $\sum_{i=1}^n \phi_{s,i}^2 = 1$ instead of $\frac{1}{n} \sum_{i=1}^n \phi_{s,i}^2 = 1$. Therefore, using the eigenvalues $\tilde{\lambda}_s$ and $\tilde{\eta}_t$ calculated from the matrix eigen-decomposition, the limiting distribution of GSU is:

$$nU \sim \sum_{t=1}^n \frac{\tilde{\eta}_t}{n} \sum_{s=1}^n \frac{\tilde{\lambda}_s}{n} (\chi_{st}^2 - 1).$$

.

REFERENCES

Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., and McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73. 3.1

Ahituv, N., Kavaslar, N., Schackwitz, W., Ustaszewska, A., Martin, J., Hebert, S., Doelle, H., Ersoy, B., Kryukov, G., Schmidt, S., Yosef, N., Ruppin, E., Sharan, R., Vaisse, C., Sunyaev, S., Dent, R., Cohen, J., McPherson, R., and Pennacchio, L. A. (2007). Medical sequencing at the extremes of human body mass. *American Journal of Human Genetics*, 80(4):779–791. 1, 4

Boyko, A. R., Williamson, S. H., Indap, A. R., Degenhardt, J. D., Hernandez, R. D., Lohmueller, K. E., Adams, M. D., Schmidt, S., Sninsky, J. J., Sunyaev, S. R., White, T. J., Nielsen, R., Clark, A. G., and Bustamante, C. D. (2008). Assessing the evolutionary impact of amino acid mutations in the human genome. *Plos Genetics*, 4(5). 1

Cohen, J. C., Kiss, R. S., Pertsemlidis, A., Marcel, Y. L., McPherson, R., and Hobbs, H. H. (2004). Multiple rare alleles contribute to low plasma levels of hdl cholesterol. *Science*, 305(5685):869–872. 1

Davies, R. B. (1980). Algorithm as 155: The distribution of a linear combination of x2 random variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 29(3):323–333. 2.5

Fay, J. C., Wyckoff, G. J., and Wu, C. I. (2001). Positive and negative selection on the human genome. *Genetics*, 158(3):1227–1234. 1

Gregory, G. G. (1977). Large sample theory for u-statistics and tests of fit. *Annals of Statistics*, 5(1):110–123. 2.3

Ji, W. Z., Foo, J. N., O'Roak, B. J., Zhao, H., Larson, M. G., Simon, D. B., Newton-Cheh, C., State, M. W., Levy, D., and Lifton, R. P. (2008). Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nature Genetics*, 40(5):592–599. 1

Kryukov, G. V., Pennacchio, L. A., and Sunyaev, S. R. (2007). Most rare missense alleles are deleterious in humans: Implications for complex disease and association studies. *American Journal of Human Genetics*, 80(4):727–739. 1

Kuonen, D. (1999). Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika*, 86(4):929–935. 2.5

Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., Christiani, D. C., Wurfel, M. M., and Lin, X. H. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *American Journal of Human Genetics*, 91(2):224–237. 3.2

Li, B. S. and Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *American Journal of Human Genetics*, 83(3):311–321. 1

Lin, D.-Y. and Tang, Z.-Z. (2011). A general framework for detecting disease associations with rare variants in sequencing studies. *American journal of human genetics*, 89(3):354–367. 1

Lindsay, B. G., Markatou, M., Ray, S., Yang, K., and Chen, S. C. (2008). Quadratic distances on probabilities: A unified foundation. *Annals of Statistics*, 36(2):983–1006. 2.3

Liu, H., Tang, Y. Q., and Zhang, H. H. (2009a). A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics & Data Analysis*, 53(4):853–856. 2.5

Liu, J., Pei, Y., Papasian, C. J., and Deng, H.-W. (2009b). Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations. *Genetic Epidemiology*, 33(3):217–227. 5

Luo, L., Zhu, Y., and Xiong, M. (2012). Quantitative trait locus analysis for next-generation sequencing with the functional linear models. *Journal of Medical Genetics*, 49(8):513–524. 1

Lynch, M. and Ritland, K. (1999). Estimation of pairwise relatedness with molecular markers. *Genetics*, 152(4):1753–1766. 2.2

Madsen, B. E. and Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *Plos Genetics*, 5(2). 1

Maity, A., Sullivan, P. E., and Tzeng, J. Y. (2012). Multivariate phenotype association analysis by marker-set kernel machine regression. *Genetic Epidemiology*, 36(7):686–695. 5

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). *Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis*, 615(1-2):28–56. 1

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 Genetics*, 7(3). 1

O'Neil, K. A. and Redner, R. A. (1993). Asymptotic distributions of weighted u-statistics of degree 2. *The Annals of Probability*, 21(2):1159–1169. 2.3

Pritchard, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? *American Journal of Human Genetics*, 69(1):124–137. 1

Romeo, S., Yin, W., Kozlitina, J., Pennacchio, L. A., Boerwinkle, E., Hobbs, H. H., and Cohen, J. C. (2009). Rare loss-of-function mutations in angptl family members contribute to plasma triglyceride levels in humans. *Journal of Clinical Investigation*, 119(1):70–79. 1

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. *American Journal of Human Genetics*, 76(5):780–793. 5

Shapiro, C. P. and Hubert, L. (1979). Asymptotic normality of permutation statistics derived from weighted sums of bivariate functions. *The Annals of Statistics*, 7(4):788–794. 2.3

Shieh, G. S. (1997). Weighted degenerate u- and v-statistics with estimated parameters. *Statistica Sinica*, 7(4):1021–1038. 2.3

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(3):822–832. 2.2

van der Vaart, A. and Wellner, J. A. (2000). *weak convergence and empirical processes*. Springer, 2 edition. Appendix B

Wei, C. S., Schaid, D. J., and Lu, Q. (2013). Trees assembling mann-whitney approach for detecting genome-wide joint association among low-marginal-effect loci. *Genetic Epidemiology*, 37(1):84–91. 5

Wu, M. C., Lee, S., Cai, T. X., Li, Y., Boehnke, M., and Lin, X. H. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*, 89(1):82–93. 1, 3.2

Zhang, H. P., Liu, C. T., and Wang, X. Q. (2010). An association test for multiple traits based on the generalized kendall's tau. *Journal of the American Statistical Association*, 105(490):473–481. 5