Geng *et al. BMC Genetics* (2019) 20:40
https://doi.org/10.1186/s12863-019-0742-z

BMC Genetics

METHODOLOGY ARTICLE Open Access

An integrative U method for joint analysis of multi-level omic data

Pei Geng[1], Xiaoran Tong[2] and Qing Lu[2*]

Abstract

Background: The advance of high-throughput technologies has made it cost-effective to collect diverse types of omic data in large-scale clinical and biological studies. While the collection of the vast amounts of multi-level omic data from these studies provides a great opportunity for genetic research, the high dimensionality of omic data and complex relationships among multi-level omic data bring tremendous analytic challenges.

Results: To address these challenges, we develop an integrative U (IU) method for the design and analysis of multi-level omic data. While non-parametric methods make less model assumptions and are flexible for analyzing different types of phenotypes and omic data, they have been less developed for association analysis of omic data. The IU method is a nonparametric method that can accommodate various types of omic and phenotype data, and consider interactive relationship among different levels of omic data. Through simulations and a real data application, we compare the IU test with commonly used variance component tests.

Conclusions: Results show that the proposed test attains more robust type I error performance and higher empirical power than variance component tests under various types of phenotypes and different underlying interaction effects.

Keywords: Non-parametric method, Functional data analysis, Integrative analysis

Background

With rapidly evolving high-throughput technologies and ever-decreasing costs, it has become feasible to systematically study diverse types of omic data in biological and clinical studies [1, 2]. The collection of multi-level omic data from these studies provides us a great opportunity to integrate information from different levels of omic data into association analysis [3–6]. Although omic-based association analysis holds great promise for discovering novel disease-associated biomarkers, the discovery process is hampered by the lack of appropriate statistical tools to consolidate and analyze multi-level omic data. The development of advanced statistical methods to address the analytical challenges faced by ongoing omic data analysis can enhance our ability to identify new disease-associated biomarkers.

Comprehensive reviews of integrative analysis on multi-level omic data are summarized in [3, 5, 7] and the references therein. Most of the existing methods for integrative analysis are developed based on score-type tests or variance component tests. For instance, in the integrative analysis of single-nucleotide variants (SNVs) and transcript expression data, [6] used the estimating equations to estimate parameters of interest, and then proposed a Wald test to evaluate the association between the outcome and a set of genetic variants, considering possible interactions. In order to efficiently test the joint effects of SNVs and gene expression with a binary phenotype, [8] developed a combined variance component test in the mixed model framework. Based on this work, [9] further investigated a variance component score test for modeling multiple genomic data including SNVs, gene expression, and methylation data, each of which can come from different samples or studies. While those methods have attractive properties under various scenarios, most of these methods are parametric-based or semi-parametric-based, which often rely on a distribution assumption (e.g., a normal distribution assumption). When this assumption is violated, these methods are subject to false positive

*Correspondence: qlu@epi.msu.edu
[2]Department of Epidemiology and Biostatistics, Michigan State University, 48824, East Lansing, MI, USA
Full list of author information is available at the end of the article

results and/or power loss [10]. The diagnostic assessments of human diseases can often be of different types (e.g., binary, ordinal and continuous) and follow known or unknown distributions. This issue is, however, paid less attention by the existing methods.

Moreover, the molecular complexity of human diseases manifests itself at the genomic, transcriptomic, epigenomic and proteomic levels [11, 12]. Different levels of omic data can interact in the disease process. By considering interactions between different levels of omic data, the power of detecting disease-associated biomarkers can be potentially enhanced. While some of existing methods consider interactions between omic data [6, 8], they commonly assume a particular interaction model (e.g., a multiplicative model), and are subject to suboptimal performance if the underlying model has different forms (e.g., a threshold model).

To address these limitations, we propose a non-parametric framework for association analysis using multi-level omic data. The IU test is a U-statistic-based test, which is constructed using the pairwise omic and phenotype similarities of subjects. It has several remarkable features worthy of attention: 1) it makes no distribution assumptions, and therefore provides a robust and powerful performance when analyzing phenotypes and omic data with unknown distributions; 2) it provides a unified framework for analyzing various types of phenotypes and omic data (binary, ordinal and continuous); and 3) it considers interactions among different levels of omic data without posing specific model assumptions.

The remaining of the paper is organized as follows. We begin with a detailed description of the proposed integrative U method in "Methods" section, and then present the simulation results of the IU method under different types of phenotypes and various genetic or interaction effects in "Simulation" section. Using the proposed method, we performed an integrative analysis of the DNA sequencing and gene expression data from a hypertension study in "An integrative analysis of gene and gene expression data of hypertension" section. "Conclusion" section summarizes the advantages and limitations of the IU test. Details of the proof of the main results can be found in the Additional file 1.

Methods

Suppose that we are interested in evaluating the joint association of M levels of omic data with a disease phenotype of interest. Without loss of generality, we illustrate the method with two levels of omic data (i.e., SNVs and gene expression data). The extension to more than 2 levels of omic data will be discussed later in "Conclusion" section. Let Y_i be a continuous or discrete disease phenotype, S_i be a scalar gene expression variable, and $\mathbf{G_i} = (G_i(t_1), G_i(t_2), ..., G_i(t_p))$ be the genotypes of p SNVs (e.g., coding variants in a gene) for the ith individual ($i = 1, ..., n$), where t_j is the SNV location and $G_i(t_j) = 0, 1, 2$ is coded as the number of minor alleles.

Genetic smoothing

In recent literature, functional data analysis has been often applied to handle the genetic data. For instance, [13] proposed a functional linear model for quantitative traits using B-spline basis functions to expand the genotype functions. Vsevolozhskaya et al. [14] proposed a functional analysis of variance method to test the association of sequence variants in a genomic region with a qualitative trait. Functional data analysis has also been developed for different types of traits and study purposes in genetic research. For instance, [15] developed a Cox proportional hazard model with functional regression for gene-based association analysis of survival traits. Moreover, [16] proposed a generalized functional linear model to perform meta-analysis of multiple studies to evaluate the association of genetic variants with dichotomous traits.

Here we adopt the functional data analysis to handle the SNVs. Rather than assuming $G_i(t_j)$ as a random variable, we assume that $G_i(t_j)$ is a discrete realization of a function $G_i(t)$ generated from a stochastic process with mean function $\eta(t)$ and covariance function $\Gamma(s, t)$. The B-spine smoothing technique is then used to model the underlying function curve $G_i(t)$. In other words, $G_i(t)$ can be written as a linear sum of specified basis functions:

$$G_i(t) = \sum_{k=1}^{K} \beta_{ik} B_k(t),$$

where $\{\beta_k(t), k = 1, ..., K\}$ is the polynomial basis functions in L_2 Hilbert space. The fitted smoothing curves are demonstrated in Fig. 1. Similar to other functional-based methods [14], we implement the smoothing by scaling the locations to the interval $[0, 1]$, and use the penalization technique to determine the appropriate number of knots (i.e., smoothness).

Test statistic

With the assumptions of Y, $G(t)$ and S mentioned above, we aim to test the hypotheses:

H_0: Y is independent of $G(t)$ and S;

H_a: Y is associated with $G(t)$ or S.

Since we do not assume any regression form of the association between Y and the genetic variables (G and S), to perform the hypothesis testing, we propose a non-parametric integrative U statistic defined as

$$U_n = \frac{1}{n(n-1)} \sum_{i,j=1, i \neq j}^{n} K_1(Y_i, Y_j) K_2(S_i, S_j) \int_0^1 G_i(t) G_j(t) dt,$$

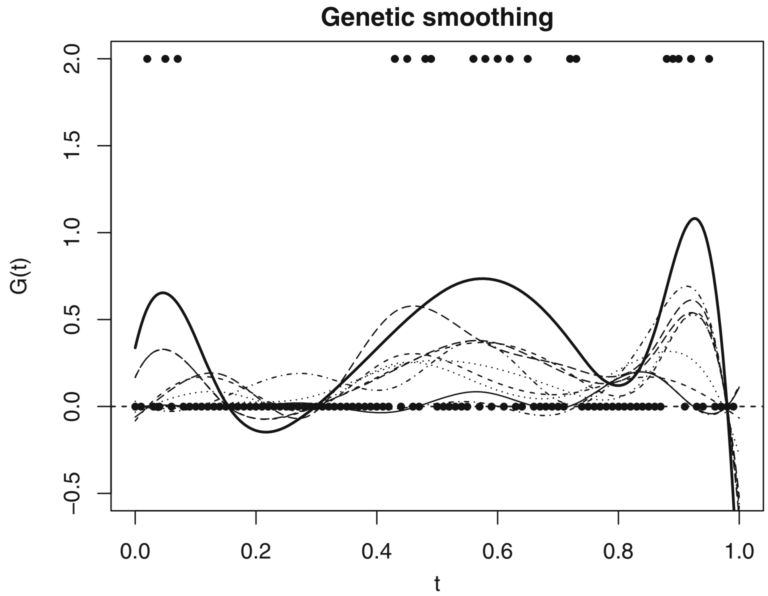

Fig. 1 A simple illustration of fitted B-spline curves for genotype sequence. The solid dots present the genotype sequence for one subject, and the solid thick curve is the corresponding smooth curve. In total, a sample of 10 subjects are simulated and the 10 corresponding smoothed curves are presented

where $K_1(\cdot,\cdot)$ and $K_2(\cdot,\cdot)$ are symmetric kernel matrices that measure the similarities of two individuals' phenotypes and gene expression values, respectively. For simplicity, we use the cross product for both kernel matrices

$$U_n = \frac{1}{n(n-1)} \sum_{i,j=1,i\neq j}^{n} Y_i Y_j S_i S_j \int_0^1 G_i(t) G_j(t) dt.$$

In addition to the cross product kernel, other kernels, such as those proposed in [10] and [17] can also be used.

From the above equation, the proposed test statistic is a U statistic defined on all possible pairs of subjects (i,j), where the genetic similarity of subjects i and j is defined as the inner product of the smooth curves of the stochastic process, i.e., $\int_0^1 G_i(t)G_j(t)dt$. The phenotype similarity and gene expression similarity between the subjects i and j are simply products of two subjects' phenotype and gene expression values, respectively.

Asymptotic property

Under the null hypothesis and the assumption $G_i(t) \sim SP(\eta(t), \Gamma(s,t))$, we obtain

$$r_i = E(U_n|Z_i) = \mu_Y \mu_S Y_i S_i \int_0^1 G_i(t)\eta(t)dt,$$
$$\mu_0 = E(U_n) = \mu_Y^2 \mu_S^2 ||\eta(t)||^2,$$

where $Z_i = (Y_i, S_i, \mathbf{G_i})$, μ_Y and μ_S are the population means of Y and S, respectively. The asymptotic result of non-degenerated U statistics in [18] implies that

$$\sqrt{n}(U_n - \mu_0) \to N\left(0, 4\sigma_1^2\right),$$

where $\sigma_1^2 = Var(r_1)$. Moreover, by applying the result of stochastic processes in Section 4.2 of [19], we can further obtain that

$$\begin{aligned} Var(r_1) &= \mu_Y^2 \mu_S^2 Var\left(Y_1 S_1 \int G_1(t)\eta(t)dt\right) \\ &= \mu_Y^2 \mu_S^2 \left(\sum_{k=1}^{m} \lambda_k \delta_k^2\right) \left(\mu_Y^2 + \sigma_Y^2\right)\left(\mu_S^2 + \sigma_S^2\right) \\ &\quad + \mu_Y^2 \mu_S^2 \left(\mu_Y^2 \sigma_S^2 + \mu_S^2 \sigma_Y^2 + \sigma_Y^2 \sigma_S^2\right) ||\eta||^4, \end{aligned}$$

where m is the number of eigenvalues of Γ, $(\lambda_k, \phi_k(t))$ are eigenvalues and eigenfunctions of the covariance function Γ, $\delta_k = \int \phi_k(t)\eta(t)dt$, σ_Y^2 and σ_S^2 are the population variances of Y and S.

Because μ_0 is unobservable, we propose to estimate it by substituting the population means of Y, S and $G(t)$ by their corresponding sample means, i.e.,

$$\hat{\mu}_0 = \bar{Y}^2 \bar{S}^2 ||\bar{G}(t)||^2. \tag{1}$$

Let s_Y^2 and s_S^2 be the sample variances of Y and S, $\hat{\lambda}_k$ and $\hat{\phi}_k$ be the eigenvalues and eigenfunctions of $\hat{\Gamma}(s,t) = \frac{1}{n}\sum_{i=1}^{n}\left(G_i(t) - \bar{G}(t)\right)\left(G_i(s) - \bar{G}(s)\right)$. By letting $\hat{\delta}_k = \int_0^1 \hat{\phi}_k(t)\bar{G}(t)dt$, we can obtain the asymptotic distribution of the test statistic under H_0:

Theorem 1. *Assume that both Y and S have finite means and variances and* $G(t) \sim SP(\eta(t), \Gamma(s,t))$. *Moreover,*

$\mu_Y\mu_S \neq 0$ and S is independent of $G(t)$. Then, under H_0,

$$T_n = \frac{\sqrt{n}(U_n - \hat{\mu}_0)}{\hat{\sigma}} \to N(0,1),$$

where $\hat{\mu}_0$ is defined in (1) and $\hat{\sigma}^2 = 4\sum\hat{\lambda}_k\hat{\delta}_k^2\left\{\bar{Y}^2\bar{S}^2\left(\bar{Y}^2s_S^2 + \bar{S}^2s_Y^2 + s_Y^2s_S^2\right)\right\} + 4\bar{Y}^2\bar{S}^2||\bar{G}||^4s_Y^2s_S^2$.

We would reject H_0 if $|\sqrt{n}(U_n - \hat{\mu}_0)/\hat{\sigma}| > z_{\alpha/2}$, where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution.

Remark 1. *The assumption of the underlying stochastic process is very general. We do not need a specific condition on the pointwise distributions such as Gaussian, which is the required assumption in [14].*

Remark 2. *The proposed test inherits the robustness property from U statistics, and is capable of handling both discrete and continuous phenotypes with various underlying distributions. Moreover, the proposed test does not need to specify any form of the regression function $\mu = E(Y|S,G)$, hence the test procedure is free of model assumptions.*

The method can also be used for different study purposes. For instance, to only test the effect of SNVs (e.g., in a genetic association study), the corresponding integrative U test statistic can be simplified as $U_G = \frac{1}{n(n-1)}\sum_{i\neq j} Y_iY_j\int_0^1 G_i(t)G_j(t)dt$ with $\hat{\mu}_G = \bar{Y}^2||\bar{G}(t)||^2$ and variance estimator $\hat{\sigma}_G^2 = 4\hat{\mu}_Y^2s_Y^2\sum\hat{\lambda}_k\hat{\delta}_k^2$.

Power and sample size

While omic-based studies become increasingly popular in human genetic research, few statistical tools are available for power and sample size calculation. In this section, we investigate the power of the proposed method under certain alternative hypotheses and provide a convenient way for power/sample size calculation.

We have studied the IU test under the null hypothesis, $E\left\{Y_iY_jS_iS_j\int_0^1 G_i(t)G_j(t)dt\right\} = \mu_0 = \mu_Y^2\mu_S^2||\eta||^2$. Under the alternative hypothesis, we assume that $E\left\{Y_iY_jS_iS_j\int_0^1 G_i(t)G_j(t)dt\right\} = \mu_1 \neq \mu_0$ and $Var\left(Y_iY_jS_iS_j\int_0^1 G_i(t)G_j(t)dt|Y_i, S_i, G_i\right) = \tau_1^2$. Without loss of generality, we assume that $\mu_1 > \mu_0$. Applying the Hoeffding projection in [18], we obtain that $\frac{\sqrt{n}(U_n-\mu_1)}{2\tau_1} \to N(0,1)$. Hence, under the alternative hypothesis, U_n can be written as

$$\frac{\sqrt{n}(U_n - \hat{\mu}_0)}{\hat{\sigma}} = \frac{2\tau_1}{\hat{\sigma}}\frac{\sqrt{n}(U_n - \mu_1)}{2\tau_1} + \frac{\sqrt{n}(\mu_1 - \hat{\mu}_0)}{\hat{\sigma}}.$$

Since both $\hat{\mu}_0$ and $\hat{\sigma}$ are consistent estimators of μ_0 and σ, for a sufficiently large n, we have

$$\frac{\sqrt{n}(U_n - \hat{\mu}_0)}{\hat{\sigma}} \sim N\left(\frac{\sqrt{n}(\mu_1 - \mu_0)}{\sigma}, \frac{4\tau_1^2}{\sigma^2}\right).$$

Therefore, the power of the proposed test can be calculated by

$$\begin{aligned}
&P\left(\left|\frac{\sqrt{n}(U_n - \hat{\mu}_0)}{\hat{\sigma}}\right| > z_{\alpha/2}\right)\\
&\geq P\left(\frac{\sqrt{n}(U_n - \hat{\mu}_0)}{\hat{\sigma}} > z_{\alpha/2}\right)\\
&\geq P\left(\frac{\sqrt{n}(U_n - \mu_1)}{2\tau_1} > \frac{\hat{\sigma}}{2\tau_1}\left(z_{\alpha/2} - \frac{\sqrt{n}(\mu_1 - \mu_0)}{\hat{\sigma}}\right)\right)\\
&= \Phi\left(\frac{\hat{\sigma}}{2\tau_1}\left(\frac{\sqrt{n}(\mu_1 - \mu_0)}{\hat{\sigma}} - z_{\alpha/2}\right)\right).
\end{aligned}$$

For desired power β, one can derive the minimal required sample size by setting the inequality $\Phi\left(\frac{\sigma}{2\tau_1}\left(\frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma} - z_{\alpha/2}\right)\right) \geq \beta$, therefore the required sample size can be calculated by

$$n = \min_{m\in\mathbb{Z}}\left\{m \geq \frac{\sigma^2}{(\mu_1 - \mu_0)^2}\left(z_{\alpha/2} + \frac{2\tau_1z_\beta}{\sigma}\right)^2\right\}.$$

Results

Simulation

Through simulations, we compared the type I error and empirical power of the proposed test with those of two variance component methods: the adjusted kernel sequencing association test and variance component test proposed by [8]. Since the original kernel sequencing test developed by [17] is proposed for only sequencing SNVs, we slightly modify the method to incorporate gene expression data in the test. Recall that the sequence kernel association test (SKAT) proposed by [17] has the form

$$Q = (\mathbf{Y} - \hat{\mu})^T\mathbf{K}(\mathbf{Y} - \hat{\mu}),$$

where $\mathbf{Y} = (Y_1, ..., Y_n)^T$ and $\mathbf{K}(G_i, G_j) = \sum_{k=1}^p w_kG_{ik}G_{jk}$. To make the methods comparable, the adjusted SKAT (Adj-SKAT) is modified as:

$$\tilde{Q} = (\mathbf{Y} - \hat{\mu})^T\tilde{\mathbf{K}}_S(\mathbf{Y} - \hat{\mu}),$$

where the elements of $\tilde{\mathbf{K}}_S$ are defined as $\tilde{\mathbf{K}}_S(G_i, G_j) = \sum_{k=1}^p w_kS_iS_jG_{ik}G_{jk}$.

Interestingly, the proposed U_n has a similar form of Q. If we define

$$\mathbf{K}_S(G_i, G_j) = S_iS_j\int_0^1 G_i(t)G_j(t)dt,$$

and further define metric $\mathbf{K}_S = (K_S(G_i, G_j))_{n\times n}$ with zeros as diagonal elements, then the proposed U statistic U_n can be written as a similar form to Adj-SKAT:

$$U_n = \mathbf{Y}^T\mathbf{K}_S\mathbf{Y}.$$

In addition to Adj-SKAT, we also compared IU test with the component variance test (VCT) developed under the generalized linear mixed model framework by [8]. VCT is proposed to test the joint effects of SNVs and gene expression, and is defined as

$$\bar{Q} = \frac{1}{n}(\mathbf{Y} - \hat{\mu})^T \left\{ a_1 \mathbf{G}\mathbf{G}^T + a_2 \mathbf{S}\mathbf{S}^T + a_3 \mathbf{C}\mathbf{C}^T \right\} (\mathbf{Y} - \hat{\mu}),$$

where (a_1, a_2, a_3) are weight parameters and $\mathbf{C}$ is the product of SNVs and gene expression. In the simulation, we used the recommended weight, the inverse of the square root of the variance, as suggested by [8].

The genetic data was simulated from the 1000 Genome Project [20]. Specifically, we used a 1Mb region of the genome (Chromosome 17: 7344328-8344327) from 1092 individuals in 1000 Genome Project. In each simulation replicate, SNVs were generated by randomly choosing a segment with $p = 100$ consecutive SNVs from the genome. Then the stochastic smoothing function curves were constructed by applying the functional data analysis to the SNV sequences. Gene expression data was generated from a normal distribution, $N(1, 1.2^2)$. The natural cubic spline smoothing with penalty parameter introduced in [14] was applied. All the results of type I error and empirical power were calculated based on 1000 simulated replicates.

Type I error performance

The phenotype assessments can often be of different types (e.g., binary and continuous phenotypes), with unknown underlying distributions. In this simulation, we evaluated the robustness of three methods against different phenotype distributions. Totally, four types of distributions: Bernoulli (binary), Gaussian, T and Double-exponential (DE), were considered in this simulation. Both T and DE have heavier tails than Gaussian. The original VCT test is developed for binary phenotype, but can be extended for other types of phenotypes by using different link functions. For instance, by using the identity function, VCT can be used to analyze continuous phenotypes. Under H_0, Y is independently generated from Bernoulli$(1/(1+e^{0.2}))$, $N(1, 1)$, T_2, T_4 and $DE(1)$, respectively. Table 1 summarizes the type I error performance of three methods. From Table 1, we find that type I error rates of both VCT and Adj-SKAT are well controlled for Bernoulli- and Gaussian-type of phenotypes, but are inflated under the heavy tailed distributions(T and DE). As we expect from the U statistic property, the IU test attains robust performance under all phenotype distributions.

We further investigated the empirical levels of the proposed test for different sample sizes. For this simulation, we considered both continuous and binary phenotypes and varied the study sample size from 100 to 500. A normal distribution was used to simulate the continuous phenotype, while balanced samples were generated for the binary phenotype. 1000 independent simulation replicates were used to obtain the empirical sizes. The empirical levels of the test for nominal sizes 0.05 and 0.01 are summarized in Table 2. As observed in Table 2, the type I error rates of the proposed test are well controlled for both types of phenotypes and different sample sizes.

Power performance

For the power comparison, we considered the scenarios with or without an interaction between SNVs and gene expression. For the scenarios with an interaction, we studied the performance of the three methods under various interaction models. Similar to the type I error simulation, the genetic data was obtained from the 1000 Genome Project and gene expression S_i was sampled from $N(1, 1.2^2)$. The binary response Y_i was then generated from a logistic regression model. In each simulation, we randomly chose 100 cases and 100 controls to form a balanced case-control sample. For continuous phenotypes, we simulated both Gaussian-distributed and T-distributed phenotypes.

Case 1: No interaction effect

We first evaluated the power of three methods under the scenario when there is no interaction between SNVs and gene expression. In the binary case, the underlying model is similar to the one assumed in [8]

$$\begin{aligned} &\text{Model 1 (a):} logit\{P(Y_i = 1|G_i, S_i)\} \\ &= -0.2 + G_i^T \beta_G + S_i \beta_S + S_i G_i^T \gamma, \end{aligned} \quad (2)$$

where β_S is a scalar parameter, β_G and γ are $p \times 1$ vector parameters. In the continuous case, a linear mixed model was used,

$$\text{Model 1 (b):} Y_i = 2 + G_i^T \beta_G + S_i \beta_S + S_i G_i^T \gamma + \varepsilon_i,$$
$$\text{Model 1 (c):} Y_i = 2 + G_i^T \beta_G + S_i \beta_S + S_i G_i^T \gamma + e_i,$$

where β_G and β_S were defined as in (2), $\varepsilon_i \sim N(0, 1)$ and $e_i \sim T(2)$, a T-distribution with 2 degrees of freedom.

Similar to [8], we assume that β_G and γ are randomly generated from probability distributions with mean 0 and variances σ_G^2 and σ_γ^2. In this simulation, the genetic effects measured by β_G were generated from a normal distribution, $N(0, \sigma_G^2)$, while the interaction effects measured by γ were all set to be zero ($\sigma_\gamma^2 = 0$) in order to study the marginal effects of genetic variables.

The power performance for the binary, Gaussian-distributed, and T-distributed phenotypes under the Model 1 is summarized in Fig. 2. The figure shows the power performance of three methods when the effect of gene expression, β_s, increases and the effects of SNVs remain the same ($\sigma_G = 0.1$). As shown in the figure, both IU and VCT have higher power than Adj-SKAT as the

Table 1 Type I error comparison of three methods for different types of phenotypes

phenotype distributions	Bernoulli	Gaussian	T_2	T_4	DE
IU	0.048	0.049	0.052	0.055	0.052
VCT	0.046	0.053	0.091	0.068	0.075
Adj-SKAT	0.035	0.055	0.129	0.062	0.064

Bernoulli, Gaussian, T, and DE correspond to Binary, Gaussian-distributed, T-distributed, and Double-exponential distributed phenotypes. The nominal size of the test is 0.05 and $n = 200$

effect of gene expression increases for binary phenotype. As expected, VCT achieves highest power for Gaussian-distributed phenotype among the three methods. For the T-distributed phenotype, IU outperforms VCT and Adj-SKAT while Adj-SKAT has little power for T-distributed phenotype. Overall, IU test is more robust than the other two methods to the phenotype distributions.

Case 2: Interaction effect

We then compared the performance of three methods under a more complex scenario when there is an interaction between SNVs and gene expression. In this simulation, we considered three types of interaction effects: multiplicative, threshold, and random interaction effects. Similar to the simulation with no interaction, we evaluated the methods under three different kinds of phenotypes.

A multiplicative interaction effect is simulated for binary, Gaussian-distributed and T-distributed phenotypes based on the model below,

$$\text{Model 2 (a)}: logit\{P(Y_i = 1|G_i, S_i)\} = -0.2 + c\, S_i G_i^T W,$$
$$\text{Model 2 (b)}: Y_i = 2 + c\, S_i G_i^T W + \varepsilon_i,$$
$$\text{Model 2 (c)}: Y_i = 2 + c\, S_i G_i^T W + e_i,$$

where c is a scale parameter, $\varepsilon_i \sim N(0, 1)$, and $e_i \sim T(2)$. With W being a $p \times 1$ vector, each element of W equals to 1 if the corresponding genetic variant is a causal SNV, and 0 if the genetic variant is non-causal. Under the null hypothesis of no association, all elements in W are 0. Let m be the number of causal SNVs, then a larger m indicates increasing interaction effects on the phenotype. Similarly, a large c corresponds to a strong interaction effect. The upper panel of Fig. 3 shows the power comparison of three methods as m increases from 0 to 8 with scale parameter $c = 1$, and the lower panel of Fig. 3 shows power performance as the scale parameter c increases from 0 to 1.5 with $m = 4$. As we observe from the figure, IU test attains higher power than VCT and Adj-SKAT for binary and T-distributed phenotypes with the increasing interaction effect. For Gaussian-distributed phenotype, Adj-SKAT has highest power among the three methods because the cross product kernel used in Adj-SKAT perfectly captures the true underlying interaction model.

Next we considered a threshold interaction model, which has the following form:

$$\text{Model 3 (a)}: logit\{P(Y_i = 1|G_i, S_i)\} = \begin{cases} -0.2 + S_i G_i^T W, & \text{if } S_i G_i^T W > d, \\ -0.2 & \text{otherwise,} \end{cases}$$
$$\text{Model 3 (b)}: Y_i = \begin{cases} 2 + S_i G_i^T W + \varepsilon_i, & \text{if } S_i G_i^T W > d, \\ 2 + \varepsilon_i & \text{otherwise,} \end{cases}$$
$$\text{Model 3 (c)}: Y_i = \begin{cases} 2 + S_i G_i^T W + e_i, & \text{if } S_i G_i^T W > d, \\ 2 + e_i & \text{otherwise,} \end{cases}$$

where d is a threshold parameter, ε_i and e_i are the same as those in Model 2. A large d corresponds to a weak interaction effect. Under this model, the effect substantially increases when the product of SNV and gene expression exceeds the threshold d. Figure 4 shows that all the three tests have decreased power as the threshold parameter d increases from 0 to 8. Moreover, the IU test achieves much higher power than the other two methods for the binary and T-distributed phenotypes while it has a similar performance as the other two methods for the Gaussian-distributed phenotype.

Table 2 Type I error of IU for different sample sizes with nominal sizes 0.05 and 0.01

Type I error with Gaussian phenotype					
α / n	100	200	300	400	500
0.05	0.049	0.051	0.048	0.047	0.052
0.01	0.01	0.011	0.009	0.011	0.010
Type I error with binary phenotype					
α / n	100	200	300	400	500
0.05	0.044	0.046	0.048	0.052	0.048
0.01	0.011	0.012	0.009	0.011	0.012

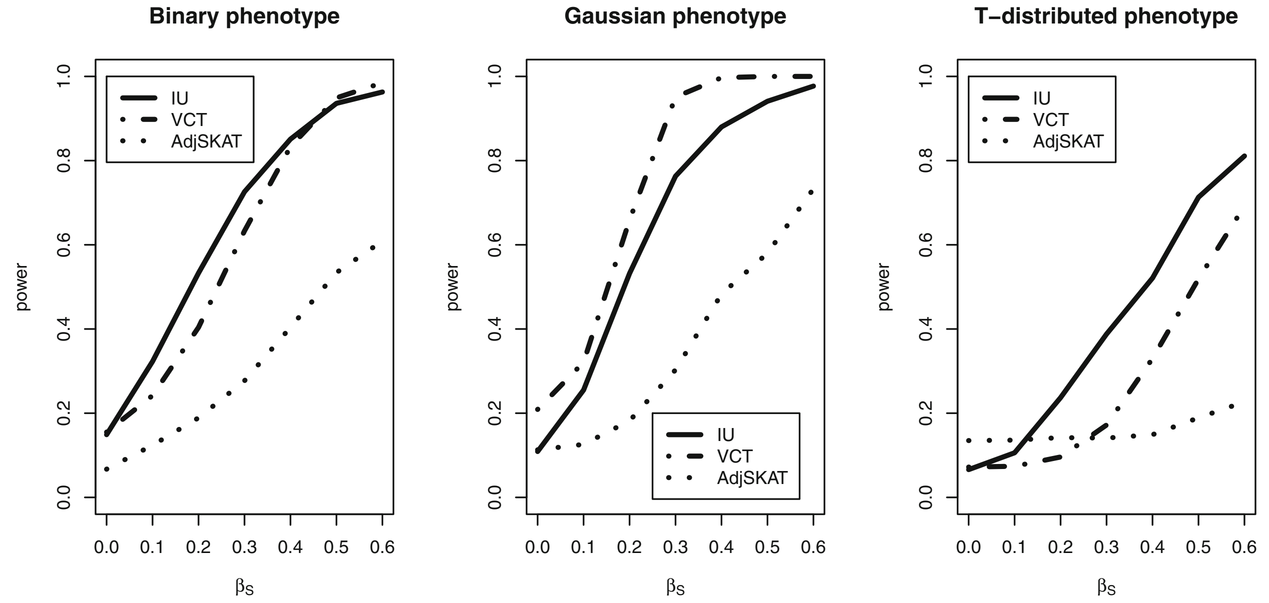

Fig. 2 Power comparison of three methods for different types of phenotypes when there is no interaction. β_S stands for the causal effect of gene expression *S*

Finally, we considered a random effect interaction model based on the model 1 described in (2) with elements of γ generated from a uniform distribution $U(-a, a), a > 0$. With no marginal effects (i.e., $\sigma_G = 0$, $\beta_S = 0$), Fig. 5 shows that IU test has a similar power performance as VCT for both binary and Gaussian-distributed phenotype. We also find that for the heavy-tailed phenotype (i.e., the T-distributed phenotype), IU attains more robust type I error and higher power than VCT and Adj-SKAT. With both marginal ($\sigma_G = 0.1$, $\beta_S = 0, 0.5, 1, 1.5, 2$) and interaction effects ($a = 0.1$), the power performance shown in Fig. 6 is similar to that of Case 1 shown in Fig. 2.

In summary, the proposed IU test obtains higher power as the marginal or interaction effects increase. Unlike VCT or Adj-SKAT, which show higher power only under some specific models (e.g., the random effect or cross-product interaction models), the IU test showed more robust and stable performance for different phenotypes and various underlying models. These features make IU more appropriate to use when we have limited knowledge on the actual underlying model.

An integrative analysis of gene and gene expression data of hypertension

Hypertension is one of the most common chronic diseases, which affects a large proportion of human population worldwide. Despite decades of research in hypertension, the genetic etiology of hypertension remains largely unknown. The successful identification of genetic variants predisposing to hypertension holds promise for providing better understanding of genetic etiology of hypertension and promoting new therapeutic targets. In this application, we performed an integrative analysis of DNA sequencing and gene expression data from the San Antonio Family Heart Study (SAFHS) and the San Antonio Family Diabetes/Gallbladder Study (SAFDGS). SAFHS and SAFDGS include standardized diagnostic assessments of hypertension (i.e., Case vs. Control). Whole-genome sequencing (WGS) data were available on the odd numbered autosomes. In addition, gene expression was measured using peripheral blood mononuclear cells collected at the first examination. In total, there are 260 subjects with WGS data, gene expression data, and the binary hypertension (HTN) phenotype measured.

Prior to the integrative analysis, we performed a quality control and data preparation process. In this process, we assembled multiple SNVs into genes based on the Genome Reference Consortium release version 38 (GRCh38) and excluded genes without gene expression data. To deal with missing values in the genetic data, we imputed the genotype values from multinomial distribution using the sample proportions as the generating probabilities. After data processing step, 2389 genes and the corresponding gene expression remained for the integrative analysis. We then applied a generalized mixed model to the binary HTN phenotype with covariates AGE, MEDS, SMOKE, SEX and the kinship matrix to remove potential confounding effects and the familial correlations. The residuals were used as the responses in this integrative analysis. Eventually, the proposed IU test

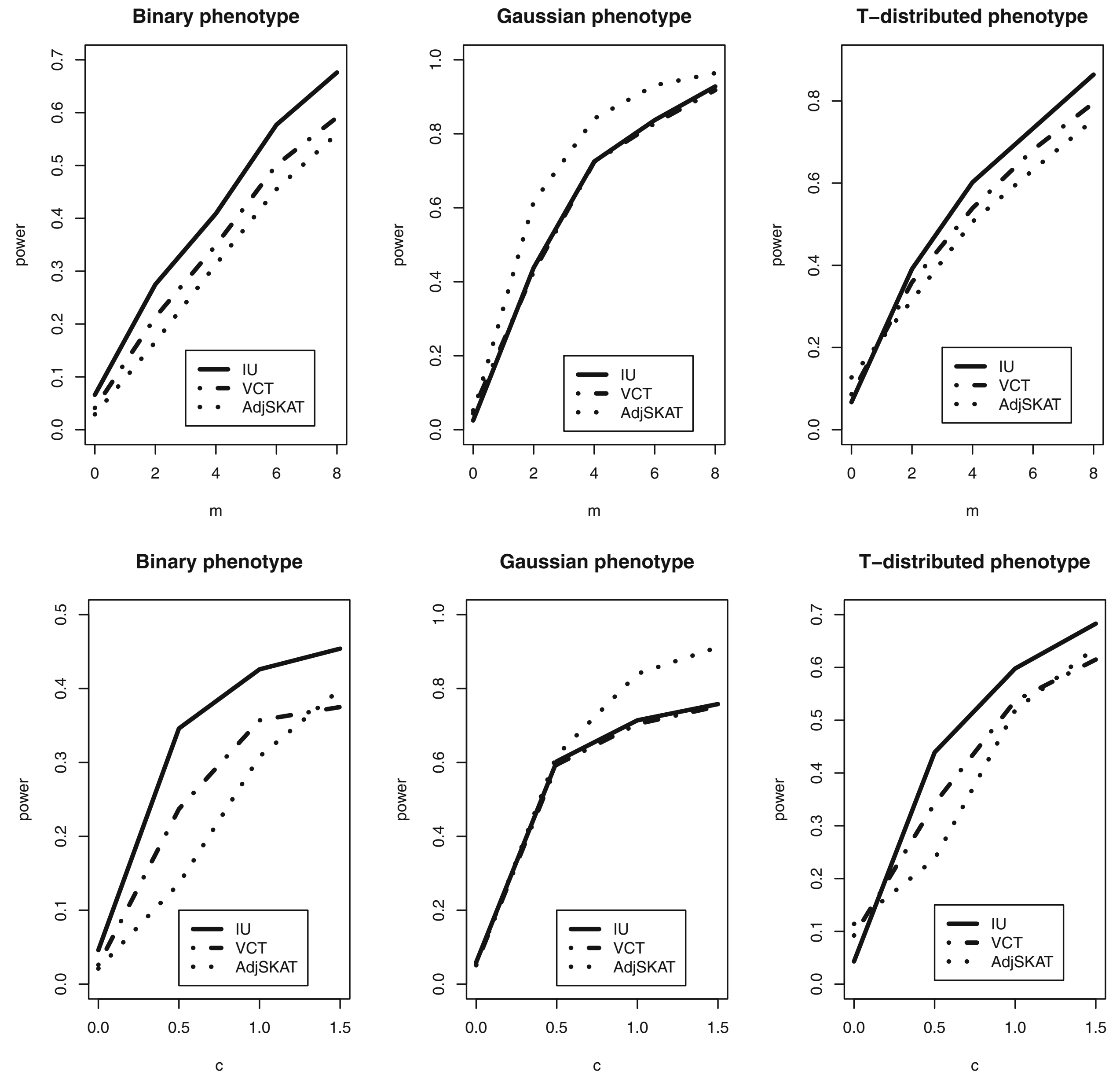

Fig. 3 Power comparison of three methods for different types of phenotypes when there is a multiplicative interaction. *m* stands for the number of causal SNVs and *c* stands for the scale of interaction effect in the Model 2

is applied to detect the joint effect of genes and gene expression data.

From the Q-Q plot of HTN in Fig. 7, we find no evidence of systematical inflation of the association result. While there is no significant findings after adjusting for multiple testing, there are a few genes reached marginal significance (e.g., *UBAC1*). Among the top findings, some genes may have biological plausibility related to hypertension. For instance, *MFGE8* has been previously reported to up-regulate the intake of Dietary Fats [21], and the Dietary Fats regulates blood pressure via Central Leptin mediated pathways [22]. Therefore, *MFGE8* could be a potential risk factor for hypertension [23]. The expression of *IFI44L* is demonstrably increased upon contact of Nickel [24] or Nickel Chloride [25], which is associated with elevated prevalence of hypertension [26]. Although previous studies suggested that some genes in Table 3 may play a role in hypertension, further studies and biological experiments are needed to confirm the association and to further investigate the potential role of these genes in hypertension.

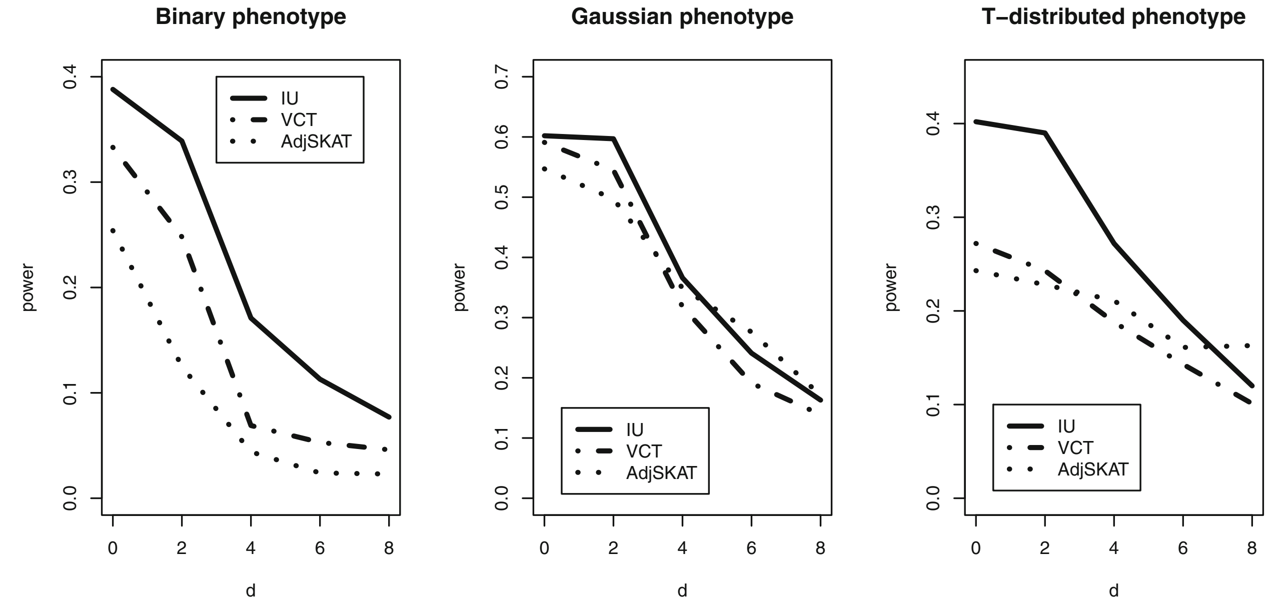

Fig. 4 Power comparison of three methods for different types of phenotypes when there is a threshold interaction. *d* is the threshold parameter in Model 3

Conclusion

To facilitate the integrative analysis of omic data, we proposed a unified non-parametric method to detect the joint association of multi-level omic data with various types of phenotypes. There are three main contributions of the proposed IU method. First, it provides robust performance for various types of phenotypes, including binary, Gaussian and heavy-tailed distributions, due to the robustness of U statistics. Second, the proposed integrative U test achieves higher or comparable power compared to existing methods (e.g., VCT) under different types of interaction models. Finally, we also provide a simple sample size/power calculation to facilitate the design of multi-level omic studies.

The connection between the proposed method and variance component tests is that all test statistics are

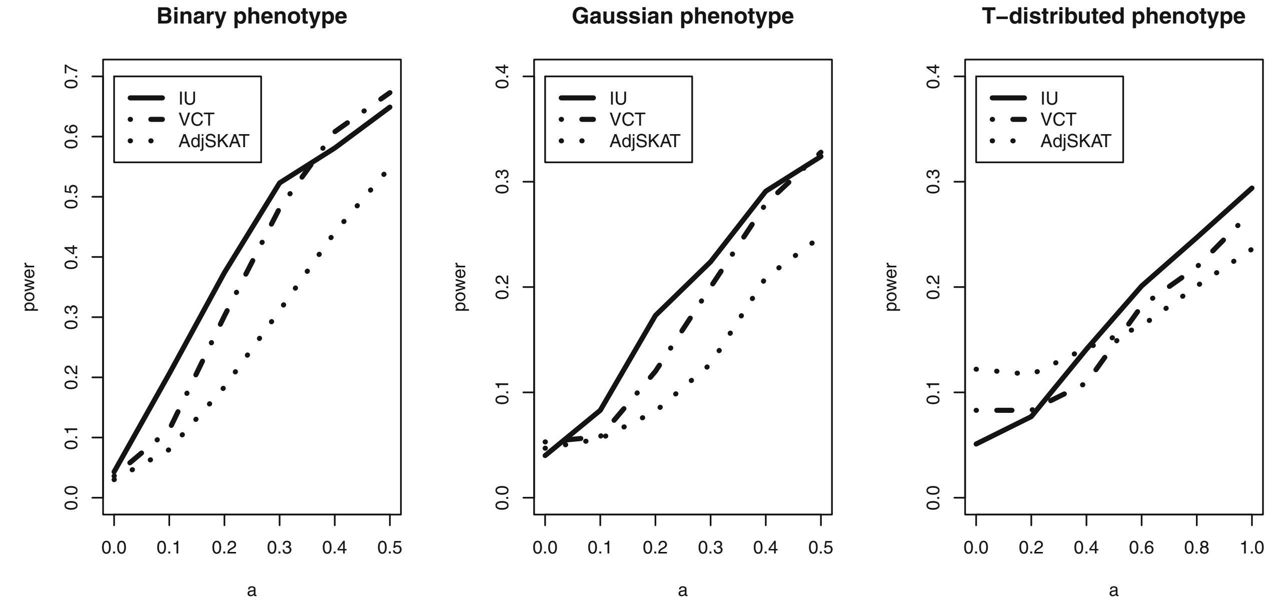

Fig. 5 Power comparison of three methods for different types of phenotypes when there is a random effect interaction in Model 1. *a* is the boundary of the uniform distribution used to generate the random effect

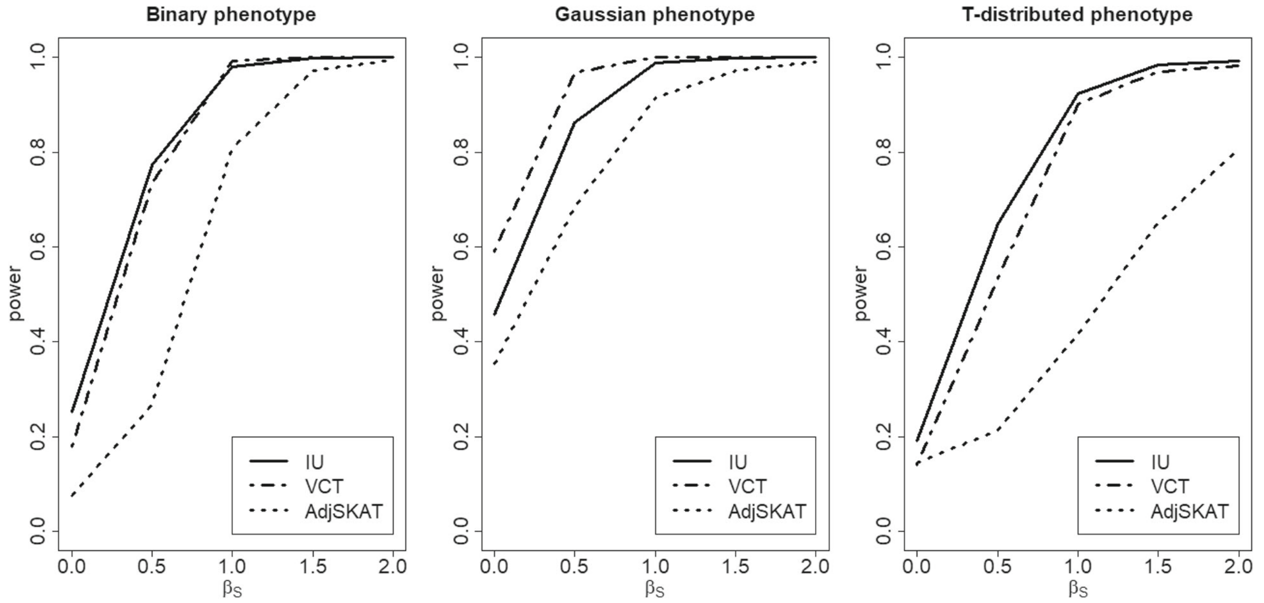

Fig. 6 Power comparison of three methods for different types of phenotypes when there are both marginal effects ($\sigma_G = 0.1$, $\beta_S = 0, 0.5, 1, 1.5$ and 2) and random interaction effect ($\gamma \sim U(-0.1, 0.1)$) in Model 1

in the form of kernel quadratic framework as seen in "Simulation" section. It also connects to several other U-statistic-based methods [10, 27]. As a similarity-based test, the IU method is proposed as a non-degenerated U statistic, which follows a normal distribution. One advantage of using a non-degenerated U statistic is the computational accuracy with no distribution approximation. If we centralize the phenotype, it becomes a degenerated U test, which follows a mixture chi-square distribution.

The IU test can be extended to handle more than 2 levels of omic data. For instance, a modified IU test can be applied for 3 levels of omic data. Besides the SNVs and gene expression, we further introduce R_i as the DNA methylation for subject i, and use kernel matrix $K_3(\cdot,\cdot)$ to measure the DNA methylation similarities. The IU test statistic can then be defined as

$$U_n = \frac{1}{n(n-1)} \sum_{i \neq j} K_1(Y_i, Y_j) K_2(S_i, S_j) K_3(R_i, R_j) \int G_i(t) G_j(t) dt.$$

The choice of K_3 is similar to K_1 and K_2 as discussed in "Methods" section. Following the same argument for Theorem 1, we can show that this modified IU test also follows an asymptotically normal distribution. In addition, with multiple genes (e.g., genes in a biological pathway) and the corresponding gene expression levels, the gene expression level S can also be modeled as a function. For such purpose, the similarity measure $K_2(S_i, S_j)$ can be modified as $\tilde{K}_2(S_i(t), S_j(t))$ where $\tilde{K}_2(\cdot,\cdot)$ measures

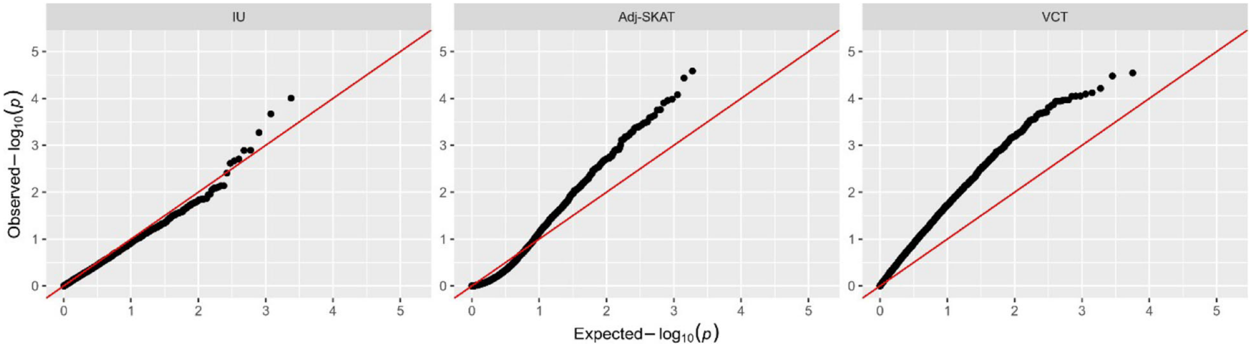

Fig. 7 QQplot of the hypertension study using the proposed IU test, Adj-SKAT and VCT

Table 3 Top 10 gene findings from the integrative analysis in a hypertension study

Name	Chromosome	Starting location	Ending location	# of SNVs	*p*-value
UBAC1	9	138823836	138854205	287	9.86×10^{-5}
MEGF11	15	66186838	66546725	2989	2.14×10^{-4}
IFI44L	1	79085201	79112428	207	5.35×10^{-4}
MFGE8	15	89440944	89457653	161	1.27×10^{-3}
ANKDD1A	15	65203490	65251983	464	1.29×10^{-3}
PDZD2	5	31798110	32111928	3655	1.96×10^{-3}
TBX4	17	59532864	59561970	221	2.15×10^{-3}
IGSF3	1	117116060	117211147	429	2.42×10^{-3}
TMEM61	1	55445562	55458886	170	3.90×10^{-3}
FAM46B	1	27330739	27340321	63	7.29×10^{-3}

the similarity between two functions. The asymptotic property can be derived based on the same argument for Theorem 1. One potential limitation of this study is that gene expression is assumed to be independent of SNVs. One technical reason of making such assumption is that, under the stochastic process setup, it is hard to model the association between the gene expression variable S and the underlying stochastic process $SP(\eta(t), \Gamma(s,t))$. Finding an appropriate way to model correlations among omic data is a challenging topic that is worth of further investigation. Nevertheless, if real data indicates correlations among different levels of omic data, one way to overcome this issue is to adopt methods introduced by [27] and [28].

Additional file

Additional file 1: The proof of Theorem 2.1 can be found in the Appendix. (PDF 94 kb)

Abbreviations

Adj-SKAT: Adjusted sequence kernel association test; DE: Double exponential; HTN: Hypertension; IU: Integrative U test; SAFDGS: San Antonio Family Diabetes/Gallbladder Study; SAFHS: San Antonio Family Heart Study; SKAT: Sequence kernel association test; VCT: Variance component test; WGS: Whole-genome sequencing

Acknowledgements

Not applicable.

Funding

This project was supported by the National Institute on Drug Abuse (Award No. R01DA043501) and the National Library of Medicine (Award No. R01LM012848).

Availability of data and materials

Data used in this article comes from the Genetic Analysis Workshop.

Authors' contributions

PG and QL participate in the design of the study. PG implemented the methods and drafted the manuscript. XT was involved in the data analysis. QL participated in the conception of the study and in editing the manuscript. All authors read and approved the final manuscripts.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

Author Qing Lu is currently acting as an Editorial Board Member for BMC Genetics. All other authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Department of Mathematics, Illinois State University, 61761, Normal, IL, USA. [2]Department of Epidemiology and Biostatistics, Michigan State University, 48824, East Lansing, MI, USA.

Received: 19 July 2018 Accepted: 20 March 2019
Published online: 10 April 2019

References

1. Collins FS, Varmus H. A new initiative on precision medicine. New Eng J Med. 2015;372(9):793–5.
2. Lappalainen T, Sammeth M, Friedlander MR, 't Hoen PA, Monlong J, Rivas MA, Gonzalez-Porta M, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013;501(7468):506–11.
3. Kristensen VN, Lingjærde OC, Russnes HG, Vollan HK, Frigessi A, Børresen-Dale A. Principles and methods of integrative genomic analyses in cancer. Nat Rev Cancer. 2014;14(5):299–313.
4. Lin W, Feng R, Li H. Regularization Methods for High-Dimensional Instrumental Variables Regression With an Application to Genetical Genomics. J Am Stat Assoc. 2015;110(509):270–88.
5. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. Nature Reviews Genetics. 2015;16(2):85–97.
6. Zhao SD, Cai TT, Li H. More powerful genetic association testing via a new statistical framework for integrative genomics. Biometrics. 2014;70(4):881–90.
7. Ainsworth HF, Shin S, Cordell HJ. A comparison of methods for inferring causal relationships between genotype and phenotype using additional biological measurements. Genet Epidemiol. 2017;41(7):577–86.
8. Huang Y-T, Vanderweele TJ, Lin X. Joint analysis of SNP and gene expression data in genetic association studies of complex diseases. Ann Appl Stat. 2014;8:352–76.
9. Huang Y-T. Integrative modeling of multiple genomic data from different types of genetic association studies. Biostatistics. 2014;15(4):587–602.
10. Wei C, Li M, He Z, Vsevolozhskaya O, Schaid DJ, Lu Q. A weighted U-statistic for genetic association analyses of sequencing data. Genet Epidemiol. 38(8):699–708.
11. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61–70.

12. Hawkins RD, Hon GC, Ren B. Next-generation genomics: an integrative approach. Nat Rev Genet. 2010;11(7):476–86.
13. Luo L, Zhu Y, Xiong M. Quantitative trait locus analysis for next-generation sequencing with the functional linear models. J Med Genet. 2012;49(8):513–24.
14. Vsevolozhskaya OA, Zaykin DV, Greenwood MC, Wei C, Lu Q. Functional analysis of variance for association studies. PLOS ONE. 2014;9(9):e105074.
15. Fan R, Wang Y, Boehnke M, Chen W, Li Y, Ren H, Lobach I, Xiong M. Gene level meta-analysis of quantitative traits by functional linear models. Genetics. 2015;200(4):1089–104.
16. Fan R, Wang Y, Chiu CY, Chen W, Ren H, Li Y, Boehnke M, Amos CI, Moore JH, Xiong M. Meta-analysis of complex diseases at gene level by generalized functional linear models. Genetics. 2015;202(2):457–70.
17. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011;89:82–93.
18. Serfling RJ. Approximation theorems of mathematical statistics. Wiley Series in Probability and Statistics. Hoboken: Wiley; 1981.
19. Zhang J-T. Analysis of Variance for Functional Data. London: Chapman & Hall; 2013.
20. 1000 Genomes Project Consortium, Abecasis GR, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061–73.
21. Kerley-Hamilton JS, Trask HW, Ridley CJ, Dufour E, Ringelberg CS, Nurinova N, Wong D, Moodie KL, Shipman SL, Moore JH, Korc M, Shworak NW, Tomlinson CR. Obesity is mediated by differential aryl hydrocarbon receptor signaling in mice fed a Western diet. Environ Health Perspect. 2012;120(9):1252–9.
22. Han C, Wu W, Ale A, Kim MS, Cai D, 2016. Central Leptin and Tumor Necrosis Factor-α (TNFα) in Diurnal Control of Blood Pressure and Hypertension. Int J Biol Chem. 291(29):15131–42.
23. BrahmaNaidu P, Nemani H, Meriga B, Mehar SK, Potana S, Ramgopalrao S. Mitigating efficacy of piperine in the physiological derangements of high fat diet induced obesity in Sprague Dawley rats. Chem Biol Interact. 2014;221:42–51.
24. Correa RJ, Malajian D, Shemer A, Rozenblit M, Dhingra N, Czarnowicki T, Khattri S, Ungar B, Finney R, Xu H, Zheng X, Estrada YD, Peng X, Suarez-Farinas M, Krueger JG, Guttman-Yassky E. Patients with atopic dermatitis have attenuated and distinct contact hypersensitivity responses to common allergens in skin. J Allergy Clin Immunol. 2015;135(3):712–20.
25. Tchou-Wong KM, Kiok K, Tang Z, Kluz T, Arita A, Smith PR, Brown S, Costa M. Effects of nickel treatment on H3K4 trimethylation and gene expression. PLOS ONE. 2011;6(3):e17728.
26. Yang AM, Bai YN, Pu HQ, Zheng TZ, Cheng N, Li JS, Li HY, Zhang YW, Ding J, Su H, Ren XW, Hu XB. Prevalence of metabolic syndrome in Chinese nickel-exposed workers. Biomed Environ Sci. 2014;27(6):475–7.
27. Wei C, Elston RC, Lu Q. A weighted U statistic for association analyses considering genetic heterogeneity. Stat Med. 2016;35(16):2802–14.
28. Jiang Y, Li N, Zhang H. Identifying Genetic Variants for Addiction via Propensity Score Adjusted Generalized Kendall's Tau. J Am Stat Assoc. 2014;109(507):905–30.