Yang *et al. BMC Bioinformatics* (2016) 17:19
DOI 10.1186/s12859-015-0868-6

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

An efficient genome-wide association test for multivariate phenotypes based on the Fisher combination function

James J. Yang[1*], Jia Li[2], L. Keoki Williams[3] and Anne Buu[4]

Abstract

Background: In genome-wide association studies (GWAS) for complex diseases, the association between a SNP and each phenotype is usually weak. Combining multiple related phenotypic traits can increase the power of gene search and thus is a practically important area that requires methodology work. This study provides a comprehensive review of existing methods for conducting GWAS on complex diseases with multiple phenotypes including the multivariate analysis of variance (MANOVA), the principal component analysis (PCA), the generalizing estimating equations (GEE), the trait-based association test involving the extended Simes procedure (TATES), and the classical Fisher combination test. We propose a new method that relaxes the unrealistic independence assumption of the classical Fisher combination test and is computationally efficient. To demonstrate applications of the proposed method, we also present the results of statistical analysis on the Study of Addiction: Genetics and Environment (SAGE) data.

Results: Our simulation study shows that the proposed method has higher power than existing methods while controlling for the type I error rate. The GEE and the classical Fisher combination test, on the other hand, do not control the type I error rate and thus are not recommended. In general, the power of the competing methods decreases as the correlation between phenotypes increases. All the methods tend to have lower power when the multivariate phenotypes come from long tailed distributions. The real data analysis also demonstrates that the proposed method allows us to compare the marginal results with the multivariate results and specify which SNPs are specific to a particular phenotype or contribute to the common construct.

Conclusions: The proposed method outperforms existing methods in most settings and also has great applications in GWAS on complex diseases with multiple phenotypes such as the substance abuse disorders.

Keywords: Genome-wide association study, Fisher combination function, Multivariate permutation, Principal component analysis

Background

In the past decade, genome-wide association studies (GWAS) have produced rich single-nucleotide polymorphism (SNP) data available to researchers. Among them, the large scale studies including the HapMap project [1] and the 1000 Genomes project [2] have provided publicly accessible databases of reference ancestral populations for imputation and quality control purposes. The idea of GWAS is to conduct fast SNP-based association tests to scan the whole genome using case-control samples. Yet, many complex diseases such as mental health disorders may have multiple phenotypic traits with continuous outcomes [3]. This pleiotropy in complex traits [4] provides several potential advantages to the direct modeling of pleiotropic associations. First, a model search for loci that are simultaneously associated with multiple phenotypes would likely have higher power than a model search that only considers each phenotype individually. Second, more exact modeling may yield more accurate prediction of either or both phenotypes. Third, pleiotropic genes may tend to have a more central role in the relevant functional pathways.

*Correspondence: jjyang@umich.edu
[1] School of Nursing, University of Michigan, Ann Arbor, Michigan
Full list of author information is available at the end of the article

Existing statistical methods for complex diseases with multivariate phenotypes can be categorized into three types of approaches. The first approach is to conduct a GWAS for each marginal phenotype and then aggregate the results. The major issue with this approach is that it does not make use of the correlation structure among phenotypes. The second approach is to summarize multiple phenotypic traits into a composite score and then conduct a GWAS on the score. This approach, however, may have difficulty in identifying proper summary scores. The third approach involves multiple phenotypic traits simultaneously. Thus, it may gain power as well as avoid the issue of multiple testing. However, it is based on stronger assumptions that may not be satisfied in some practical settings.

In this study, we provide a comprehensive review of existing statistical methods for conducting GWAS on complex diseases with multiple phenotypic traits. We also propose a new statistical method based on the Fisher combination function. The performance of competing methods is evaluated by a simulation study. In order to demonstrate applications of the proposed method, we conduct statistical analysis on the database of the Study of Addiction: Genetics and Environment (SAGE).

Methods

Let $X_i(= 0, 1, 2)$ be the number of reference alleles corresponding to a candidate SNP and $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{im})'$ be the measures of multiple phenotypes for the individual i. In this study, we conduct a comprehensive review of existing statistical methods that can be used to test the association between X_i and $\boldsymbol{Y}_i$.

Existing methods

Multivariate analysis of variance (MANOVA)

When the phenotype is univariate, we can use the one-way analysis of variance (ANOVA) with three levels of the genotype for GWAS. When we have correlated multivariate phenotypic traits, the natural extension of the one-way ANOVA is the one-way multivariate analysis of variance (MANOVA) [5]. Similar to ANOVA, MANOVA tests the equality of mean phenotypic vectors by comparing the within genotypes and between genotypes variance-covariance matrices. The strength of MANOVA is that the multivariate normal distribution provides many good statistical properties for testing and estimation [6]. However, in practice, multivariate phenotype data are very unlikely to meet the multivariate normal assumption. Furthermore, MANOVA is most powerful when the phenotypes are negatively correlated and yet this situation is also unlikely in practice, especially when the number of phenotypes is larger than 2. With respect to its relevant applications, this method has been used in GWAS on dose-response [7] and facial morphology [8].

Principal component analysis (PCA)

The principal component analysis (PCA) [9, 10] is another classical statistical method for multivariate analysis. The primary objective of PCA is to find a small set of linear combinations of the original variables (i.e. principal components) that account for the most variability in the original variables. Thus, it can be employed to reduce the dimension of multivariate phenotypes. The PCA has been used in gene-based studies to increase the power of statistical testing [11, 12]. Furthermore, He et al. [13] has used PCA to combine four highly correlated obesity phenotypes for a whole genome linkage scan. When the phenotypes are highly correlated, the first principal component (corresponding to the largest eigenvalue) contains most information about the phenotype data. Thus, testing the association between a SNP and the first principal component is a commonly adopted approach to effectively change the multivariate setting associated with multiple phenotypes in GWAS to the univariate setting (e.g. Zhang et al. [14] and Karasik et al. [15]). In this study, we investigate the statistical properties of this approach.

Generalized estimating equations (GEE)

The method of generalized estimating equations (GEE) [16] was developed for analyzing correlated multivariate outcomes primarily from longitudinal studies. It can be applied to test the association between a candidate SNP and multivariate phenotypes. The GEE only requires specification of the link function and the working correlation matrix. The former depends on the measurement scale of the outcomes (e.g. the identify link for continuous outcomes). The latter assumes the correlation structure among multivariate outcomes. The estimation of GEE is usually robust against this assumption. GEE was widely used in GWAS. For example, GEE was proposed as one of the multivariate approaches in Solovieff et al. [4]. For another example, Liu et al. [17] proposed to use GEE for bivariate association analyses for the mixture of continuous and binary traits. However, to the best of our knowledge, none of the existing studies have conducted simulations to investigate whether GEE can control the type I error rate when multivariate traits are involved in GWAS. To fill in this knowledge gap, we conduct a simulation study to examine the statistical properties of this approach. In this study, we only consider the identity link because we are mainly interested in continuous phenotypic traits. We also assume the working correlation matrix to be compound symmetry because it only requires us to estimate one additional parameter.

Trait-based association test involving the extended Simes procedure

Recently, van der Sluis et al. [18] developed a trait-based association test involving the extended Simes procedure

(TATES). The TATES calculates a global p-value based on individual p-values of association tests for marginal phenotypes. Specifically, for m-variate phenotypic traits, one can conduct m tests of the association between a candidate SNP and each marginal phenotypic trait and derive m p-values: $p_1, \ldots p_m$. Let $p_{(1)}, \ldots p_{(m)}$ be the ordered p-values from the smallest to the largest. The Simes multiple procedure declares significance between a SNP and multivariate phenotypic traits at the α level if any of the p-values satisfy $p_{(j)} < j\alpha/m$ [19]. Hence, the global p-value based on the Simes procedure is $p_{\text{traits}} = \min\{mp_{(j)}/j, j = 1, \ldots, m\}$. The TATES improves this procedure by replacing m and j with the effective number of independent traits, m_e and j_e, which are estimated from the eigenvalues of the correlation matrix [20, 21]. Since $m_e \leq m$, this new adjusted global p-value, defined as $p_{\text{traits}} = \min\{m_e p_{(j)}/j_e, j = 1, \ldots, m\}$ is smaller than the Simes global p-value. Therefore, the TATES is more powerful than the Simes. A simulation study also showed that the TATES is more powerful than MANOVA. In this study, we conduct a comprehensive simulation study to compare this method with not only the classical methods reviewed above but also the proposed methods.

Proposed methods

The methods based on the Fisher combination function

Combining independent tests of significance to form a join statistic has been used as an alternative approach to tackling complex multivariate location problems [22]. This approach is quite popular in practice because it is much easier to develop a univariate association test statistic than a multivariate association test statistic. Birnbaum [23] discussed various combination functions among which the Fisher combination function has been proven to be asymptotically Bahadur optimal [24, 25]. Thus, we focus on the Fisher combination function in this study.

Fisher combination test with the independence assumption

Based on the Fisher combination test [22], to test the association between a SNP and multivariate phenotypic traits, we only need to test the association between the SNP and each marginal phenotypic trait individually. Thus, for m-multivariate phenotypes, we have m marginal p-values: $p_1, \ldots, p_m$. The Fisher combination statistic is defined as

$$T = \sum_{j=1}^{m} -2\log(p_j). \quad (1)$$

T is used to infer the association between the SNP and multivariate phenotypic traits. When the marginal p-values are independent, the statistic T follows a chi-squared distribution with $2m$ degrees of freedom so the p-value of T can be obtained straightforwardly. In reality, however, the phenotypic traits are always correlated so the chi-squared distribution with $2m$ degrees of freedom tends to underestimate the variance of the T statistic. The resulting chi-squared test is, therefore, too liberal.

The permutation method

Because of the negative consequence of the independence assumption, it is desirable to conduct the Fisher combination test without the assumption. Ideally, we could calculate the exact p-value of the T statistic in Eq. (1) using the permutation method, which does not require the unrealistic assumption and also controls for the type I error [26]. Yet, the permutation method is a very time-consuming procedure, particularly in a genome-wide context. Thus, an improvement in computational efficiency is warranted.

The proposed efficient method

For correlated phenotypic traits, T is the sum of dependent chi-squared statistics. Brown [27] and Yang [28] have shown that, under the null hypothesis of no association between a SNP and multivariate phenotypic traits, the distribution of T statistic follows a scale chi-squared distribution ($\gamma\chi^2_\nu$), or equivalently, a gamma distribution with the shape parameter $\nu/2$ and the scale parameter 2γ. Therefore, to calculate the global p-value of T statistic, we only need to estimate the parameters ν and γ. Suppose that the mean of T is μ and the variance of T is σ^2. Using the first and second moments of T, the values of ν and γ can be calculated as $\nu = 2\mu^2/\sigma^2$ and $\gamma = \sigma^2/(2\mu)$. The following are technical details of the derivation of the mean and variance of T statistic when the marginal p-values are based on two-sided tests (see Brown [27] and Yang [28] for the case of one-sided marginal tests):

Without loss of generality, we assume that the association test statistic for the jth phenotypic trait is z_j where $j = 1, \ldots, m$. The corresponding two-sided p-value is defined as $p_j = 2\Phi(-|z_j|)$, where Φ is the standard Gaussian distribution function. Under the null hypothesis of no association between a SNP and multivariate phenotypic traits, the distribution of T statistic is approximated by a Gaussian distribution with the mean of T as

$$\mu = E[T] = 2m$$

and the variance of T as

$$\begin{aligned}
\sigma^2 &= \text{Var}[T] \\
&= \text{Var}\left\{\sum_{j=1}^{m} -2\log(p_j)\right\} \\
&= \sum_{j=1}^{m} \text{Var}\{-2\log(p_j)\} + \sum_{j\neq k} \text{cov}\{-2\log(p_j), -2\log(p_k)\} \\
&= 4m + \sum_{j\neq k} \text{cov}\{-2\log(p_j), -2\log(p_k)\}.
\end{aligned}$$

Therefore, in order to calculate the variance of T, we need to calculate the covariance for each pair (j,k) which can be expressed as

$$\begin{aligned}
&\operatorname{cov}\{-2\log(p_j), -2\log(p_k)\} \\
&= E\{[-2\log(p_j)]\,[-2\log(p_k)]\} - E\{-2\log(p_j)\}E\{-2\log(p_k)\} \\
&= 4\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\log\{2\Phi(-|z_j|)\}\log\{2\Phi(-|z_k|)\}dF(z_j,z_k) - 4,
\end{aligned}$$

where F is the standard bivariate Gaussian distribution. Let

$$\delta_{jk} = \operatorname{cov}\{-2\log(p_j), -2\log(p_k)\}.$$

Thus, δ_{jk} is a function of the correlation between z_j and z_k: $\rho_{j,k}$. We explore the relationship between δ_{jk} and $\rho_{j,k}$ by calculating δ_{jk} numerically for the values of $\rho_{j,k}$ from -0.99 to 0.99 with the step of 0.01. The results are shown in Fig. 1. Since the curve of Fig. 1 is a convex curve symmetric about the y-axis, we can approximate the relationship between δ_{jk} and $\rho_{j,k}$ using a tenth-order polynomial:

$$\delta_{jk} \doteq c_1\rho_{j,k}^2 + c_2\rho_{j,k}^4 + c_3\rho_{j,k}^6 + c_4\rho_{j,k}^8 + c_5\rho_{j,k}^{10}. \quad (2)$$

Using the adapt function in the R package fCopulae [29], we obtained the following estimates: $c_1 = 3.9081$, $c_2 = 0.0313$, $c_3 = 0.1022$, $c_4 = -0.1378$ and $c_5 = 0.0941$; the maximum residual was less than 0.0001.

To estimate δ_{jk} in Eq. (2) accurately, we have taken two steps to remove potential biases. First, since the sample correlation $\hat{\rho}_{j,k}$ is not an unbiased estimator of $\rho_{j,k}$ [30], we estimate $\rho_{j,k}$ by the bias-corrected sample correlation $\hat{r}_{j,k}$:

$$\hat{r}_{j,k} = \hat{\rho}_{j,k}\left(1 + \frac{1-\hat{\rho}_{j,k}^2}{2(n-3)}\right), \quad (3)$$

where n is the samples size used to calculate $\hat{\rho}_{j,k}$. Now, let's define the right hand side of Eq. (2) as

$$f(r) = c_1r^2 + c_2r^4 + c_3r^6 + c_4r^8 + c_5r^{10}.$$

The estimate $f(\hat{r}_{j,k})$ is still a biased estimator of δ_{jk}. Thus, we propose a second step to remove the bias. Using the Taylor series expansion, we can estimate the bias as

$$\frac{c_1}{n}\left(1-\hat{r}_{j,k}^2\right)^2.$$

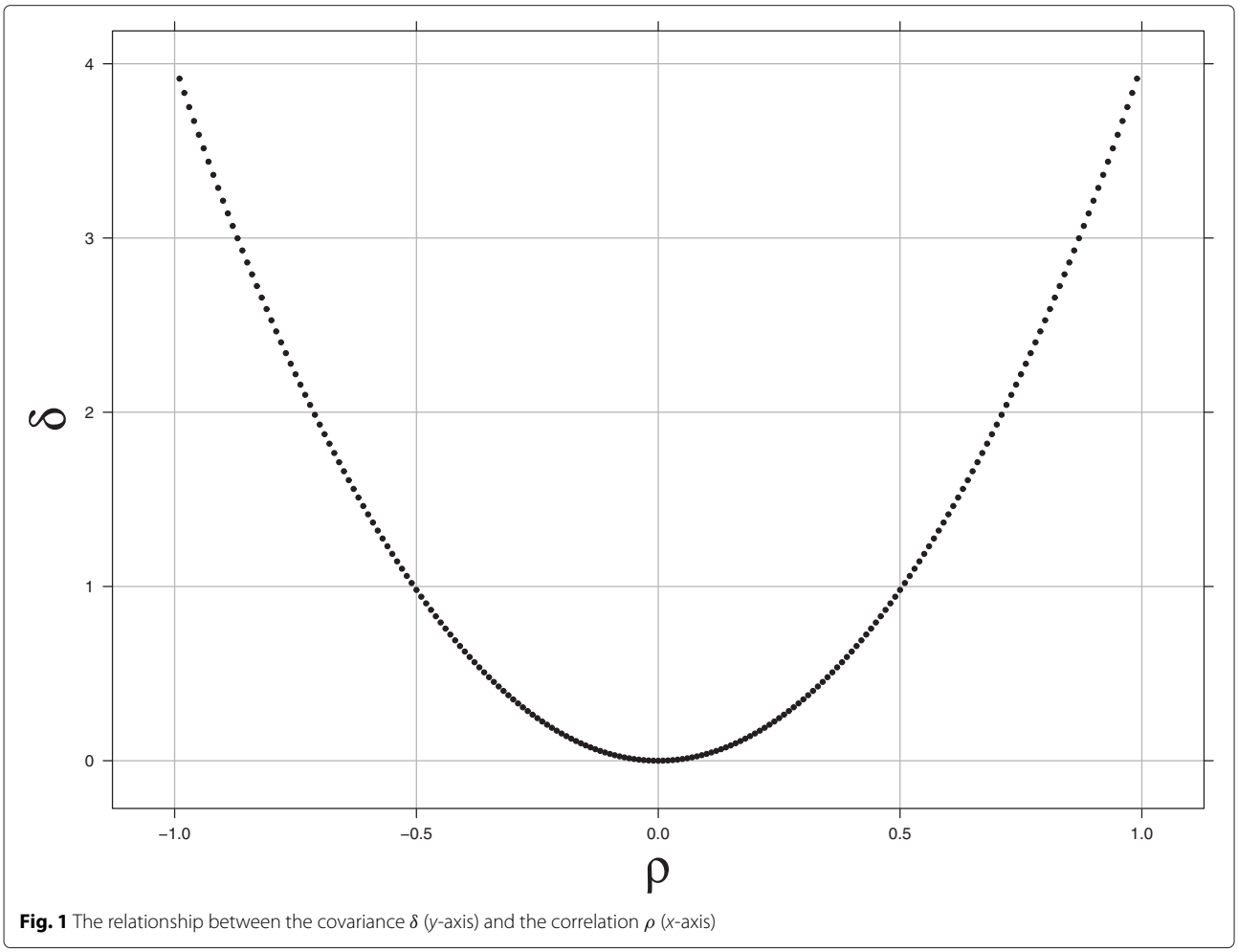

Fig. 1 The relationship between the covariance δ (y-axis) and the correlation ρ (x-axis)

Therefore, the proposed unbiased estimator of δ_{jk} is

$$f(\hat{r}_{j,k}) - \frac{c_1}{n}\left(1-\hat{r}_{j,k}^2\right)^2. \quad (4)$$

Hence, based on Eqs. (3) and (4), the variance of T can be estimated as

$$\sigma^2 = \mathrm{Var}(T) \doteq 4m + \sum_{j\neq k}\left(f(\hat{r}_{j,k}) - \frac{c_1}{n}\left(1-\hat{r}_{j,k}^2\right)^2\right).$$

Given the proposed estimators of μ and σ^2, the global p-value of T statistic can be computed efficiently using the gamma distribution function as follows:

$$\text{the global } p\text{-value} = 1 - \Gamma\left(\mu^2/\sigma^2, \sigma^2/\mu\right),$$

where $\Gamma(\nu/2, 2\gamma)$ is the gamma distribution function with the shape parameter $\nu/2$ and the scale parameter 2γ.

In this study, we compare two alternative methods to calculate $\hat{\rho}_{j,k}$: the Pearson sample correlation coefficient and the rank correlation coefficient of Kendall's τ. Kendall and Gibbons [31] have shown the relation between ρ and τ as

$$\rho = \sin\left(\frac{\pi\tau}{2}\right).$$

Thus, we can use Kendall's τ to derive $\hat{\rho}_{j,k}$ as $\sin\left(\frac{\pi\hat{\tau}_{j,k}}{2}\right)$. In the simulation study, we evaluate the robustness of Kendall's τ in comparison to Pearson's sample correlation coefficient.

Results

Simulation study

We conducted a simulation study to evaluate the performance of the four methods reviewed (MANOVA, PCA, GEE, and TATES) as well as the methods based on the Fisher combination function. In the simulation, we adopted four different approaches to calculating the p-value of the Fisher combination test:

1. FC-χ^2_{2m}: the chi-squared distribution with $2m$ degrees of freedom under the independence assumption.
2. FC-Permutation: the permutation method based on 1000 permutes.
3. FC-Pearson: the proposed method with the correlation $\hat{\rho}_{j,k}$ being estimated by the Pearson's sample correlation coefficient.
4. FC-Kendall: the proposed method with $\hat{\rho}_{j,k}$ being estimated by the Kendall's τ.

Simulation configurations

For each subject, the relationship between the SNP, x, and the multivariate phenotypes, $\boldsymbol{Y}$, was defined as

$$\boldsymbol{Y} = x\beta + \epsilon,$$

where β is the effect of the SNP on the phenotypes and ϵ is the error term. In this simulation study, we simulated x based on a minor allele frequency of 0.25. We evaluated the performance of competing methods under three different settings of effect sizes:

Null hypothesis (no effects) $\beta = (0,0,0,0,0)'$.
Moderate equal effect sizes $\beta = (0.3,0.3,0.3,0.3,0.3)'$.
Varied effect sizes $\beta = (0.1,0.2,0.3,0.4,0.5)'$.

For the error term, we considered two cases. In the first case, the error term was simulated from a multivariate normal distribution with the mean $\mathbf{0}$ and the variance-covariance $\boldsymbol{\Sigma}$, which is a compound symmetry matrix with the value of 1 on the diagonal and the value of $\varrho = 0, 0.25, 0.5$, or 0.75 on the off-diagonal. In the second case, the error term was simulated from a mixture of two multivariate normal distributions: 90 % from the same multivariate normal distribution in the first case and 10 % from the multivariate normal distribution with the mean $\mathbf{0}$ and the variance-covariance matrix $5\boldsymbol{\Sigma}$. The purpose of the second case was to simulate long tailed distributions of phenotypic traits which are common in real data. We generated simulated data of 100 subjects under each configuration. In addition, each configuration was repeated 10,000 times. The nominal type I error rate was set at 0.05 and the power was calculated as the proportion of p-values less than 0.05.

Simulation results

Table 1 presents the simulation results when the multivariate phenotypes come from a multivariate normal distribution with the value of the correlation ϱ varied from 0 to 0.75. The numbers in each cell are the mean (standard deviation) of the indicator variable for p-value < 0.05 among the 10,000 replications. The top panel corresponds to the case of $\beta = (0,0,0,0,0)'$ (i.e. when the null hypothesis is true) and thus can be used to evaluate if each of the competing methods was able to control the type I error. The results indicate that GEE and the Fisher combination test with χ^2_{2m} did not control the type I error rate when $\varrho > 0$ while all the other methods did quite well under all values of ϱ. We, thus, did not find it meaningful to further compare these two methods with the other methods in terms of the statistical power.

The middle panel of Table 1 compares the power of competing methods under the situation that the SNP has the same level of association with each of the phenotypic traits: $\beta = (0.3,0.3,0.3,0.3,0.3)'$. The power of MANOVA decreased rapidly as the correlation ϱ increased. When $\varrho = 0.75$, for instance, all the other methods had the power of at least 0.4 but the power of MANOVA was only 0.2. Further, PCA and TATES had higher power than MANOVA when $\varrho > 0$. Yet, none of these three methods can beat the three Fisher combination tests

Table 1 Simulation results when the multivariate phenotypes come from a multivariate normal distribution

ϱ	MANOVA	PCA	GEE	TATES	FC-χ^2_{2m}	FC-Permutation	FC-Pearson	FC-Kendall
				$\beta = (0,0,0,0,0)'$				
0	0.0477	0.0514	0.0109	0.0487	0.0468	0.0455	0.0455	0.0451
	(0.0021)	(0.0022)	(0.0010)	(0.0022)	(0.0021)	(0.0021)	(0.0021)	(0.0021)
0.25	0.0477	0.0499	0.0763	0.0498	0.0631	0.0488	0.0482	0.0477
	(0.0021)	(0.0022)	(0.0027)	(0.0022)	(0.0024)	(0.0022)	(0.0021)	(0.0021)
0.5	0.0477	0.0496	0.1518	0.0506	0.0942	0.0473	0.0482	0.0484
	(0.0021)	(0.0022)	(0.0036)	(0.0022)	(0.0029)	(0.0021)	(0.0021)	(0.0021)
0.75	0.0477	0.0496	0.2202	0.0494	0.1263	0.0467	0.0489	0.0485
	(0.0021)	(0.0022)	(0.0041)	(0.0022)	(0.0033)	(0.0021)	(0.0022)	(0.0021)
				$\beta = (0.3,0.3,0.3,0.3,0.3)'$				
0	0.7595	0.5679	0.9333	0.7359	0.9067	0.9058	0.9047	0.9040
	(0.0043)	(0.0050)	(0.0025)	(0.0044)	(0.0029)	(0.0029)	(0.0029)	(0.0029)
0.25	0.4086	0.7075	0.8570	0.6406	0.8076	0.7748	0.7749	0.7749
	(0.0049)	(0.0045)	(0.0035)	(0.0048)	(0.0039)	(0.0042)	(0.0042)	(0.0042)
0.5	0.2655	0.5295	0.8113	0.5668	0.7411	0.6314	0.6420	0.6421
	(0.0044)	(0.0050)	(0.0039)	(0.0050)	(0.0044)	(0.0048)	(0.0048)	(0.0048)
0.75	0.2011	0.4144	0.7827	0.4949	0.6927	0.5169	0.5272	0.5278
	(0.0040)	(0.0049)	(0.0041)	(0.0050)	(0.0046)	(0.0050)	(0.0050)	(0.0050)
				$\beta = (0.1,0.2,0.3,0.4,0.5)'$				
0	0.8550	0.6646	0.9272	0.8731	0.9457	0.9454	0.9448	0.9445
	(0.0035)	(0.0047)	(0.0026)	(0.0033)	(0.0023)	(0.0023)	(0.0023)	(0.0023)
0.25	0.6334	0.7243	0.8500	0.8237	0.8864	0.8604	0.8631	0.8621
	(0.0048)	(0.0045)	(0.0036)	(0.0038)	(0.0032)	(0.0035)	(0.0034)	(0.0034)
0.5	0.6203	0.5437	0.8043	0.7758	0.8283	0.7252	0.7334	0.7333
	(0.0049)	(0.0050)	(0.0040)	(0.0042)	(0.0038)	(0.0045)	(0.0044)	(0.0044)
0.75	0.8177	0.4227	0.7756	0.7512	0.7721	0.5821	0.5942	0.5941
	(0.0039)	(0.0049)	(0.0042)	(0.0043)	(0.0042)	(0.0049)	(0.0049)	(0.0049)

The three different effect sizes are: no effect $\beta = (0,0,0,0,0)'$; moderate effects $\beta = (0.3,0.3,0.3,0.3,0.3)'$; and varied effects $\beta = (0.1,0.2,0.3,0.4,0.5)'$. The correlation between genes is ϱ ranging from 0 to 0.75. The competing methods are MANOVA (Multivariate analysis of variance), PCA (Principal component analysis), GEE (Generalized estimating equations), TATES (Trait-based association test involving the extended Simes procedure), FC-χ^2_{2m} (the chi-squared distribution with $2m$ degrees of freedom under the independence assumption), FC-Permutation (the permutation method based on 1,000 permutes), FC-Pearson (the proposed method with the correlation $\hat{\rho}_{j,k}$ being estimated by the Pearson's sample correlation coefficient), and FC-Kendall (the proposed method with $\hat{\rho}_{j,k}$ being estimated by the Kendall's τ). The numbers in each cell are the mean (standard deviation) of the indicator variable for p-value < 0.05 among the 10,000 replications

(FC-Permutation, FC-Pearson, and FC-Kendall) that performed equally well.

The bottom panel of Table 1 evaluates the performance of competing methods under the situation that the strength of association between the SNP and phenotypic traits varies from 0.1 to 0.5. Similar to the previous situation, the three Fisher combination tests had almost identical performance and their power decreased as the correlation ϱ increased. Furthermore, the Fisher combination tests had higher power than the other three methods (MANOVA, PCA, and TATES) in most conditions. MANOVA only beat the Fisher when $\varrho = 0.75$; TATES had higher power than the Fisher when $\varrho \geq 0.5$.

Table 2 shows the simulation results when the multivariate phenotypes come from a mixture of two multivariate normal distributions. In comparison to the corresponding settings under the multivariate normal distributions in Table 1, all the competing methods tended to have lower power under these long tailed distributions. Yet, Table 2 demonstrates similar patterns to the ones observed in Table 1 in general.

Real data analysis

The Study of Addiction: Genetics and Environment(SAGE)

The National Center for Biotechnology Information (NCBI) has been managing and distributing the large

Table 2 Simulation results when the multivariate phenotypes come from a mixture of two multivariate normal distributions

ϱ	MANOVA	PCA	GEE	TATES	FC-χ^2_{2m}	FC-Permutation	FC-Pearson	FC-Kendall
				$\beta = (0,0,0,0,0)'$				
0	0.0535	0.0543	0.0135	0.0481	0.0487	0.0482	0.0461	0.0477
	(0.0023)	(0.0023)	(0.0012)	(0.0021)	(0.0022)	(0.0021)	(0.0021)	(0.0021)
0.25	0.0553	0.0514	0.0771	0.0496	0.0627	0.0465	0.0458	0.0469
	(0.0023)	(0.0022)	(0.0027)	(0.0022)	(0.0024)	(0.0021)	(0.0021)	(0.0021)
0.5	0.0537	0.0501	0.1505	0.0522	0.0895	0.0480	0.0491	0.0501
	(0.0023)	(0.0022)	(0.0036)	(0.0022)	(0.0029)	(0.0021)	(0.0022)	(0.0022)
0.75	0.0525	0.0538	0.2206	0.0481	0.1296	0.0493	0.0526	0.0513
	(0.0022)	(0.0023)	(0.0041)	(0.0021)	(0.0034)	(0.0022)	(0.0022)	(0.0022)
				$\beta = (0.3,0.3,0.3,0.3,0.3)'$				
0	0.5943	0.3299	0.8172	0.5683	0.7677	0.7633	0.7595	0.7619
	(0.0049)	(0.0047)	(0.0039)	(0.0050)	(0.0042)	(0.0043)	(0.0043)	(0.0043)
0.25	0.3038	0.5414	0.7487	0.5003	0.6779	0.6330	0.6333	0.6332
	(0.0046)	(0.0050)	(0.0043)	(0.0050)	(0.0047)	(0.0048)	(0.0048)	(0.0048)
0.5	0.2073	0.3981	0.7135	0.4402	0.6168	0.4989	0.5083	0.5082
	(0.0041)	(0.0049)	(0.0045)	(0.0050)	(0.0049)	(0.0050)	(0.0050)	(0.0050)
0.75	0.1601	0.3135	0.6847	0.3870	0.5779	0.4038	0.4111	0.4116
	(0.0037)	(0.0046)	(0.0046)	(0.0049)	(0.0049)	(0.0049)	(0.0049)	(0.0049)
				$\beta = (0.1,0.2,0.3,0.4,0.5)'$				
0	0.6972	0.4002	0.8087	0.7328	0.8451	0.8425	0.8379	0.8408
	(0.0046)	(0.0049)	(0.0039)	(0.0044)	(0.0036)	(0.0036)	(0.0037)	(0.0037)
0.25	0.4766	0.5579	0.7427	0.6698	0.7656	0.7269	0.7236	0.7259
	(0.0050)	(0.0050)	(0.0044)	(0.0047)	(0.0042)	(0.0045)	(0.0045)	(0.0045)
0.5	0.4728	0.4083	0.7073	0.6237	0.7036	0.5766	0.5855	0.5862
	(0.0050)	(0.0049)	(0.0046)	(0.0048)	(0.0046)	(0.0049)	(0.0049)	(0.0049)
0.75	0.6576	0.3172	0.6799	0.5976	0.6394	0.4532	0.4624	0.4617
	(0.0047)	(0.0047)	(0.0047)	(0.0049)	(0.0048)	(0.0050)	(0.0050)	(0.0050)

The three different effect sizes are: no effect $\beta = (0,0,0,0,0)'$; moderate effects $\beta = (0.3,0.3,0.3,0.3,0.3)'$; and varied effects $\beta = (0.1,0.2,0.3,0.4,0.5)'$. The correlation between genes is ϱ ranging from 0 to 0.75. The competing methods are MANOVA (Multivariate analysis of variance), PCA (Principal component analysis), GEE (Generalized estimating equations), TATES (Trait-based association test involving the extended Simes procedure), FC-χ^2_{2m} (the chi-squared distribution with $2m$ degrees of freedom under the independence assumption), FC-Permutation (the permutation method based on 1,000 permutes), FC-Pearson (the proposed method with the correlation $\hat{\rho}_{j,k}$ being estimated by the Pearson's sample correlation coefficient), and FC-Kendall (the proposed method with $\hat{\rho}_{j,k}$ being estimated by the Kendall's τ). The numbers in each cell are the mean (standard deviation) of the indicator variable for p-value < 0.05 among the 10,000 replications

database of Genotypes and Phenotypes (dbGaP) for scientific investigation of various human diseases [32]. In order to demonstrate the application of the proposed method, we conducted statistical analysis on the Study of Addiction: Genetics and Environment (SAGE) data [33], http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000092.v1.p1. The institutional review board of the University of Michigan has approved this secondary data analysis project (HUM00084927). The SAGE is a case-control study that aggregated together the data from three large scale studies in the substance abuse field: the Collaborative Study on the Genetics of Alcoholism (COGA), the Family Study of Cocaine Dependence (FSCD), and the Collaborative Genetic Study of Nicotine Dependence (COGEND). The total number of individuals with individual level data available is 4121. Each individual was genotyped using the Illumina Human 1M-Duo beadchip which contains over 1 million SNP markers.

We selected unrelated individuals that passed the quality control measures according to the Gene Environment Association Studies Initiative (GENEVA) quality control report. The final number of unrelated individuals is 3,741 (1,732 male, 2,079 female) and the total number of SNP markers is 917,694. Because the purpose of our analysis is to identify the genes that are associated with addiction, we used the symptomatology variables

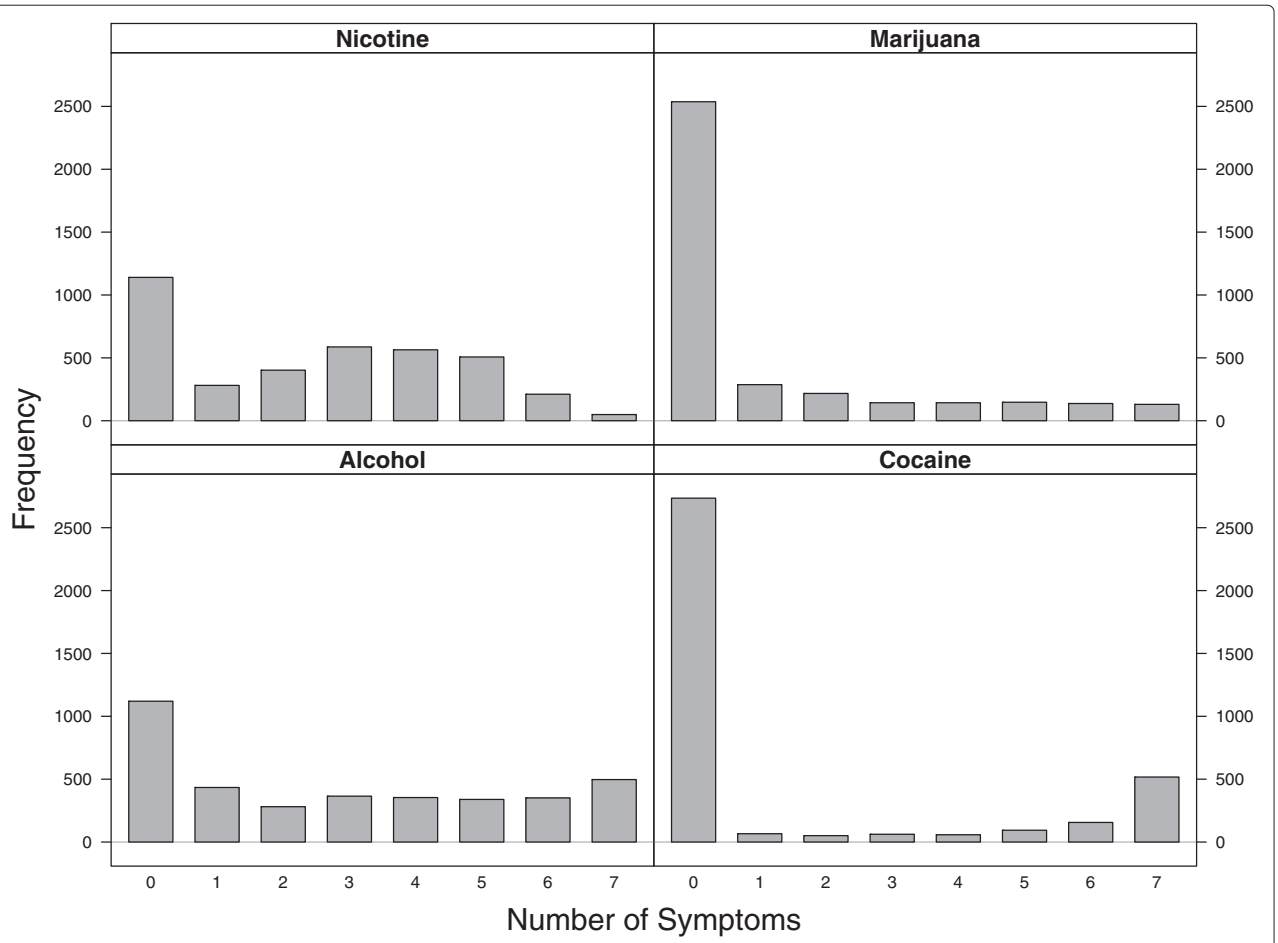

Fig. 2 The distributions of phenotypes for alcohol, nicotine, marijuana and cocaine dependence. The *x*-axis is the number of symptoms, and the *y*-axis is the frequency

of four highly comorbid substance use disorders as the phenotype outcomes: the number of alcohol dependence symptoms endorsed (`alc_sx_tot`), the number of nicotine dependence symptoms endorsed (`nic_sx_tot`), the number of marijuana dependence symptoms endorsed (`mj_sx_tot`), and the number of cocaine dependence symptoms endorsed (`coc_sx_tot`).

SAGE data analysis results

The values of phenotype variables range from 0 to 7. Figure 2 shows the frequency distributions of the 4 phenotype variables. Since they are not normally distributed, we calculated their correlations using the Kendall rank correlation. Table 3 shows moderate correlations ranged from 0.34 to 0.51.

For each trait, we conducted a genome-wide association test using the hurdle model [34] because of the discrete nature and excess zero values associated with the symptom counts. The hurdle regression model assumes the observed data are generated from two processes: one generates zero and the other generates positive values. Since our interest is in the severity of symptomatology, the *p*-values from the positive component of the hurdle model were used for further analysis. For each phenotype of addiction, the estimated *p*-values are summarized using QQ-plots in Fig. 3. The diagonal straight lines have the slope 1 and intercept 0. When the curve of the p-values deviate far away from the diagonal line, it indicates that there are many SNPs significantly associated with the corresponding phenotype trait. By examining the 4 plots in Fig. 3, we obtain the following two findings: (1) For the nicotine symptoms, the *p*-values fall on the diagonal line

Table 3 The Kendall rank pairwise correlations between alcohol, nicotine, marijuana, and cocaine outcomes

	Alcohol	Nicotine	Marijuana	Cocaine
Alcohol	1	0.4554	0.4236	0.5029
Nicotine		1	0.3373	0.3375
Marijuana			1	0.5067
Cocaine				1

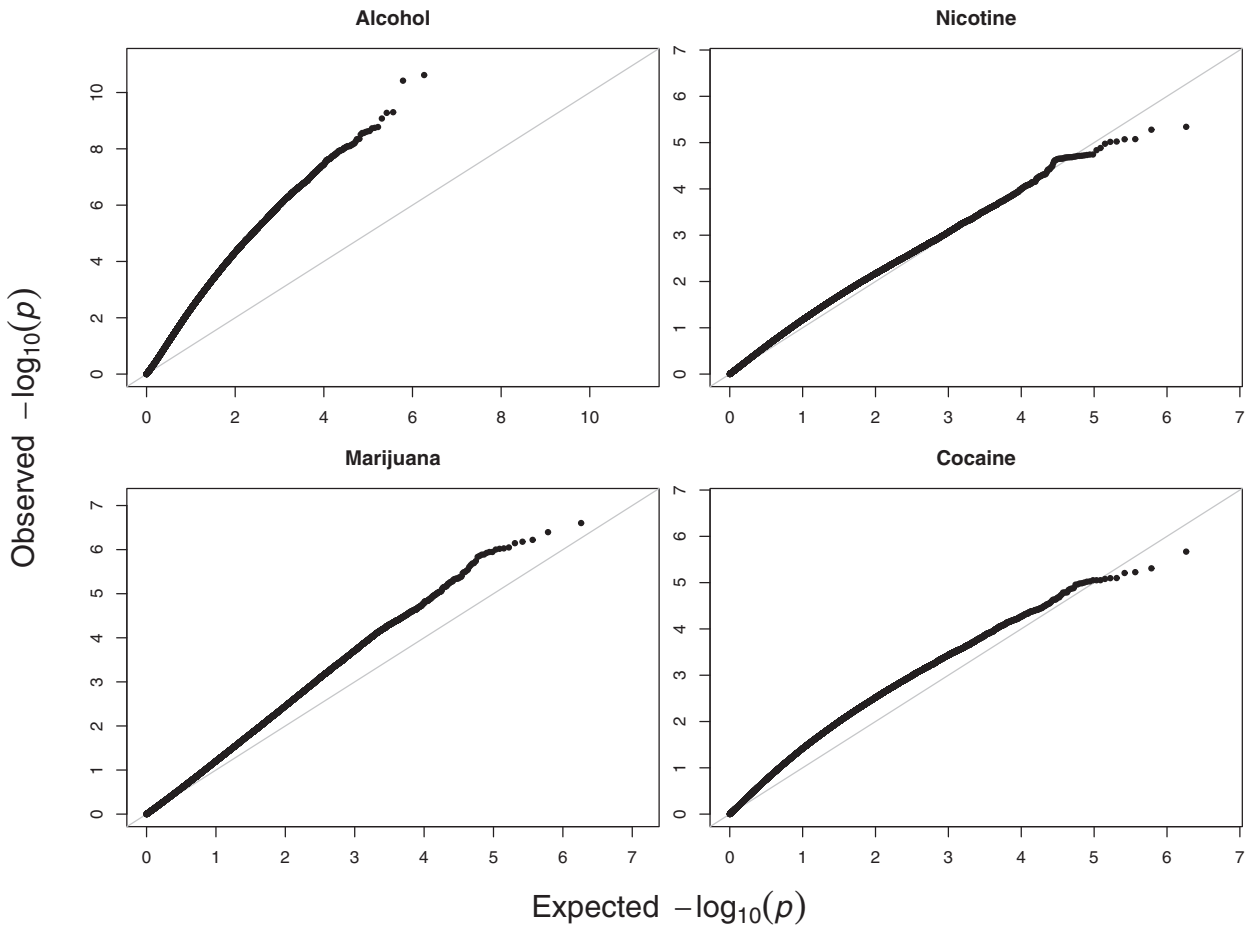

Fig. 3 The QQ-plots of *p*-values for the marginal tests of association between the SNPs and each of the four addiction symptomatology variables. The *x*-axis is the expected $-\log_{10}(p\text{-value})$, and the *y*-axis is the observed $-\log_{10}(p\text{-value})$. The diagonal gray straight lines have the slope 1 and intercept 0

and this indicates that there is no SNP associated with nicotine symptoms; (2) For the symptoms of the other three substances, all the *p*-values deviate from the diagonal lines (except for 0 and 1), with the *p*-values of alcohol symptoms furthest away from the diagonal line and the *p*-values of the remaining two symptoms closer to the diagonal line. Since the four symptomatology variables are moderately correlated and they all measure the common construct of addiction, we used them as the multivariate phenotype and applied the proposed Fisher combination approach (FC-Kendall) to identify the SNPs associated with it. The QQ-plot of the *p*-values for this multivariate analysis is shown in Fig. 4 indicating that some SNPs are associated with addiction across substances.

To identify the SNPs associated with the phenotypes, we adopted the commonly used significance level for GWAS of 10^{-6} to account for multiplicity. Based on the results of the marginal tests, the numbers of SNPs identified to be associated with the individual phenotypes are 917 for alcohol dependence symptoms, 0 for nicotine symptoms, 9 for marijuana symptoms, and 0 for cocaine symptoms. Using the proposed Fisher combination method (FC-Kendall), on the other hand, we identified 6 SNPs associated with the multivariate phenotype of addiction. Among them, 5 SNPs were also identified by the marginal test for alcohol symptoms. This implies that if we ignore the correlations among the 4 phenotypic traits and conduct the marginal tests, we would identify many SNPs that may be specific to alcohol dependence. Thus, if our goal is to identify the genes associated with the construct of addiction that contributes to the 4 types of substance dependence symptomatology, the proposed method is a better approach.

Discussion

In GWAS for complex diseases, the association between a SNP and each phenotype is usually weak. Combining multiple related phenotypic traits can increase the power

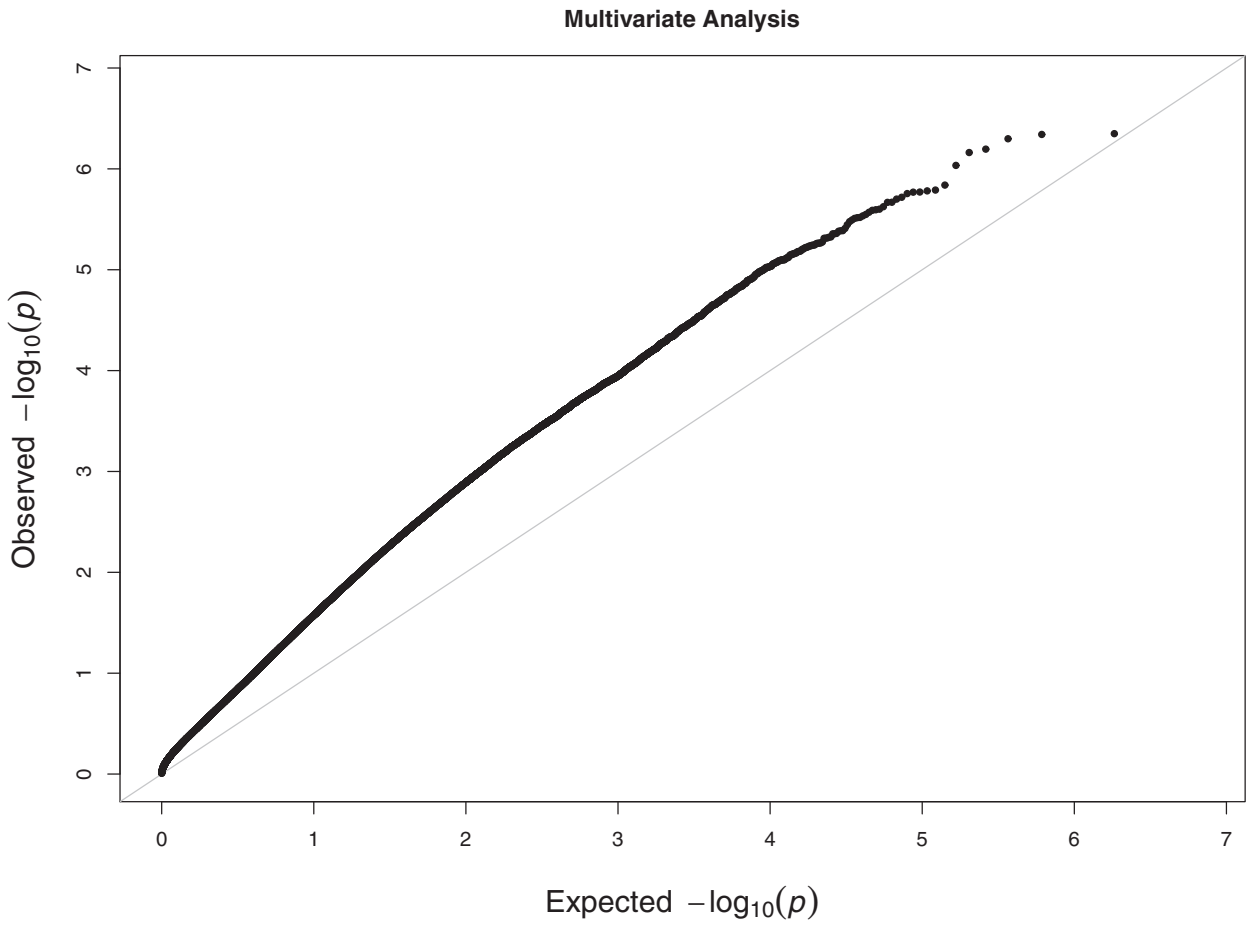

Fig. 4 The QQ-plot of *p*-values for the Fisher combination test of association between the SNPs and the multivariate phenotype of addiction. The *x*-axis is the expected $-\log_{10}$(*p*-value), and the *y*-axis is the observed $-\log_{10}$(*p*-value). The diagonal gray straight lines have the slope 1 and intercept 0

of gene search and thus is a practically important area that requires methodology work. This study provides a comprehensive review of existing methods for conducting GWAS on complex diseases with multiple phenotypes including MANOVA, PCA, GEE, TATES, and the classical Fisher combination test. Built upon the Fisher combination test, we proposed a new method that relaxes the unrealistic independence assumption and is also computationally efficient. Particularly, in an exploratory study where multiple sets of phenotypes may be of interest, when the set is changed, our proposed methods only require recalculation of the correlation between phenotypes and then the available marginal p-values for each SNP can be re-used. The competing methods which do not involve marginal *p*-values such as the PCA, MANOVA, and GEE, on the other hand, would require a complete re-analysis.

We conducted a simulation study to compare the performance of the competing methods. The GEE and the Fisher combination test with the independence assumption did not control the type I error rate and thus are not recommended. In general, the power of the methods decreased as the correlation between phenotypes increased. Furthermore, all the competing methods tended to have lower power when the multivariate phenotypes come from long tailed distributions. The proposed method (with the correlation being estimated by the Pearson's sample correlation coefficient or the Kendall's τ) performed as well as the permutation method and yet only required 10^{-2} computational time. In most settings of the simulation, these three Fisher combination tests outperformed the other methods. The real data analysis also demonstrated that the Fisher combination tests allow us to compare the marginal results with the multivariate results and specify which SNPs are specific to a particular phenotype or contribute to the common construct.

In our simulation study, we only considered continuous multivariate phenotypes. Future studies may extend the methodology work to the case of correlated discrete phenotypes. For example, in the substance abuse field, many outcomes are zero-inflated count data [35] or ordinal data

[36]. A future direction that is particularly challenging is how to analyze multivariate phenotypes with different measurement scales.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
JJY and JL conducted the literature review and simulation study; LWK and AB provided funding and obtained the approval of using the data; JJY, LKW, and AB performed real data analysis and interpreted the results; JJY and AB wrote the manuscript with input from the other authors. All authors read and approved the final manuscript.

Acknowledgement
This research was supported by National Institutes of Health (NIH) grants, R01 AI079139 & R01 HL118267 to L. K. Williams, and K01 AA016591 & R01 DA035183 to A. Buu. The funding agencies had no role in study design, analysis, interpretation of results, decision to publish, or preparation of the manuscript. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. We also acknowledge the Study of Addiction: Genetics and Environment (SAGE), which is part of the Gene Environment Association Studies (GENEVA) initiative supported by the National Human Genome Research Institute (dbGaP study accession phs000092.v1.p1).

Author details
[1]School of Nursing, University of Michigan, Ann Arbor, Michigan. [2]Department of Public Health Sciences, Henry Ford Health System, Detroit, Michigan. [3]Center for Health Policy and Health Services Research, Henry Ford Health System, Detroit, Michigan. [4]Department of Systems, Populations and Leadership, University of Michigan, Ann Arbor, Michigan.

Received: 28 October 2015 Accepted: 22 December 2015
Published online: 05 January 2016

References

1. The International HapMap Consortium. The international hapmap project. Nature. 2003;426(6968):789–96.
2. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061–73.
3. McGue M, Zhang YW, Miller MB, Basu S, Vrieze S, Hicks B, et al. A genome-wide association study of behavioral disinhibition. Behav Genet. 2013;43(5):363–73.
4. Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. Pleiotropy in complex traits: challenges and strategies. Nat Rev Genet. 2013;14(7):483–95.
5. Smith H, Gnanadesikan R, Hughes JB. Multivariate analysis of variance (manova). Biometrics. 1962;18(1):22–41.
6. Morrison DF. Multivariate Statistical Methods. McGraw-Hill series in probability and statistics. New York: McGraw-Hill; 1967.
7. Brown CC, Havener TM, Medina MW, Krauss RM, McLeod HL, Motsinger-Reif AA. Multivariate methods and software for association mapping in dose-response genome-wide association studies. Biodata Mining. 2012;5:21.
8. Liu F, van der Lijn F, Schurmann C, Zhu G, Chakravarty MM, Hysi PG, et al. A genome-wide association study identifies five loci influencing facial morphology in europeans. Plos Genetics. 2012;8(9):1002932.
9. Hotelling H. Analysis of a complex of statistical variables into principal components. J Educ Psychol. 1933;24:417–41.
10. Jolliffe IT. Principal Component Analysis, 2nd edn. New York: Springer; 2002.
11. Brown CC, Havener TM, Medina MW, Krauss RM, McLeod HL, Motsinger-Reif AA. Principal components regression: Multivariate, genebased tests in imaging genomics. Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on. 2011;289–293.
12. Wang K, Abbott D. A principal components regression approach to multilocus genetic association studies. Genet Epidemiol. 2008;32(2):108–18.
13. He LN, Liu YJ, Xiao P, Zhang L, Guo Y, Yang TL, et al. Genomewide linkage scan for combined obesity phenotypes using principal component analysis. Ann Hum Genet. 2008;72:319–26.
14. Zeger SL, Zeger SL, Zhang F, Guo X, Wu S, Han J, et al. Genome-Wide Pathway Association Studies of Multiple Correlated Quantitative Phenotypes Using Principle Component Analyses. Plos One. 2012;7(12):e53320.
15. Zeger SL, Karasik D, Cheung CL, Zhou Y, Cupples LA, Kiel DP, et al. J Bone Miner Res. 2012;27(2):319–30.
16. Zeger SL, Liang KY, Albert PS. Models for longitudinal data - a generalized estimating equation approach. Biometrics. 1988;44(4):1049–60.
17. Liu J, Pei Y, Papasian CJ, Deng H. Bivariate Association Analyses for the Mixture of Continuous and Binary Traits With the Use of Extended Generalized Estimating Equations. Genet Epidemiol. 2009;33(3):217–27.
18. van der Sluis S, Posthuma D, Dolan CV. Tates: Efficient multivariate genotype-phenotype analysis for genome-wide association studies. PLoS Genet. 2013;9(1):1003235.
19. Simes RJ. An improved bonferroni procedure for multiple tests of significance. Biometrika. 1986;73(3):751–4.
20. Galwey NW. A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests. Genet Epidemiol. 2009;33(7):559–68.
21. Li MX, Yeung JMY, Cherny SS, Sham PC. Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. Hum Genet. 2012;131(5):747–56.
22. Fisher RA. Statistical Methods for Research Workers. Edinburgh: Oliver & Boyd; 1932.
23. Birnbaum A. Combining independent tests of significance. J Am Stat Assoc. 1954;49(267):559–74.
24. Littell RC, Folks JL. Asymptotic optimality of fishers method of combining independent tests. J Am Stat Assoc. 1971;66(336):802.
25. Littell RC, Folks JL. Asymptotic optimality of fishers method of combining independent tests ii. J Am Stat Assoc. 1973;68(341):193–4.
26. Good P. Permutation, Parametric and Bootstrap Tests of Hypotheses. 3rd edn, Springer Series in Statistics. New York: Springer; 2005.
27. Brown MB. Method for combining non-independent, one-sided tests of significance. Biometrics. 1975;31(4):987–92.
28. Yang JJ. Distribution of fisher's combination statistic when the tests are dependent. J Stat Comput Simul. 2010;80(1–2):1–12.
29. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2014. R Foundation for Statistical Computing. http://www.R-project.org.
30. Olkin I, Pratt JW. Unbiased estimation of certain correlation-coefficients. Ann Math Stat. 1958;29(1):201–11.
31. Kendall M, Gibbons JD. Rank Correlation Methods, 5th edn. London: Oxford; 1990.
32. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, et al. The ncbi dbgap database of genotypes and phenotypes. Nat Genet. 2007;39(10):1181–6.
33. Bierut LJ, Agrawal A, Bucholz KK, Doheny KF, Laurie C, Pugh E, et al. A genome-wide association study of alcohol dependence. Proc Natl Acad Sci U S A. 2010;107(11):5082–7.
34. Mullahy J. Specification and Testing of Some Modified Count Data Models. J Econ. 1986;33(3):341–65.
35. Buu A, Li RZ, Tan XM, Zucker RA. Statistical models for longitudinal zero-inflated count data with applications to the substance abuse field. Stat Med. 2012;31(29):4074–86.
36. Hedeker D, Mermelstein RJ. Analysis of longitudinal substance use outcomes using ordinal random-effects regression models. Addiction. 2000;95:381–94.