Published in final edited form as:

Behav Genet. 2018 January ; 48(1): 55–66. doi:10.1007/s10519-017-9883-x.

Adaptive SNP-set Association Testing in Generalized Linear Mixed Models with Application to Family Studies

Jun Young Park¹, Chong Wu¹, Saonli Basu¹, Matt McGue², and Wei Pan^{*,1}

¹Division of Biostatistics, University of Minnesota, Minneapolis, MN

²Department of Psychology, University of Minnesota, Minneapolis, MN

Abstract

In genome-wide association studies (GWAS), it has been increasingly recognized that, as a complementary approach to standard single SNP analyses, it may be beneficial to analyze a group of functionally related SNPs together. Among the existent population-based SNP-set association tests, two adaptive tests, the aSPU test and the aSPU_{path} test, offer a powerful and general approach at the gene- and pathway-levels by data-adaptively combining the results across multiple SNPs (and genes) such that high statistical power can be maintained across a wide range of scenarios. We extend the aSPU and the aSPU_{path} test to familial data under the framework of the generalized linear mixed models (GLMMs), which can take account of both subject relatedness and possible population structure. As in population-based GWAS, the proposed aSPU and aSPU_{path} tests require only fitting a single and common GLMM (under the null hypothesis) for all the SNPs, thus are computationally efficient and feasible for large GWAS data. We illustrate our approaches in identifying genes and pathways associated with alcohol dependence in the Minnesota Twin Family Study. The aSPU test detected a gene associated with the trait, in contrast to none by the standard single SNP analysis. Our aSPU test also controlled Type I errors satisfactorily in a small simulation study. We provide R code to conduct the aSPU and aSPU_{path} tests for familial and other correlated data.

Keywords

Alcohol Dependence; aSPU; GEE; GLMM; GWAS; Score Test

*Corresponding Author Address: A460 Mayo Building, MMC 303, 420 Delaware St. SE, Minneapolis, MN 55455 Phone: (612) 626-2705 Fax: (612) 626-0660, weip@biostat.umn.edu.

Compliance with Ethical Standards

Conflict of interest: Jun Young Park, Chong Wu, Saonli Basu, Matt McGue, and Wei Pan declare that they have no conflict of interest.

Human and Animal Rights and Informed Consent: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study.

Introduction

There has been increasing interest in SNP set analysis in genome-wide association studies (GWAS). It typically includes association testing on a group of related SNPs such as in a gene or a pathway. Compared to the most popular approach that tests on individual SNPs one by one, tests based on a set of SNPs not only offer a data-driven approach to analyze each functional unit of SNPs, but also may improve statistical power by aggregating information in the presence of multiple causal or associate SNPs and by reducing the number of hypotheses tested.

Since there is no uniformly most powerful multivariate association test with an unknown truth, a number of SNP-set association tests have been proposed. The most familiar tests could be the score test based on the asymptotic normality of the score vector and the UminP test taking the minimum of the p-values from the single SNP-based tests. In addition, the sequence kernel association test (SKAT) (Wu et al. 2010) and the adaptive sum of powered score (aSPU) test (Pan et al. 2014) use the score vector to construct gene-level score-based tests. The SKAT and the aSPU tests were shown to outperform other gene-level score-based tests in many situations. The aSPU test is a flexible association test that data-adaptively chooses the best test among a class of score-based tests. For example, it is possible that multiple SNPs in or near a gene are causal with similar effect sizes and the association direction, which is ideal to using a burden test. On the other hand, it is also possible that only one or few of them are causal, for which the UminP test is high powered. The score-based tests used in the aSPU test are very general, including the burden test, the SKAT (with the linear kernel), and a test similar to the UminP test. For population-based GWAS, the aSPU test has been extended to aSPU_{path} for pathway-level association testing (Pan et al. 2015). Going beyond the population-based single-trait studies (with independent data), the GEE-aSPU test was also proposed to test the association between multiple traits and a single SNP or multiple SNPs under the framework of the generalized estimating equations (GEE) (Zhang et al. 2014; Kim et al. 2016).

This paper extends the existent population-based aSPU and aSPU_{path} tests to familial data. The standard regression approach cannot be used because familial data inherently involve correlated observations. We use the generalized linear mixed models (GLMMs), instead of generalized linear models (GLMs), to account for within-family correlations. A GLMM can also account for other types of dependencies such as population stratification and cryptic relatedness, while other approaches, such as using principal components or genomic control, may not suffice (Devlin and Roeder 1999; Price et al. 2006). Fitting a general GLMM efficiently for large genetic data has only become available recently with an R package called GMMAT (Chen et al. 2016). It has been mainly used to perform the score test on each single SNP with a binary trait, showing that it effectively controlled Type I errors in the presence of population stratification and relatedness while the usual linear mixed models (LMM) failed (Chen et al. 2016). It is noted that, in contrast to that for a GLM, it is difficult from scratch to construct the score vector for a GLMM. The main purpose of this paper is to extend the powerful aSPU and aSPU_{path} tests to GLMMs with the use of GMMAT to automatically and simply extract the score vector and its covariance matrix.

The major methodological contribution of this paper is to show how several gene- and pathway-level score-based tests, including the aSPU test and the aSPUpath test, are conducted using the score vector extracted in the GMMAT package of a GLMM with multiple random components, then compare them with those under GEE. We illustrate our approach in identifying genes associated with alcohol dependence from the Minnesota Twin Family Study (MTFS) (Iacono and McGue 2002). Our proposed methods preserve the advantages of the original aSPU and aSPUpath tests in population-based studies. The tests require fitting only a single model under the null hypothesis, rather than fitting a separate model for each set of SNPs to be tested, which would not be computationally feasible for complex GLMMs and large GWAS data. Also, the tests showed the desired data-adaptiveness in analysis of the real data while controlling Type I errors satisfactorily.

Methods

Notation

We let $y_i = (y_{i1}, \dots, y_{in_i})'$ denote the traits of n_i individuals in family $i = 1, \dots, F$, $X_i = (X_{i1}, \dots, X_{in_i})'$ is a $n_i \times q$ matrix on q covariates including the intercept, and $G_i = (G_{i1}, \dots, G_{in_i})'$ is a $n_i \times p$ matrix for p SNPs (additively coded 0, 1, and 2); in this paper we consider only the common SNPs (with minor allele frequencies greater than 0.05) that are between 10Kbp

upstream and downstream the gene coding region of each gene being tested. Let $n = \sum_{i=1}^F n_i$ be the total number of individuals in the study, $b_i = (b_{i1}, \dots, b_{in_i})'$ be the vector of random intercepts for y_i in GLMM. Extending this, we denote

$G = (G'_1, \dots, G'_F)'$, $X = (X'_1, \dots, X'_F)'$, $y = (y'_1, \dots, y'_F)'$, and $b = (b'_1, \dots, b'_F)'$. In addition, we let $g(\bullet)$ be the link function in GLM, GLMM and GEE.

Score-based Tests and the aSPU/aSPUpath Tests

Among the most popular and asymptotically equivalent likelihood ratio test, the score test, and the Wald test, the score test is advantageous in GWAS due to its computational simplicity: only a single null model under the null hypothesis is to be fitted and there is no need to fit a model for each SNP or SNP set to be tested. Once the score vector $U = (U_1, \dots, U_p)'$ and its covariance matrix V under the null hypothesis are obtained, the score test, or more generally any score-based test, typically requires only matrix operations to compute its test statistics. Note that the score vector preserves the direction of the effect of the genotype and, when variance of each score is close to each other, it provides the relative effect sizes. In contrast, the Wald test or likelihood ratio test requires to fit a full model for each SNP or SNP set to be tested, which will be problematic in GWAS with GLMM since fitting a GLMM may take several minutes to several hours for each of a half million SNPs in a typical GWAS.

We briefly describe a few commonly used score-based tests here. Suppose that p variants are encoded in a gene. The score test takes $U'V^{-1}U$ as the test statistic and it asymptotically follows a $\chi^2_{df=p}$ distribution under the null hypothesis. The UminP test takes $\max_j (U_j^2/V_{jj})$ as the test statistic, where V_{jj} is a j^{th} diagonal element of V . The idea of the UminP test is to take the minimum p-value among the single SNP-based tests applied to each SNP separately.

The burden test adds up all elements of the score vector, taking $\sum_{j=1}^p U_j$ as the test statistic. Note that the burden test loses power if the directions and the effect sizes of the SNPs are not the same. Lastly, the SSU test is a variance component test, taking $U'U = \sum_{j=1}^p U_j^2$ as the test statistic and is not affected by varying association direction. It was shown (Pan 2011) that the SSU test is equivalent to kernel machine regression (KMR) with a linear kernel (Kwee et al. 2008) and genomic distance-based regression (GDBR) with the Euclidean distance (Wessel and Schork 2006).

The sum of powered score test, say $\text{SPU}(\gamma)$, is one of a family of the score-based association tests; each test is indexed by a constant $\gamma \in \{0, 1, \dots, \infty\}$. The test statistic for a $\text{SPU}(\gamma)$ test

is $T_{\text{SPU}(\gamma)} = T_{\text{SPU}(\gamma)}(\mathbf{U}) = \sum_{j=1}^p U_j^\gamma$ for $0 < \gamma < \infty$, and $= \max_j |U_j|$ when $\gamma = \infty$. The p-values of the $\text{SPU}(\gamma)$ test can be quickly computed based on asymptotics for $\gamma = 1$ or 2, but not otherwise. Pan et al proposed using Monte Carlo simulations or permutations to calculate the p-values (Pan et al. 2014). Since the score vector $\mathbf{U}$ is asymptotically distributed as $\mathcal{N}(\mathbf{0}, \mathbf{V})$ under the null hypothesis, the p-value for $\text{SPU}(\gamma)$ is calculated as

$$P_{\text{SPU}(\gamma)} = \left\{ 1 + \sum_{b=1}^B I \left(\left| T_{\text{SPU}(\gamma)}^{(b)} \right| \geq \left| T_{\text{SPU}(\gamma)} \right| \right) \right\} / (1+B), \quad (1)$$

where B is the number of simulated score vectors (denoted as $\mathbf{U}^{(b)} \sim \mathcal{N}(\mathbf{0}, \mathbf{V})$ independently for the b^{th} simulation), and $T_{\text{SPU}(\gamma)}^{(b)} = T_{\text{SPU}(\gamma)}(\mathbf{U}^{(b)})$ is the $\text{SPU}(\gamma)$ statistic based on $\mathbf{U}^{(b)}$.

One advantage of the $\text{SPU}(\gamma)$ test is that various types of score-based tests are indexed by a single parameter γ . It is easily observed that the burden test is equal to $\text{SPU}(1)$, while the SSU test is equivalent to $\text{SPU}(2)$ (Pan 2011). Moreover, the univariate UminP test is very similar to $\text{SPU}(\infty)$ (when the variances of the score elements are close to each other).

Alternatively, one can regard γ as a value determining the weight on each SNP, $w_j = U_j^{\gamma-1}$, which are combined for a gene-level analysis. Even though it is possible to calculate the p-values of the SPU tests for every $\gamma \in \mathbb{N}$, to save time and maintain power, practical experiences suggested using $\gamma \in \{1, 2, \dots, 8, \infty\}$ often suffices (Pan et al. 2014).

One drawback of any non-adaptive test is that it may not be powerful in some situations, even if it is in others. It is possible that there are only one or few causal SNPs with large effect sizes in some causal genes, while in others there may be more causal SNPs with smaller effect sizes; depending on the unknown truth, the identity of the more powerful $\text{SPU}(\gamma)$ or other tests may change. Thus, using a single test (e.g. a single $\text{SPU}(\gamma)$ test for a single value of γ) for all the genes might not be powerful to detect some truly associated genes. This motivates the use of the adaptive SPU (aSPU) test, which takes the minimum p-value of the multiple $\text{SPU}(\gamma)$ tests as the test statistic T_{aSPU} . The idea is to take the best $\text{SPU}(\gamma)$ test available for each gene. The p-value for the aSPU test using simulations is calculated as

$$P_{aSPU} = \left\{ 1 + \sum_{b=1}^B I \left(T_{aSPU} \geq T_{aSPU}^{(b)} \right) \right\} / (1+B), \quad (2)$$

and

$$T_{aSPU}^{(b)} = \min_{\gamma} \left[\left\{ 1 + \sum_{k \neq b} I \left(\left| T_{SPU(\gamma)}^{(k)} \right| \geq \left| T_{SPU(\gamma)}^{(b)} \right| \right) \right\} / B \right]. \quad (3)$$

The aSPU test can be further extended to pathway analysis. A pathway-level analysis considers all the genetic variants in a group of functionally related genes, such as a biological pathway in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa et al. 2010). The gene-level statistic for the SPU_{path}(γ , γ_G) test is constructed similar to the SPU(γ) test. The SPU_{path} statistic depends on two values γ and γ_G , where a large γ is effective when there are a small number of associated SNPs within each gene and a large γ_G is effective when there are a small number of associated genes in a pathway. The test statistic for the adaptive aSPU_{path} is the minimum p-value of the SPU_{path}(γ , γ_G) tests, where $\gamma \in \{1, 2, 3, \dots, 8\}$ and $\gamma_G \in \{1, 2, 4, 8\}$ are often used based on their empirical good performance; for more details, see Pan et al. (2015).

Score Vector in Generalized Linear Mixed Models

A generalized linear mixed model (GLMM) is a conditional and parametric approach to correlated phenotypes with the use of random effects, in addition to the usual fixed effects in a GLM. It requires a full specification of the covariance structures. A GLMM with multiple random intercepts is specified as

$$g(\mu_i^b) = g(E[Y_i | b_i]) = X_i \alpha + G_i \beta + b_i \text{ with } b \sim \mathcal{N} \left(0, \sum_{k=1}^K \tau_k \Psi_k \right), \quad (4)$$

where Ψ_k is the k^{th} $n \times n$ pre-specified positive-definite correlation matrix, and τ_k is the corresponding non-negative variance component parameter. For example, when Ψ_G is a genetic relationship matrix (GRM), $\Psi_{G(ij, i'j')}$ gives the genetic relationship between the j^{th} member of the i^{th} family and the j'^{th} member of the i'^{th} family. Lastly, $Var(y_{ij} | b_{ij}) = \phi \cdot v(\mu_{ij})$, where ϕ is an overdispersion parameter and $v(\bullet)$ is a pre-specified variance function. In particular, for the two most common types of phenotypes, their distributions (conditional on random effects b_j) are

$$y_{ij}|b_{ij} \sim \begin{cases} N(\mu_{ij}^b, \varphi) & \text{for quantitative traits;} \\ \text{Bernoulli}(\mu_{ij}^b) & \text{for binary traits.} \end{cases} \quad (5)$$

Fitting GLMM using maximum likelihood is computationally challenging because the marginal log-likelihood involves a multiple integration over the random effects:

$$\mathcal{L} = \log(f(y|\alpha, \beta, \tau_1, \dots, \tau_k)) = \log \int \prod_{i=1}^F \prod_{j=1}^{n_i} f(y_{ij}|\alpha, \beta, b) f(b|\tau_1, \dots, \tau_k) db, \quad (6)$$

where $f(\bullet|\bullet)$ is a generic notation for conditional probability distribution for the trait and for the random effects, as shown in ((4) and (5). Accordingly, there is no closed-form for the marginal log-likelihood and thus its corresponding score vector for non-quantitative traits. Xu and Pan approximated the score vector based on the maximum likelihood estimator (Xu and Pan 2015). However, it requires fitting a different GLMM for each SNP and thus is computationally demanding for large GWAS data. Instead of using the marginal likelihood, the penalized quasi-likelihood (PQL) has been used as a simpler and effective approximation, which also gives an approximate score vector and its covariance matrix in closed forms (Breslow and Clayton 1993). Also, the PQL estimates are the same as the maximum likelihood estimates for normally distributed quantitative traits. An R package **GMMAT** has recently become available, from which the (approximate) score vector and its covariance matrix can be obtained (Chen et al. 2016). The GLMM specified in **GMMAT** is exactly in (4) and (5).

For a canonical link function $g(\bullet)$, the score vector and its covariance matrix under null hypothesis $H_0: \beta = 0$ are

$$U = G' \begin{pmatrix} y - \hat{\mu}^b \\ \hat{\phi} \end{pmatrix} \text{ and } \mathbf{V} = \text{Cov}(\mathbf{U}|H_0) = \mathbf{G}' \hat{\mathbf{P}} \mathbf{G}, \quad (7)$$

where $\hat{\phi} = 1$ for binary traits, $\hat{\mu}^b$ is the vector of the fitted trait values under the null hypothesis, and $\hat{\mathbf{P}}$ is an $n \times n$ projection matrix defined as (Breslow and Clayton 1993; Harville 1977),

$$\hat{\mathbf{P}} = \hat{\Sigma}^{-1} - \hat{\Sigma}^{-1} X (X' \hat{\Sigma}^{-1} X)^{-1} X' \hat{\Sigma}^{-1}, \quad (8)$$

$$\hat{\Sigma} = \text{diag}(\hat{\phi} \cdot u(\hat{\mu}_{ij}))^{-1} + \sum_{k=1}^K \hat{\tau}_k \Psi_k. \quad (9)$$

U and V can be obtained directly from GMMAT.

Score Vector in Generalized Estimating Equations

The generalized estimating equations (GEE) is a marginal and semi-parametric approach to correlated data analysis. It does not require a specification of a full distribution for the trait, offering flexibility. Suppose $\mathbf{R}_i$ is a correlation matrix for traits within cluster i . In general, $\mathbf{R}_i$ is unknown and can be hard to model; accordingly, we use a possibly incorrect $\mathbf{R}_{wi}(\delta)$, a working correlation matrix depending on parameter δ . Typical choices of working correlation matrix include the compound symmetry (CS) model (assuming an equal correlation within each group), and the independence model. One advantage of GEE is its consistent estimation and inference even when the working correlation matrix is misspecified. A marginal regression model in GEE is specified as $g(\mu_i) = g(E[y_i]) = \mathbf{X}_i\alpha + \mathbf{G}_i\beta$, where the link function $g(\bullet)$ is pre-specified.

Under $H_0 : \beta = 0$, the score vector $\mathbf{U} \equiv \mathbf{U}(\beta)$ and its variance $\mathbf{V} = \text{Cov}(\mathbf{U}|H_0) = \mathbf{V}_{\beta\beta} - \mathbf{V}_{\beta\alpha}\mathbf{V}_{\alpha\alpha}^{-1}\mathbf{V}_{\alpha\beta}$ can be extracted using

$$\begin{pmatrix} \mathbf{U}(\alpha) \\ \mathbf{U}(\beta) \end{pmatrix} = \sum_{i=1}^F \mathbf{U}_i \text{ and } \begin{pmatrix} \mathbf{V}_{\alpha\alpha} & \mathbf{V}_{\alpha\beta} \\ \mathbf{V}_{\beta\alpha} & \mathbf{V}_{\beta\beta} \end{pmatrix} = \sum_{i=1}^F \mathbf{U}_i \mathbf{U}_i', \quad (10)$$

and each $\mathbf{U}_i$ is obtained by

$$\mathbf{U}_i = (\nabla \mu_i') \left[\hat{\phi}_i^{1/2} \mathbf{R}_{wi}(\hat{\delta}) \hat{\phi}_i^{1/2} \right]^{-1} (\mathbf{y}_i - \hat{\mu}_i), \quad (11)$$

where all parameters are estimated under the null hypothesis and $\nabla \mu_i'$ is the partial derivative of μ_i with respect to $(\alpha', \beta')'$ (Liang and Zeger 1986). For binary traits with the canonical logit link function and $\hat{\phi} = 1$, we have

$$\mathbf{U}_i = \begin{pmatrix} \mathbf{X}_i' \\ \mathbf{G}_i' \end{pmatrix} \hat{\Delta}_i \left[\hat{\Delta}_i^{1/2} \mathbf{R}_{wi}(\hat{\delta}) \hat{\Delta}_i^{1/2} \right]^{-1} (\mathbf{y}_i - \hat{\mu}_i), \quad (12)$$

$$\hat{\Delta}_i = \text{diag}(\hat{\mu}_{ij}(1 - \hat{\mu}_{ij}), \dots, \hat{\mu}_{in_i}(1 - \hat{\mu}_{in_i})). \quad (13)$$

For quantitative traits with the canonical identity link function, we have $\hat{\gamma}_j$ as an identity matrix and $\mathbf{U}_j$ reduces to

$$\mathbf{U}_i = \begin{pmatrix} \mathbf{X}'_i \\ \mathbf{G}'_i \end{pmatrix} [\hat{\phi} \mathbf{R}_{w_i}(\hat{\delta})]^{-1} (\mathbf{y}_i - \hat{\mu}_i). \quad (14)$$

Results

Real Data

Data and Model Setup—Data were obtained from ongoing longitudinal studies on the development of substance abuse undertaken by the Minnesota Center for Twin and Family Research (MCTFR) (Iacono and McGue 2002). MCTFR participants include monozygotic and dizygotic twins, as well as adopted and non-adopted siblings and their parents. Analyses reported here are based on the alcohol dependence scores whose development is described by Hicks et al. (Hicks et al. 2011). Briefly, Hicks et al. used a factor analytic approach to derive a composite alcohol dependence symptoms scale that loaded primarily on DSM-III-R (the diagnostic system that was current at the time the MCTFR studies were initiated) symptoms of withdrawal and tolerance, social and occupational complications and compulsive drinking. Biometric analysis of the alcohol dependence scores yielded an additive genetic heritability estimate of 66% (95% CI of 60%-71%), suggesting that this trait is moderately to strongly heritable and so suitable for additional genetic analysis. Offsprings were assessed multiple times in MCTFR research but parents were assessed only once. Consequently, analysis reported here are based on offspring alcohol dependence scores obtained closest to their 17th birthday. Details of the genotyping of the MCTFR samples are provided by Miller et al. (Miller et al. 2012).

Using alcohol dependence as a trait for the study, we pre-processed our data as follows by ruling out the individuals with missing trait or covariates. As a result, a total of 7,230 individuals from 2,301 families were used for the analysis. Furthermore, we also dichotomized the trait using the 1st quartile (−0.028) as a cutoff (see Fig. 1). For covariates, we used gender, age, age², and an indicator of being a parent. The descriptive summary statistics of the covariates are provided in Table 1. For genetic variants, we excluded the SNPs if their minor allele frequencies were less than 0.05, and leading to 350,615 SNPs. For gene and pathway-level analyses, we used hg18 as a reference genome and included SNPs that are within 10Kbp upstream and downstream the gene coding region for each gene. As a result, a total of 21,613 genes were included for gene-level analysis. We used the KEGG database for pathway analysis and a total of 56,199 SNPs in 4698 genes out of 214 pathways were included. We added two variance components into the models. First, we estimated the genetic relationship matrix (GRM) using 20,000 randomly selected SNPs. We also considered a within-family (genetic and environmental) effect, which we call familial relationship throughout this paper. The familial relationship matrix was block diagonal, where each block consisted of 1s for each family.

As a primary analysis, we conducted the aSPU test and the aSPU_{path} test under GLMM. We used $\gamma \in \{1, 2, \dots, 8, \infty\}$ for the aSPU test and $\gamma \in \{1, 2, \dots, 8\}$ and $\gamma_G \in \{1, 2, 4, 8\}$ for the aSPU_{path} test. Our results were then compared with other gene-level score-based tests. We fitted two GEE models treating each family as a cluster under CS and independence working correlation matrices. As a reference, we also fitted GLMM and GEE models using the original quantitative trait (alcohol dependence prior to being dichotomized). For each model, we used a number of score-based tests: the score test and the UminP test as well as other SPU(γ) tests. Finally, we also conducted the SNP-level score test to see if any SNP was significant.

Gene-level Association Tests under GLMM

We first performed the aSPU test using GLMM. We fitted the null model using GMMAT and extracted the score vector for each gene. We report that the estimated variance component for genetic relatedness turned out to be 0, suggesting that genetic relationship was fully captured by the familial relationship. Using the family-wise significance level of 0.05 with Bonferroni adjustment, we used $0.05/21,613 \approx 2.3 \times 10^{-6}$ as a genome-wide significance level.

The aSPU test under GLMM detected one significant gene (MCHR1: Melanin-concentrating hormone receptor 1) in chromosome 22 as shown in Fig. 2. A number of previous articles reported the relatedness of MCH1 to stress, depression and anxiety (Hervieu 2003; 2006; Roy et al 2007). It turned out that SPU(1), SPU(2), SPU(3) and SPU(4) all detected its significance while the SPU tests with $\gamma > 4$ did not, suggesting that the association signals were not dominated by one or few SNPs. It is also illustrative in Fig. 3 that the SNP rs133074 retained the lowest p-value among the SNPs near the MCHR1 coding region but did not meet the genome-wide SNP significance-level (see below). We also underscore the adaptability of the aSPU test by noting three other genes that had p-values less than 1.00×10^{-4} (HSD17B14 in chromosome 19, CRYBB3 and KIAA1671 in chromosome 22), as these genes were detected by the SPU(γ) tests with γ greater than 3 but not by the SPU(1) or the SPU(2) test.

We also compared the aSPU test to the score test and the UminP test. Both the latter two tests did not detect any significant gene. Also, the p-values of the four genes mentioned above were larger in the UminP test. The score test notably yielded gene OR2S2 (on chromosome 9) with a p-value (1.44×10^{-5}) near the genome-wide significance level, while none of the SPU tests and the aSPU test showed any signal for the gene.

Gene-level Association Tests under GEE

The GLMM-based tests could be compared directly with GEE-based tests, as the variance components for genetic relatedness was 0 in our data. We thus fitted GEE models with two different working correlation matrices (CS and independence) treating each family as a cluster and compared the p-values to those from the GLMM.

The p-values under the GEE models were close to the p-values under the GLMM model for all the SPU(γ) tests and the aSPU test, as shown in Fig. 4. The gene MCHR1 was also

detected by the aSPU test under the GEE (CS) model, as well as the SPU(γ) tests when $1 \leq \gamma \leq 6$. In the GEE (Ind) model, only the SPU(1) test detected the gene MCHR1 while the other tests, including the aSPU test, did not. The similarity of the p-values under GLMM and GEE (CS) attributes to the same correlation structure within each family, because the estimated variance component for genetic related was 0.

SNP-level Association Tests under GLMM and GEE

We also performed the score test for each SNP, using the null GLMM model. Using the standard 5.00×10^{-8} as the genome-wide significance level for single SNP, the score test did not detect any significant SNP. However, it partially explained why the aSPU test, or generally the gene-level score-based tests, can be more useful. As shown in Table 2, four out of six SNPs encoded in MCHR1 had relatively low p-values, suggesting more significant results for the SPU(γ) tests with smaller values of γ . For genes CRYBB3 and KIAA1671, including 12 and 35 SNPs respectively, the SPU(γ) test with $\gamma = 8$ or ∞ (or the UminP test) showed more significant results, as small fractions of SNPs show high associations. The similar p-values of CRYBB3 and KIAA1671 in SPU(∞) and the UminP test were due to the fact that SNP rs1547435 is included in both genes. Similar arguments can be made to the GEE models, as the p-values for the SNP-level were close to the p-values under the GLMM models.

Pathway-level Association Test under GLMM

We conducted a pathway-level analysis using the aSPU_{path} test. We used a genome-wide significance level of $0.05/214 = 2.33 \times 10^{-4}$ with a Bonferroni adjustment for 214 KEGG pathways to be tested. We found one pathway (hsa04110) significantly associated with alcohol dependence (p-value = 8.00×10^{-5} by the aSPU_{path} test). Among the individual SPU_{path}(γ, γ_G) tests, only the SPU_{path}(1,1) test gave a p-value (7.00×10^{-5}) less than the significance level while all the others did not, suggesting a possible dense signal scenario (i.e. with many weakly associated SNPs/genes). It was also noticeable that the only pathway (hsa04080) containing gene MCHR1 had a p-value 5.41×10^{-3} . It can be partially explained by the weak or null associations from the 232 other genes in the pathway, because only 4 genes had a p-value less than 0.01 in the aSPU test. It suggests a sparse signal situation (i.e. with only few associated genes), as also confirmed by the results of the individual SPU_{path} tests, among which the SPU_{path}(8,8) test showed the lowest p-value.

Comparison with Tests using Quantitative Traits

We fitted the GLMM and the GEE models (with the same working correlation matrices) using the original non-dichotomized alcohol dependence as a trait. We conducted the SPU and the aSPU tests as well as the score and the UminP tests accordingly for each model. All the tests did not detect any significant gene except for the score test with GLMM; however, there was some evidence of the inflated Type I error, as shown in the QQ plot for the SNP level score test (Fig. 6) and its inflation factor $\lambda = 1.05$. Compared to GEE, GLMM depends heavily on the distributional (i.e. Gaussian) assumptions on the quantitative trait (see Fig. 6). Since the alcohol dependence score may not be normally distributed (Fig. 1), conducting a test based on the Gaussian assumption on the data could be questionable.

Simulation Studies

We conducted a small simulation study to compare the performances of the aSPU test under GLMM and GEE. It was based on the genes MCHR1 and KIAA1671 mentioned before, but we present the results based on MCHR1 here for clarity as the results based on KIAA1671 did not differ qualitatively. Using MCHR1 gene detected by the aSPU test above, we generated a simulated binary trait as follows:

$$\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \tau_G \mathbf{\Psi}_G + \tau_F \mathbf{\Psi}_F), \quad (15)$$

$$\mu_i^b = g^{-1}(\mathbf{X}_i \alpha + \theta \cdot \mathbf{G}_i \beta + \mathbf{b}_i), \quad (16)$$

$$y_{ij} \sim \text{Bernoulli}(\mu_{ij}^b). \quad (17)$$

We first simulated the random intercepts as described in (15). We used the same genetic relationship matrix ($\mathbf{\Psi}_G$) and the familial relationship matrix ($\mathbf{\Psi}_F$) as used in the real data analysis. We fixed the variance component for the familial relationship effects to be the same as the estimated values from the full model ($\tau_F = 1.34$). We also compared the results using five different values of variance components for genetic relatedness ($\tau_G \in \{0, 0.5, 1, 1.5, 5\}$).

We fixed the nominal significance level at 0.05 and decreased the simulation sample size by taking random samples of 1,000 families out of 2,301. We also controlled the effect size by introducing the rescaling factor θ as described in (16) to the effect sizes of the gene, where $\theta \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. Note that the empirical power becomes the empirical Type I error when $\theta = 0$. Finally, we used the replicates of simulations at 2,000 for estimating empirical Type I error and 200 for empirical power.

The simulation results are shown in Table 3. In GLMM all the tests controlled Type I error effectively regardless of the values of θ or τ_G . In GEE, there were slightly inflated Type I error rates. The estimated power was slightly higher in most GEE models than in GLMM, but the differences were marginal and partially due to the slightly inflated Type I error rates in GEE.

Computing

We used R extensively in this paper. Fitting a null GLMM in *GMMAT* took 1.8 hours for binary traits and 0.5 hour for quantitative traits. We used the `geeglm()` function from the *geepack* package for fitting a null GEE model, and it took less than 3 seconds. Time for computing the p-value of the aSPU test depends on the simulation size B and the length of the score vector (i.e. number of SNPs used for a gene or a pathway). Assuming the usual 0.05 family-wise significance level with a Bonferroni adjustment for less than 30,000 genes,

$B = 10^7$ would be sufficient to yield p-values small enough to detect statistically significant genes. Since most genes would not be significant, to save computing time, we followed Pan et al (2014) with a step-up procedure to data-adaptively determine the values of B : we started with a small $B = 10^3$ as an initial simulation size; if a p-value $> B/10$, then stop; if a calculated p-value was less than $10/B$, we increased B by 10 times before calculating its p-value again; we repeated the process until B reached 10^7 for gene-level analysis or $B = 10^5$ for pathway-level analysis. Using parallel computing with 200 cores and a 2.5GB memory limit per core, it took 0.3 hour and 8.9 hours to compute the p-values for all genes and pathways under GLMM, respectively.

Discussions

A main goal of this paper is to demonstrate how to perform gene- and pathway-level score-based association tests in a GLMM for family-based GWAS. In particular, we illustrated the use of the aSPU test and the aSPUpath test for binary and quantitative traits in GLMM and GEE. In the application to the Minnesota Twin Family Study data, we detected gene MCHR1 (melanin concentrating hormone receptor 1) to be significantly associated with alcohol dependence, which of course needs to be validated in the future.

We showed that the gene-level score-based tests can be more effective in detecting significant associations than SNP-level tests. In our application to the MTFs data, several gene-level score-based tests could detect a significant association under both GLMM and GEE, while the SNP-level association testing failed to uncover it. In particular, the aSPU test offers a powerful and general approach to gene-level analysis to family-based GWAS and other multiple-trait or longitudinal data (Wang et al 2017).

For the real data with the binary trait, we found that results under GLMM were pretty similar to the results under GEE with a compound symmetric working correlation matrix. It was likely due to the same covariance structure as the estimated variance component for genetic relatedness in GLMM ($\hat{\tau}_G$) was exactly zero (i.e. $\hat{\tau}_G = 0$), which can be explained by a few possibilities. First, the effect itself was actually small or did not exist as the estimated variance component without using a separate familial correlation structure was close to 0.001 (not shown here). It might be reasonable to argue that in many situations the effects of cryptic relatedness is not a major factor compared to the shared familial genetic and environmental effects. Note that the major part of the genetic relatedness could be captured by the familial relatedness. It is also possible due to estimation errors. It is known that the PQL estimates for variance components are biased, especially for correlated binary data (Breslow and Clayton 1993). One can alleviate the problem by using corrected PQL or second-order Laplace methods as discussed by Breslow and Lin (1995) and Lin and Breslow (1996), though neither is yet available in the GMMAT software. We conjecture that the non-optimal PQL estimates might have contributed to the diminishing advantage of using GLMM over GEE in our real data analysis and simulations.

We also compared the results using the quantitative trait. The score test under GLMM was the only test that detected at least one significant gene, but one has to be cautious due to the strong Gaussian assumptions on the trait that might be violated in the real data, which was

partially confirmed by the slightly inflated genomic control factor in the QQ plot of the SNP-level score test.

To summarize, we believe the aSPU and the aSPU_{path} tests under GLMM is competitive for family-based GWAS. First, the GLMM approach offers a general framework to account for cryptic relatedness as well as other population structures. We here emphasize the importance of controlling cryptic relatedness in GWAS, as its importance was pointed out by Devlin and Roeder (Devlin and Roeder 1999). The GLMM approach can be even adopted in population-based GWAS when the relatedness is not explicit and suspected. Also, conducting the aSPU and the aSPU_{path} tests along with fitting a single null GLMM model with GMMAT was computationally feasible, and the data-adaptiveness of the aSPU and the aSPU_{path} can boost the power in detecting more associated genes and pathways while controlling Type I error rates effectively. The GEE approach, on the other hand, only considers within-family correlations and cannot account for cryptic relatedness. Also, the GLMM can be applied to large pedigree data with large families and a small number of families, for which the GEE approach will not be effective.

Finally, the aSPU and aSPU_{path} tests are implemented in R package aSPU2 that is publicly available at <https://github.com/ChongWu-Biostat/aSPU2>; the example R code, including extracting the score vector and its covariance matrix from the GMMAT, is available on WP's website at <http://www.biostat.umn.edu/weip/prog.html>.

Acknowledgments

The authors are grateful to two reviewers and an editor for many helpful comments.

Funding: This research was supported by National Institutes of Health grants R37DA05147, R01AA09357, R01AA11886, R01DA13240, U01DA024417, R01MH066140, R01GM113250, R01HL105397, R01HL116720, and by the Minnesota Supercomputing Institute.

References

- Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *J Amer Statist Assoc.* 1993; 88(421):9–25.
- Breslow NE, Lin X. Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika.* 1995; 82(1):81–91.
- Chen H, Wang C, Conomos MP, Stilp AM, Li Z, Sofer T, Szpiro AA, Chen W, Brehm JM, Celedn JC, Redline S, Papanicolaou GJ, Thornton TA, Laurie CC, Rice K, Lin X. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *Am J Hum Genet.* 2016; 98(4):653–666. [PubMed: 27018471]
- Devlin B, Roeder K. Genomic control for association studies. *Biometrics.* 1999; 55(4):997–1004. [PubMed: 11315092]
- Harville DA. Maximum likelihood approaches to variance component estimation and related problems. *J Amer Statist Assoc.* 1977; 72(358):320–340.
- Hervieu G. Melanin-concentrating hormone functions in the nervous system: food intake and stress. *Expert Opin Ther Targets.* 2003; 7(4):495–511. [PubMed: 12885269]
- Hervieu GJ. Further insights into the neurobiology of melanin-concentrating hormone in energy and mood balances. *Expert Opin Ther Targets.* 2006; 10(2):211–229. [PubMed: 16548771]
- Hicks BM, Schalet BD, Malone SM, Iacono WG, McGue M. Psychometric and Genetic Architecture of substance use disorder and behavioral disinhibition measures for gene association studies. *Behav Genet.* 2011; 41(4):459–475. [PubMed: 21153693]

- Iacono WG, McGue M. Minnesota twin family study. *Twin Res.* 2002; 5(5):482–487. [PubMed: 12537881]
- Jiang YD, Chiu CY, Yan Q, Chen W, Gorin MB, Conley YP, Lakhal-Chaieb ML, Cook RJ, Amos CI, Wilson AF, Bailey-Wilson JE, Xiong MM, Weeks DE, Fan RZ. Gene-based association testing of dichotomous traits with generalized functional linear mixed models using extended pedigrees. Under review. 2017
- Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* 2010; 38:D355D360. [PubMed: 19880382]
- Kim J, Zhang Y, Pan W. Powerful and adaptive testing for multi-trait and multi-SNP associations with GWAS and sequencing data. *Genetics.* 2016; 203(2):715–731. [PubMed: 27075728]
- Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP. A powerful and flexible multilocus association test for quantitative traits. *Am J Hum Genet.* 2008; 82(2):386–397. [PubMed: 18252219]
- Liang K, Zeger S. Longitudinal data analysis using generalized linear models. *Biometrika.* 1986; 73(1):13–22.
- Lin X, Breslow NE. Bias correction in generalized linear mixed models with multiple components of dispersion. *J Amer Statist Assoc.* 1996; 91(435):1007–1016.
- Miller MB, Basu S, Cunningham J, Eskin E, Malone SM, Oetting WS, Schork N, Sul JH, Iacono WG, McGue M. The Minnesota center for twin and family research genome-wide association study. *Twin Research and Human Genetics.* 2012; 15(6):767–774. [PubMed: 23363460]
- Pan W. Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. *Genet Epidemiol.* 2011; 35(4):211–216. [PubMed: 21308765]
- Pan W, Kim J, Zhang Y, Shen X, Wei P. A powerful and adaptive association test for rare variants. *Genetics.* 2014; 197(4):1081–1095. [PubMed: 24831820]
- Pan W, Kwak I, Wei P. A powerful and pathway-based adaptive test for genetic association with common or rare Variants. *Am J Hum Genet.* 2015; 97(1):86–98. [PubMed: 26119817]
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* 2006; 38(8): 904–909. [PubMed: 16862161]
- Roy M, David N, Cueva M, Giorgetti M. A study of the involvement of melanin-concentrating hormone receptor 1 (MCHR1) in murine models of depression. *Biol Psychiatry.* 2007; 61(2):174–180. [PubMed: 16934771]
- Wang Z, Xu K, Zhang X, Wu X, Wang Z. Longitudinal SNP-set association analysis of quantitative phenotypes. *Genet Epidemiol.* 2017; 41:81–93. [PubMed: 27859628]
- Wessel J, Schork NJ. Generalized genomic distance-based regression methodology for multilocus association analysis. *Am J Hum Genet.* 2006; 79(5):792–806. [PubMed: 17033957]
- Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet.* 2010; 86(6):929–942. [PubMed: 20560208]
- Xu Z, Pan W. Approximate score-based testing with application to multivariate trait association analysis. *Genet Epidemiol.* 2015; 39(6):469–479. [PubMed: 26198454]
- Zhang Y, Xu Z, Shen X, Pan W. Alzheimer's Disease Neuroimaging Initiative. Testing for association with multiple traits in generalized estimation equations, with application to neuroimaging data. *Neuroimage.* 2014; 96:309–325. [PubMed: 24704269]

Fig. 1. Histogram of the alcohol dependence score in the Minnesota Twin Family Study.

Fig. 2. Manhattan plots of gene-level score-based tests using the binary trait and GLMM

Fig. 3. LocusZoom plot for the region near MCHR1 using the single SNP-based score test under GLMM.

(a) Comparison between GLMM and GEE (CS)

(b) Comparison between GLMM and GEE (Ind)

Fig. 4.
Scatter plots of the $-\log_{10}(p\text{-value})$ of the aSPU test under GLMM versus that under GEE.

Fig. 5. Manhattan plots of the SNP-level score tests using the (a) binary trait or (b) quantitative trait under GLMM.

(a) GLMM, binary trait ($\lambda = 0.96$) (b) GLMM, quantitative trait ($\lambda = 1.05$) (c) GEE (CS), binary trait ($\lambda = 1.01$) (d) GEE (CS), quantitative trait ($\lambda = 1.02$) (e) GEE (Ind), binary trait ($\lambda = 1.01$) (f) GEE (Ind), quantitative trait ($\lambda = 1.03$)**Fig. 6.**

QQ plots of the SNP-level score tests using the (a) binary trait or (b) quantitative trait under and GEE.

Table 1

Descriptive summary of the covariates.

	Lower Dependence	Higher Dependence	Total
Participants	1616	5614	7230
Parents (%)	423 (26.2)	3529 (62.9)	3952 (54.7)
Female (%)	1114 (68.9)	2753 (49.0)	3867 (53.5)
Age (SD)	24.68 (12.21)	33.72 (13.21)	31.70 (13.53)

Table 2
Summary of the 15 lowest p-values from the SNP-level score test under GLMM on chromosome 22.

SNP	Position	Corresponding gene(s)	Aref	Aalt	p-value		
					GLMM	GEE (CS)	GEE (Ind)
rs1547435	23919792	CRYBB3, KIAA1671	G	A	1.28×10 ⁻⁶	7.28×10 ⁻⁷	3.45×10 ⁻⁶
rs133074	39408419	MCHR1	A	G	8.47×10 ⁻⁶	3.03×10 ⁻⁶	9.55×10 ⁻⁶
rs133076	39411507	MCHR1	A	G	3.66×10 ⁻⁵	2.18×10 ⁻⁵	6.05×10 ⁻⁵
rs763279	23907688	KIAA1671	A	G	5.97×10 ⁻⁵	4.63×10 ⁻⁵	1.06×10 ⁻⁴
rs138312	39532082	MIR4766, SLC25A17	A	G	8.28×10 ⁻⁵	4.89×10 ⁻⁵	5.64×10 ⁻⁴
rs4822568	23915047	KIAA1671	A	G	2.07×10 ⁻⁴	1.52×10 ⁻⁴	2.55×10 ⁻⁴
rs165793	19477293	PI4KA, SERPIND1	A	G	4.55×10 ⁻⁴	4.32×10 ⁻⁴	6.73×10 ⁻⁴
rs739314	23928101	CRYBB3, KIAA1671	C	A	4.79×10 ⁻⁴	3.43×10 ⁻⁴	4.04×10 ⁻⁴
rs136070	45724074	TBC1D22A	A	G	5.71×10 ⁻⁴	6.11×10 ⁻⁴	4.98×10 ⁻⁴
rs4820378	38199155	MGAT3	G	A	7.50×10 ⁻⁴	4.48×10 ⁻⁴	4.21×10 ⁻⁴
rs2064165	45703857	TBC1D22A	G	A	8.32×10 ⁻⁴	9.08×10 ⁻⁴	7.67×10 ⁻⁴
rs80533	39415915	MCHR1	A	G	9.79×10 ⁻⁴	5.51×10 ⁻⁴	1.86×10 ⁻³
rs929020	46349705	(NA)	A	C	1.03×10 ⁻³	4.71×10 ⁻⁴	6.58×10 ⁻⁴
rs1569517	45699784	TBC1D22A	A	G	1.05×10 ⁻³	9.57×10 ⁻⁴	1.05×10 ⁻³
rs133067	39404114	MCHR1	G	A	1.05×10 ⁻³	4.85×10 ⁻⁴	7.36×10 ⁻⁴

Table 3

The empirical Type I error rate (Scale=0) and power (Scale 0) of the aSPU test and the SSU (i.e. SPU(2)) test using different values of variance components for genetic relatedness and of scaled effect sizes of the MCHR1 gene, based on 2,000 (Scale=0) or 200 (Scale 0) independent replicates.

		aSPU					SSU				
		τ_G					τ_G				
Scale (θ)		0	0.5	1	1.5	5	0	0.5	1	1.5	5
GLMM	0	0.047	0.047	0.045	0.047	0.051	0.045	0.045	0.046	0.046	0.050
	0.2	0.105	0.100	0.090	0.075	0.085	0.100	0.100	0.065	0.065	0.080
	0.4	0.190	0.210	0.185	0.225	0.140	0.165	0.190	0.175	0.230	0.120
	0.6	0.390	0.410	0.375	0.380	0.320	0.345	0.390	0.355	0.355	0.280
	0.8	0.590	0.725	0.570	0.560	0.465	0.565	0.680	0.540	0.555	0.425
	1.0	0.870	0.795	0.835	0.770	0.610	0.845	0.790	0.800	0.765	0.580
GEE (CS)	0	0.057	0.054	0.051	0.058	0.055	0.052	0.047	0.055	0.055	0.050
	0.2	0.125	0.110	0.095	0.085	0.090	0.095	0.110	0.075	0.075	0.090
	0.4	0.200	0.220	0.210	0.245	0.170	0.200	0.190	0.190	0.215	0.145
	0.6	0.425	0.455	0.390	0.415	0.325	0.395	0.420	0.360	0.365	0.310
	0.8	0.615	0.740	0.580	0.585	0.475	0.615	0.715	0.550	0.575	0.425
	1.0	0.890	0.820	0.840	0.785	0.625	0.870	0.795	0.825	0.770	0.615
GEE (Ind)	0	0.054	0.054	0.053	0.057	0.057	0.051	0.050	0.055	0.053	0.054
	0.2	0.120	0.120	0.105	0.085	0.095	0.115	0.110	0.075	0.070	0.085
	0.4	0.215	0.225	0.205	0.250	0.145	0.200	0.195	0.205	0.225	0.135
	0.6	0.425	0.435	0.380	0.370	0.340	0.375	0.410	0.365	0.360	0.310
	0.8	0.605	0.715	0.590	0.565	0.475	0.570	0.705	0.550	0.555	0.430
	1.0	0.855	0.820	0.830	0.785	0.600	0.835	0.800	0.805	0.775	0.585