Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2012, Article ID 652569, 13 pages
doi:10.1155/2012/652569

Review Article

Methods for Analyzing Multivariate Phenotypes in Genetic Association Studies

Qiong Yang[1] and Yuanjia Wang[2]

[1] *Department of Biostatistics, Boston University School of Public Health, 810 Mass Avenue, Boston, MA 02118, USA*
[2] *Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10027, USA*

Correspondence should be addressed to Qiong Yang, qyang@bu.edu

Received 30 March 2012; Accepted 21 May 2012

Academic Editor: Yongzhao Shao

Multivariate phenotypes are frequently encountered in genetic association studies. The purpose of analyzing multivariate phenotypes usually includes discovery of novel genetic variants of pleiotropy effects, that is, affecting multiple phenotypes, and the ultimate goal of uncovering the underlying genetic mechanism. In recent years, there have been new method development and application of existing statistical methods to such phenotypes. In this paper, we provide a review of the available methods for analyzing association between a single marker and a multivariate phenotype consisting of the same type of components (e.g., all continuous or all categorical) or different types of components (e.g., some are continuous and others are categorical). We also reviewed causal inference methods designed to test whether the detected association with the multivariate phenotype is truly pleiotropy or the genetic marker exerts its effects on some phenotypes through affecting the others.

1. Introduction

Association studies, where the correlation between a genetic marker and a phenotype is assessed, are useful for mapping genes influencing complex diseases. With reduction of genotyping cost, completion of the HapMap Project [1], and more recently the 1000 Genomes Project [2], genome-wide association studies (GWAS) with several hundred thousands to tens of millions genotyped and/or imputed single nucleotide polymorphisms (SNPs) have become a common approach nowadays to search for genetic determination of complex traits.

In the study of complex diseases, several correlated phenotypes, or a multivariate (MV) phenotype with several components, may be measured to study a disorder or trait. For example, hypertension is evaluated using systolic and diastolic blood pressures;

a person's cognitive ability is usually measured by tests in domains including memory, intelligence, language, executive function, and visual-spatial function. The tests within and between domains are correlated. Most published GWAS only analyzed each individual phenotype separately, although results on related phenotypes may be reported together. Published single phenotype GWAS have successfully identified a large number of novel genetic variants predisposing to a variety of complex traits [3, 4]. However, majority of the identified genetic variants only explain a small fraction of total heritability defined as between individual phenotype variability attributable to genetic factors [4, 5]. It has been hypothesized that current GWAS may be underpowered to detect many genetic variants of moderate-to-small effects. Joint analysis of correlated phenotypes can exploit the correlation among the phenotypes, which may lead to better power to detect additional genetic variants with small effects across multiple traits or pleiotropy effects. Furthermore, joint analysis avoids multiple testing penalty incurred in analyzing each phenotype separately. Therefore, it is important to identify appropriate methods that fully utilize information in multivariate phenotypes to detect novel genetic loci in genetic association studies.

In addition to discovery of novel loci of potential pleiotropy effects, it is also important to detangle the complex relationship between phenotype components and genetic variants One of the frequently asked questions is whether a genetic variant affects multiple phenotypes simultaneously (pleiotropy) or affects one phenotype through affecting another phenotype. In this paper, we review methods for both purposes.

2. Methods for Detecting Association Using Multivariate Phenotypes

For all the methods mentioned in this section, the null hypothesis is no association between a single genetic marker and any components of a multivariate (MV) phenotype; the alternative hypothesis is the genetic maker associated with at least one phenotype component. Here we review methods for an MV phenotype consisting of all continuous, all categorical, or all time-to-event components, and methods for MV phenotypes consisting of a mixture of different types of components.

2.1. Regression Models

Regression models for clustered observations such as linear and generalized mixed effects models, generalized estimating equations, and frailty models can be used to analyze the association of a genetic marker with all continuous, categorical, or survival multivariate phenotypes.

2.1.1. Mixed Effects Models

Mixed effects models such as linear mixed effects model (LME) and generalized linear mixed effects model (GLMM) involve using fixed effects for the genetic marker effect and random effects to account for correlation among multivariate phenotypes [6, 7].

Let y_{jk} denote the kth ($k = 1, \ldots, K$) continuous component of the K-dimensional phenotype of the jth ($j = 1, \ldots, J$) individual. Let g_j be the genotype of a genetic marker of

the jth individual, and $X(g_j)$ a score of the genotype. The linear mixed effects model takes the following form:

$$y_{jk} = \beta_0 + \beta_k X(g_j) + \eta_{jk} + e_{jk}, \tag{2.1}$$

where β_0 is the intercept or other genetic or environmental fixed effects; β_k is the fixed effect size of $X(g_j)$ on the kth phenotype; $\eta_{jk}(k = 1,\ldots,K) \sim N(0,\Sigma)$ are the random effects correlated within jth person; e_{jk} is the random errors iid. $\sim N(0,\sigma_e^2)$. Between any two individuals, $\eta_{jk}, k = 1,\ldots,K$ are independent. Within a person, $\eta_{jk}, k = 1,\ldots,K$ are correlated. The null hypothesis that the genetic marker is not associated with any phenotype component corresponds to $H_0 : \beta_1 =,\ldots,\beta_K = 0$. The estimation of variance parameters and fixed effect parameters can be obtained using restricted maximum likelihood method (REML) [8, 9].

When y_{jk} is categorical, it can be modeled with generalized mixed effects model (GLMM) as follows:

$$E(y_{jk} \mid \eta_k) = \mu^{-1}(\beta_0 + \beta_k X(g_j) + \eta_{jk}), \tag{2.2}$$

where μ is a link function and μ^{-1} is its inverse. For Gaussian distributed traits, μ is the identity link, thus (2.2) is identical to the linear mixed effects model (2.1); for binary traits, μ is the logit link $\mu(x) = \ln(x/1-x)$. For links other than identity function, the likelihood for this model contains integrals without a close form solution. All existing algorithms for likelihood maximization are either based on theoretical or numerical approximation [10, 11].

The null hypothesis under the LME or GLMM can be tested using the likelihood ratio test or Wald chi-squared test. They can be implemented using SAS PROC Mixed or *R* lme4 package function *lmer()*. The Wald chi-squared test statistic takes the form $\boldsymbol{\beta}^T \mathrm{cov}(\boldsymbol{\beta})^{-1}\boldsymbol{\beta} \sim \chi_K^2$, where $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_K)$ is estimated using (2.1) or (2.2). For example, Kraja et al. [12] have employed a model similar to (2.1) to the analyses of bivariate continuous metabolic traits. We can also fit a model assuming $\beta_1 = \cdots = \beta_K = \beta$, that is, $E(y_{jk} \mid \eta_k) = \mu^{-1}(\beta_0 + \beta X(g_j) + \eta_{jk})$, where a single degree-of-freedom (df) test $\widehat{\beta}/\mathrm{se}(\widehat{\beta})$ can be used to test the null hypothesis. This test can be more powerful than the multi-df Wald chi-squared test if the effect sizes are in the same direction and not very different. It, however, may lack power if the $\beta_1,\ldots,\beta_K$ are very different, especially have different signs and cancel each other out.

2.1.2. Frailty Models

When the phenotypes are correlated survival times, frailty models can be used to fit the association model. Suppose the survival or censoring times are t_{kj} for the kth $(k = 1,\ldots,K)$ phenotype of the jth $(j = 1,\ldots,J)$ individual. Let g_j be the genotype of a genetic marker of the jth individual, and $X(g_j)$ a score of the genotype as follows:

$$h(t_{kj}; X(g_j)) = h_0(t_{kj}) \exp(\beta_0 + \beta X(g_j) + \eta_{kj}), \tag{2.3}$$

where $\eta_{kj}(j = 1,\ldots,J)$ are subject specific random effects following $N(0,\Sigma)$, and Σ is a K-dimensional correlation matrix. This is the Gaussian frailty model. There is another class of frailty models where $\exp(\eta_{kj})$ follows a gamma distribution. A Gaussian or gamma frailty

model assuming an exchangeable correlation within a person can be fitted using *coxph()* in the survival package of *R* by including a *frailty()* term in the regressor. In addition, including a *cluster()* term in *coxph()* fits generalized estimating equations (GEE) type of model that assumes an independent working correlation matrix [13]. Frailty models with an arbitrary prespecified Σ can be fitted with the *coxme()* in *R* coxme package for Gaussian random effects model.

Fitting a mixed effects (frailty) model requires predetermining the correlation matrix Σ of random effects η_{jk} within jth person. The correlation between the phenotypes y_{jk} within a person is attributable to the random effects η_{jk} and the fixed effects of the genetic marker. However, since the fixed effects are unknown, it is impossible to directly infer the correlation among the random effects. Misspecifying the correlation among random effects may result in bias in the inference on fixed effects. But the bias seems to be small for genetic association studies [14, 15].

2.1.3. Generalized Estimating Equations

Different from mixed effects model is a class of models called marginal models. Instead of having random effects as regressors in addition to random errors to model correlation in multivariable response, marginal models collapse the random effects and random residual errors in the model. Generalized estimating equations (GEE) [16] solve the quasi-likelihood score function as follows:

$$\sum_{j=1}^{n} \left(\frac{\partial \mu_j}{\partial \boldsymbol{\beta}} \right)^t V_j^{-1} (Y_j - \mu_j) = 0, \tag{2.4}$$

where $V_j = A_j^{1/2} R(\alpha) A_j^{1/2}$, and $R(\alpha)$ is the working correlation matrix for the residual correlation. The variance and covariance of $\boldsymbol{\beta}$ is estimated with the so-called robust variance estimator [16]. Similar to the LME, single- or multi-df Wald test statistic can be usually used to test that the genetic marker is not associated with any of the phenotypes.

In our experience, GEE results are inflated with low minor frequency SNPs and not as powerful as LME in general [15, 17]. However, GEE is robust to misspecification of response distribution or association model and thus can be used when the LME shows bias or inflation due to these reasons.

2.2. Variable Reduction Method

Variable reduction approaches are in general only applicable to MV phenotype consisting of all continuous phenotypes that are approximately normal distributed. It derives a single or a few new phenotypes that are linear combinations of the original phenotypes, for example,

$$\widetilde{Y} = a_1 Y_1 + a_2 Y_2 + \cdots + a_K Y_K. \tag{2.5}$$

Existing methods include principal components analysis (PCA) where for the first component, $a_i, i = 1, \ldots, K$ are coefficients that maximize the variance of $\widetilde{Y}$; principal component of heritability (PCH) with coefficients maximizing the total heritability of $\widetilde{Y}$ [18]

and penalized PCH applicable to high-dimensional data [19, 20]; and principal components of heritability with coefficients maximizing the quantitative trait locus (QTL) heritability (PCQH) of $\widetilde{Y}$ [21–24], that is, the variance explained by the genetic marker. The PCQH approaches are designed to maximize the individual phenotype variation explained by the genetic marker and thus may be more powerful than PCA and PCH in genetic association studies.

2.2.1. PCQH Approaches

The approaches proposed by Lange et al. [21, 25] and Klei et al. [23] involve using a subset of the sample to estimate the coefficients in (2.5) that maximize the correlation between $\widetilde{Y}$ and the genetic marker. Specifically, in the estimation sample, the total phenotype variance is partitioned into QTL variance and residual variance as follows:

$$V_p = V_q + V_\varepsilon, \tag{2.6}$$

where V_p is the $K \times K$ total phenotype variance-covariance matrix, V_q the QTL variance matrix, and V_ε the residual variance matrix. Let $A = (\alpha_1, \ldots, \alpha_K)$, then the variance of $\widetilde{Y} = A^t Y$ explained by the genetic marker is

$$h_A^2 = \frac{A^t V_q A}{A^t V_p A}. \tag{2.7}$$

A that maximizes h_A^2 can be obtained by solving the following generalized eigen system [18]:

$$V_q A = \lambda V_p A. \tag{2.8}$$

$V_q = \mathrm{var}(\beta_1 X, \ldots, \beta_K X)$ can be approximated by $\Gamma 1 1^t \Gamma$, where $\Gamma = \mathrm{diag}(|\beta_1|\sigma_x, \ldots, |\beta_K|\sigma_x)$, σ_x is the sample standard deviation of the score of genotype $X(g)$ across all individuals, β_i is estimated using the least squared estimator of $Y_i = \alpha + \beta_i X(g) + \varepsilon$, and $1 = (\mathrm{sign}(\beta_1), \ldots, \mathrm{sign}(\beta_K))$.

Lange et al. [21, 25] approaches are only applicable to family-based association design. They suggest using the noninformative families or parental genotypes to estimate A because these data will not contribute directly to the family-based association tests (FBAT). Then perform FBAT of $\widetilde{Y}$ on $X(g)$. However, FBAT has low power in the absence of population stratification [26] compared to population based approaches. Klei et al's. [23] is a population-based association approach where they randomly split the sample into two subsets: one used to estimate A, the other used to test the association of $\widetilde{Y}$ with $X(g)$ via a linear regression model: $\widetilde{Y} = \alpha + \beta X(g) + \varepsilon$. This ensures valid P value in the association test.

2.2.2. Canonical Correlation Analysis

Canonical correlation analysis seeks coefficients so that the squared correlation between $\widetilde{Y}$ in (2.5), and the score of genetic marker, $X(g)$, is maximized. Here $\widehat{\rho} = \mathrm{corr}(\widetilde{Y}, X)$ is called

Table 1: Relationship between MANOVA test statistics and canonical correlation for association test of multivariate phenotype and a genetic marker.

MANOVA test	$f(\hat{\rho})$
Roy's largest root	$\hat{\rho}^2$
Hotelling-Lawley trace	$\hat{\rho}^2/1-\hat{\rho}^2$
Wilks lambda	$1-\hat{\rho}^2$
Pillai-Bartlett trace	$\hat{\rho}^2$

estimated canonical correlation. To obtain $\hat{\rho}$, the covariance matrix of Y and X is partitioned as follows:

$$\operatorname{cov}\begin{bmatrix} Y \\ X \end{bmatrix} = \begin{bmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{bmatrix}, \tag{2.9}$$

where Σ_{YY} is the $K \times K$ matrix of the variance-covariance matrix of Y, Σ_{YX} and its transpose Σ_{XY} are $K \times 1$ and $1 \times K$matrix of the covariance matrix between Y and X, Σ_{XX} is the variance of X, a scalar. All these submatrices can be estimated using the respective sample co-variance matrix. The canonical correlation, $\hat{\rho} = \Sigma_{XY}A/\left(A^t\Sigma_{YY}A\Sigma_{XX}\right)^{1/2}$, is solved as the squared root of the largest eigenvalue of $\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$, and the corresponding eigenvector A contains the coefficients for constructing $\widetilde{Y}$. multivariate analysis of variance (MANOVA) tests correspond to evaluating canonical correlation. Table 1 details the relationship between $\hat{\rho}$ and commonly reported test statistics in MANOVA of a multivariate phenotype Y on $X(g)$ [27].

These tests are implemented in SAS PROC GLM and *R* function *summary.manova()*. As part of the PLINK package specifically developed for genetic analysis, Ferreira et al. [24] implemented the Wilks lambda, and its P value is obtained from F-approximation $F = (\hat{\rho}^2/K)/((1-\hat{\rho}^2)/(n-K-1))$.

Canonical correlation analysis shares similarity with PCQH [23] in that both estimate a linear combination of original phenotypes, so that the genotype score explains most of the variation (in terms of percent of total variance and squared correlation, resp.) of the new phenotype. The difference between the two approaches is that the canonical correlation analysis evaluates squared correlation using whole sample, while PCQH estimates the loadings using a subset of the sample and test the association in the rest of the sample. Extensive simulation studies performed in [28]. The author of [28] showed that MANOVA via Wilk's lambda was substantially more powerful than PCQH [23] with $K = 5$ phenotypes.

2.3. Combining Test Statistics from Univariate Analysis

An alternative way to analyze multivariate phenotypes is to perform univariate phenotype-genotype association test for each phenotype individually and then combine the test statistics from the univariate analysis. The advantage of such approach is the simplicity, that is, the methods to deal with univariate phenotypes are generally simpler than methods for MV phenotypes. It is especially useful for analyzing multivariate phenotype consisting of components of different types of distributions such as continuous, dichotomous, and survival. Regression methods for analyzing such multivariate phenotype are generally

complicated and not trivial to implement for MV phenotype with dimension > 2, see for example, [29, 30].

In recent years, researchers have generated large amount of univariate GWAS results for a variety of complex traits. Methods that combine the univariate results of multiple traits to detect genetic markers associated with multiple phenotypes are appealing.

2.3.1. Methods for Homogeneous Genetic Effects across Phenotypes

Assume that $\mathbf{T} = [T_1, T_2, \ldots, T_K]^T$ is a vector of K test statistics obtained from association analyses of each individual component phenotype against the genetic marker. Assume that $\mathbf{T}$ follows a multivariate normal distribution with mean $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_K)^T$ and a nonsingular covariance matrix $\boldsymbol{\Sigma}$. For example, $\mathbf{T}$ can be the β coefficients from least squared regression model for individual components or the t-test statistics from the regression models. The null hypothesis of no association to any phenotypes is H_0: $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_K)^T = \mathbf{0}$. O'Brien [31–33] suggested the following linear combination of $T_1, T_2, \ldots, T_K$, with weight $\mathbf{e} = (1, 1, \ldots, 1)^T$ of length K:

$$S = \mathbf{e}^T \Sigma^{-1} \mathbf{T} \tag{2.10}$$

when $\tau_1 = \tau_2 = \cdots = \tau_K \neq 0$ (2.10) is the most powerful test among a class of tests statistics that are linear combinations of $T_1, T_2, \ldots, T_K$. Under the null hypothesis, S follows the normal distribution with mean 0 and variance $\mathbf{e}^T \Sigma^{-1} \mathbf{e}$. To estimate $\boldsymbol{\Sigma}$ with GWAS results, Yang et al. [34] suggested using the sample covariance matrix of the statistics on a large number of SNPs genomewide with little or no linkage disequilibrium among them (say HapMap $r^2 < 0.1$).

The power of O'Brien's method depends on the assumption $\tau_1 = \tau_2 = \cdots = \tau_K$. When the means are very different or with opposite signs, O'Brien's method may not be efficient. Yang et al. proposed a sample splitting approach that replaces the uniform weight $\mathbf{e}^T$ by weights $\mathbf{w}$ estimated using a portion of the sample and only used the remaining sample to estimate $\mathbf{T}$ in (2.10), that is, $S = \mathbf{w}^T \Sigma^{-1} \mathbf{T}$. To overcome the variability introduced by a random sample splitting, Yang et al. also evaluated a cross-validation approach that averages the test statistics of 10 random splitting samples. The results showed that when $\tau_1, \tau_2, \ldots, \tau_K$ are of different magnitude or in opposite directions, O'Brien's method is less powerful than Yang et al., which indicates room for improvements for O'Brien's method. However, the sample splitting and cross-validation methods are less powerful than O'Brien's method with homogeneous effect sizes.

2.3.2. Methods for Heterogeneous Genetic Effects across Phenotypes

The limitation of O'Brien statistic is that it is not powerful for heterogeneous effects across multiple phenotypes, especially if some effects are of opposite directions. Another class of statistics that takes a quadratic form of the vector of the individual association statistic may overcome the limitation. For example, the following Wald chi-squared type test statistic was mentioned in Xu et al. [32].

$$S_w = \mathbf{T}^T \Sigma^{-1} \mathbf{T}. \tag{2.11}$$

The difference between (2.10) and (2.11) is that the vector $\mathbf{e} = (1, 1, \ldots, 1)^T$ is replaced by the $\mathbf{T}$ in (2.11). S follows a chi-squared distribution with degree of freedom equal to the number of the phenotypes K or rank of Σ if it is not full rank. Due to the "curse of dimensionality," power of (2.11) is diminishing with the increased number of phenotypes. Similar problem has been extensively studied and discussed in high-dimensional data analysis field and most recently in the analyses of multiple rare variants. Borrowing ideas from these fields, we propose the following test statistic that may be more powerful than (2.10) and (2.11) with heterogeneous effects.

$$S_{\text{sq}} = \mathbf{T}^T\mathbf{T} = \sum_{i=1}^{K} t_i^2. \tag{2.12}$$

The difference between (2.12) and (2.11) is that there is no variance-covariance matrix in (2.12). This statistic was first proposed by Pan [35] to analyze multiple rare or common variants against a single phenotype, where the t_i is the beta coefficient for the ith genetic variant. Different from Pan [35], here t_i is the association statistic for the ith phenotype with a single marker. Based on the groundwork of Zhang [36], Pan [35] pointed out that the distribution of (2.11) is a mixture of single degree-of-freedom chi-squared variates, $\sum_{i=1}^{K} c_i\chi_1^2$ where c_is are the eigen values of Σ, that is, the variance-covariate matrix of t_i. The distribution of (2.12) can be well approximated by $a\chi_d^2 + b$ with

$$a = \frac{\sum_{i=1}^{K} c_i^3}{\sum_{i=1}^{K} c_i^2}, \qquad b = \sum_{i=1}^{K} c_i - \frac{\left(\sum_{i=1}^{K} c_i^2\right)^2}{\sum_{i=1}^{K} c_i^3}, \qquad d = \frac{\left(\sum_{i=1}^{K} c_i^2\right)^3}{\left(\sum_{i=1}^{K} c_i^3\right)^2}. \tag{2.13}$$

The P value is calculated as $p(\chi_d^2 > (S_{\text{Sq}} - b)/a)$. The degree of freedom of the S_{sq} may be less than K with highly correlated phenotypes. In addition, (2.12) does not have the problem of instability observed for (2.10) and (2.11) when some of the components are highly correlated (in one of our applications, a correlation ~0.7 has resulted in inflated results for (2.10) and (2.11)). We have developed an R package CUMP (combining univariate results of multivariate phenotypes) that have implemented all the aforementioned combining statistics approaches. The software can be downloaded at (http://people.bu.edu/qyang/), and a short report of this software is submitted [37].

3. Identifying Pleiotropy

All the aforementioned methods can be used to detect association that is potentially due to pleiotropy. But they do not answer the question if the detected association is truly pleiotropy, that is, the marker locus affects all components of the MV phenotype directly. The detect association can affect some of the phenotypes and/or mediate through these phenotypes to affect the other phenotypes. Vansteelandt et al. [38] illustrated potential confounding mechanism between the genotype of a genetic marker and a phenotype using a causal diagram (Figure 1): the association between the genotype, denoted as G, and the response phenotype Y can occur through the paths connecting the two variables along all unbroken sequences of edges regardless of the direction of the arrows, given that there are no colliders

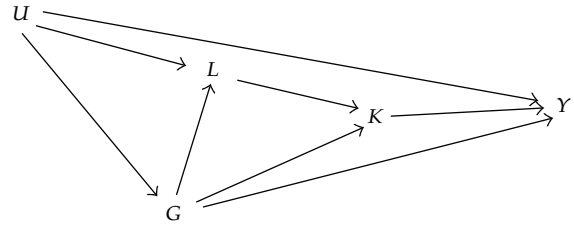

Figure 1: Causal diagram showing potential confounding mechanisms for the association between the genotype of a genetic marker G and the phenotype Y. The variable K denotes the intermediate phenotype, L the collection of known environmental and genetic risk factors, and U the unknown environmental and genetic risk factors such as population stratification and unknown genetic variants in linkage disequilibrium with G.

(i.e., variables in which two arrows converge, e.g., variables K and L in Figure 1) in the sequence [39].

The genotype G may be associated with Y due to (1) direct causal effect, that is, $G \rightarrow Y$; (2) through intermediate phenotype or risk factors, that is, $G \rightarrow K \rightarrow Y$or $G \rightarrow L \rightarrow K \rightarrow Y$; (3) because of confounding factors, that is, $G \leftarrow U \rightarrow Y$ or $G \leftarrow U \rightarrow L \rightarrow K \rightarrow Y$.

The authors showed that two commonly used approaches to detangle the complex relationship between phenotypes, genotype, and traditional risk factors are flawed. The first commonly used approach derives the residuals of Y regressing on K, say $\widetilde{Y} = Y - \beta K$, and then the association between G and $\widetilde{Y}$ is tested. The disadvantage of this approach is that not only the direct causal effect of K on Y is removed but also any indirect effect of K on Y through G (e.g., $K \leftarrow G \rightarrow Y$ and $Y \leftarrow U \leftarrow L \leftarrow G \rightarrow K$) and other factors (e.g., $K \leftarrow L \rightarrow U \rightarrow Y$). Therefore β may be biased in the presence of confounding factors which leads to biased test of G with $\widetilde{Y}$.

The second commonly used approach tests the direct effect of G on Y in a regression model including K and L as covariates. Adjustment of K removes the relationship between G and Y through $G \rightarrow K \rightarrow Y$; however, because K is a collider (Figure 1), the adjustment of K induces a spurious association [39, 40] along the path $G \rightarrow K \rightarrow L \leftarrow U \rightarrow Y$. Additionally, adjusting for L induces spurious associated through the path $G \rightarrow L \leftarrow U \rightarrow Y$.

To overcome the limitation of the two commonly adapted approaches, Vansteelandt et al. [38] proposed a least squared regression model to estimate the direct effect size of K on Y. This regression model includes the suspected intermediate phenotype, the score of the genetic marker genotype, $X(G)$, and other common risk factors between the two phenotypes as regressors:

$$E(Y_i) = \gamma_0 + \gamma_1 K_i + \gamma_2 X_i + \gamma_3 L_i. \tag{3.1}$$

The estimated effect size of the phenotype represents the direct effect of the K on Y, that is, not confounded by the effect of X mediated through any of the covariates. Then, a new phenotype is created as the residual of the response subtract the effect of K only $\widetilde{Y}_i = Y_i - \overline{y} - \widehat{\gamma}_1(K_i - \overline{k})$. Then, whether the G only exerts its effect on Y through K can be tested using any standard association test statistic between the residual and the X. A negative result indicates that G only exerts its effect on Y through K while a positive result indicates that the G has a direct

effect on Y and/or a spurious effect through other confounders. Extensions of the method to dichotomous and time-to-event outcomes have been proposed [41, 42].

4. Discussion

In this paper, we reviewed methods available for joint analyzing correlated phenotypes in genetic association studies. Some of these methods are designed to detect potential association with multiple phenotypes (pleiotropy), while the others are designed to test whether the detected association with the MV phenotype is truly pleiotropy or the genetic marker exerts its effects on some phenotypes through affecting the others.

For methods designed to detect association, each method has its own pros and cons. Random effects model requires knowledge of residual correlation, and misspecifying the correlation may incur inflation or power loss. Generalized estimating equations are robust to misspecification of residual correlation, but it is inflated for low-frequency variants and less powerful than random effects model in our experience. Variable reduction approaches are appealing because correlated outcomes are reduced to a single or fewer number of uncorrelated outcomes. However, in the presence of missing data in the outcomes, individuals with missing data do not contribute to the analysis, which may result in power loss. The approaches combining univariate association results are more flexible than the other methods especially when MV phenotypes consist of a mixture of continuous, discrete, and/or time-to-events data. Regression approaches have been developed to deal with such phenotypes. But they are generally complicated and few available software implements these methods. Since univariate association results are used, individuals with incomplete observations still contribute to the analysis of available phenotypes. Simulations on all continuous phenotypes indicated that the power of O'Brien's method, one of the approaches combining univariate association results is similar to regression and variable reduction methods when the effects size are similar across multiple phenotypes [34].

All the approaches introduced here for population based approaches assume unrelated individuals. When there are related individuals in the data, not accounting for family structure can result in inflation or power loss. Extension of introduced methods to account for family data are possible. For example, one may add a random effect in mixed effects model to account for family structure. For approaches combining univariate association results, a model that account for family structure need be used in the univariate analyses.

In terms of computational cost, mixed effects models may be most time consuming since maximization of likelihood is required.

Finally, it has been shown that traditional causal inference is useful in distinguishing true pleiotropy from other mechanisms that also result in genetic association with multiple phenotypes. A related causal inference in recent genetic literature is Mendelian randomization test [43–45]. This approach can be used to infer whether an intermediate phenotype has a causal effect on an outcome phenotype, using genetic marker(s) in association with the intermediate phenotype. Unlike a phenotype that is subject to the influence of uncontrolled environmental factors and/or reverse causation of another phenotype, genotype(s) of genetic marker(s) is(are) free of influence of environmental factors and reverse causation. For this approach, marker genotype(s) is(are) used as an instrument variable. This test requires that there is no pleiotropy effect of the genetic marker on outcome phenotype. Association of the genotype and outcome phenotype indicates that the intermediate phenotype may causally affect the outcome phenotype.

5. URLs to Software Mentioned in This Paper

SAS: http://www.sas.com/,

R: http://www.r-project.org/,

CUMP: http://cran.r-project.org/web/packages/CUMP/index.html,

coxme: http://cran.r-project.org/web/packages/coxme/index.html,

gee: http://cran.r-project.org/web/packages/gee/index.html,

survival: http://cran.r-project.org/web/packages/survival/index.html,

lme4: http://cran.r-project.org/web/packages/lme4/index.html,

PLINK: http://pngu.mgh.harvard.edu/~purcell/plink/.

Acknowledgments

Q. Yang's work is supported by the National Heart, Lung, and Blood Institute's Framingham Heart Study (Contract no. N01-HC-25195) and Grant no. R01HL093328 and R01HL093029. Y. Wang's work is supported by NIH Grants nos. R03AG031113-01A2 and 1R01NS073671-01 1.

References

[1] International HapMap Consortium, K. A. Frazer, D. G. Ballinger et al., "A second generation human haplotype map of over 3.1 million SNPs," *Nature*, vol. 449, pp. 851–861, 2007.

[2] 1000 Genomes Project Consortium, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, pp. 1061–1073, 2010.

[3] L. A. Hindorff, H. A. Junkins, P. N. Hall, J. P. Mehta, and T. A. Manolio, "A catalog of published genome-wide association studies," *National Human Genome Research Institute*, 2011, http://www.genome.gov/gwastudies/.

[4] T. A. Manolio, F. S. Collins, N. J. Cox et al., "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, pp. 747–753, 2009.

[5] E. E. Eichler, J. Flint, G. Gibson et al., "Missing heritability and strategies for finding the underlying causes of complex disease," *Nature Reviews Genetics*, vol. 11, no. 6, pp. 446–450, 2010.

[6] N. M. Laird and J. H. Ware, "Random-effects models for longitudinal data," *Biometrics*, vol. 38, no. 4, pp. 963–974, 1982.

[7] G. M. Fitzmaurice and N. M. Laird, "A likelihood-based method for analysing longitudinal binary responses," *Biometrika*, vol. 80, no. 1, pp. 141–151, 1993.

[8] H. D. Patterson and R. Thompson, "Recovery of inter-block information when block sizes are unequal," *Biometrika*, vol. 58, pp. 545–554, 1971.

[9] D. A. Harville, "Maximum likelihood approaches to variance component estimation and to related problems," *Journal of the American Statistical Association*, vol. 72, no. 358, pp. 320–340, 1977.

[10] N. E. Breslow and D. G. Clayton, "Approximate inference in generalized linear mixed models," *Journal of the American Statistical Association*, vol. 88, pp. 9–25, 1993.

[11] D. M. Bates and S. DebRoy, "Linear mixed models and penalized least squares," *Journal of Multivariate Analysis*, vol. 91, no. 1, pp. 1–17, 2004.

[12] A. T. Kraja, D. Vaidya, J. S. Pankow et al., "A bivariate genome-wide approach to metabolic syndrome: STAMPEED Consortium," *Diabetes*, vol. 60, no. 4, pp. 1329–1339, 2011.

[13] T. M. Therneau, P. M. Grambsch, and V. S. Pankratz, "Penalized survival models and frailty," *Journal of Computational and Graphical Statistics*, vol. 12, no. 1, pp. 156–175, 2003.

[14] R. M. Pfeiffer, A. Hildesheim, M. H. Gail et al., "Robustness of inference on measured covariates to misspecification of genetic random effects in family studies," *Genetic Epidemiology*, vol. 24, no. 1, pp. 14–23, 2003.

[15] M. H. Chen, X. Liu, F. Wei et al., "A comparison of strategies for analyzing dichotomous outcomes in genome-wide association studies with general pedigrees," *Genetic Epidemiology*, vol. 35, no. 7, pp. 650–657, 2011.
[16] K. Y. Liang and S. L. Zeger, "Longitudinal data analysis using generalized linear models," *Biometrika*, vol. 73, no. 1, pp. 13–22, 1986.
[17] L. A. Cupples, H. T. Arruda, E. J. Benjamin et al., "The Framingham Heart Study 100K SNP genome-wide association study resource: overview of 17 phenotype working group reports," *BMC Medical Genetics*, vol. 8, supplement 1, 2007.
[18] J. Ott and D. Rabinowitz, "A principal-components approach based on heritability for combining phenotype information," *Human Heredity*, vol. 49, no. 2, pp. 106–111, 1999.
[19] Y. Wang, Y. Fang, and M. Jin, "A ridge penalized principal-components approach based on heritability for high-dimensional data," *Human Heredity*, vol. 64, no. 3, pp. 182–191, 2007.
[20] Y. . Wang, Y. Fang, and S. Wang, "Clustering and principal-components approach based on heritability for mapping multiple gene expressions," *BMC Proceedings*, vol. 1, supplement 1, p. S121, 2007.
[21] C. Lange, K. van Steen, T. Andrew et al., "A family-based association test for repeatedly measured quantitative traits adjusting for unknown environmental and/or polygenic effects," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, pp. 1544–6115, 2004.
[22] C. Lange, D. L. DeMeo, and N. M. Laird, "Power and design considerations for a general class of family-based association tests: quantitative traits," *American Journal of Human Genetics*, vol. 71, no. 6, pp. 1330–1341, 2002.
[23] L. Klei, D. Luca, B. Devlin, and K. Roeder, "Pleiotropy and principal components of heritability combine to increase power for association analysis," *Genetic Epidemiology*, vol. 32, no. 1, pp. 9–19, 2008.
[24] M. A. R. Ferreira and S. M. Purcell, "A multivariate test of association," *Bioinformatics*, vol. 25, no. 1, pp. 132–133, 2009.
[25] C. Lange, E. K. Silverman, X. Xu, S. T. Weiss, and N. M. Laird, "A multivariate family-based association test using generalized estimating equations: FBAT-GEE," *Biostatistics*, vol. 4, no. 2, pp. 195–206, 2003.
[26] Y. S. Aulchenko, D. J. de Koning, and C. Haley, "Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis," *Genetics*, vol. 177, no. 1, pp. 577–585, 2007.
[27] K. E. Muller and B. L. Peterson, "Practical methods for computing power in testing the multivariate general linear hypothesis," *Computational Statistics and Data Analysis*, vol. 2, no. 2, pp. 143–158, 1984.
[28] H. Wu, *Methods for genetic association studies using longitudinal and multivariate phenotypes in families [Ph.D. thesis]*, Boston University, Boston, Mass, USA, 2009.
[29] G. M. Fitzmaurice and N. M. Laird, "Regression models for mixed discrete and continuous responses with potentially missing values," *Biometrics*, vol. 53, no. 1, pp. 110–122, 1997.
[30] J. Liu, Y. Pei, C. J. Papasian, and H. W. Deng, "Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations," *Genetic Epidemiology*, vol. 33, no. 3, pp. 217–227, 2009.
[31] P. C. O'Brien, "Procedures for comparing samples with multiple endpoints," *Biometrics*, vol. 40, no. 4, pp. 1079–1087, 1984.
[32] X. Xu, L. Tian, and L. J. Wei, "Combining dependent tests for linkage or association across multiple phenotypic traits," *Biostatistics*, vol. 4, no. 2, pp. 223–229, 2003.
[33] L. J. Wei and W. E. Johnson, "Combining dependent tests with incomplete repeated measurements," *Biometrika*, vol. 72, no. 2, pp. 359–364, 1985.
[34] Q. Yang, H. Wu, C. Y. Guo, and C. S. Fox, "Analyze multivariate phenotypes in genetic association studies by combining univariate association tests," *Genetic Epidemiology*, vol. 34, no. 5, pp. 444–454, 2010.
[35] W. Pan, "Asymptotic tests of association with multiple SNPs in linkage disequilibrium," *Genetic Epidemiology*, vol. 33, no. 6, pp. 497–507, 2009.
[36] J.-T. Zhang, "Approximate and asymptotic distributions of chi-squared-type mixtures with applications," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 273–285, 2005.
[37] X. Liu and Q. Yang, "CUMP: an R package for analyzing multivariate phenotypes in genetic association studies".
[38] S. Vansteelandt, S. Goetgeluk, S. Lutz et al., "On the adjustment for covariates in genetic association analysis: a novel, simple principle to infer direct causal effects," *Genetic Epidemiology*, vol. 33, no. 5, pp. 394–405, 2009.

[39] J. Pearl, "Causal diagrams for empirical research," *Biometrika*, vol. 82, no. 4, pp. 669–688, 1995.
[40] J. M. Robins, "Data, design, and background knowledge in etiologic inference," *Epidemiology*, vol. 12, no. 3, pp. 313–320, 2001.
[41] P. J. Lipman, K. Y. Liu, J. D. Muehlschlegel, S. Body, and C. Lange, "Inferring genetic causal effects on survival data with associated endo-phenotypes," *Genetic Epidemiology*, vol. 35, no. 2, pp. 119–124, 2011.
[42] S. Vansteelandt, "Estimation of controlled direct effects on a dichotomous outcome using logistic structural direct effect models," *Biometrika*, vol. 97, no. 4, pp. 921–934, 2010.
[43] G. D. Smith and S. Ebrahim, "'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?" *International Journal of Epidemiology*, vol. 32, no. 1, pp. 1–22, 2003.
[44] D. A. Lawlor, R. M. Harbord, J. A. C. Sterne, N. Timpson, and G. D. Smith, "Mendelian randomization: using genes as instruments for making causal inferences in epidemiology," *Statistics in Medicine*, vol. 27, no. 8, pp. 1133–1163, 2008.
[45] P. M. McKeigue, H. Campbell, S. Wild et al., "Bayesian methods for instrumental variable analysis with genetic instruments ("Mendelian randomization"): example with urate transporter SLC2A9 as an instrumental variable for effect of urate levels on metabolic syndrome," *International Journal of Epidemiology*, vol. 39, no. 3, pp. 907–918, 2010.