Published in final edited form as:

Stat Med. 2016 July 20; 35(16): 2802–2814. doi:10.1002/sim.6877.

A Weighted U Statistic for Association Analyses Considering Genetic Heterogeneity

Changshuai Wei^{a,*}, Robert C. Elston^b, and Qing Lu^{c,*}

^aDepartment of Biostatistics and Epidemiology, University of North Texas Health Science Center

^bDepartment of Epidemiology and Biostatistics, Case Western Reserve University

^cDepartment of Epidemiology and Biostatistics, Michigan State University

Abstract

Converging evidence suggests that common complex diseases with the same or similar clinical manifestations could have different underlying genetic etiologies. While current research interests have shifted toward uncovering rare variants and structural variations predisposing to human diseases, the impact of heterogeneity in genetic studies of complex diseases has been largely overlooked. Most of the existing statistical methods assume the disease under investigation has a homogeneous genetic effect and could, therefore, have low power if the disease undergoes heterogeneous pathophysiological and etiological processes. In this paper, we propose a heterogeneity weighted U (HWU) method for association analyses considering genetic heterogeneity. HWU can be applied to various types of phenotypes (e.g., binary and continuous) and is computationally efficient for high-dimensional genetic data. Through simulations, we showed the advantage of HWU when the underlying genetic etiology of a disease was heterogeneous, as well as the robustness of HWU against different model assumptions (e.g., phenotype distributions). Using HWU, we conducted a genome-wide analysis of nicotine dependence from the Study of Addiction: Genetics and Environments (SAGE) dataset. The genome-wide analysis of nearly one million genetic markers took 7 hours, identifying heterogeneous effects of two new genes (i.e., *CYP3A5* and *IKBKB*) on nicotine dependence.

Keywords

High-dimensional Data; Non-parametric Statistic; Nicotine Dependence

1. Introduction

Benefiting from high-throughput technology and ever-decreasing genotyping cost, large-scale genome-wide and sequencing studies have become commonplace in biomedical

*Correspondence to: Qing Lu, Department of Epidemiology and Biostatistics, Michigan State University. qlu@epi.msu.edu; Changshuai Wei, Department of Biostatistics and Epidemiology, University of North Texas Health Science Center. changshuai.wei@unthsc.edu.

Disclosure Declaration: We declare no conflict of interest.

Supplementary Material: Materials include Supplementary Appendix A to C, Supplementary Simulation I to II, Supplementary Real Data Analysis, Supplementary Table S1 to S9, and Supplementary Figure S1 to S4.

research. From these large-scale studies, thousands of genetic variants have been identified as associated with complex human diseases, some with compelling biological plausibility for a role in the disease pathophysiology and etiology. Despite such success, for most complex diseases the identified genetic variants account for only a small proportion of the heritability. While substantial efforts have shifted toward finding rare variants, gene-gene/gene-environment interactions, structural variations, and other genetic variants accounting for the missing heritability [1], there is a considerable lack of attention being paid to genetic heterogeneity in the analysis of complex human diseases.

Genetic heterogeneity represents genetic causes of a disease being heterogeneous for different group of people. Traditional models of genetic heterogeneity, such as allelic heterogeneity and locus heterogeneity, originate from linkage analysis[2]. In allelic heterogeneity or locus heterogeneity, multiple alleles or loci are independently associated with disease. For example, the causes of a disease will exhibit allelic heterogeneity, if a risk allele exists in one population but not in another population. In other words, the focus of allelic or locus heterogeneity is on the heterogeneity of the distributions of a variant in different groups of people, while in this paper we focus on heterogeneous effects of a genetic variant. In particular, we define genetic heterogeneity as a genetic variant having different effects on individuals or on subgroups of a population (e.g., gender and ethnic groups). For instance, the effect size and the effect direction of the genetic variant can be different according to the individuals' genetic background, personal/demographic characteristics and/or the phenotype sub-groups they belong to. Furthermore, the genetic heterogeneity defined here is different from an interaction effect. The key difference is that we assume there is a latent population structure that acts in some joint fashion with the genetic variants, while in the usual interaction effect model the genetic variants are assumed to interact with known variables.

Substantial evidence from a wide range of diseases suggests that complex diseases are characterized by remarkable genetic heterogeneity [3, 4, 5]. Despite the strong evidence of genetic heterogeneity in human disease etiology, investigating genetic variants with heterogeneous effects remains a great challenge, primarily because: i) the commonly used study designs (e.g. the case-control design) may not be optimal for studying heterogeneous effects; ii) there is a lack of prior knowledge that can be used to infer the latent similarity (i.e., heterogeneous subgroups in the population); iii) replication studies are more challenging and need to be carefully designed; and iv) computationally efficient and flexible statistical methods for high-dimensional data analysis, taking into account genetic heterogeneity, have not been well developed. Most of the existing methods assume that the disease under investigation is a unified phenotype with homogeneous genetic causes. When genetic heterogeneity is present, the current methods will likely yield attenuated estimates for the effects of genetic variants, leading to low power of the study.

To account for genetic heterogeneity in association analyses, we propose a heterogeneity weighted U, referred to as HWU. Because the new method is based on a weighted U statistic, it assumes no specific distribution of phenotypes; it can be applied to both qualitative and quantitative phenotypes with various types of distributions. Moreover, HWU

is computationally efficient and has been implemented in a C++ package for high-dimensional data analyses (<https://www.msu.edu/~changs18/software.html#HWU>).

2. Method

2.1. Motivation from a Gaussian random effect model

To motivate the idea of the heterogeneity weighted U, we first introduce a Gaussian random effect model to test genetic association when considering genetic heterogeneity. Assume the following random effect model,

$$Y_i = \mu + Z_i^T \alpha + g_i \beta_i + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2),$$

where Y_i , Z_i and g_i represent the phenotype, the covariates to be adjusted and the single-locus genotype of individual i , respectively. Z_i can be confounding covariates as well as principle components used to adjust for population stratification. g_i can be coded as 0, 1, and 2 (i.e., the additive model), or 0 and 1 (e.g., the dominant/recessive model); β_i is normally distributed, $\beta_i \sim N(\mu_b, \sigma_b^2)$, and ε_i is the iid random error. Let $\kappa_{i,j}$ represent the background similarity or the latent similarity for individuals i and j . We assume that the more similar two individuals are, the more similar are their genetic effects, i.e., $\text{cov}(\beta_i, \beta_j) = \kappa_{i,j} \sigma_b^2$.

We define $\beta = (\beta_1, \dots, \beta_n)^T$, $Y = (Y_1, \dots, Y_n)^T$, $Z = (Z_1, \dots, Z_n)^T$, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$, $G = \{\text{diag}(g_1, \dots, g_n)\}_{n \times n}$, and $K = \{\kappa_{i,j}\}_{n \times n}$. The model can then be written as: $Y = \mu + Za + G\beta + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 I)$, $\beta \sim N(\mu_b \mathbf{1}, \sigma_b^2 K)$. We denote $\delta = G\beta$, and rewrite the model as: $Y = \mu + Za + \delta + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 I)$, $\delta \sim N(\mu_b G \mathbf{1}, \sigma_b^2 GKG)$. A variance component score test statistic can be formed to test the $\delta = \mathbf{0}$,

$$T = \tilde{Y}^T GKG \tilde{Y},$$

where $\tilde{Y}_i = (Y_i - \mu - Z_i^T \alpha) / \sigma$ is the standardized residual under the null. We can partition the test statistic T into two parts, $T = \sum_{i \neq j} \kappa_{i,j} g_i g_j \tilde{Y}_i \tilde{Y}_j + \sum_{i=1}^n g_i^2 \tilde{Y}_i^2$, where the first summation is closely related to the weighted U statistic introduced below.

2.2. Heterogeneity weighted U

The Gaussian random effect model assumes a normal distribution. In order to consider phenotypes with various distributions and modes of inheritance, we develop a heterogeneity weighted U with rank-based U kernels and flexible weight functions. We first order the subjects according to their phenotypic values Y_i and assign subject scores based on their ranks, denoted by R_i , $i = 1, \dots, n$. When there are ties in the sample, we assign the averaged rank. For example, in a case-control study with N_0 controls ($Y_i = 0$) and N_1 cases ($Y_i = 1$), all the controls are assigned a score $(N_0 + 1)/2$, while all cases are assigned $N_0 + (N_1 + 1)/2$. The phenotypic similarity between subjects i and j can be defined as,

$$S_{i,j}=h(R_i, R_j),$$

where $h(\cdot, \cdot)$ is a two degree mean zero symmetric kernel function (i.e., $h(R_i, R_j) = h(R_j, R_i)$ and $E_H(h(R_i, R_j)) = 0$) that satisfies the finite second moment condition, $E_H(h^2(R_i, R_j)) < \infty$, and the degenerate kernel condition, $\text{var}(E(h(R_i, R_j)|R_j)) = 0$. In this paper, we choose

$h(R_i, R_j) = \sigma_R^{-2}(R_i - \mu_R)(R_j - \mu_R)$, where $\mu_R = E(R)$ and $\sigma_R^2 = \text{var}(R)$. Let $G_i = (g_{i,1}, \dots, g_{i,Q})$ denote the multiple genetic variants for individual i . We further define a weight function to measure the genetic similarity under the latent population structure $\kappa_{i,j}$

$$w_{i,j} = \kappa_{i,j} f(G_i, G_j),$$

where $f(G_i, G_j)$ represents the genetic similarity calculated based on the genetic variants of interest. We can then form the heterogeneity weighted U, referred to as HWU,

$$U = 2 \sum_{1 \leq i < j \leq n} w_{i,j} S_{i,j},$$

to evaluate the association between the phenotype and the genetic variants, considering the latent population structure.

Thus HWU is a summation, over all pairs of individuals, of their phenotypic similarities weighted by their genetic similarities. Under the null hypothesis of no association, the phenotypic similarity is unrelated to the genetic similarity. Because the phenotypic similarity has mean 0 (i.e., $E_H(h(R_i, R_j)) = 0$), the expectation of HWU is 0. Under the alternative, the phenotypic similarities should increase as the genetic similarities increases. The positive phenotypic similarities are more heavily weighted and the negative phenotypic similarities are more lightly weighted, leading to a positive value of HWU under the alternative.

2.3. Asymptotic distribution of heterogeneity weighted U

To assess the significance of the association, a permutation test can be used to calculate a p-value for HWU. However, for high-dimensional data, the permutation test could be computationally intensive. Therefore, we derive the asymptotic distribution of HWU under the null hypothesis.

The asymptotic properties of the un-weighted U statistic (i.e., $w_{i,j} \equiv 1$) are well established [6, 7]. When the kernel is non-degenerate ($\text{var}(E(h(R_i, R_j)|R_j)) > 0$), the limiting distribution is normal. When the kernel is degenerate ($\text{var}(E(h(R_i, R_j)|R_j)) = 0$), the limiting distribution is a sum of independent chi-square variables. However, the limiting distribution of the weighted U statistic depends on both the weight function and the kernel function [8]. Because non-normality also occurs for a non-degenerate kernel with certain weight functions, we use the degenerate kernel for HWU to obtain a unified form of limiting distribution, as shown in the following derivation.

We first expand the kernel function $h(\cdot, \cdot)$ as the sum of products of its eigenfunctions. Let $\{\alpha_t\}$ and $\{\phi_t(\cdot)\}$ denote the eigenvalues and the corresponding ortho-normal eigenfunctions of the kernel. We can write $h(\cdot, \cdot)$ as $h(R_i, R_j) = \sum_{t=1}^{\infty} \alpha_t \phi_t(R_i) \phi_t(R_j)$, and the weighted U as,

$$U = \sum_{i \neq j} w_{i,j} \sum_{t=1}^{\infty} \alpha_t \phi_t(R_i) \phi_t(R_j),$$

Where

$$E(\phi_t(R_i) \phi_s(R_j)) = \begin{cases} 1, & t=s \text{ and } i=j \\ 0, & \text{otherwise.} \end{cases}$$

By exchanging the two summations, $(\sum_{t=1}^{\infty} \alpha_t \sum_{i \neq j} w_{i,j} \phi_t(R_i) \phi_t(R_j))$, the weighted U statistic is an infinite sum of quadratic forms and can be approximated by a linear combination of chi-square random variables [9, 10]. Letting $W = \{w_{i,j}\}_{n \times n}$ be the weight matrix with all diagonal element equal to 0, the limiting distribution can be written as

$$U \sim \sum_{t=1}^{\infty} \alpha_t \sum_{s=1}^n \lambda_s (\chi_{1,ts}^2 - 1),$$

where $\{\lambda_s\}$ are the eigenvalues of the weight matrix and $\{\chi_{1,ts}^2\}$ are iid chi-square random variables with 1 df. In this paper, we use a cross product kernel,

$h(R_i, R_j) = \sigma_R^{-2} (R_i - \mu_R)(R_j - \mu_R)$. In this case, the expansion of $h(\cdot, \cdot)$ can be simplified to $h(R_i, R_j) = \alpha_1 \phi_1(R_i) \phi_1(R_j)$, where $\alpha = 1$ and $\phi_1(R) = \sigma_R^{-1} (R - \mu_R)$. Using this

representation and the fact that $\sum_{s=1}^n \lambda_s = 0$, the limiting distribution can be simplified to $U \sim \sum_{s=1}^n \lambda_s \chi_{1,s}^2$. We also note that the parameters μ_R and σ_R^2 are unknown and need to be estimated from the data, which influences the limiting distribution of HWU [11, 12]. Taking the parameter estimation into account, the limiting distribution can be expressed as a

weighted sum of independent chi-squared variables, $U \sim \sum_{s=1}^n \lambda_{1,s} \chi_{1,s}^2$ (Appendix A.1), where $\{\lambda_{1,s}\}$ are the eigenvalues of the matrix $(I - J)W(I - J)$, in which I is an identity matrix and J is a matrix with all elements equal to $1/n$.

The HWU described above can also be modified to allow for covariate adjustment. Suppose $Z_{n \times p} = (1, z_1, \dots, z_p)$ is the covariate matrix. In the cross product kernel of HWU, we can calculate the estimators of μ_R and σ_R^2 as $\hat{\mu}_R = PR$ and $\hat{\sigma}_R^2 = (R - \hat{\mu}_R)^T (R - \hat{\mu}_R) / (n - p - 1)$,

where $P = Z(Z^T Z)^{-1} Z^T$. The limiting distribution can then be written as $U \sim \sum_{s=1}^n \lambda_{1,s}^* \chi_{1,s}^2$, where $\{\lambda_{1,s}^*\}$ are the eigenvalues of the matrix $(I - P)W(I - P)$.

Davies' method [13] can be used to calculate the p-value for the association test. When the calculation involves large matrix eigen-decomposition, we use the state-of-the-art algorithm nu-TRLan [14] to improve the computational efficiency.

2.4. Weighting schemes

The weight function comprises two components, $\kappa_{i,j}$ and $f(G_i, G_j)$. The definition takes into account genetic variants of interests as well as latent heterogeneity. Two individuals, even with same genetic variants of interests ($G_i = G_j$), can have different risk of disease if they come from a heterogeneous population, measured by $\kappa_{i,j}$. Depending on the type of data, different functions can be used to calculate $\kappa_{i,j}$. A common way of calculating genetic similarity is termed Identity By State (IBS). Let x_i be the vector of D genetic variants for individual i , then the IBS for subject i and j can be calculated as

$\kappa_{i,j} = \sum_{d=1}^D |x_{i,d} - x_{j,d}| / (2D)$. We can apply the genome-wide averaged IBS function on GWAS data and the genome-wide weighted average IBS (WIBS) function on sequencing data to calculate $\kappa_{i,j}$ [15]. Therefore, $\kappa_{i,j}$ contains the information across the whole genome and represents the similarity of genome profiles for two individuals. Using HWU, we expect to detect heterogeneous effects of a variant if it lies in different genome profiles. We can also use environmental covariates to calculate $\kappa_{i,j}$ based on Euclidian distance [16]. Given environmental covariates, we first standardize each covariate according to its mean and standard deviation, denoted by x_d ($x_d = (x_{d,1}, \dots, x_{d,n})^T$, $d = 1, 2, \dots, D$), and then calculate $\kappa_{i,j} = \exp(-(x_i - x_j)^T R (x_i - x_j))$, where R is used to reflect the relative importance (e.g., $R = \{diag(\omega_d)\}_{D \times D}$ in which ω_d measures the importance) or inner correlation

(e.g., $R = (\frac{1}{n} \sum_{i=1}^n x_i^T x_i)^{-1}$) of the covariates.

$f(G_i, G_j)$ measures the genetic similarity at genetic variants of interest. For a single-locus model, we can use the cross product $f(G_i, G_j) = f(g_i, g_j) = g_i g_j$ when the effect is additive. Otherwise, we can use $f(G_i, G_j) = f(g_i, g_j) = 1(g_i = g_j)$ for an unspecified mode of inheritance, where $1(\cdot)$ is the indicator function. The above measurements can be easily

extended to handle Q multiple markers by using $f(G_i, G_j) = \sum_{q=1}^Q g_{q,i} g_{q,j}$.

The weight function $w_{i,j}$ can also be specified for different purposes. For instance, if we choose $w_{i,j} = f(G_i, G_j)$ (i.e., $\kappa_{i,j} \equiv 1$), then the weighted U tests the association without consideration of genetic heterogeneity. We refer to this statistic as the non-heterogeneity weighted U (NHWU). Furthermore, we can construct a statistic to examine the difference between HWU and NHWU, by setting $w_{i,j}^* = (\kappa_{i,j} - \bar{\kappa}) f(G_i, G_j)$, where

$$\bar{\kappa} = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \kappa_{i,j}.$$

3. Results

3.1. Simulations

In simulation I and simulation II, we simulated various cases of genetic heterogeneity and compared the proposed HWU test with two other tests, NHWU and the likelihood ratio test using the conventional generalized linear model (GLM). In simulation III, we investigated the robustness of HWU to non-normal distributions and mis-specified weight functions. In all sets of simulations, unless otherwise specified we used Euclidian-distance-based $\kappa_{i,j}$ by setting $R = I$ and cross-product-based $f(g_i, g_j)$ to form the weight function. For each simulation setting, we simulated 1000 replicate datasets, each having a sample size of 1000. Power and type 1 error of the methods were first calculated based on the proportion of p-values in the 1000 replicates smaller than or equal to 0.05.

3.1.1. Simulation I—In this simulation, we assumed two sub-populations, and considered both continuous and binary phenotypes. We simulated binary phenotypes using the logistic model,

$$\text{logit}(P(y_{i,j(i)}=1))=\mu+g_{i,j(i)}\beta_i,$$

where i and $j(i)$ represented respectively the i -th subpopulation and j -th individual in the i -th sub-population. A set of numbers $\{a_i\}$ was used to represent the sub-population structure. However, in reality, the sub-population is latent or unobserved. Thus, we generated a covariate $x_{i,j(i)} = a_i + \delta_{i,j(i)}$, $\delta_{i,j(i)} \sim N(0, \sigma_c^2)$, from which we inferred the latent sub-populations. Continuous phenotypes were simulated similarly by using a linear regression model. The values of the regression coefficients β_i for different models are listed in Table 1, while the details of the simulation are described in Supplementary Appendix A.

No substantial inflation of type I error was detected for any of the three methods (Table 1). In the presence of genetic heterogeneity (i.e., T1, T3, and T4 in Table 1), HWU outperformed NHWU and GLM, especially when the genetic effects for the two sub-populations were in the opposite direction (i.e., T1). In such a case, NHWU and GLM could barely detect any genetic effect, while HWU had high statistical power to detect the association. In the absence of genetic heterogeneity (i.e., T2 in Table 1), HWU remained comparable in performance to NHWU and GLM. We also noted that, in the absence of genetic heterogeneity (i.e., T2), the non-parametric NHWU had almost identical power to GLM.

We also investigated the performance of the three methods when the underlying phenotype distribution and the modes of inheritance were unknown (Supplementary Simulation I). Overall, HWU outperformed the other two methods. In particular, when the phenotype was non-normal, both HWU and NHWU had higher power than GLM (Supplementary Table S1). By using $f(g_i, g_j) = 1(g_i = g_j)$, HWU was robust to the disease model when the mode of inheritance was unknown, e.g., heterozygote effect (Supplementary Tables S2 and S3).

3.1.2. Simulation II—In simulation II, we used the same simulation model as in simulation I, but considered a more complicated latent population structure by increasing the number of sub-populations to 20, and sampling β_i ($i = 1, 2, \dots, 20$) from a uniform distribution with mean μ_β and variance σ_β^2 . We simulated 25 covariates, $x_{i,j(i)}^d = a_i^d + \delta_{i,j(i)}$, $\delta_{i,j(i)} \sim N(0, \sigma_c^2)$, ($d = 1, 2, \dots, 25$), to generate the latent population structure (Supplementary Appendix B). No substantial inflation of type I error was detected for any of the three methods at the 0.05 level (Supplementary Table S4). Through simulation, we demonstrated that HWU outperformed NHWU and GLM for both binary (Figure 1) and continuous (Figure 2) phenotypes. In the presence of genetic heterogeneity (i.e., when σ_β/μ_β is large), HWU attained higher power than NHWU and GLM. When the genetic heterogeneity was negligible (i.e., when σ_β/μ_β is small), HWU had comparable performance to NHWU and GLM. When the average genetic effect (μ_β) increased, all three methods gained power. Nevertheless, when the variance of the genetic effect (σ_β) increased, only HWU gained substantial increase in power. We also investigated the performance of HWU when the covariates could not accurately infer the latent population structure. For this purpose, we investigated the power of HWU as the noise parameter σ_c^2 changed. The result showed that the power of HWU decreased as the “noise” increased (Supplementary Table S5).

In practice, the nature of the latent population structure may not be “categorical”. Therefore, we also simulated genetic effects using a random effect model, where effects were different for each subject (Supplementary Simulation II). The three methods had comparable power when the genetic heterogeneity was negligible. Nevertheless, as the genetic heterogeneity increased, there was a clear advantage of HWU over NHWU and GLM (Supplementary Figures S1 and S2).

3.1.3. Simulation III—In simulation III, we first investigated the robustness of HWU against different non-normal phenotype distributions. In order to separate the influence of heterogeneity and phenotype distribution, we compared HWU with its “parametric alternative”, the variance component score test (VCscore), instead of GLM. We simulated the phenotype using a random effect model,

$$y_i = \mu + Z_i\alpha + g_i\beta_i + \varepsilon_i, \varepsilon_i \sim F,$$

where Z_i denotes covariates for subject i , α denotes covariate effects and F followed a non-normal distribution (Supplementary Appendix C). We simulated three types of non-normal distribution for F : 1) t distributions with $df = 2$, 2) Cauchy distribution, and 3) a mixture of normal and chi-squared distribution. For each distribution, we simulated the model with and without confounding effects, where the confounding effect is simulated by generating Z_i that is correlated with g_i . Furthermore, Z_i is also correlated with y_i since $\alpha \neq 0$. We included Z in the analysis for both HWU and VCscores, and summarize the Type I errors in Table 2. No substantial inflation of type I error was detected for HWU for 3 non-normal distributions, regardless of whether there were confounding effects. VCscores is robust against a mixture of normal and chi-squared distributions, but have inflated type I error for heavy tailed

distribution (e.g., Cauchy distribution). If we did not include Z in the analysis, both methods showed inflated type I error when there were confounding effects (Supplementary Table S6). Further investigations on power performance showed slightly more advantage of HWU over VCscore for non-normal distributions (Supplementary Table S7).

We also investigated the performance of HWU when the weight function was mis-specified (Table 3). In this simulation, we considered 4 different scenarios, either with mis-specified $f(g_i, g_j)$ or mis-specified $\kappa_{i,j}$. Type I error rates were well controlled when the weight function was mis-specified. However, we found the power of HWU with a mis-specified weight function was lower than that with a correct weight function, especially when $\kappa_{i,j}$ was mis-specified (Table 3).

Moreover, we investigated the performance of HWU, when the covariates used to infer $\kappa_{i,j}$ are partially mis-selected (Supplementary Table S9). We found that the power of HWU is decreased when some covariates are missing or when some noise covariates are added for the calculation of $\kappa_{i,j}$. When some covariates used to infer $\kappa_{i,j}$ are correlated with the phenotype, the power is also decreased, but the type I error is well controlled.

In GWAS, usually a large number of genetic variants are tested, which requires a more stringent p-value cut-off than 0.05. We therefore performed additional simulation from 1 million simulated data replicates, and the results show HWU can control type I error at more stringent p-value cut-off (Supplementary Table S10). For example, at 5×10^{-5} cut-off, the type I error of HWU is 4.0×10^{-5} .

Though the focus of the paper is single locus association, in reality, multiple genetic variants often jointly act on the disease phenotype. We simulated multi-gene models that contained 5

genes (Supplementary Table S11), where $f(G_i, G_j) = \sum_{q=1}^Q g_{q,i} g_{q,j}$ was used to calculate genetic similarity. Type I error is well controlled for HWU (0.046) and HWU has slightly higher power than VCscore for a non-normal distribution (e.g., 0.648 for HWU v.s. 0.521 for VCscore).

3.2. Genome-wide association analysis of Nicotine Dependence

We applied our methods to the Genome-wide association study (GWAS) dataset from the Study of Addiction: Genetics and Environments (SAGE). SAGE is one of the largest and most comprehensive case-control studies conducted to date aimed at discovering new genetic variants contributing to addiction. Multiple phenotypes were measured in the SAGE studies, including DSM-IV Nicotine Dependence diagnostic criteria (DSM-IV ND) and Fagerstrom Test for Nicotine Dependence (FTND). While DSM-IV ND is generally recognized as a valid and reliable diagnostic criteria for ND, the number of cigarettes smoked per day (CPD), the 4th item in FTND, is considered as a measurement capturing genetic aspects of ND, which has been popularly used as a phenotype for gene mapping[17]. Thus we used CPD as phenotype, which has been categorized into 4 classes (0 for less than 10 cigarettes, 1 for 11-to-20 cigarettes, 2 for 21-to-30 cigarettes, and 3 for more than 31 cigarettes). After undertaking a careful quality control process (i.e., removing samples with missing phenotype data and low-quality genetic markers), 2845 subjects and 949,658 single-

nucleotide polymorphisms (SNPs) remained for the analysis. The SAGE comprises samples from both Caucasian and African-American populations. To make the association analysis robust against this confounding effects, we adjusted for the first 20 principal components from the available genome-wide genetic markers, as well as for gender and race, in the analysis.

Considering that the etiology of nicotine dependence has been shown to be heterogeneous for gender [18], we used gender to infer the latent population structure (i.e., $\kappa_{ij} = \exp(-(x_i - x_j)^2)$, where, $x_i = 0, 1$ for male and female respectively) and assumed an additive effect for genetic variants (i.e., $f(g_i, g_j) = g_i g_j$). Using HWU, the genome-wide scanning of 949,658 SNPs on the SAGE dataset was completed in about 7 hours by parallel computation on 19 cores. The top 10 SNPs having the strongest association with nicotine dependence are listed in Table 4. Among the 10 SNPs, 3 SNPs (rs4560769, rs9694574, and rs9694958) are located within the gene *IKBKB*, while another 3 SNPs (rs4646437, rs4646457, rs776746) are located within or near the gene *CYP3A5*. The 3 SNPs related to gene *IKBKB* are in high linkage disequilibrium (LD), with the estimated correlations ranging from 0.736 to 0.853. The highest association signal was from rs4560769 (p-value = 1.93×10^{-8}). The 3 SNPs related to gene *CYP3A5* were also in high LD (correlations from 0.781 to 0.913), among which rs4646437 had the strongest association with nicotine dependence (p-value = 2.91×10^{-8}). To evaluate the sensitivity of the results, we performed association tests using other weight functions (Table 4). Using a homogeneity weight $w_{ij} = g_i g_j$ (NHWU), none of the 10 SNPs had a p-value smaller than 0.01. The difference between HWU and NHWU indicated heterogeneous effects of the two genes on nicotine dependence in males and females. We also performed interaction analysis and stratified analysis using linear regression. Most of the 10 SNPs have p-values around 10^{-8} by either interaction test or stratified test, which is consistent with the findings from HWU (Table 4). Though the phenotype does not fit the distribution assumption of linear regression, to some extent these provide further hints on the heterogeneous effects of the top 10 SNPs. Additional results (i.e. coefficients estimates, standard errors, and confidence intervals) from stratified analysis also suggested the heterogeneous effect of the two genes in males and females (Supplementary Real Data Analysis). For example, rs4560769 (in gene *IKAKB*) has coefficients estimate of -0.545 with p-value 4.90×10^{-6} in males and coefficients estimate of 0.223 with p-value 0.0721 in females; rs776746 (in gene *CYP3A5*) has coefficients estimate of -0.177 with p-value 0.88 in males and coefficients estimate of 0.619 with p-value 1.97×10^{-9} in females (Supplementary Table S11).

In addition to gender, we also investigated potential genetic heterogeneity due to different ethnic and genetic backgrounds. In these analyses, we considered the same covariates as those used in the gender heterogeneity analysis. However, the results suggested there was no strong evidence of genetic heterogeneity due to different ethnic and genetic backgrounds (Supplementary Real Data Analysis).

4. Discussion

In recent years, U-statistic based methods have been gaining popularity in genetic association studies due to their robustness and flexibility [19, 20]. Yet, few methods have

been developed to model genetic heterogeneity, especially under the weighted U framework. In this paper, we have proposed a flexible and computationally efficient method, HWU, for high-dimensional genetic association analyses allowing for genetic heterogeneity. With HWU, we were able to integrate the latent population similarity (inferred from genetic background or environmental covariates) into a weight function and test heterogeneous effects without stratifying the sample. Simulation studies were conducted to compare the power of the proposed HWU method with methods that do not model genetic heterogeneity (i.e., NHWU and GLM). In the presence of genetic heterogeneity, HWU attained higher power than NHWU and GLM. In the absence of genetic heterogeneity, HWU still had comparable performance to NHWU and GLM. Unlike conventional methods, such as GLM, our method was developed based on a nonparametric U statistic, and therefore offers robust performance when the underlying phenotype distribution and mode of inheritance are unknown.

For association analysis considering genetic heterogeneity, current methods mainly target the heterogeneous distributions of genetic variants across different groups, where the effects of a genetic variant are assumed to be the same. In HWU, we allow the effects to be different. We use genome profiles or environmental covariates to build the latent similarity (i.e., the latent structure $\kappa_{i,j}$) and combine it with the genetic similarity to form the weight function $f(G_i, G_j)$. We then evaluate its relationship with a phenotype by using a weighted U statistic. Our method is different from testing an interaction effect. The key difference is that, for HWU, we assume there is a latent variable that acts in some joint fashion with the genetic variants, while in the usual interaction effect model the genetic variants are assumed to interact with known variables. For example, HWU can test the joint action of a genetic variant with a latent variable inferred from genome profile, while usual interaction tests the interaction of a genetic variant with other known genetic variants. Furthermore, HWU has fewer degrees of freedom than the usual interaction tests. The common approach to consider latent population structure in genetic heterogeneity is either to stratify the sample, or to create latent variables (i.e., surrogate variables or principle components) for adjustment.[21, 22] Yet, these approaches assume the effects are homogeneous after stratification or covariate adjustment. Using HWU, we can test the heterogeneous effects of a genetic variant after stratifying or adjustment. HWU is based on the idea that the more similar two subjects are, the more similar are their genetic effects. The idea of relating phenotype similarity to genotype similarity is not new. For example, Tzeng et.al proposed a gene-trait similarity regression for multi-locus association analysis [23]. However, their method is based on the usual regression framework and does not consider genetic heterogeneity.

In this paper, we focus on a single-locus test with consideration of genetic heterogeneity and assume an additive model. By modifying the weight function, HWU can easily be extended to model a multi-locus effect and other modes of inheritance (e.g., dominant/recessive effects). The weight function also offers flexibility for constructing latent population structure. Various similarity-based or distance-based functions can be applied to informative environmental and genetic covariates to infer the latent population structure. Although type I error is generally controlled for a variety of weight functions, the choice of an appropriate function to construct the latent population structure could impact the power of HWU. In this article, we suggest a Euclidian-distance based function, $\kappa_{i,j} = \exp(-(x_i - x_j)R(x_i - x_j)^T)$, in

which prior knowledge can be incorporated for potential power improvement. Nevertheless, a cross product kernel (i.e., $\kappa_{i,j} = \frac{1}{D} x_i x_j^T$) can also be used if the underlying model favors linearity. In the scenario where multiple functions can be used to construct the latent population structure, we can choose the optimum function to improve power. One simple approach is to calculate p-values of HWU using different functions for calculating $\kappa_{i,j}$, choose the function with smallest p-value, and adjust the p-value. Alternatively, we can employ methods such as kernel averaging or perturbation-based inference to provide an optimum function^[24].

Another advantage of our method is its computational efficiency. For the analysis of high-dimensional data, we derived the asymptotic distribution of the weighted U statistic and optimized the computational algorithm (e.g. using efficient eigen-decomposition). The genome-wide analysis of 949,658 SNPs took 7 hours and identified two genes, *IKBKB* and *CYP3A5*. *IKBKB* functions as a central coordinator in inflammatory responses and plays a key role in numerous physiological processes. *CYP3A5* encodes a protein involved in drug metabolism and synthesis of cholesterol, steroids and other lipids. Neither genes has been reported associated with nicotine dependence. Although our analysis suggests that these two genes are associated with nicotine dependence and have heterogeneous effects according to gender, further study and biological experiments are needed to confirm the association and to further investigate the potential function of these two genes in nicotine dependence.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Institute on Drug Abuse under Award Number K01DA033346 and by the National Institute of Dental & Craniofacial Research under Award Number R03DE022379. The datasets used for the analyses was obtained from dbGaP through dbGaP accession number [phs000092.v1.p1].

A. Appendix

A.1. Asymtotic distribution of HWU with parameter estimation

As shown in the main text, the limiting distribution of the weighted U with a cross product kernel can be simplified to $U \sim \sum_{s=1}^n \lambda_s \chi_{1,s}^2$. Taking the parameter estimation into account^[11, 12], the limiting distribution becomes:

$$U \sim \sum_{s=1}^n \lambda_s (\phi_s + c_s \phi_0)^2,$$

where $\{\lambda_s\}$ are the eigenvalues from the eigen-decomposition of $W = B\Lambda B^T$, in which $\Lambda = \{\text{diag}(\lambda_s)\}_{n \times n}$ and $B = \{b_{i,j}\}_{n \times n}$, $\{\phi_s\}$ are i.i.d. standard normal random variables. ϕ_0 is a

standard normal random variable with $\text{cov}(\varphi_s, \varphi_0) = c_s$, where c_s is defined as

$$c_s = \frac{1}{\sqrt{n}} \sum_{i=1}^n b_{i,s}.$$

Let $\gamma = (\varphi_1 + c_1\varphi_0, \dots, \varphi_n + c_n\varphi_0)$ be a random vector, where $\gamma \sim MVN(0, \Sigma_{n \times n})$. Letting I be the $n \times n$ identity matrix and J be the $n \times n$ matrix with all elements equal to $1/n$, we can easily show that $\Sigma = I - B^T J B$ and $\Sigma \Sigma = \Sigma$. Letting ξ be a random vector, $\xi \sim MVN(0, I_{n \times n})$, we have $\Sigma \xi \sim MVN(0, \Sigma)$ and

$$\sum_{s=1}^n \lambda_s (\phi_s + c_s \phi_0)^2 = \gamma^T \Lambda \gamma = \xi^T \Sigma \xi = \xi^T B^T (B \Sigma B^T) B \Lambda B^T (B \Sigma B^T) B \xi = \xi^T B^T (I - J) W (I - J) B \xi.$$

Because $B \xi \sim MVN(0, I_{n \times n})$, the limiting distribution of the weighted U is the weighted sum of independent chi-squares, $U \sim \sum_{s=1}^n \lambda_{1,s} \chi_{1,s}^2$, where $\{\lambda_{1,s}\}$ are the values of the matrix $(I - J) W (I - J)$.

References

1. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. Viewpoint missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*. 2010; 11(6):446–450.
2. Thornton-Wells TA, Moore JH, Haines JL. Genetics, statistics and human disease: analytical retooling for complexity. *Trends in Genetics*. 2004; 20(12):640–647. [PubMed: 15522460]
3. McClellan J, King MC. Genetic heterogeneity in human disease. *Cell*. 2010; 141(2):210–217. [PubMed: 20403315]
4. Galvan A, Ioannidis JPA, Dragani TA. Beyond genome-wide association studies: genetic heterogeneity and individual predisposition to cancer. *Trends in Genetics*. 2010; 26(3):132–141. [PubMed: 20106545]
5. Wang K, Baldassano R, Zhang H, Qu HQ, Imielinski M, Kugathasan S, Annese V, Dubinsky M, Rotter JI, Russell RK, et al. Comparative genetic analysis of inflammatory bowel disease and type 1 diabetes implicates multiple loci with opposite effects. *Human molecular genetics*. 2010; 19(10):2059–2067. [PubMed: 20176734]
6. Hoeffding W. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*. 1948; 19(3):293–325.
7. Serfling, R. *Approximation Theorems of Mathematical Statistics* (Wiley Series in Probability and Statistics). Wiley-Interscience; 1981.
8. O'Neil KA, Redner RA. Asymptotic distributions of weighted u-statistics of degree 2. *The Annals of Probability*. 1993; 21(2):1159–1169.
9. Dewet T, Venter JH. Asymptotic distributions for quadratic forms with applications to tests of fit. *Annals of Statistics*. 1973; 1(2):380–387.
10. Shieh GS, Johnson RA, Frees EW. Testing independence of bivariate circular data and weighted degenerate u-statistics. *Statistica Sinica*. 1994; 4(2):729–747.
11. Dewet T, Randles RH. On the effect of substituting parameter estimators in limiting chi-2 u-statistics and v-statistics. *Annals of Statistics*. 1987; 15(1):398–412.
12. Shieh GS. Weighted degenerate u- and v-statistics with estimated parameters. *Statistica Sinica*. 1997; 7(4):1021–1038.
13. Davies RB. Algorithm as 155: The distribution of a linear combination of chi square random variables. *Journal of the Royal Statistical Society Series C (Applied Statistics)*. 1980; 29(3):323–333.

14. Wu K, Simon H. Thick-restart lanczos method for large symmetric eigenvalue problems. *SIAM Journal on Matrix Analysis and Applications*. 2000; 22(2):602–616.
15. Astle W, Balding DJ. Population structure and cryptic relatedness in genetic association studies. *Statistical Science*. 2009; 24(4):451–471.
16. Jiang Y, Zhang H. Propensity score-based nonparametric test revealing genetic variants underlying bipolar disorder. *Genetic epidemiology*. 2011; 35(2):125–132. [PubMed: 21254220]
17. Furberg H, Kim Y, Dackor J, Boerwinkle E, Franceschini N, et al. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature genetics*. 2010; 42(5):441–447. [PubMed: 20418890]
18. Li MD, Cheng R, Ma JZ, Swan GE. A meta-analysis of estimated genetic and environmental effects on smoking behavior in male and female adult twins. *Addiction*. 2003; 98(1):23–31. [PubMed: 12492752]
19. Schaid DJ, McDonnell SK, Hebring SJ, Cunningham JM, Thibodeau SN. Nonparametric tests of association of multiple genes with human disease. *American Journal of Human Genetics*. 2005; 76(5):780–793. [PubMed: 15786018]
20. Zhang HP, Liu CT, Wang XQ. An association test for multiple traits based on the generalized kendall's tau. *Journal of the American Statistical Association*. 2010; 105(490):473–481. [PubMed: 20711441]
21. Zhou H, Pan W. Binomial mixture model-based association tests under genetic heterogeneity. *Annals of Human Genetics*. 2009; 73:614–630. [PubMed: 19725835]
22. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*. 2007; 3(9):1724–1735. [PubMed: 17907809]
23. Tzeng JY, Zhang D. Haplotype-based association analysis via variance-components score test. *American Journal of Human Genetics*. 2007; 81(5):927–938. [PubMed: 17924336]
24. Wu MC, Maity A, Lee S, Simmons EM, Harmon QE, Lin X, Engel SM, Molldrem JJ, Armistead PM. Kernel machine snp-set testing under multiple candidate kernels. *Genetic epidemiology*. 2013; 37(3):267–275. [PubMed: 23471868]

Figure 1. Power comparison of three methods for a binary phenotype with 20 heterogeneous sub-populations

*the genetic effect, β_i , for the i -th sub-population was sampled from a uniform distribution with mean μ_β and variance σ_β^2 .

Figure 2. Power comparison of three methods for a continuous phenotype with 20 heterogeneous sub-populations

*the genetic effect, β_i , for the i -th sub-population was sampled from a uniform distribution with mean μ_β and variance σ_β^2 .

Table 1
Type I error and power comparison of three methods when there are two heterogeneous sub-populations

Model ¹	Binary Phenotype				Continuous Phenotype					
	Effect ²		Type I error/Power		Effect		Type I error/Power			
	β_1	β_2	HWU	NHWU	GLM	β_1	β_2	HWU	NHWU	GLM
Null	0	0	0.046	0.051	0.051	0	0	0.052	0.049	0.049
T1	-0.1	0.1	0.085	0.07	0.07	-0.1	0.1	0.241	0.062	0.07
	-0.3	0.3	0.455	0.069	0.068	-0.3	0.3	0.972	0.157	0.173
	-0.5	0.5	0.899	0.084	0.084	-0.5	0.5	1	0.376	0.417
T2	0.1	0.1	0.122	0.149	0.15	0.1	0.1	0.284	0.368	0.384
	0.3	0.3	0.601	0.743	0.75	0.3	0.3	0.999	1	1
	0.5	0.5	0.978	0.993	0.993	0.5	0.5	1	1	1
T3	0	0.2	0.107	0.102	0.102	0	0.2	0.32	0.273	0.282
	0	0.4	0.378	0.307	0.31	0	0.4	0.902	0.767	0.803
	0	0.6	0.718	0.582	0.586	0	0.6	0.999	0.978	0.984
T4	-0.1	0.3	0.239	0.091	0.092	-0.1	0.3	0.718	0.161	0.18
	0.1	0.3	0.304	0.377	0.381	0.1	0.3	0.822	0.89	0.906
	-0.3	0.5	0.72	0.069	0.069	-0.3	0.5	0.997	0.059	0.076

¹Various scenarios of heterogeneity were considered in the simulation, including no genetic effect for both sub-populations (Null), the same effect size but with different directions (T1), the same effect size with the same direction (T2), no genetic effect for one sub-population but having a genetic effect for the other (T3), and different effect sizes with the same or different directions (T4).

²Single-locus effects for the 2 sub-populations, where the effect for the sub-population 1 is denoted by β_1 and the effect for the sub-population 2 is denoted by β_2 .

Table 2
Type I error comparisons of HWU and VCscore under non-normal distributions

Confounding Effect	Model	Distribution		
		Mixture	$t_{df=2}$	Cauchy
No	HWU	0.041	0.057	0.057
	VCscore	0.039	0.070	0.095
Yes	HWU	0.053	0.060	0.062
	VCscore	0.059	0.146	0.250

* the mixture distribution follows $a\chi^2_{df=1} + (1 - a)N(5, 1)$, where $a \sim \text{Bernoulli}(0.6)$.

Table 3

Performance of HWU with a mis-specified weight function

Component	Mis-specification ¹		Model ²	Method ³	
	Mis	True		HWU(mis)	HWU(true)
$f(g_i, g_j)$	$\mathcal{D}(g_i, g_j)$	$g_i g_j$	Null	0.047	0.049
			Alt	0.459	0.481
	$1(g_i = g_j)$	$g_i g_j$	Null	0.048	0.052
			Alt	0.451	0.51
$\kappa_{i,j}$	$\frac{1}{D} x_i x_j^T$	$\mathcal{D}(x_i, x_j)$	Null	0.05	0.058
			Alt	0.174	0.505
	$\mathcal{D}(x_i, x_j)$	$\frac{1}{D} x_i x_j^T$	Null	0.046	0.053
			Alt	0.072	0.515

¹“Mis” represents the misspecified $f(g_i, g_j)$ or $\kappa_{i,j}$ when analyzing simulated data, while “True” represent the true $f(g_i, g_j)$ or $\kappa_{i,j}$ in the corresponding simulation setting. Here, $\mathcal{D}(\cdot, \cdot)$ represents the euclidian distance based weight, i.e., $\mathcal{D}(g_i, g_j) = \exp(-(g_i - g_j)^2)$ and $\mathcal{D}(x_i, x_j) = \exp(-\frac{1}{D} \sum_{d=1}^D (x_{d,i} - x_{d,j})^2)$.

²The error distribution was set as t distribution with $df=2$. “Null” represents the null model with $\mu\beta=0$ and $\sigma^2_\beta=0$, “Alt” represents the heterogeneous effect model with $\mu\beta=0$ and $\sigma^2_\beta=0$.

³HWU(mis) represents the HWU model with a mis-specified weight function, while HWU(true) represents the HWU model with the true weight function.

Table 4
Top 10 nicotine dependence associated SNPs from the GWAS analysis considering gender heterogeneity

Name	Chr	Position	Gene	MAF	p-value		
					HWU	NHWU	Stratified ²
rs17078660	3	46160432	NA, near <i>FLT1P1</i>	0.025	9.98×10^{-9}	0.017	4.89×10^{-6}
MitoA15302G	26	15302	NA	0.226	1.69×10^{-8}	0.688	6.23×10^{-9}
rs7753843	6	67055504	NA	0.413	1.86×10^{-8}	0.571	3.23×10^{-8}
rs10493279	1	60368804	NA, near <i>C1orf87</i>	0.033	1.88×10^{-8}	0.014	1.97×10^{-8}
rs4560769	8	42259961	<i>IKBKB</i>	0.255	1.93×10^{-8}	0.062	1.40×10^{-8}
rs9694958	8	42275203	<i>IKBKB</i>	0.242	2.82×10^{-8}	0.122	1.47×10^{-8}
rs776746	7	99108475	<i>CYP3A5</i>	0.245	2.91×10^{-8}	0.241	1.15×10^{-8}
rs4646437	7	99203019	<i>CYP3A4</i>	0.278	4.03×10^{-8}	0.512	4.22×10^{-8}
rs9694574	8	42279609	<i>IKBKB</i>	0.242	4.63×10^{-8}	0.144	1.33×10^{-8}
rs4646457	7	99083016	<i>ZSCAN25</i>	0.253	4.74×10^{-8}	0.138	2.50×10^{-8}

¹ A test for interaction between genetic variant and gender ($H_0 : \beta_{gs} = 0$) is performed using linear regression model, $y_i = \mu + Z_i\alpha + g_i\beta_g + s_i\beta_s + g_i s_i\beta_{gs} + \varepsilon_i$, for each SNP, where y_i , g_i , s_i , and Z_i represents response variable, genetic variant, gender and other covariates respectively.

² A test stratifying genetic variant based on gender ($H_0 : \beta_g M = \beta_g F = 0$) is performed using linear regression model, $y_i = \mu + Z_i\alpha + s_i\beta_s + g_i1\{s_i = Male\}\beta_g M + g_i1\{s_i = Female\}\beta_g F + \varepsilon_i$, for each SNP, where y_i , g_i , s_i , and Z_i represents response variable, genetic variant, gender and other covariates respectively.