Statistics and Its Interface Volume 10 (2017) 379–386

Genome-wide association test of multiple continuous traits using imputed SNPs

Baolin Wu* and James S. Pankow

More and more large cohort studies have conducted or are conducting genome-wide association studies (GWAS) to reveal the genetic components of many complex human diseases. These large cohort studies often collected a broad array of correlated phenotypes that reflect common physiological processes. By jointly analyzing these correlated traits, we can gain more power by aggregating multiple weak effects and shed light on the mechanisms underlying complex human diseases. The majority of existing multi-trait association test methods are based on jointly modeling the multivariate traits conditional on the genotype as covariate, and can readily accommodate the imputed SNPs by using their imputed dosage as a covariate. An alternative class of multi-trait association tests is based on the inverted regression, which models the distribution of genotypes conditional on the covariate and multivariate traits, and has been shown to have competitive performance. To our knowledge, all existing inverted regression approaches have implicitly used the "best-guess" genotypes, which is not efficient and known to lead to dramatic power loss, and there have not been any proposed methods of incorporating imputation uncertainty into inverted regressions. In this work, we propose a general and efficient framework that can account for the imputation uncertainty to further improve the association test power of inverted regression models for imputed SNPs. We demonstrate through extensive numerical studies that the proposed method has competitive performance. We further illustrate its usefulness by application to association test of diabetes-related glycemic traits in the Atherosclerosis Risk in Communities (ARIC) Study.

Keywords and phrases: GEE, GWAS, Pleiotropy, Imputation.

1. INTRODUCTION

Genetic studies often collect multiple phenotypes, which could be analyzed jointly to increase power by aggregating multiple weak effects and provide additional insights into the etiology of complex human diseases (Solovieff *et al.*, 2013).

Existing multi-trait association test methods (see, e.g., Ferreira and Purcell, 2009; Liu *et al.*, 2009; Yang *et al.*, 2010; Rasmussen-Torvik *et al.*, 2010; O'Reilly *et al.*, 2012; Tang and Ferreira, 2012; van der Sluis *et al.*, 2013; He *et al.*, 2013; Schifano *et al.*, 2013; Stephens, 2013; Seoane *et al.*, 2014) can be broadly classified into two categories. The first one is based on jointly modeling the multiple correlated outcomes with some multivariate regression models. Another novel approach is based on the inverted regression model, where the genotypes are regressed on the covariates and multivariate outcomes to estimate and test the multi-trait associations and typically some ordinal multinomial regression model is used. For example, O'Reilly *et al.* (2012) adopted the proportional odds model (POM), and Wu and Pankow (2015) proposed the adjacent category logit (ACL) model. For the multivariate regression based approach, it is straightforward to accommodate the imputed SNPs by using their imputation dosages as the covariate. While for the inverted regression approach, to our knowledge, all existing methods have implicitly used the "best-guess" genotypes, which is not efficient and known to lead to dramatic power loss, and there have not been any proposed methods in the literature that can incorporate the imputation uncertainty into inverted regressions. We propose a general and efficient GEE modeling approach to extending the inverted regression model to multi-trait association test of imputed SNPs.

*Corresponding author.

2. MATERIALS AND METHODS

2.1 Genotype based multinomial regression model

Consider a collection of continuous traits $Y = (y_1, \ldots, y_m)^T$, a p-vector of covariates X to be adjusted (which could contain both ancestry and non-ancestry covariates, e.g., ancestry principal components, age and gender), and a genotype score G (number of minor alleles). Assume the multivariate normal trait model, $(Y|G, X) \sim N(\gamma_0 + \gamma_X X + \gamma G, \Sigma)$, where γ_0 is a m-vector, γ_X is a $m \times p$ matrix, γ is a m-vector, and Σ is a $m \times m$ covariance matrix. The null hypothesis of multi-trait association is $H_0 : \gamma = 0$. When modeling the population genotype distribution $\Pr(G|X)$ with a logistic regression model (it holds when, e.g., the genotypes follow the Hardy-Weinberg equilibrium within each ancestry population), we can derive an adjacent-category logit model (ACL) (Wu and Pankow, 2015)

$$\log \frac{\phi_{g+1}}{\phi_g} = \beta_{0g} + X^T \beta_X + Y^T \beta, \quad g = 0, 1, \tag{1}$$

where $\phi_g = \Pr(G = g|X, Y)$ is the conditional genotype distribution probability, β_X is a p-vector, and β is a m-vector (specifically $\beta = \Sigma^{-1}\gamma$). The multi-trait association amounts to testing $H_0 : \beta = 0$. A closely related approach is the MultiPhen method (O'Reilly *et al.*, 2012), which assumed the proportional odds model (POM) for analyzing the three genotypes

$$(2) \quad \log \frac{\sum_{k=0}^{g} \phi_k}{\sum_{k=g+1}^{2} \phi_k} = \tilde{\beta}_{0g} + X^T \tilde{\beta}_X + Y^T \tilde{\beta}, \quad g = 0, 1.$$

The multi-trait association amounts to testing $H_0 : \tilde{\beta} = 0$. In general the POM provides a good approximation to the ACL, and two approaches have similar performance for directly genotyped/observed SNPs (Wu and Pankow, 2015). We want to remark that the inverted regression approach has assumed that the genotypes are directly observed, and all existing methods have implicitly used the "best-guess" genotypes for imputed SNPs. For both the inverted regression and multivariate regression models, the main parameters of interest are a vector of length m. The inverted regression model has smaller number of nuisance parameters, $p + 2$, compared to the multivariate regression model, $m + mp + m(m + 1)/2$.

2.2 Genotype imputation

To facilitate SNP association studies and across studies meta-analysis, many ungenotyped SNPs are typically imputed based on outside reference panel of existing samples, e.g., the HapMap and 1000 genome project (Browning and Browning, 2009; Howie *et al.*, 2009; Li *et al.*, 2010). These imputation approaches rely on the intuition that individuals can share short stretches of haplotypes inherited from distant common ancestors. Once these stretches are identified using those genotyped SNPs, alleles for intervening SNPs that are not genotyped in the individuals can then be imputed based on those individuals with measured SNPs (i.e., reference panel samples) (Li *et al.*, 2009, 2010). The typical imputation takes as input those haplotypes for polymorphic markers in the reference panel (e.g., the phased HapMap or 1000 genome chromosomes), and those directly genotyped markers in the individuals to be imputed. The sequence of markers are modeled as a mosaic of the set of reference haplotypes based on a Hidden Markov Model (HMM) (Li and Stephens, 2003; Stephens and Scheet, 2005). In the HMM, the reference haplotypes are treated as the hidden states, and the genotyped markers are treated as the observed signals. The HMM parameters are estimated iteratively and missing genotypes are sampled at each iteration based on the current HMM estimates. The sampled genotype counts over all iterations are aggregated together to give an indication of the relative probability of observing each possible genotypes (Li *et al.*, 2010). The relative fractions of three genotypes comprise the imputation scores for an imputed SNP.

In the following, we develop two modeling approaches to incorporating the imputation scores into the inverted regression. The first approach is rooted in the weighted multinomial regression approach with robust GEE covariance estimates (Lipsitz *et al.*, 1994; Preisser *et al.*, 2002). The second approach is based on the fractional multinomial regression modeling (Murteira and Ramalho, 2016), which is very suited to model the imputed genotype proportions. We will further show that these two modeling approaches are equivalent.

2.3 Association test of imputed SNPs: weighted multinomial regression

We develop a computationally fast weighted regression approach, where the imputation scores are treated as weights. Since the same sample will be used three times (for the three genotype scores), we need to take into account their dependence in the estimation of parameter covariance. The model-based covariance estimate from the independent weighted regression will under-estimate the variation. We propose to use the robust GEE sandwich covariance (Liang and Zeger, 1986). Specifically here we adopt the approach of Lipsitz *et al.* (1994) for modeling the multinomial outcomes, and the modeling framework of Preisser *et al.* (2002) for incorporating weights in the GEE.

For a collection of n unrelated individuals, denote X_i as the covariate, and Y_i as the m-vector of outcomes for sample $i = 1, \ldots, n$. Consider testing association of an imputed SNP. For the i-th sample, denote (p_{i0}, p_{i1}, p_{i2}) as the imputation scores (posterior probabilities) of genotype $0, 1, 2$. Denote $\Pr(G_i = k|X_i, Y_i) = \phi_{ik}$, $i = 1, \ldots, n, k = 0, 1, 2$. We convert the genotype score into a bivariate indicator of being the first two genotypes: the genotype scores $0/1/2$ are coded as (1,0), (0,1), (0,0) respectively. For the i-th sample, the three imputed genotypes (0,1,2) are represented by the working vector $\boldsymbol{G}_i = (1, 0, 0, 1, 0, 0)^T$. We define a probability vector $\boldsymbol{\mu}_i = (\phi_{i0}, \phi_{i1}, \phi_{i0}, \phi_{i1}, \phi_{i0}, \phi_{i1})^T$. Denote the imputation score matrix $W_i = \mathrm{diag}(p_{i0}, p_{i0}, p_{i1}, p_{i1}, p_{i2}, p_{i2})$. Assume a block-diagonal working covariance matrix V_i with the 2×2 diagonal blocks equal to $\mathrm{diag}(\phi_{i0}, \phi_{i1}) - (\phi_{i0}, \phi_{i1})^T(\phi_{i0}, \phi_{i1})$, which is the multinomial covariance matrix. Denote $\boldsymbol{\theta}$ as the collection of all model parameters. We use the following estimating equations for model estimation and inference

$$(3) \quad \sum_{i=1}^{n} U_i = 0, \quad U_i = D_i^T V_i^{-1} W_i (\boldsymbol{G}_i - \boldsymbol{\mu}_i), \quad D_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}}.$$

The robust sandwich covariance of $\hat{\boldsymbol{\theta}}$ can then be computed as $B^{-1}\Big(\sum_{i=1}^{n} U_i U_i^T\Big)B^{-1}$, where $B = \sum_{i=1}^{n} D_i^T V_i^{-1} W_i D_i$ and we plugin the estimated $\hat{\boldsymbol{\theta}}$.

Let $\hat{\theta}_2$ denote the m-vector of estimated regression parameters of main interest for the multivariate traits Y (i.e.,

β in the ACL and $\tilde{\beta}$ in the POM). Denote the corresponding covariance of $\hat{\theta}_2$ as V. The statistic $Q = \hat{\theta}_2^T V^{-1} \hat{\theta}_2$, which asymptotically has a null m-DF chi-square distribution, can be used to test the multi-trait association. When genetic effects are similar across traits, we can further improve the multi-trait association test power using a 1-DF statistic to test linear combinations of θ_2 following the line of O'Brien (1984). To test the similar or similar scaled effects across different traits, we propose the test statistics: $T = \mathbf{1}_m^T \hat{\Sigma}_0^{-1} V^{-1} \hat{\theta}_2 / (\mathbf{1}_m^T \hat{\Sigma}_0^{-1} V^{-1} \hat{\Sigma}_0^{-1} \mathbf{1}_m)^{1/2}$, $T' = S^T \hat{\Sigma}_0^{-1} V^{-1} \hat{\theta}_2 / (S^T \hat{\Sigma}_0^{-1} V^{-1} \hat{\Sigma}_0^{-1} S)^{1/2}$, where $\mathbf{1}_m$ is a column vector of m ones, $S = [\mathrm{diag}(\hat{\Sigma}_0)]^{1/2}$, and $\hat{\Sigma}_0$ is computed as the sample covariance matrix of residual vector of regressing Y on X (see Appendix for details). Their significance p-values can be computed based on the standard normal distribution. This generic GEE modeling approach can be readily generalized to analyze imputed SNPs using any inverted regression methods.

In the following, we show that the proposed GEE modeling approach is equivalent to a fractional multinomial regression model, which provides more intuitive justifications to model the imputed genotype scores.

2.4 Genotype based fractional multinomial regression model

For the ith individual, note that its imputation scores (p_{i0}, p_{i1}, p_{i2}) tell the relative fractions of three genotypes *under ideal repeated sampling*: for N individuals with the same characteristics (including covariate values) as the ith individual, $N(p_{i0}, p_{i1}, p_{i2})$ will be the observed counts of three genotypes, and naturally we can model them with a three-category multinomial distribution. Thus we can model the imputation scores with a multinomial distribution based quasi-likelihood, $\ell_i = \sum_{k=0}^{2} p_{ik} \log(\phi_{ik})$, and study the following quasi-likelihood for parameter estimation, $L = \sum_{i=1}^{n} \ell_i$. This model is also known as the fractional multinomial regression model (Murteira and Ramalho, 2016). We maximize L to obtain the quasi-maximum likelihood estimates (QMLE) for parameters, $\tilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L$, and compute its asymptotic covariance matrix based on the GEE as follows. Denote $\tilde{U}_i = \partial \ell_i / \partial \boldsymbol{\theta} = \sum_{k=0}^{2} \frac{p_{ik}}{\phi_{ik}} \frac{\partial \phi_{ik}}{\partial \boldsymbol{\theta}}$. The estimator $\tilde{\boldsymbol{\theta}}$ is obtained by solving estimating equations $\sum_{i=1}^{n} \tilde{U}_i = 0$, and its robust sandwich covariance matrix can then be computed as $\tilde{V} = \tilde{B}^{-1} \Omega \tilde{B}^{-1}$, where $\Omega = \sum_{i=1}^{n} \tilde{U}_i \tilde{U}_i^T$, and $\tilde{B} = \sum_{i=1}^{n} \partial^2 \ell_i / (\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T)$ and we plugin the estimated $\tilde{\boldsymbol{\theta}}$. We can show that this QMLE will lead to the same estimates as the previous weighted GEE approach. Specifically we can show that $\tilde{U}_i = U_i$ (see appendix for technical derivations). This QMLE can be cast into a weighted multinomial regression model and can be readily and quickly solved using existing software.

Previous derivations have assumed the additive genetic model, and they can be easily extended to recessive and dominant genetic models (see Supplementary materials http://intlpress.com/site/pub/pages/journals/items/sii/content/vols/0010/0003/s001).

In the following we conduct simulation studies to investigate the performance of the proposed methods for testing the multi-trait association of imputed SNPs.

3. SIMULATION STUDY

We simulate a standard normal covariate X_1 and an ancestry Bernoulli covariate X_2 with probability of 0.5 (population indicator). The SNP genotype G is simulated from a Binomial distribution, Binom$(2, f_0)$, where the minor allele frequency (MAF) $f_0 = p_0 + p_1 X_2$. We conducted simulations for testing m related traits of 1,000 unrelated individuals. Each time we simulate the m traits from a multivariate normal distribution with a compound symmetry correlation matrix with correlation ρ. The first trait has a variance of 2 and all the other traits have unit variance, $\sigma_1^2 = 2, \sigma_{k>1}^2 = 1$. We set $E(Y_k) = 1 + 0.5X_1 + 0.5X_2 + \gamma_k G$ for odd index k, and $E(Y_k) = 1 + X_1 + X_2 + \gamma_k G$ for even index k. For a given SNP G, we simulate its imputation probabilities from the Dirichlet distribution with parameters $(\alpha_0, \alpha_1, \alpha_2)$, where $\alpha_G = \tau$ and $\alpha_g = (1-\tau)/2$ for $g \neq G$, with larger τ reflecting higher imputation accuracy. We used 10^6 experiments to evaluate the type I error, and 10^4 experiments to evaluate the power under various combinations of $(\gamma_1, \ldots, \gamma_m)$. We conducted simulations for various parameter settings. Here we reported the results for $m = 4$, $p_0 = 0.3$, $p_1 = 0.1$, $\rho = 0.2, 0.5$, and $\tau = 0.8, 0.95$. The conclusions remain the same for other settings.

We studied the two inverted regression methods, the ACL and POM based GEE tests. For comparison we included the multiple linear regression model (MLM) based efficient GEE score tests (Avery *et al.*, 2011; He *et al.*, 2013), which have been shown to appropriately control the type I errors and have the overall best performance compared to the other methods (e.g., TATES of van der Sluis *et al.*, 2013 and other univariate test based methods) in extensive numerical studies. All methods reported three p-values based on the m-DF omnibus test and two 1-DF tests assuming common or common scaled effects. Denote the respective three tests as (Q_a, T_a, T_a') for ACL GEE test, (Q_o, T_o, T_o') for POM GEE test, and (Q_s, T_s, T_s') for the MLM GEE test. We use the imputed dosage as a covariate in the MLM GEE tests. In the appendix, we technically show that the MLM GEE tests are essentially based on a joint model of the multivariate traits with the imputation dosage as a covariate. As a by-product, we derive very fast numerical algorithms for genome-wide association test. For illustration, we also include the naive approach of modeling the "best-guess" genotypes for the two inverted regression methods, denoted as $(\tilde{Q}_a, \tilde{T}_a, \tilde{T}_a')$ for the ACL, and $(\tilde{Q}_o, \tilde{T}_o, \tilde{T}_o')$ for the POM respectively.

Table 1 summarizes the estimated type I errors. Overall we can see that for the two inverted regression (ACL and POM) based tests, using the "best-guess" genotypes leads

Table 1. Type I error of testing four continuous traits. The MAFs of SNP are 0.3 and 0.4 in the two populations. Q is the 4-DF omnibus test, T and T' are the 1-DF tests assuming common or common scaled effect. (Q_a, T_a, T'_a) are the ACL GEE tests. (Q_o, T_o, T'_o) are the POM GEE tests. (Q_s, T_s, T'_s) are the MLM GEE tests. $(\tilde{Q}_a, \tilde{T}_a, \tilde{T}'_a)$ are the ACL tests using the "best-guess" genotypes. $(\tilde{Q}_o, \tilde{T}_o, \tilde{T}'_o)$ are the POM tests using the "best-guess" genotypes. The type I errors have been scaled by the nominal significance level α

$\rho = 0.2, \tau = 0.8$															
α	Q_a	T_a	T'_a	$\tilde{Q}_a$	$\tilde{T}_a$	$\tilde{T}'_a$	Q_o	T_o	T'_o	$\tilde{Q}_o$	$\tilde{T}_o$	$\tilde{T}'_o$	Q_s	T_s	T'_s
10^{-4}	1.13	1.08	1.07	0.76	0.82	0.80	1.32	1.31	1.29	0.86	0.96	0.98	0.89	1.01	1.05
10^{-3}	1.09	1.07	1.06	0.84	0.89	0.89	1.15	1.13	1.17	0.94	1.01	1.02	0.96	1.00	1.00
10^{-2}	1.05	1.03	1.04	0.92	0.95	0.95	1.08	1.05	1.05	0.99	0.99	0.99	0.98	0.99	1.01
$\rho = 0.5, \tau = 0.8$															
α	Q_a	T_a	T'_a	$\tilde{Q}_a$	$\tilde{T}_a$	$\tilde{T}'_a$	Q_o	T_o	T'_o	$\tilde{Q}_o$	$\tilde{T}_o$	$\tilde{T}'_o$	Q_s	T_s	T'_s
10^{-4}	1.03	0.94	1.09	0.65	0.69	0.74	1.26	1.18	1.22	0.76	0.79	0.89	0.81	0.84	0.96
10^{-3}	1.10	1.03	1.02	0.80	0.88	0.88	1.15	1.11	1.11	0.94	0.96	0.96	0.92	0.95	0.96
10^{-2}	1.06	1.02	1.03	0.92	0.94	0.94	1.09	1.06	1.04	0.99	0.98	0.99	0.98	0.98	0.99
$\rho = 0.2, \tau = 0.95$															
α	Q_a	T_a	T'_a	$\tilde{Q}_a$	$\tilde{T}_a$	$\tilde{T}'_a$	Q_o	T_o	T'_o	$\tilde{Q}_o$	$\tilde{T}_o$	$\tilde{T}'_o$	Q_s	T_s	T'_s
10^{-4}	0.95	1.00	0.99	0.72	0.83	0.81	1.48	1.25	1.20	0.92	0.94	0.96	0.69	0.92	0.93
10^{-3}	1.05	1.03	1.05	0.82	0.84	0.90	1.17	1.09	1.10	0.96	0.92	0.97	0.97	0.98	1.02
10^{-2}	1.03	1.02	1.01	0.94	0.95	0.95	1.06	1.05	1.04	0.99	1.00	1.00	0.98	1.01	1.00
$\rho = 0.5, \tau = 0.95$															
α	Q_a	T_a	T'_a	$\tilde{Q}_a$	$\tilde{T}_a$	$\tilde{T}'_a$	Q_o	T_o	T'_o	$\tilde{Q}_o$	$\tilde{T}_o$	$\tilde{T}'_o$	Q_s	T_s	T'_s
10^{-4}	1.04	0.97	0.96	0.76	0.73	0.71	1.51	1.27	1.25	0.96	0.87	0.84	0.94	0.94	0.88
10^{-3}	1.02	0.99	1.00	0.82	0.88	0.88	1.16	1.10	1.04	0.96	0.95	0.96	0.93	0.95	0.96
10^{-2}	0.99	1.01	1.01	0.90	0.94	0.95	1.05	1.02	1.04	0.97	0.99	0.99	0.96	0.99	1.00

Table 2. Power of testing four traits at significance level $\alpha = 10^{-5}$. The MAFs of SNP are 0.3 and 0.4 in the two populations. The SNP imputation uncertainty parameter $\tau = 0.8$. γ_i is the SNP coefficient

$\rho = 0.2, \tau = 0.8$															
$(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$	Q_a	T_a	T'_a	$\tilde{Q}_a$	$\tilde{T}_a$	$\tilde{T}'_a$	Q_o	T_o	T'_o	$\tilde{Q}_o$	$\tilde{T}_o$	$\tilde{T}'_o$	Q_s	T_s	T'_s
(0.3,0,0,0)	0.066	0	0.001	0.031	0	0	0.078	0	0.002	0.041	0	0.001	0.057	0	0.001
(0.3,0.2,0.1,0)	0.274	0.051	0.111	0.152	0.026	0.059	0.307	0.063	0.130	0.183	0.034	0.074	0.248	0.047	0.102
(.25,.18,.18,.18)	0.293	0.496	0.528	0.168	0.335	0.367	0.325	0.531	0.563	0.202	0.380	0.413	0.268	0.480	0.509
(0.2,0.2,0.2,0.2)	0.379	0.623	0.591	0.226	0.452	0.422	0.415	0.659	0.626	0.267	0.500	0.472	0.351	0.606	0.573
$\rho = 0.5, \tau = 0.8$															
$(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$	Q_a	T_a	T'_a	$\tilde{Q}_a$	$\tilde{T}_a$	$\tilde{T}'_a$	Q_o	T_o	T'_o	$\tilde{Q}_o$	$\tilde{T}_o$	$\tilde{T}'_o$	Q_s	T_s	T'_s
(0.3,0,0,0)	0.216	0	0.001	0.116	0	0	0.246	0	0.001	0.143	0	0.001	0.195	0	0.001
(0.3,0.2,0.1,0)	0.341	0.004	0.028	0.200	0.002	0.013	0.375	0.005	0.034	0.236	0.002	0.018	0.313	0.004	0.026
(.25,.18,.18,.18)	0.085	0.176	0.218	0.040	0.101	0.129	0.102	0.196	0.242	0.052	0.119	0.150	0.073	0.166	0.206
(0.2,0.2,0.2,0.2)	0.137	0.309	0.250	0.070	0.190	0.149	0.157	0.336	0.274	0.088	0.220	0.172	0.121	0.293	0.236

to slightly conservative type I errors compared to their corresponding GEE tests that properly account for the imputation uncertainty. The ACL "best-guess" tests are generally more conservative compared to the corresponding POM "best-guess" tests, which control the type I error rate reasonably well. All ACL based tests appropriately control the type I errors. The POM GEE tests have slightly inflated type I errors at small significance level. The MLM GEE tests have well-controlled type I errors.

Table 2 and 3 summarize the power under $\tau = 0.8$ and $\tau = 0.95$ respectively. The 1-DF tests are the most powerful when either γ_j or γ_j/σ_j are close to each other. Not surprisingly using the "best-guess" genotypes leads to power loss for the two inverted regression (ACL and POM) based tests especially under lower imputation accuracy. The ACL GEE tests have comparable performance as the MLM GEE tests under relatively high imputation accuracy ($\tau = 0.95$). For imputed SNPs with less accuracy ($\tau = 0.8$), the ACL GEE tests have improved power compared to the MLM GEE tests. Overall the POM GEE tests have the largest power among all methods, which need to be interpreted with caution since POM GEE tests have slightly inflated type I errors as we have shown in Table 1. Under the same imputation uncertainty, when multiple traits have similar genetic effects, all tests have larger power under $\rho = 0.2$ compared to $\rho = 0.5$; while when genetic effects are different across

Table 3. Power of testing four traits at significance level $\alpha = 10^{-5}$. The MAFs of SNP are 0.3 and 0.4 in the two populations. The SNP imputation uncertainty parameter $\tau = 0.95$. γ_i is the SNP coefficient

$\rho = 0.2, \tau = 0.95$															
$(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$	Q_a	T_a	T'_a	$\tilde{Q}_a$	$\tilde{T}_a$	$\tilde{T}'_a$	Q_o	T_o	T'_o	$\tilde{Q}_o$	$\tilde{T}_o$	$\tilde{T}'_o$	Q_s	T_s	T'_s
(0.3,0,0,0)	0.242	0	0.003	0.206	0	0.002	0.256	0	0.004	0.222	0	0.002	0.229	0	0.003
(0.3,0.2,0.1,0)	0.658	0.140	0.286	0.601	0.108	0.245	0.674	0.157	0.309	0.622	0.131	0.272	0.643	0.148	0.291
(.25,.18,.18,.18)	0.684	0.850	0.872	0.634	0.819	0.843	0.699	0.859	0.879	0.657	0.832	0.852	0.672	0.848	0.870
(0.2,0.2,0.2,0.2)	0.781	0.924	0.906	0.732	0.904	0.879	0.793	0.928	0.913	0.752	0.911	0.890	0.768	0.922	0.903
$\rho = 0.5, \tau = 0.95$															
$(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$	Q_a	T_a	T'_a	$\tilde{Q}_a$	$\tilde{T}_a$	$\tilde{T}'_a$	Q_o	T_o	T'_o	$\tilde{Q}_o$	$\tilde{T}_o$	$\tilde{T}'_o$	Q_s	T_s	T'_s
(0.3,0,0,0)	0.571	0	0.001	0.515	0	0	0.588	0	0.001	0.543	0	0.001	0.557	0	0.001
(0.3,0.2,0.1,0)	0.742	0.010	0.084	0.693	0.006	0.063	0.756	0.012	0.102	0.717	0.009	0.079	0.731	0.013	0.091
(.25,.18,.18,.18)	0.278	0.442	0.518	0.239	0.396	0.470	0.292	0.455	0.526	0.255	0.418	0.486	0.268	0.436	0.514
(0.2,0.2,0.2,0.2)	0.417	0.666	0.580	0.367	0.617	0.526	0.432	0.677	0.589	0.390	0.633	0.547	0.403	0.662	0.577

traits, all tests have larger power under $\rho = 0.5$ compared to $\rho = 0.2$. Here joint multi-trait association test works well when combining highly correlated traits with heterogeneous genetic effects or lowly correlated traits with similar genetic effects.

We also performed simulation studies for less frequent and rare MAF (0.1, 0.05, and 0.01). The complete results are available at the supplementary materials. The overall conclusions remain the same.

4. ARIC GWAS OF DIABETES-RELATED GLYCEMIC TRAITS

The Atherosclerosis Risk in Communities (ARIC) study (The ARIC Investigators, 1989) is a multi-center prospective investigation of atherosclerotic disease in men and women aged 45–64 years at baseline. They were recruited from four U.S. communities: Forsyth County, North Carolina; Jackson, Mississippi; suburban areas of Minneapolis, Minnesota; and Washington County, Maryland. A total of 15,792 individuals participated in the baseline examination in 1987–1989. The vast majority of ARIC participants are of European (73%) or African ancestry (26%). Among 15,792 ARIC participants, we jointly analyzed the four fasting glucose levels of 5947 genotyped ARIC white participants who were non-diabetic at four visits measured approximately three years apart. Excluded from the analysis are a total of 9845 participants due to the following reasons: (1) 4314 participants are non-white; (2) 2751 participants do not complete all four visits; (3) 1556 participants have diabetes diagnosis or unknown diabetes status at any of the four visits; (4) 373 participants have no fasting glucose measurements for at least one of the four visits; (5) 851 participants do not have GWAS data. All ARIC participants have complete information on age, gender, and study center. The ARIC Study design, plasma glucose measurement, genotyping and other covariates have been described previously (Rasmussen-Torvik *et al.*, 2010). The glucose levels had an average correlation of 0.55 between visits. We applied an additive genetic model and adjusted for age, gender and study center (population indicators).

For illustration, we analyze those typed and imputed SNPs in chromosome 1 and 2. We test those common SNPs with MAF ≥ 0.05 and imputation $R^2 \geq 0.3$, which leads to 163,048 and 189,023 SNPs in chromosome 1 and 2 respectively. There were no identified genome-wide significant SNPs (p-value $\leq 5 \times 10^{-8}$) for chromosome 1, and multiple significant SNPs for chromosome 2. Specifically for the three tests: the m-DF omnibus test, and two 1-DF tests assuming common or common scaled effects, the ACL GEE tests (Q_a, T_a, T'_a) identified 56, 60, and 60 significant SNPs, the POM GEE tests (Q_o, T_o, T'_o) identified 56, 56, 59 SNPs, and the MLM GEE tests (Q_s, T_s, T'_s) identified 56, 59, 60 SNPs. All the identified SNPs are genome-wide significant in a meta-analyses of 21 fasting glucose GWAS with around 46,186 non-diabetic participants conducted by the MAGIC Consortium (Dupuis *et al.*, 2010). Compared to the MLM test T'_s, the ACL test T'_a identified one additional genome-wide significant SNP, rs1260326, with p-value of 3.3×10^{-8}. The p-value reported by the MAGIC meta-analysis of fasting glucose was 4.3×10^{-13}. Compared to the POM test T_o, the ACL test T_a identified four additional genome-wide significant SNPs, rs1260326, rs574981, rs549410 and rs550151, with p-values of 3.3×10^{-8}, 9.1×10^{-9}, 9.1×10^{-9}, and 7.5×10^{-9} respectively. Their respective p-values reported by the MAGIC meta-analysis of fasting glucose were 4.3×10^{-13}, 8.6×10^{-14}, 1.7×10^{-13}, and 1.2×10^{-13}.

All identified significant SNPs in chromosome 2 are imputed with imputation R^2 in the range of 0.90 to 0.9998. To our knowledge, all previous inverted regression approaches have implicitly used the "best-guess" genotypes. When using the "best-guess" genotypes, T'_a missed one significant SNP, rs1260326, T'_o missed three significant SNPs, rs574981, rs549410 and rs550151, and T'_s missed one SNP, rs1260326, at the genome-wide significance level, compared to their corresponding GEE tests using the imputation scores.

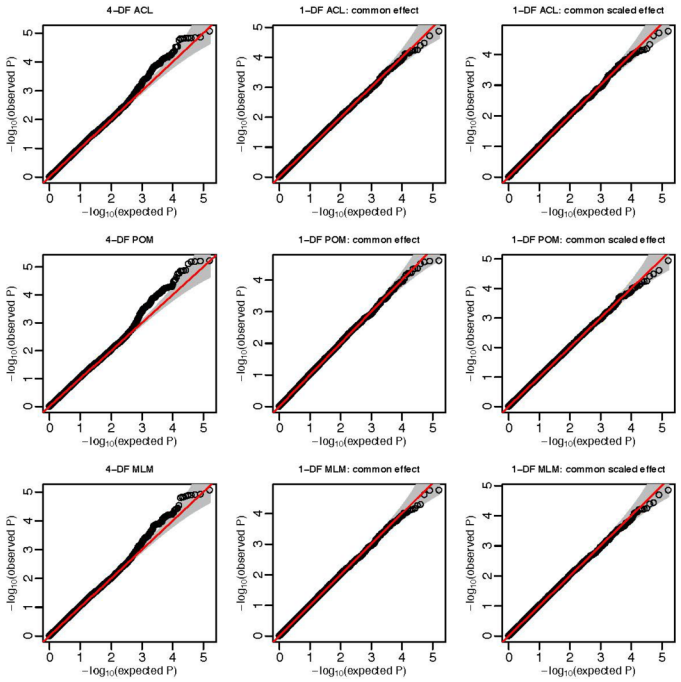

Figure 1. QQ-plot for SNPs in chromosome 1.

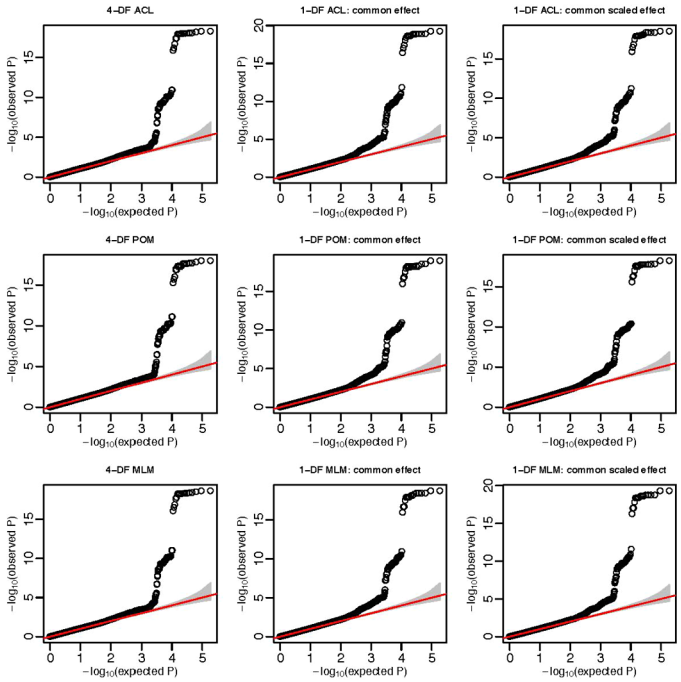

Figure 2. QQ-plot for SNPs in chromosome 2.

The QQ-plots of p-values for chromosome 1 and 2 SNPs are shown in figure 1 and 2 respectively. We also compute the genomic control (GC) parameters, which are the mean of the 1-DF chi-square test statistics, and the mean of the 4-DF chi-square test statistics scaled by four. The three methods have similar GC values: 1.01–1.02 for chromosome 1 and 1.04–1.10 for chromosome 2.

5. DISCUSSION

Most existing GWAS have primarily focused on testing single trait associations, which have led to discovery of many genome-wide significant variants for many human diseases and traits. However for most complex human diseases and traits, the explained heritability or trait variance by these identified variants still remain very small, which indicates significant "missing heritability" and yet more variants with small or moderate effects to be discovered. Recently there have been many efforts of conducting joint association test of correlated traits that reflect common physiological processes to identify more interesting genetic variants and provide additional insights into the disease etiology. Testing multiple correlated traits can aggregate weak variant effects to improve the genetic association test power. Among the existing multi-trait association test methods, the inverted regression approach models the conditional distribution of genotypes on covariates and multivariate traits, and provides a convenient and powerful approach with competitive performance. However it is not straightforward to analyze imputed SNPs for the inverted regression models in contrast to the trait based regression modeling approach, which can readily use the imputed dosage as covariate. In this paper, we proposed a general GEE based inverted regression modeling method to appropriately and efficiently test the multi-trait association of imputed SNPs. We show that the naive approach of analyzing "best-guess" genotypes could lead to dramatic power loss, while the proposed GEE based modeling approach offers much improved power and has comparable or larger power compared to the dosage based trait regression modeling approach.

For genome-wide association analyses, speed and robustness are both key issues. The proposed GEE modeling approach is robust and computationally fast. It is worthwhile to explore the likelihood based approach (e.g., mixed effects modeling approach or more generally likelihood ratio test based approach), which could bring more power under correct model assumptions than the typically Wald test based GEE modeling approach.

In this paper, we have focused on the multiple continuous traits association test of single variants. It is worthwhile to extend the inverted regression methods to association test at the gene level (Guo *et al.*, 2013; van der Sluis *et al.*, 2015), and generally to joint association test of mixed outcomes.

APPENDIX A. EQUIVALENCE OF QMLE AND WEIGHTED GEE ESTIMATES

In the weighted GEE approach, the three imputed genotypes (0,1,2) are represented by the working vector $\boldsymbol{G}_i =$

$(1,0,0,1,0,0)^T$ for the i-th sample. Denote a probability vector $\boldsymbol{\mu}_i = (\phi_{i0}, \phi_{i1}, \phi_{i0}, \phi_{i1}, \phi_{i0}, \phi_{i1})^T$. Denote the imputation score matrix $W_i = \text{diag}(p_{i0}, p_{i0}, p_{i1}, p_{i1}, p_{i2}, p_{i2})$. Assume a block-diagonal working covariance matrix V_i with the 2×2 diagonal blocks equal to $A_i = \text{diag}(\phi_{i0}, \phi_{i1}) - (\phi_{i0}, \phi_{i1})^T(\phi_{i0}, \phi_{i1})$, which is the multinomial covariance matrix. Denote $\boldsymbol{\theta}$ as the collection of all model parameters. The weighted GEE for the i-th sample is defined as $U_i = D_i^T V_i^{-1} W_i(\boldsymbol{G}_i - \boldsymbol{\mu}_i)$, where $D_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}}$. First note that

$$A_i^{-1} = \frac{1}{\phi_{i0}\phi_{i1}(1-\phi_{i0}-\phi_{i1})} \begin{bmatrix} \phi_{i1}(1-\phi_{i1}) & \phi_{i0}\phi_{i1} \\ \phi_{i0}\phi_{i1} & \phi_{i0}(1-\phi_{i0}) \end{bmatrix}$$

and we can check that $A_i^{-1}(1-\phi_{i0}, -\phi_{i1})^T = (\phi_{i0}^{-1}, 0)^T$, $A_i^{-1}(-\phi_{i0}, 1-\phi_{i1})^T = (0, \phi_{i1}^{-1})^T$, and $A_i^{-1}(-\phi_{i0}, -\phi_{i1})^T = (-\phi_{i2}^{-1}, -\phi_{i2}^{-1})^T$. Therefore we have

$$U_i = \sum_{k=0}^{1} \frac{p_{ik}}{\phi_{ik}} \frac{\partial \phi_{ik}}{\partial \boldsymbol{\theta}} - \frac{p_{i2}}{\phi_{i2}} \left\{ \sum_{k=0}^{1} \frac{\partial \phi_{ik}}{\partial \boldsymbol{\theta}} \right\}.$$

Note that $\phi_{i2} = 1 - \phi_{i0} - \phi_{i1}$, and hence we have

$$U_i = \sum_{k=0}^{2} \frac{p_{ik}}{\phi_{ik}} \frac{\partial \phi_{ik}}{\partial \boldsymbol{\theta}} = \tilde{U}_i.$$

APPENDIX B. GEE SCORE TEST OF MULTIPLE CONTINUOUS TRAITS

Here we show that for multiple continuous traits, the MLM GEE test of He *et al.* (2013) is essentially based on a joint model of the multivariate traits with imputation dosage as a covariate. Given the observations, denote the $n \times p$ covariate matrix as $\boldsymbol{X}$ (intercept included), the genotype dosage vector as G, and the kth outcome vector as Y_k, $k = 1, \ldots, K$. Consider the following joint multivariate linear regression model, $Y_k = \boldsymbol{X}\boldsymbol{\alpha}_k + G\beta_k + E_k$, where $E_k = (\epsilon_{k1}, \ldots, \epsilon_{kn})^T$. We model the error vector with a zero-mean multivariate normal distribution with $Var(\epsilon_{ki}) = \sigma_k^2$ and $Corr(\epsilon_{ki}, \epsilon_{li}) = \rho_{kl}$. Denote the $n \times n$ projection matrix $H = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. The scaled score statistic for testing β_k is $u_k = (Y_k - HY_k)^T G/\hat{\sigma}_k$. Denote the score vector $U = (u_1, \ldots, u_K)^T$. Note that we can equivalently write $u_k = Y_k^T(G - HG)/\hat{\sigma}_k$. Hence we have asymptotically $Var(u_k) = \|G - HG\|^2$, and $Cov(u_k, u_l) = \|G - HG\|^2\rho_{kl}$. The MLM GEE test is based on U and its estimated covariance, $U^T\widehat{Cov}(U)^{-1}U$, which asymptotically follows a K-DF chi-square distribution under null. He *et al.* (2013) consistently estimated $\widehat{Cov}(U)$ based on the efficient score vectors (Lin, 2005*a*,*b*). We can easily verify that the efficient score vectors are $Z_k = (Y_k - HY_k) \circ (G - HG)/\hat{\sigma}_k$, where $\circ$ is the Hadamard product (matrix element-wise product). Note that the residual vectors $Y_k - HY_k$ and H can be precomputed, and we just need to compute $G - HG$ to test the genome-wide multi-trait associations.

APPENDIX C. 1-DF MULTI-TRAIT ASSOCIATION TEST

Consider $U = a^T\hat{\theta}_2$. U asymptotically follows a normal distribution, $U \sim N(a^T\eta, a^TVa)$, where η is the true value of θ_2. For the ACL (1), we have $\theta_2 = \Sigma^{-1}\gamma$, where Σ is the covariance matrix of Y and γ is the corresponding marginal genetic effects. For the POM (2), we assume $\theta_2 \approx \Sigma^{-1}\gamma$ since the POM approximates the ACL. Assuming a common genetic effect for all traits, we have $\eta = \nu\Sigma^{-1}\mathbf{1}_m$. The effect size of U is then proportional to $\nu(a^T\Sigma^{-1}\mathbf{1}_m)/(a^TVa)^{1/2} = \nu b^TV^{-1/2}\Sigma^{-1}\mathbf{1}_m$, where $b = V^{1/2}a/(a^TVa)^{1/2}$ (note $b^Tb = 1$). Taking $b \propto V^{-1/2}\Sigma^{-1}\mathbf{1}_m$ will maximize the effect size. Therefore we use the following statistic $T = \mathbf{1}_m^T\Sigma^{-1}V^{-1}\hat{\theta}_2/(\mathbf{1}_m^T\Sigma^{-1}V^{-1}\Sigma^{-1}\mathbf{1}_m)^{1/2}$. With a common scaled genotype effect for all traits, we have $\eta = \nu\Sigma^{-1}S$, where $S = [\text{diag}(\Sigma)]^{1/2}$. Similarly we can derive $T' = S^T\Sigma^{-1}V^{-1}\hat{\theta}_2/(S^T\Sigma^{-1}V^{-1}\Sigma^{-1}S)^{1/2}$. In practice we estimate Σ by $\hat{\Sigma}_0$, the sample covariance matrix of $\tilde{Y}$, the residual vector of regressing Y on X.

ACKNOWLEDGEMENTS

This research was supported in part by NIH grant GM083345 and CA134848. We are grateful to the University of Minnesota Supercomputing Institute for assistance with the computations. We want to thank the associate editor and reviewers for their constructive comments which have greatly improved the presentation of the paper.

The ARIC Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN-268201100011C, and HHSN268201100012C), R01HL087641, R01HL59367 and R01HL086694; National Human Genome Research Institute contract U01HG004402; and National Institutes of Health contract HHSN268200625226C. The authors thank the staff and participants of the ARIC study for their important contributions. Infrastructure was partly supported by Grant Number UL1RR025005, a component of the National Institutes of Health and NIH Roadmap for Medical Research.

Received 23 June 2015

REFERENCES

Avery, C. L., He, Q., North, K. E., Ambite, J. L., Boerwinkle, E., Fornage, M., Hindorff, L. A., Kooperberg, C., Meigs, J. B., Pankow, J. S., Pendergrass, S. A., Psaty, B. M., Ritchie, M. D., Rotter, J. I., Taylor, K. D., Wilkens, L. R., Heiss, G. and Lin, D. Y. (2011) A phenomics-based strategy identifies loci on APOC1, BRAP, and PLCG1 associated with metabolic syndrome phenotype domains. *PLoS Genet*, **7** (10), e1002322.

Browning, B. L. and Browning, S. R. (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American Journal of Human Genetics*, **84** (2), 210–223.

Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A. U. and others. (2010) New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature Genetics*, **42** (2), 105–116.
Ferreira, M. A. R. and Purcell, S. M. (2009) A multivariate test of association. *Bioinformatics*, **25** (1), 132–133.
Guo, X., Liu, Z., Wang, X. and Zhang, H. (2013) Genetic association test for multiple traits at gene level. *Genetic Epidemiology*, **37** (1), 122–129.
He, Q., Avery, C. L. and Lin, D. Y. (2013) A general framework for association tests with multivariate traits in large-scale genomics studies. *Genetic Epidemiology*, **37** (8), 759–767.
Howie, B. N., Donnelly, P. and Marchini, J. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*, **5** (6), e1000529.
Klei, L., Luca, D., Devlin, B. and Roeder, K. (2008) Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genetic Epidemiology*, **32** (1), 9–19.
Li, N. and Stephens, M. (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, **165** (4), 2213–2233.
Li, Y., Willer, C., Sanna, S. and Abecasis, G. (2009) Genotype imputation. *Annual Review of Genomics and Human Genetics*, **10**, 387–406.
Li, Y., Willer, C. J., Ding, J., Scheet, P. and Abecasis, G. R. (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, **34** (8), 816–834.
Liang, K. Y. and Zeger, S. L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73** (1), 13–22. MR0836430
Lin, D. Y. (2005*a*) On rapid simulation of P values in association studies. *American Journal of Human Genetics*, **77** (3), 513–514.
Lin, D. Y. (2005*b*) An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics*, **21** (6), 781–787.
Lipsitz, S. R., Kim, K. and Zhao, L. (1994) Analysis of repeated categorical data using generalized estimating equations. *Statistics in Medicine*, **13** (11), 1149–1163.
Liu, J., Pei, Y., Papasian, C. J. and Deng, H. W. (2009) Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations. *Genetic Epidemiology*, **33** (3), 217–227.
Murteira, J. M. R. and Ramalho, J. J. S. (2016) Regression analysis of multivariate fractional data. *Econometric Reviews*, **35** (4), 515–552. MR3464348
O'Brien, P. C. (1984) Procedures for comparing samples with multiple endpoints. *Biometrics*, **40** (4), 1079–1087. MR0786180
O'Reilly, P. F., Hoggart, C. J., Pomyen, Y., Calboli, F. C. F., Elliott, P., Jarvelin, M. R. and Coin, L. J. M. (2012) MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PLoS ONE*, **7** (5), e34861.
Preisser, J. S., Lohman, K. K. and Rathouz, P. J. (2002) Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. *Statistics in Medicine*, **21** (20), 3035–3054.
Rasmussen-Torvik, L. J., Alonso, A., Li, M., Kao, W., Kattgen, A., Yan, Y., Couper, D., Boerwinkle, E., Bielinski, S. J. and Pankow, J. S. (2010) Impact of repeated measures and sample selection on genome-wide association studies of fasting glucose. *Genetic Epidemiology*, **34** (7), 665–673.
Schifano, E., Li, L., Christiani, D. and Lin, X. (2013) Genome-wide association analysis for multiple continuous secondary phenotypes. *The American Journal of Human Genetics*, **92** (5), 744–759.
Seoane, J. A., Campbell, C., Day, I. N. M., Casas, J. P. and Gaunt, T. R. (2014) Canonical correlation analysis for gene-based pleiotropy discovery. *PLoS Comput Biol*, **10** (10), e1003876.
Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. and Smoller, J. W. (2013) Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics*, **14** (7), 483–495.
Stephens, M. (2013) A unified framework for association analysis with multiple related phenotypes. *PLoS ONE*, **8** (7), e65245.
Stephens, M. and Scheet, P. (2005) Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *American Journal of Human Genetics*, **76** (3), 449–462.
Tang, C. S. and Ferreira, M. A. R. (2012) A gene-based test of association using canonical correlation analysis. *Bioinformatics*, **28** (6), 845–850. MR2934904
The ARIC Investigators. (1989) The atherosclerosis risk in communities (ARIC) study: design and objectives. *American Journal of Epidemiology*, **129** (4), 687–702.
van der Sluis, S., Dolan, C. V., Li, J., Song, Y., Sham, P., Posthuma, D. and Li, M. X. (2015) MGAS: a powerful tool for multivariate gene-based genome-wide association analysis. *Bioinformatics*, **31** (7), 1007–1015.
van der Sluis, S., Posthuma, D. and Dolan, C. V. (2013) TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS Genet*, **9** (1), e1003235.
Wu, B. and Pankow, J. S. (2015) Statistical methods for association tests of multiple continuous traits in genome-wide association studies. *Annals of Human Genetics*, **79** (4), 282–293.
Yang, Q., Wu, H., Guo, C. Y. and Fox, C. S. (2010) Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. *Genetic Epidemiology*, **34** (5), 444–454.

Baolin Wu
Division of Biostatistics
School of Public Health
University of Minnesota
MN
USA
E-mail address: baolin@umn.edu

James S. Pankow
Division of Epidemiology and Community Health
School of Public Health
University of Minnesota
MN
USA
E-mail address: panko001@umn.edu