Published in final edited form as:

Genet Epidemiol. 2009 April ; 33(3): 217–227. doi:10.1002/gepi.20372.

Bivariate Association Analyses for the Mixture of Continuous and Binary Traits with the Use of Extended Generalized Estimating Equations

Jianfeng Liu^{1,2}, Yufang Pei^{1,2,4}, Chris J. Papasian², and Hong-Wen Deng^{1,2,3,4}

¹Department of Orthopedic Surgery, School of Medicine, University of Missouri-Kansas City, 2411 Holmes Street, Kansas City, MO 64108

²Department of Basic Medical Science, School of Medicine, University of Missouri-Kansas City, 2411 Holmes Street, Kansas City, MO 64108

³Laboratory of Molecular and Statistical Genetics, College of Life Sciences, Hunan Normal University, Changsha, Hunan 410081, P. R. China

⁴The Key Laboratory of Biomedical Information Engineering of Ministry of Education and Institute of Molecular Genetics, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an, Shanxi 710049, P.R. China

Abstract

Genome-wide association (GWA) study is becoming a powerful tool in deciphering genetic basis of complex human diseases/traits. Currently, the univariate analysis is the most commonly used method to identify genes associated with a certain disease/phenotype under study. A major limitation with the univariate analysis is that it may not make use of the information of multiple correlated phenotypes, which are usually measured and collected in practical studies. The multivariate analysis has proven to be a powerful approach in linkage studies of complex diseases/traits, but it has received little attention in GWA. In this study, we aim to develop a bivariate analytical method for GWAS, which can be used for a complex situation that a continuous trait and a binary trait measured are under study. Based on the modified extended generalized estimating equation (EGEE) method we proposed herein, we assessed the performance of our bivariate analyses through extensive simulations as well as real data analyses. In the study, to develop an EGEE approach for bivariate genetic analyses, we combined two different generalized linear models corresponding to phenotypic variables using a Seemingly Unrelated Regression (SUR) model. The simulation results demonstrated that our EGEE-based bivariate analytical method outperforms univariate analyses in increasing statistical power under a variety of simulation scenarios. Notably, EGEE-based bivariate analyses have consistent advantages over univariate analyses whether or not there exists a phenotypic correlation between the two traits. Our study has practical importance, as one can always use multivariate analyses as a screening tool when multiple phenotypes are available, without extra costs of statistical power and false positive rate. Analyses on empirical GWA data further affirm the advantages of our bivariate analytical method.

INTRODUCTION

Genome-wide association (GWA) studies offer a powerful tool for identifying genes conferring modest disease risks [Risch and Merikangas 1996]. Currently, statistical methods developed for GWA data analyses were largely for single phenotypes. However, in practice, investigators often collect a set of correlated phenotypes that may share common environmental and/or genetic factors. The multivariate analysis is a natural way to take all phenotypes into consideration simultaneously. Multivariate analyses have been widely adopted in linkage studies [Allison, et al. 1998; Huang and Jiang 2003; Jiang and Zeng 1995; Lange and Whittaker 2001; Liu, et al. 2007a; Williams, et al. 1999], and there is a general consensus that multivariate analyses outperform univariate analyses in increasing statistical power and precision of parameter estimation. However, in GWA, multivariate analyses have not received sufficient attention, although there are some sporadic reports on methodology development of multi-trait analyses [Lange, et al. 2003; Lange, et al. 2002; Verzilli, et al. 2005]. In particular, as to the common situation when both binary (*e.g.*, disease status) and continuous data (*e.g.*, quantitative measures that are risk factors for the disease under study) are involved, little work has been done to evaluate the performance of the multivariate analytical method.

For a mixture of correlated phenotypes involving non-normal traits (such as binary or ordinal data), it is not straightforward to specify a full probability model, making it difficult to directly implement likelihood-based approaches [Lange and Whittaker 2001; Lange, et al. 2002]. To surmount this problem, several methods were developed [Catalano 1997; Gueorguieva and Agresti 2001; Gueorguieva and Sanacora 2006; Regan and Catalano 2000] based on the threshold model assumption, which transform a discrete phenotype to a “latent” normal variable. However, the potential concerns with these methods are the arbitrary and untestable distributional assumptions on the “latent” variable, and the difficulty with computational features of model fitting [Prentice and Zhao 1991].

Since the seminal work of Liang and Zeger [1986], a suite of methods [Hall 2001; Hall and Severini 1998; Liang, et al. 1992; Prentice and Zhao 1991; Zhao and Prentice 1990] based on the Generalized Estimating Equation (GEE) have been developed to address the situation of non-normally distributed variables in multivariate data analyses. GEE can be regarded as a multivariate version of generalized linear model (GLM), a method with reasonable statistical efficiency to analyze correlated data with non-normal distributions. A vast of literature on GEE theory and application has arisen over the past two decades [Zeger and Liang 1992; Ziegler, et al. 1998; Zorn 2001]. The most commonly used GEE-based approaches, referred to as GEE1 and GEE2, were introduced by Liang and Zeger [1986] and Zhao and Prentice [1990], respectively.

GEE1 and GEE2 have their respective limitations. In GEE1, regression and association parameters are presumed to be orthogonal which may not necessarily true, and the association parameter is treated as a nuisance [Liang and Zeger 1986; Liang, et al. 1992]. Because the relationship between the first- and second-order moment parameters is ignored, GEE1 has less efficient estimators than GEE2 if the working covariance model is correctly specified [Hall and Severini 1998; Liang, et al. 1992]. In contrast to GEE1, GEE2 removes the assumption of “orthogonality” initially imposed on GEE1. Thus, GEE2 can perform estimation for regression and association parameters simultaneously under a unified framework, leading to a significant increase in efficiency of parameter estimation. However, a limitation of GEE2 is that it may lose the consistency of regression parameter estimators if the covariance structure is misspecified [Hall 2001; Hall and Severini 1998; Liang, et al. 1992; Zhao and Prentice 1990].

To improve the performance of GEE1 and GEE2, an alternative GEE-based approach, named Extended Generalized Estimating Equation (EGEE), was proposed by Hall and Severini [1998]. EGEE was developed using the ideas of extended quasi-likelihood [McCullagh and Nelder 1989; Nelder and Pregibon 1987]. It retains many of the advantages of the original GEE1 and GEE2 methods while avoiding their potential limitations: (1) EGEE can offer consistent estimations of both regression and association parameters even under an incorrect covariance model; (2) EGEE is not required to assume third- and fourth-order moment “working” models which are necessary in GEE2; and (3) the estimation efficiency of EGEE is comparable to GEE2 and often outperforms GEE1 [Hall 2001; Hall and Severini 1998].

In the present study, using EGEE, being an improved GEE-based approach, we aim to develop bivariate gene/phenotype association analyses for population-based data when both continuous and binary traits are involved in the same studies. Since the original EGEE [Hall 2001; Hall and Severini 1998] cannot deal with mixed continuous and binary outcomes in a straightforward fashion, we modified the EGEE method for such situations. Our strategy is motivated by the usage of Seemingly Unrelated Regression (SUR) [Zellner 1962]. SUR procedure is an extension of ordinary least squares analyses for dealing with the system of linear equations with correlated error terms, which can be treated as a special case of generalized least squares analyses. Based on the idea of SUR models, two GLMs with different link functions can be incorporated into a unified equation system. Consequently, we developed a “specific” form of EGEE, which is well suited for the mixture of two phenotypes with different distributions.

We applied our modified EGEE method to assess the performance of bivariate analyses through extensive simulation studies. The results demonstrated that, compared with univariate analyses, bivariate analyses can substantially improve power while having comparable false-positive rates under almost all the scenarios simulated. The advantages of our EGEE-based bivariate analyses method was further validated in empirical data analyses using our genome-wide association scan data for osteoporosis and obesity. Given the emerging widespread GWA, our proposed method may have practical significance in the genetic epidemiology field in general.

METHODS

We begin with the situation in which N unrelated random subjects are recruited in a population-based genetic association study, each having observations on two different types of phenotypes: one is in normal distribution (named T_1) and the other is a binary trait (named T_2). A 2×1 random observation vector for individual i ($i = 1, \dots, N$) is denoted as $\mathbf{y}_i = (y_{i1}, y_{i2})'$ with the mean vector $\boldsymbol{\mu}_i = E(\mathbf{y}_i) = (\mu_{i1}, \mu_{i2})'$ and the 2×2 (co)variance matrix $\text{var}(\mathbf{y}_i) = \mathbf{V}_i$. Let $\mathbf{x}_i$ be a $L \times 1$ vector of explanatory variables including genetic makers as well as other fixed effects (*e.g.*, age, sex) corresponding to individual i . We assume $\mathbf{x}_i$ affects both phenotypes through corresponding L -dimensional regression parameter vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. Similarly, we set a $2L \times 1$ vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$. Here, the subscripts 1 and 2 relate the corresponding variables to traits T_1 and T_2 respectively. Within the GLM framework [McCullagh and Nelder 1989], using an identity link (for T_1) and a logit link (for T_2), we model the relationship between the explanatory variables and the marginal means of the two phenotypes as follows:

$$g_1(\mu_{i1}) = \mu_{i1} = \mathbf{x}_i' \boldsymbol{\beta}_1, \quad [1]$$

$$g_2(\mu_{i2}) = \ln\left(\frac{\mu_{i2}}{1 - \mu_{i2}}\right) = \mathbf{x}_i' \beta_2. \quad [2]$$

Here, for ease of incorporating the two different types of data into an overall EGEE analysis, we employ SUR model [Zellner 1962] to combine Equations [1] and [2] into a unified framework:

$$\mathbf{g}(\mu_i) = \mathbf{X}_i' \beta, \quad [3]$$

where $\mathbf{X}_i = \begin{pmatrix} \mathbf{x}_i & 0 \\ 0 & \mathbf{x}_i \end{pmatrix}$, and $\mathbf{g}(\mu_i)$ is a compound function vector in the form of $(g_1(\mu_{i1}), g_2(\mu_{i2}))'$.

Following the GLM theory, the respective variance functions of two distributions, *i.e.*, the normal and the binary, are given as:

$$v_1(\mu_1) = 1, \quad \text{and} \quad v_2(\mu_{i2}) = \mu_{i2}(1 - \mu_{i2}), \quad [4]$$

where μ_{i1} and μ_{i2} are variables of the two variance functions corresponding to the two phenotypes. Note that we have $v_1(\bullet) \equiv 1$ here because of a normal distribution chosen by the trait T_1 .

In the framework of GLM, probability distributions of the random variables are usually parameterized in terms of not only the means (e.g., μ_{i1} and μ_{i2} herein) but also dispersion parameters. Accordingly, we denote the dispersion parameters for different distributions of T_1 and T_2 as ψ_1 and ψ_2 , respectively. Since in the binary outcome context there exists no overdispersion [Hall 2001], we will always have $\psi_2 \equiv 1$.

For specifying the marginal (co)variance model $\mathbf{V}_i$, we further define the “working”

correlation matrix $\mathbf{R}(\xi) = \begin{pmatrix} 1 & \xi \\ \xi & 1 \end{pmatrix}$, where ξ is the unknown association parameter falling within $(-1.0, 1.0)$. It is noted that the regressors involved in the SUR model are allowed to have possibly different effects on either of the two regressands considered, with residuals from each regression assumed to be correlated [Verzilli, et al. 2005].

Given the variance functions, dispersion parameters and working correlation matrix defined above, the (co)variance matrix $\mathbf{V}_i$ between the two phenotypes for individual i has the form similar to those given by Rochon [1996] and Lange *et al.* [2002].

$$\mathbf{V}_i = \begin{pmatrix} \psi_1 & 0 \\ 0 & \psi_2 \end{pmatrix}^{\frac{1}{2}} \begin{pmatrix} v_1(\mu_{i1}) & 0 \\ 0 & v_2(\mu_{i2}) \end{pmatrix}^{\frac{1}{2}} \begin{pmatrix} 1 & \xi \\ \xi & 1 \end{pmatrix} \begin{pmatrix} v_1(\mu_{i1}) & 0 \\ 0 & v_2(\mu_{i2}) \end{pmatrix}^{\frac{1}{2}} \begin{pmatrix} \psi_1 & 0 \\ 0 & \psi_2 \end{pmatrix}^{\frac{1}{2}}. \quad [5]$$

On the basis of Equation [4] and the equality $\psi_2 \equiv 1$, Equation [5] can be fully described as:

$$\mathbf{V}_i = \begin{pmatrix} \psi_1 & \xi \sqrt{\psi_1 \mu_{i2}(1 - \mu_{i2})} \\ \xi \sqrt{\psi_1 \mu_{i2}(1 - \mu_{i2})} & \mu_{i2}(1 - \mu_{i2}) \end{pmatrix}. \quad [6]$$

From Equation [6], it is obvious that $\mathbf{V}_i$ is determined by $\boldsymbol{\beta}$, ζ and ψ_1 . Since ψ_1 should be restricted as positive, we make a “squared” transform, $\psi_1 = \varphi^2$, such that the estimate of ψ_1 can fall into its space. Accordingly, Equation [6] is rewritten as:

$$\mathbf{V}_i = \begin{pmatrix} \varphi^2 & \xi\varphi\sqrt{\mu_{i2}(1-\mu_{i2})} \\ \xi\varphi\sqrt{\mu_{i2}(1-\mu_{i2})} & \mu_{i2}(1-\mu_{i2}) \end{pmatrix}. \quad [7]$$

For ease of presentation, we combine the second-order moment parameters in a 2×1 vector fashion, $\tilde{\alpha} = (\xi, \varphi)'$.

To examine whether the marker locus is associated with either of the two phenotypes within the EGEE framework, we take two steps (named in turn **Estimation Step** and **Testing Step**) on the basis of the SUR regression model and the marginal (co)variance model given in Equations [3] and [7] respectively.

Estimation Step

From the EGEE perspective, the unknown parameters, including the regression vector $\boldsymbol{\beta}$ and the association vector $\tilde{\alpha}$ aforementioned, can be estimated in a set of estimating equations. Here, we directly adopt the same form of the EGEEs as given in the original study [Hall 2001]:

$$\sum_{i=1}^N U_i(\boldsymbol{\beta}, \tilde{\alpha}) = \sum_{i=1}^N \begin{pmatrix} \mathbf{D}_i' & 0 \\ 0 & \mathbf{F}_i' \end{pmatrix} \begin{pmatrix} \mathbf{V}_i^{-1} & 0 \\ 0 & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{y}_i & -\mu_i \\ \mathbf{s}_i & -\sigma_i \end{pmatrix} = 0, \quad [8]$$

where, the $2 \times 2L$ matrix $\mathbf{D}_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}'$, and the 4×2 matrix $\mathbf{F}_i = \partial \{\text{vec} \mathbf{V}_i^{-1}\} / \partial \tilde{\alpha}'$. The operator $\text{vec}(\bullet)$ here is a function for generating a column vector via stacking the columns (or rows) of the argument matrix. The 4×1 vectors $\mathbf{s}_i$ and $\boldsymbol{\sigma}_i$ are defined for estimating the second-order moment parameters, where $\mathbf{s}_i = \text{vec}\{(\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)'\}$ and $\boldsymbol{\sigma}_i = E(\mathbf{s}_i) = \text{vec} \mathbf{V}_i$.

Obviously, Equation [8] regarding the unknowns $\boldsymbol{\beta}$ and $\tilde{\alpha}$ is hard to resolve straightforwardly due to its nonlinear property. Alternatively, a Fisher scoring algorithm may offer a natural way for obtaining the estimates of $\boldsymbol{\beta}$ and $\tilde{\alpha}$. Starting with initial values $\boldsymbol{\beta}^{(0)}$ and $\tilde{\alpha}^{(0)}$, the updated parameter estimates $\boldsymbol{\beta}^{(l)}$ and $\tilde{\alpha}^{(l)}$ in the l th iteration are given by

$$\begin{pmatrix} \boldsymbol{\beta}^{(l)} \\ \tilde{\alpha}^{(l)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}^{(l-1)} \\ \tilde{\alpha}^{(l-1)} \end{pmatrix} + \left(\sum_{i=1}^N \begin{pmatrix} \mathbf{D}_i'^{(l-1)} & 0 \\ 0 & \mathbf{F}_i'^{(l-1)} \end{pmatrix} \begin{pmatrix} \mathbf{V}_i^{-1(l-1)} & 0 \\ 0 & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{D}_i^{(l-1)} & 0 \\ 0 & \mathbf{F}_i^{(l-1)} \end{pmatrix} \right)^{-1} \left(\sum_{i=1}^N U_i(\boldsymbol{\beta}^{(l-1)}, \tilde{\alpha}^{(l-1)}) \right). \quad [9]$$

The above iteration is repeated until convergence. In our study, we set a sufficiently rigid convergence criterion as

$$\left[\left(\boldsymbol{\beta}^{(l)} - \boldsymbol{\beta}^{(l-1)}, \tilde{\alpha}^{(l)} - \tilde{\alpha}^{(l-1)} \right) \left(\boldsymbol{\beta}^{(l)} - \boldsymbol{\beta}^{(l-1)}, \tilde{\alpha}^{(l)} - \tilde{\alpha}^{(l-1)} \right)' \right] / (2L+2) \leq 10E-12.$$

Accordingly, the updated values of $\boldsymbol{\beta}$ and $\tilde{\alpha}$ in the latest iteration will be treated as the solution of the EGEEs given by Equation [8], which are denoted as $\tilde{\boldsymbol{\beta}}$ and $\tilde{\tilde{\alpha}}$, respectively.

Accordingly, a sandwich-type estimate of the asymptotic (co)variance matrix of $\tilde{\beta}$ and $\tilde{\alpha}$, $\text{var}(\tilde{\beta}, \tilde{\alpha})$, is given by

$$\text{var}(\tilde{\beta}, \tilde{\alpha}) = \left\{ U^* \left(\tilde{\beta}, \tilde{\alpha} \right) \right\}^{-1} \left\{ \sum_{i=1}^N U_i \left(\tilde{\beta}, \tilde{\alpha} \right) U_i \left(\tilde{\beta}, \tilde{\alpha} \right)' \right\} \left\{ U^* \left(\tilde{\beta}, \tilde{\alpha} \right) \right\}^{-1}, \quad [10]$$

$$\text{where } U^* \left(\beta, \tilde{\alpha} \right) = \sum_{i=1}^N \begin{pmatrix} \mathbf{D}_i' & 0 \\ 0 & \mathbf{F}_i' \end{pmatrix} \begin{pmatrix} \mathbf{V}_i^{-1} & 0 \\ 0 & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{D}_i & 0 \\ \partial \sigma_i / \partial \beta' & \partial \sigma_i / \partial \tilde{\alpha} \end{pmatrix}.$$

Testing Step

We employ a Wald chi-square statistic to test whether the marker considered has an effect on either the continuous phenotype or the binary trait. With the assumption that the distribution of genotypes at the SNP locus is in Hardy-Weinberg equilibrium, the EGEEs for individual i can include the allele-based index variable $x_{i,\mathbf{m}}$ with the regression coefficients $\beta_{m,1}$ and $\beta_{m,2}$ on traits 1 and 2 respectively. Under the null hypothesis that the SNP has no effect on either of the two traits, an asymptotic chi-square statistic with 2 degree of freedom (df) can be expressed as

$$\begin{aligned} \chi_{SNP-EGEE}^2 &= \left(\widehat{\beta}_{m,1}, \widehat{\beta}_{m,2} \right) \text{var} \left(\widehat{\beta}_{m,1}, \widehat{\beta}_{m,2} \right)^{-1} \left(\widehat{\beta}_{m,1}, \widehat{\beta}_{m,2} \right)' \\ &= \left(\widehat{\beta}_{m,1}, \widehat{\beta}_{m,2} \right) \begin{pmatrix} \text{Var}(\widehat{\beta}_{m,1}) & \text{Cov}(\widehat{\beta}_{m,1}, \widehat{\beta}_{m,2}) \\ \text{Cov}(\widehat{\beta}_{m,1}, \widehat{\beta}_{m,2}) & \text{Var}(\widehat{\beta}_{m,2}) \end{pmatrix}^{-1} \left(\widehat{\beta}_{m,1}, \widehat{\beta}_{m,2} \right)' \sim \xi_2^2, \end{aligned} \quad [11]$$

where $\widehat{\beta}_{m,1}$ and $\widehat{\beta}_{m,2}$ are the estimates of $\beta_{m,1}$ and $\beta_{m,2}$ respectively, which are obtained from the estimate vector $\widehat{\beta}$ in the **Estimation Step**. Similarly, $\text{var}(\widehat{\beta}_{m,1}, \widehat{\beta}_{m,2})$, the (co)variance matrix between $\widehat{\beta}_{m,1}$ and $\widehat{\beta}_{m,2}$, is a sub-matrix of the matrix $\text{var}(\tilde{\beta}, \tilde{\alpha})$ given by Equation [10], where only those rows and columns regarding $\widehat{\beta}_{m,1}$ and $\widehat{\beta}_{m,2}$ are selected.

Given a significant threshold such as $\alpha = 0.05$, if the calculated p -value for $\chi_{SNP-EGEE}^2$ is lower than α , we can declare that the SNP locus affects either or both of the two phenotypes. It is worth noting that we mainly focus on identifying the potential SNP effects on the traits through constructing the statistic $\chi_{SNP-EGEE}^2$ based on the estimates of regression parameters (*i.e.*, $\widehat{\beta}_{m,1}$ and $\widehat{\beta}_{m,2}$). A similar statistic can be easily developed in the same way to test the significance of the association parameter ζ if one is substantially interested in further exploring the correlation between two traits.

For comparison with univariate analyses, we separately detect SNP-phenotype associations based on traditional linear regression analysis (for T_1) and logistic regression analysis (for T_2). The respective p -values are denoted as p_1 and p_2 . Under this situation, we calculated the adjusted significance level (denoted as α_c) for multiple testing by two different methods, Bonferroni correction and M_{eff} method [Nyholt 2004]. The originally M_{eff} method is proposed to adjust for multiple testing of SNPs in linkage disequilibrium (LD) based on the

spectral decomposition (SpD) of matrices of pairwise LD between SNPs. Here we directly apply M_{eff} method to correct for multiple testing of two correlated phenotypes. Accordingly, we declare a significant association if $\min(p_1, p_2) < \alpha_c$.

SIMULATION STUDY

DESIGN OF THE SIMULATION EXPERIMENTS

To evaluate the performance of bivariate association analyses based on our developed EGEE method, we conduct extensive simulations for joint analyses of a range of mixtures of one continuous trait and one binary trait in different scenarios.

Under an additive model, we assume a biallelic locus under consideration has effects on both traits, *i.e.*, T_1 with normal distribution and T_2 in a binary pattern, within a random population. Firstly, for ease of describing the genetic effect on the binary phenotype, we adopt the penetrance function f_i as used in Liu *et al* [2007b], which is the penetrance of the genotype with i copies of causal allele ($i = 0, 1$ or 2). Similarly, we define the ratio of f_1/f_0 as heterozygote genotype relative risk (GRR). Then we generate the marker genotypes and the two types of phenotypic observations for each individual under a suite of parameter settings, including sample size N (150 vs. 300), frequency of the liability allele q (0.05 vs. 0.15), the residual error correlation r (-0.45, -0.15, 0, 0.15 vs. 0.45) between two traits, QTL

heritability h_1^2 (0, 0.01 vs. 0.03) for trait 1, and heterozygote GRR (1.0, 1.5 vs. 2.0) for trait 2. Under the above settings, a total of 180 simulation scenarios are generated by combining these settings at different levels. For a simple illustration, we set the residual error variance $\sigma_{e,1}^2 = 1.0$ for trait 1. As to trait 2, we assume f_0 , *i.e.*, the population incidence among individuals carrying zero copy of liability allele, being 0.1, and the ratio of cases and controls of the sample is set to 1:1 for each simulation scenario. According to the above parameter settings, two related phenotypes, both affected by the same locus as well as influenced by the environmental factors, can be generated via the following steps:

1. Two alleles are sampled and paired to form a random individual according to the liability allele frequency q . Following the theory of quantitative genetics [Falconer and Mackay 1996], under an additive model, the simulated genotype effect on T_1 can be expressed as

$$a_{1,j} = \begin{cases} -\sqrt{\frac{h_1^2}{2q(1-q)(1-h_1^2)}} & (j=0) \\ 0 & (j=1) \\ \sqrt{\frac{h_1^2}{2q(1-q)(1-h_1^2)}} & (j=2) \end{cases} \quad [12]$$

where j is the number of liability alleles assigned to the simulated subject.

2. We assume the locus simulated above also underlies a latent normal variable z_2 with mean $a_{2,j}$ and variance $\sigma_{e,2}^2 = 1$, and the residual error correlation between T_1 and z_2 is set to r .

The random variable z_2 can be truncated to generate the binary trait T_2 using a threshold of t . Based on our earlier study [Deng, et al. 2000], $a_{2,j}$ and the threshold t can be determined via the aforementioned simulation parameters, including heterozygote GRR and the penetrance function f_0 . To be concrete, we provide a detailed description in the APPENDIX section on how to bridge the gap between

normal variable z_2 and binary data T_2 through heterozygote GRR and f_0 as well as the threshold t .

3. For the random individual i , we draw the related variables T_1 and z_2 conditionally on its genotype $\mathbf{G}_j$ from the bivariate normal distribution

$$\begin{pmatrix} T_1 \\ z_2 \end{pmatrix} \sim N \left(\begin{pmatrix} a_{1,j} \\ a_{2,j} \end{pmatrix}, \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} \right).$$
 Accordingly, the respective binary phenotype T_2 is truncated to a case ($z_2 \geq t$) or a control ($z_2 < t$).
4. Through steps (1) ~ (3), we can obtain a random individual with a specific marker genotype as well as both the normal and binary phenotypic observations with correlation coefficient r . To ensure the ratio of cases and controls involved in the binary data T_2 , we repeat these processes till $0.5N$ cases and $0.5N$ controls are formed.

Under each of the specific simulation scenarios, we generate 20,000 independent samples for obtaining the robust estimates of statistical power and false-positive rates by the EGEE-based bivariate association analyses as well as by traditional univariate association analyses. For both bivariate and univariate analyses, we define the identical null hypothesis $\mathbf{H}_0$ as: the SNP marker has no association with either T_1 or T_2 , i.e., $h_1^2=0$ and heterozygote $GRR=1$. The corresponding alternative hypothesis $\mathbf{H}_a$ is designed as: the SNP marker is associated with either or both of the two traits, i.e., $h_1^2>0$ or/and heterozygote $GRR>1$. Accordingly, the power and the false-positive rate can be calculated as the proportions of significant SNP/phenotype associations reported in 20,000 independent replicates with the true $\mathbf{H}_a$ and $\mathbf{H}_0$, respectively.

RESULTS OF THE SIMULATION EXPERIMENTS

According to the combinations of various parameter settings for simulations, we perform extensive simulations involving 160 scenarios under the hypothesis $\mathbf{H}_a$ as well as 20 scenarios under the hypothesis $\mathbf{H}_0$, defined above. To study the performance of detecting liability alleles based on the bivariate test, we conduct three association tests for each set of simulated phenotypes employing the EGEE-based bivariate analytical method (denoted as **B_EGEE**) and separate univariate analyses adjusted for multiple testing by the Bonferroni correction (denoted as **U_b**) and the M_{eff} method [Nyholt 2004] (denoted as **U_e**). For all tests, we set the experiment-wise significance threshold $\alpha = 0.05$.

Tables 1 and 2 present the power estimates of gene-phenotype associations with sample sizes of 150 and 300, respectively, for both bivariate and univariate association analytical methods. Under each setting, we highlight the maximal power in bold to emphasize the best performance among the three tests. It is obvious that the largest increases in power are consistently observed with **B_EGEE**, with the exception of only 3 data sets. This shows that the EGEE-based bivariate association test is highly preferable to the individual univariate tests adjusted either by the Bonferroni correction or by the M_{eff} method.

Comparing the power of **B_EGEE** across different levels of correlation coefficient r between the two traits, we can see that the performance of bivariate analyses varies along with r in the wide range of situations investigated. The effects of residual error correlation on the performance of **B_EGEE** are twofold. On one hand, for the situation in which only one phenotype, either T_1 or T_2 , is associated with the locus, the estimated power of **B_EGEE** has a trend of increasing with a rising absolute value of r from 0 to 0.45, regardless of its direction. Note that we set $r > 0$ or $r < 0$ herein, corresponding to the situations where it has either the identical or opposite direction, respectively, as that of correlation induced by the simulated pleiotropic QTL. On the other hand, under the scenario

in which both phenotypes are associated with the gene considered, the performance of **B_GEE** is obviously influenced by both degree and direction of correlation r . Specifically, when the residual error correlation is opposite in sign to the correlation induced by QTL effects, **B_EGEE** has an increasing gain in power as the degree of r increases; otherwise, the power of **B_EGEE** decreases with the increase of r . The above trend of the power of **B_EGEE** is repeated over various design combinations of liability allele frequency and effect of QTL as well as sample size.

In Tables 1 and 2, comparing the performance of univariate tests with the M_{eff} method and the Bonferroni correction, we can clearly see that the former consistently outperforms the latter across all those simulation scenarios with $r > 0$. Furthermore, the higher degree r has, the more obvious the advantage of M_{eff} . Since M_{eff} may take non-independence between correlated phenotypes into consideration, it reasonably overcomes the conservativeness of the Bonferroni correction while having the ease of computation.

Results from Tables 1 and 2 further demonstrate that the highest power levels for both bivariate and univariate association analyses are achieved under the combinational design of the higher liability allele frequency at $q = 0.15$, a larger sample size at $N = 300$ and the highest QTL effects on both phenotypes. This trend logically complies with the routine power profiles in genetic association studies.

Table 3 offers the results of false-positive rate calculations for the three tests, *i.e.*, **B_EGEE**, **U_b** and **U_e**, under various design combinations at the significance level of 0.05. Overall, all estimated levels of false positive rates for both bivariate and univariate analyses are close to the threshold value of 0.05. This aspect further suggests that the increased power in **B_EGEE** is due primarily to the fact that it considers all phenotype data within a unified test framework rather than focusing on information of each single phenotype separately.

Table 3 also shows that **B_EGEE** has a more stringent false-positive rate control for lower levels of both allele frequency r and sample size N . This result demonstrates that the convergence aspect of **B_EGEE** estimates toward their asymptotic distribution is affected, to some extent, by the sample characteristic mentioned above. For the two univariate analyses, it is obvious that **U_b** is slightly more conservative than **U_e** (when $r \neq 0$) since the former ignores the correlation between tests of the two phenotypes.

ANALYSES OF EMPIRICAL GWA DATA

THE DATA

We perform real data analyses using our available GWA datasets for osteoporosis and obesity. Osteoporosis is a major public health problem characterized by low bone mineral density (BMD) and increased risk of fracture. Obesity is a condition of extra body weight defined as body mass index (BMI, weight/height^2) $\geq 30 \text{ kg/m}^2$. The relationship between osteoporosis and obesity has been widely studied, with conflicting results [Guney, et al. 2003; Radak 2004; Zhao, et al. 2008; Zhao, et al. 2007]. We recently found that osteoporosis and obesity share common genetic factors that may have inverse effects on BMD and BMI. To date, we are not aware of any studies investigating the relationship between obesity and low-trauma osteoporotic fractures (OF), the endpoint of osteoporosis.

Our GWA data are obtained using the Affymetrix 500K arrays on 700 Chinese Han subjects (350 with osteoporotic hip fractures and 350 controls) (Y. Guo, et. al., unpublished data). Among these, 598 also have BMI data.

To further validate the proposed method by assessing its empirical performance, we perform bivariate GWA using **B_EGEE** on the 598 subjects for which information is available for

both OF and BMI. Before the association test, a suite of quality control procedures are performed. First, for the initial full-set of 500,568 SNPs, missing genotypes at each locus are imputed using HelixTree 5.3.1 (Golden Helix, Bozeman, MT), based on Expectation-Maximization (EM) algorithm. Second, we discard SNPs with a call rate $< 95\%$ (10,467 SNPs), those deviating from Hardy-Weinberg equilibrium (HWE) ($P < 0.001$, 21,274 SNPs), and those having a minor allele frequency (MAF) < 0.05 in the total sample (141,945 SNPs). Eventually, 326,882 SNPs are available for subsequent analyses. At each locus, the allele-based association test is performed, and age, sex and age² are included as covariates to adjust their respective effects on raw observations in the models for the **B_EGEE** bivariate analyses as well as respective univariate analyses.

ANALYSES RESULTS

We test whether either of the two phenotypes is associated with SNPs across the whole genome. For ease of presentation, we only select the top 100 SNPs with the strongest association signals, *i.e.*, the lowest p -values, from the **E_GEE**-based bivariate analyses. The profile of the p -values (in $-\log$ scale) is plotted in **Fig. 1**. For comparison, for these 100 SNPs, the respective raw p -values of univariate analyses and the p -values adjusted by Bonferroni correction for multiple traits are also presented in **Figs. 1a** and **1b** respectively. We also scrutinize the top 100 most significant SNPs based on the adjusted p -values of two univariate analyses. We find that there is a high overlap between these two sets of top 100 SNPs (89 out of 100), suggesting that the comparisons between the two methods are reasonable in terms of the top 100 SNPs chosen by **B_EGEE**.

Fig. 1a depicts the p -value profiles of three different analyses, *i.e.*, two separate traditional univariate analyses for the continuous trait BMI and the binary trait OF and **EGEE**-based bivariate analyses for the mixture of the both traits. It can be seen that the majority of the SNPs (66 out of 100) show the strongest association signals in **B_EGEE** analyses even compared with the raw p -values in univariate analyses. For the remaining SNPs (34 out of 100), the p -values of **B_EGEE** fall between the two raw p -values of the univariate analyses, but are closer to the smaller and more significant ones. The trend from **Fig. 1a** suggests that bivariate association analyses have practical significance. Specifically, when two or more traits are involved in GWA, we may experience the situations shown in **Fig. 1a**, *i.e.*, some causal variants with modest effects may show moderate signals (e.g., $p = 1E-3$) in univariate analyses while showing much stronger signals (such as $p = 1E-7$) in bivariate analyses. If only univariate associations are performed, these signals will probably not reach genome-wide significance. However, if a bivariate GWA screen is performed initially, some of these loci may achieve a genome-wide significance level. In the second stage, using separate univariate analyses to further confirm the association identified in bivariate GWA, we only need to adjust raw p -values of each univariate analysis for the number of phenotypes and specific handful markers screened out in the initial bivariate analyses. Therefore, those SNPs with moderate signals in univariate analyses may remain significant as the issue of multiple testing is minimized.

Fig. 1b further provides comparisons between the p -values of **E_GEE** statistic in bivariate analyses and the corresponding adjusted p -values for multiple testing of two phenotypes in univariate analyses. It shows that, with the exception of three loci, the p -value of **B_EGEE** statistic at each SNP locus is apparently lower than the corresponding adjusted p -value of univariate analyses. This further supports the conclusion that the performance of **B_EGEE** proposed herein greatly outperforms univariate analyses in studies in which multiple phenotypes are involved and multiple testing is adjusted for the number of phenotypes involved.

In addition, we selected 10 most significant SNPs based on p -values of **B_EGEE** statistic for our bivariate GWA data in Table 4. These identified SNPs could be treated as the potential genetic factors which are associated with the two traits investigated, i.e., BMI and OF. Due to the relatively small sample size of our GWA data, all of the association signals in our analysis did not reach the genome-wide significance threshold, such as 4.2×10^{-7} proposed by Lencz, *et al.* [2007]. However, our preliminary findings herein may offer evidence for further follow-up genetic studies on osteoporosis and obesity.

DISCUSSION

In this study, we have developed a bivariate association analytical method based on EGEE for GWA studies. To the best of our knowledge, this is the first study to develop specific methods designed for investigating the performance of bivariate genetic association analyses for population-based GWA data. The extensive simulation experiments and GWA real data analyses consistently demonstrated the advantages of our method over univariate analyses.

Here we consider a complex but common situation where both continuous and discrete phenotypic observations are available, for which, however, no existing analytical methods or public software are readily available. In this study, using a SUR model, we incorporated two GLMs, corresponding to different phenotypic distributions, into the same set of estimating equations; this was followed by parameter estimation and statistical testing based on the EGEE method. As pointed out by Hall and colleagues [Hall 2001; Hall and Severini 1998], EGEE generally performs better in estimation of both regression and association parameters compared to either GEE1 or GEE2 approaches. However, the original EGEE was developed for analyzing multiple related phenotypes with the same type of distributions that have the same dispersion parameter, such as longitudinal data obtained in clinical trials. The novelty of our modified EGEE is that it relaxes the constraint of the original EGEE and can handle response variables having different distributions or dispersion parameters. Hence, our modified EGEE method is well suited for bivariate association analyses on the mixture of a normal distributed continuous trait and a binary trait.

The advantages of multivariate analyses over univariate analyses may lie in two general aspects: (1) Multivariate analyses can incorporate information of correlation between traits into a single test statistic, which renders a gain in power. (2) Multivariate analyses can avoid or alleviate the issue of multiple testing. This becomes particularly important in circumstances in which the traits under investigation can be divided into several subgroups [Lange, et al. 2003]. Another similar situation is that statistical significance (especially when modest) of a potential causal locus may diminish if adjusted for the number of loci in a univariate GWA. However, it is likely that this locus will remain significant if genome-wide multivariate association analysis is performed and the genome-wide significance can be achieved, as described in the section of ANALYSES OF EMPIRICAL GWA DATA.

Two important aspects of this study worth further emphasis. First, in the circumstance in which both phenotypes are affected by the locus, the power of bivariate analyses is affected by both degree and direction of the residual correlation r ; a larger power increase is always favored by the residual correlation r with opposite direction to the QTL-induced correlation. This is in agreement with studies of Jiang and Zeng [1995] and Allison *et al* [1998] for bivariate linkage studies. Further theoretical interpretations on our finding herein can trace to a recently report [Evans 2002], which pinpointed why the sign of correlations between variables affect the power of bivariate analyses. The rationale lies in that the opposite sign between QTL-induced and residual correlations can result in a corresponding positive term involved in the expression of non-centrality parameter (NCP), increasing the magnitude of NCP and the power; otherwise a negative term of NCP arises, decreasing the NCP and the

power. As a result, the opposite direction of the correlations always contributes to the greater increase in power in contrast to the correlations with the same direction. Although the above interpretations are originally for bivariate linkage analyses, it should work in the similar way in association studies from statistical perspective. Second, even in the circumstance in which two phenotypes do not share genetic determinations at the locus investigated or residual effects, bivariate analyses still have obvious advantages over univariate analyses across various simulated scenarios due to alleviating the issue of multiple testing. This has also been observed in multivariate family-based association studies [Lange, et al. 2003]. These two features make the **B_EGEE** procedure an ideal screening tool for detecting gene-phenotype association no matter whether the phenotypic correlation between traits exists or not.

In **B_EGEE** method, we employed a Fisher scoring algorithm to estimate the unknown parameters due to nonlinear property of the equation system. The program implementing **B_EGEE** method is written in FORTRAN language, which is available on our website <http://l.web.umkc.edu/liujian/>. A potential concern is that our method may entail prohibitively high computational demand for conducting the iterative calculation. To illustrate that the proposed method is computationally practical for GWA studies, we assessed the CPU time required by the program in simulation and empirical GWA data analyses. All the analyses were carried out on a computer with Intel® Pentium® 4 3.4GHz dual processors and 2.0GB RAM. It took 55hrs 16mins to complete simulation analyses for all 180 simulated scenarios and 14hrs 22mins for GWA data analyses. This indicates that the computation time required for simulation and empirical data analyses is acceptable, and thus our method is practical for bivariate analyses in the field of candidate gene or GWA studies.

For simplicity, we only conducted bivariate analyses for two correlated traits, with the purpose of demonstrating the advantages of **B_EGEE** over univariate analyses. It should be noted that our method can be easily extended to the situation of multiple correlated traits (>2) with more complex distributions, *e.g.*, Poisson distribution, Gamma distribution, etc. The basic idea of this extension is rather straightforward. One only needs to set up the respective GLM for each phenotypic variable via a specific link function according to its distribution feature [McCullagh and Nelder 1989]. Similar to **B_EGEE**, these GLMs then can be combined into a unified framework through the SUR model.

Another potential extension of **B_EGEE** is to incorporate haplotype information across multiple loci in the models. It has been shown that the haplotype-based strategy may be preferable in some situations [Akey, et al. 2001; Liu, et al. 2007b]. To test phenotype-haplotype association for multiple traits, the substantial modification needed for **B_EGEE** is to construct a new design matrix $\mathbf{X}_i$ which includes the index variables corresponding to the two haplotypes carried by individual i . For uncertainty of haplotypes from phase-unknown data, we may borrow the idea of haplotype trend regression [Zaykin, et al. 2002] in univariate analysis. We may incorporate the posterior probabilities of all uncertainties into the haplotype design matrix rather than most likely haplotype configurations. This will help to avoid a loss of information and potential bias in subsequent analyses [Liu, et al. 2007b].

As verified in simulation studies, Bonferroni correction is more conservative than the M_{eff} method when performing multiple testing corrections. Even so, in univariate analyses of real GWA data, we did not present the M_{eff} -adjusted p -values for multiple testing of the number of traits involved. The main reasons are as follows. On one hand, the M_{eff} method requires the correlation coefficients among phenotypic variables to estimate an effective number of independent variables [Nyholt 2004]. However, it is unfeasible to calculate the correlation between the continuous observation and the binary disease status using the traditional approach for our real data set, although the correlation coefficient is directly given as a

parameter setting in the simulation studies; On the other hand, we borrowed the estimate of the association parameter ζ obtained in **B_EGEE** as an alternative of correlation coefficient between two phenotypic variables. At 100 most significant loci, the estimates of ζ vary from 0.016 through 0.023. Given these estimates of ζ , we implemented the M_{eff} method to estimate the effective number of independent tests, and the results range within the interval [1.9995, 1.9997]. This means the number of independent phenotypes is quite close to the true number of phenotype. This suggests that M_{eff} -adjusted p -values are almost the same as those by Bonferroni correction. Hence, to simplify our presentation, we just presented p -value comparisons between the proposed **B_EGEE** test and the adjusted univariate analyses using Bonferroni correction rather than using the M_{eff} method in Fig. 1b.

For population-based association studies, population stratification may cause false positive/negative results [Weiss, et al. 2001]. In both simulation and real data analyses conducted herein using **B_EGEE**, we did not consider or control population stratification. However, we point out that this can be done using the methods developed and commonly used for univariate analysis, including Structure [Pritchard, et al. 2000] and Principal Component methods [Price, et al. 2006]. For example, we can incorporate estimates of individual ancestry [Price, et al. 2006; Pritchard, et al. 2000] as covariates in each GLM. It has been shown that use of individual ancestry as a covariate is a promising way for adjusting the effects due to population stratification at most marker loci except those with very strong differentiation even among closely related populations and ethnic groups [Chen and Abecasis 2007]. In univariate association analysis, another way to control population stratification is to adjust test statistics using the genomic control method [Roeder, et al. 2006]; however, the direct use of this method for multiple dependent variables is not straightforward. We will explore the feasibility of extending the genomic control approach to multivariate analyses in further endeavors.

Acknowledgments

We are grateful to the anonymous reviewers for their insightful comments and constructive suggestions that greatly improved our manuscript. We thank Dr. Yongjun Liu for his help in revising our earlier manuscript and valuable discussions which led to improvements in the paper. Investigators of this work were partially supported by grants from NIH (R01 AR050496-01, R21 AG027110, R01 AG026564, and P50 AR055081). The study also benefited from grants from National Science Foundation of China, Huo Ying Dong Education Foundation, HuNan Province, Xi'an Jiaotong University, and the Ministry of Education of China.

APPENDIX

How a binary observation is determined by a latent normal distribution variable

Assume a SNP locus with two alleles: A_1 and A_2 , where A_2 is the risk allele for the binary trait T_2 . To keep consistent with previous description, we adopt the same symbols $a_{2,j}$ ($j = 0, 1, 2$) as before to denote the genotype values corresponding to three genotypes A_1A_1 , A_1A_2 and A_2A_2 . Under the additive model, we have

$$a_{2,1}=0 \quad \text{and} \quad a_{2,0} = -a_{2,2}. \quad [\text{A1}]$$

For the latent normal variable z_2 , we assume it follows a mixture of normal distribution $z_2 \sim N(a_{2,j}, 1)$ ($j = 0, 1, 2$) conditionally on the genotype carried by each individual. Accordingly, we have the following expressions according to the theory of threshold theory:

$$f_0 = \Pr(T_2 = 1 | A_1 A_1) = \Pr(z_2 > t | A_1 A_1) = \Phi(a_{2,0} - t), \quad [A2]$$

$$f_1 = f_0 \cdot GRR = \Pr(T_2 = 1 | A_1 A_2) = \Pr(z_2 > t | A_1 A_2) = \Phi(a_{2,1} - t) = \Phi(-t), \quad [A3]$$

$$f_2 = \Pr(T_2 = 1 | A_2 A_2) = \Pr(z_2 > t | A_2 A_2) = \Phi(a_{2,2} - t), \quad [A4]$$

where, $\Phi(\bullet)$ is the cumulative distribution function of a standard normal variable.

Given the value of penetrance function f_0 for the genotype $A_1 A_1$ and heterozygote GRR , the status of the binary variable T_2 can be determined conditionally by the genotype of each individual through the respective penetrance function f_i based on the above Equations $A1 \sim A4$.

REFERENCES

- Akey J, Jin L, Xiong M. Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *Eur J Hum Genet.* 2001; 9(4):291–300. [PubMed: 11313774]
- Allison DB, Thiel B, Jean P, Elston RC, Infante MC, Schork NJ. Multiple phenotype modeling in gene-mapping studies of quantitative traits: power advantages. *Am J Hum Genet.* 1998; 63(4): 1190–201. [PubMed: 9758596]
- Catalano PJ. Bivariate modelling of clustered continuous and ordered categorical outcomes. *Stat Med.* 1997; 16(8):883–900. [PubMed: 9160486]
- Chen WM, Abecasis GR. Family-based association tests for genomewide association scans. *Am J Hum Genet.* 2007; 81(5):913–26. [PubMed: 17924335]
- Deng HW, Chen WM, Recker RR. QTL fine mapping by measuring and testing for Hardy-Weinberg and linkage disequilibrium at a series of linked marker loci in extreme samples of populations. *Am J Hum Genet.* 2000; 66(3):1027–45. [PubMed: 10712216]
- Evans DM. The power of multivariate quantitative-trait loci linkage analysis is influenced by the correlation between variables. *Am J Hum Genet.* 2002; 70(6):1599–602. [PubMed: 11992269]
- Falconer, DS.; Mackay, TFC. *Introduction to Quantitative Genetics.* Longmans Green; Harlow, Essex, UK: 1996.
- Georguevieva RV, Agresti A. A Correlated Probit Model for Joint Modeling of Clustered Binary and Continuous Responses. *Journal of the American Statistical Association.* 2001; 96(455):1102–1112.
- Georguevieva RV, Sanacora G. Joint analysis of repeatedly observed continuous and ordinal measures of disease severity. *Stat Med.* 2006; 25(8):1307–22. [PubMed: 16217846]
- Guney E, Kisakol G, Ozgen G, Yilmaz C, Yilmaz R, Kabalak T. Effect of weight loss on bone metabolism: comparison of vertical banded gastroplasty and medical intervention. *Obes Surg.* 2003; 13(3):383–8. [PubMed: 12841898]
- Hall DB. On the Application of Extended Quasi-Likelihood to the Clustered Data Case. *The Canadian Journal of Statistics.* 2001; 29(1):77–97.
- Hall DB, Severini TA. Extended generalized estimating equations for clustered data. *Journal of the American Statistical Association.* 1998; 93(444):1365–1375.
- Huang J, Jiang Y. Genetic linkage analysis of a dichotomous trait incorporating a tightly linked quantitative trait in affected sib pairs. *Am J Hum Genet.* 2003; 72(4):949–60. [PubMed: 12644967]
- Jiang C, Zeng ZB. Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics.* 1995; 140(3):1111–27. [PubMed: 7672582]

- Lange C, Silverman EK, Xu X, Weiss ST, Laird NM. A multivariate family-based association test using generalized estimating equations: FBAT-GEE. *Biostatistics*. 2003; 4(2):195–206. [PubMed: 12925516]
- Lange C, Whittaker JC. Mapping quantitative trait Loci using generalized estimating equations. *Genetics*. 2001; 159(3):1325–37. [PubMed: 11729173]
- Lange C, Whittaker JC, Macgregor AJ. Generalized estimating equations: A hybrid approach for mean parameters in multivariate regression models. *Statistical Modelling*. 2002; 2(3):163–181.
- Lencz T, Morgan TV, Athanasiou M, Dain B, Reed CR, Kane JM, Kucherlapati R, Malhotra AK. Converging evidence for a pseudoautosomal cytokine receptor gene locus in schizophrenia. *Mol Psychiatry*. 2007; 12(6):572–80. [PubMed: 17522711]
- Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986; 73(1):13–22.
- Liang K-Y, Zeger SL, Qaqish B. Multivariate Regression Analyses for Categorical Data. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1992; 54(1):3–40.
- Liu J, Liu Y, Liu X, Deng HW. Bayesian mapping of quantitative trait loci for multiple complex traits with the use of variance components. *Am J Hum Genet*. 2007a; 81(2):304–20. [PubMed: 17668380]
- Liu J, Papasian C, Deng HW. Incorporating single-locus tests into haplotype cladistic analysis in case-control studies. *PLoS Genet*. 2007b; 3(3):e46. [PubMed: 17381242]
- McCullagh, PJ.; Nelder, JA. *Generalized Linear Models*. Chapman and Hall; London: 1989.
- Nelder JA, Pregibon D. An extended quasi-likelihood function. *Biometrika*. 1987; 74:221–231.
- Nyholt DR. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet*. 2004; 74(4):765–9. [PubMed: 14997420]
- Prentice RL, Zhao LP. Estimating Equations for Parameters in Means and Covariances of Multivariate Discrete and Continuous Responses. *Biometrics*. 1991; 47(3):825–839. [PubMed: 1742441]
- 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):904–9. [PubMed: 16862161]
- Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. *Am J Hum Genet*. 2000; 67(1):170–81. [PubMed: 10827107]
- Radak TL. Caloric restriction and calcium's effect on bone metabolism and body composition in overweight and obese premenopausal women. *Nutr Rev*. 2004; 62(12):468–81. [PubMed: 15648822]
- Regan MM, Catalano PJ. Regression models and risk estimation for mixed discrete and continuous outcomes in developmental toxicology. *Risk Anal*. 2000; 20(3):363–76. [PubMed: 10949415]
- Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science*. 1996; 273(5281):1516–7. [PubMed: 8801636]
- Rochon J. Analyzing Bivariate Repeated Measures for Discrete and Continuous Outcome Variables. *Biometrics*. 1996; 52(2):740–750. [PubMed: 8672710]
- Roeder K, Bacanu SA, Wasserman L, Devlin B. Using linkage genome scans to improve power of association in genome scans. *Am J Hum Genet*. 2006; 78(2):243–52. [PubMed: 16400608]
- Verzilli CJ, Stallard N, Whittaker JC. Bayesian modelling of multivariate quantitative traits using seemingly unrelated regressions. *Genet Epidemiol*. 2005; 28(4):313–25. [PubMed: 15789447]
- Weiss ST, Silverman EK, Palmer LJ. Case-control association studies in pharmacogenetics. *Pharmacogenomics J*. 2001; 1(3):157–8. [PubMed: 11908747]
- Williams JT, Van Eerdewegh P, Almasy L, Blangero J. Joint multipoint linkage analysis of multivariate qualitative and quantitative traits. I. Likelihood formulation and simulation results. *Am J Hum Genet*. 1999; 65(4):1134–47. [PubMed: 10486333]
- Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum Hered*. 2002; 53(2):79–91. [PubMed: 12037407]
- Zeger SL, Liang KY. An overview of methods for the analysis of longitudinal data. *Stat Med*. 1992; 11(14-15):1825–39. [PubMed: 1480876]

- Zellner A. An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias. *Journal of the American Statistical Association*. 1962; 57(298):348–368.
- Zhao LJ, Jiang H, Papasian CJ, Maulik D, Drees B, Hamilton J, Deng HW. Correlation of obesity and osteoporosis: effect of fat mass on the determination of osteoporosis. *J Bone Miner Res*. 2008; 23(1):17–29. [PubMed: 17784844]
- Zhao LJ, Liu YJ, Liu PY, Hamilton J, Recker RR, Deng HW. Relationship of obesity with osteoporosis. *J Clin Endocrinol Metab*. 2007; 92(5):1640–6. [PubMed: 17299077]
- Zhao LP, Prentice RL. Correlated binary regression using a quadratic exponential model. *Biometrika*. 1990; 77(3):642–648.
- Ziegler A, Kastner C, Blettner M. The Generalised Estimating Equations: An Annotated Bibliography. *Biometrical Journal*. 1998; 40(2):115–139.
- Zorn CJW. Generalized Estimating Equation Models for Correlated Data: A Review with Applications. *American Journal of Political Science*. 2001; 45(2):470–490.

Fig. 1. Profiles of the smallest p -values of top 100 SNPs in the EGEE-based bivariate genome-wide association study for the continuous trait BMI and the binary trait OF

For comparison, both raw and adjusted p -values for separate univariate association analyses are also presented. For bivariate and univariate analyses, $-\log_{10}$ of p -values are plotted against the ranked SNPs by respective p -values in a descending fashion.

1-a: Profiles of the p -values corresponding to three different analyses, including two separate traditional univariate analyses for the continuous trait BMI and the binary trait OF and the **EGEE**-based bivariate analysis for the mixture of the both traits;

1-b: Profiles of p -values from two different sources, *i.e.*, **E_GEE** statistic in bivariate analyses as well as the corresponding adjusted p -value for multiple testing for two phenotypes in univariate analyses. At each locus, the adjusted p -value is obtained from the raw p -values of two univariate analyses by Bonferroni correction.

Table 1

Power estimates of association tests for the mixture of the continuous and binary traits with the use of EGEE-based bivariate analyses and univariate analyses under the significance level $\alpha = 0.05$ and the sample size $N = 150$.

Allele Frequency q	Correlation Coefficient r	Analytical Method ^a	QTL Effects (h_1^2 & heterozygote GRR) ^b											
			0 & 1.5	0 & 2.0	0.01 & 1.0	0.03 & 1.0	0.01 & 1.5	0.01 & 2.0	0.03 & 1.5	0.03 & 2.0				
0.05	0.45	B_EGEE	0.190	0.345	0.245	0.560	0.274	0.370	0.530	0.573				
		U_b	0.120	0.235	0.168	0.456	0.219	0.304	0.475	0.515				
		U_e	0.129	0.249	0.180	0.475	0.232	0.323	0.492	0.533				
	0.15	B_EGEE	0.180	0.325	0.238	0.523	0.303	0.419	0.560	0.633				
		U_b	0.118	0.239	0.174	0.455	0.227	0.322	0.490	0.547				
		U_e	0.118	0.241	0.176	0.457	0.228	0.324	0.491	0.548				
0.15	0	B_EGEE	0.171	0.314	0.237	0.519	0.332	0.444	0.595	0.674				
		U_b	0.118	0.237	0.176	0.452	0.231	0.332	0.497	0.552				
		U_e	0.118	0.237	0.176	0.452	0.231	0.332	0.497	0.552				
	-0.15	B_EGEE	0.181	0.313	0.239	0.532	0.355	0.486	0.628	0.716				
		U_b	0.119	0.232	0.168	0.460	0.231	0.336	0.504	0.568				
		U_e	0.119	0.234	0.169	0.462	0.232	0.338	0.506	0.571				
0.45	-0.45	B_EGEE	0.191	0.343	0.269	0.581	0.443	0.592	0.722	0.806				
		U_b	0.117	0.232	0.169	0.457	0.244	0.352	0.515	0.592				
		U_e	0.126	0.246	0.179	0.476	0.259	0.371	0.535	0.611				
	0.45	B_EGEE	0.300	0.608	0.206	0.517	0.322	0.577	0.536	0.698				
		U_b	0.244	0.534	0.173	0.460	0.311	0.558	0.518	0.680				
		U_e	0.259	0.553	0.188	0.478	0.329	0.574	0.527	0.697				
0.15	0.15	B_EGEE	0.270	0.570	0.191	0.479	0.3713	0.619	0.603	0.777				
		U_b	0.240	0.537	0.174	0.461	0.3276	0.578	0.555	0.721				
		U_e	0.241	0.538	0.175	0.463	0.3288	0.579	0.556	0.723				
	0	B_EGEE	0.272	0.560	0.196	0.473	0.4043	0.661	0.640	0.814				
		U_b	0.243	0.530	0.180	0.459	0.3366	0.591	0.564	0.734				
		U_e	0.243	0.530	0.1809	0.459	0.3366	0.591	0.564	0.734				
-0.15	-0.15	B_EGEE	0.273	0.567	0.199	0.482	0.4487	0.700	0.692	0.854				
		U_b	0.244	0.534	0.178	0.456	0.3463	0.600	0.578	0.749				

Allele Frequency q	Correlation Coefficient r	Analytical Method ^a	QTL Effects (h_1^2 & heterozygote GRR) ^b							
			0 & 1.5	0 & 2.0	0.01 & 1.0	0.01 & 1.5	0.01 & 2.0	0.03 & 1.5	0.03 & 2.0	0.750
-0.45		U_e	0.245	0.535	0.180	0.458	0.3477	0.602	0.581	0.750
		B_EGEE	0.297	0.613	0.216	0.532	0.5580	0.807	0.8118	0.932
		U_b	0.237	0.537	0.168	0.461	0.3600	0.624	0.606	0.779
		U_e	0.253	0.555	0.181	0.478	0.3820	0.644	0.629	0.796

^aNOTE.— **B_EGEE** is the EGEE-based bivariate analytical method; **U_b** and **U_e** are two different univariate analytical methods adjusted for multiple testing by Bonferroni correction and the M_{eff} method.[Nyholt 2004] respectively.

^bQTL effects are denoted by h_1^2 (for the continuous trait T_1) and heterozygote GRR (for the binary trait T_2) simultaneously. Simulation scenarios with $h_1^2=0$ or $GRR = 1$ mean only one of the two traits is affected by the QTL.

Table 2

Power estimates of association tests for the mixture of the continuous and binary traits with the use of EGEE-based bivariate analyses and univariate analyses under the significance level $\alpha = 0.05$ and the sample size $N = 300$.

Allele Frequency q	Correlation Coefficient r	Analytical Method ^a	QTL Effects (h_1^2 & heterozygote GRR) ^b											
			0 & 1.5	0 & 2.0	0.01 & 1.0	0.03 & 1.0	0.01 & 1.5	0.01 & 2.0	0.03 & 1.5	0.03 & 2.0				
0.05	0.45	B_EGEE	0.258	0.517	0.386	0.823	0.412	0.575	0.785	0.831				
		U_b	0.191	0.419	0.324	0.776	0.389	0.534	0.785	0.816				
		U_e	0.202	0.435	0.341	0.790	0.407	0.551	0.796	0.828				
	0.15	B_EGEE	0.238	0.479	0.357	0.786	0.470	0.638	0.815	0.883				
		U_b	0.194	0.417	0.323	0.776	0.409	0.562	0.797	0.847				
		U_e	0.195	0.419	0.325	0.777	0.411	0.565	0.799	0.848				
0	0	B_EGEE	0.235	0.476	0.351	0.780	0.511	0.678	0.847	0.904				
		U_b	0.192	0.417	0.320	0.773	0.421	0.574	0.808	0.850				
		U_e	0.192	0.417	0.320	0.773	0.421	0.574	0.808	0.850				
	-0.15	B_EGEE	0.232	0.480	0.356	0.788	0.552	0.737	0.876	0.929				
		U_b	0.189	0.421	0.319	0.776	0.430	0.602	0.817	0.864				
		U_e	0.191	0.423	0.321	0.777	0.431	0.605	0.819	0.865				
0.15	-0.45	B_EGEE	0.255	0.523	0.397	0.834	0.669	0.835	0.933	0.973				
		U_b	0.189	0.420	0.322	0.777	0.442	0.622	0.833	0.897				
		U_e	0.201	0.437	0.336	0.790	0.462	0.642	0.844	0.907				
	0.45	B_EGEE	0.506	0.884	0.362	0.826	0.572	0.871	0.831	0.943				
		U_b	0.440	0.843	0.320	0.781	0.552	0.860	0.831	0.934				
		U_e	0.458	0.853	0.335	0.794	0.559	0.865	0.843	0.940				
0.15	0.15	B_EGEE	0.466	0.850	0.333	0.784	0.628	0.900	0.886	0.970				
		U_b	0.444	0.838	0.320	0.779	0.578	0.877	0.859	0.951				
		U_e	0.446	0.839	0.322	0.782	0.580	0.879	0.860	0.951				
	0	B_EGEE	0.463	0.850	0.325	0.781	0.683	0.923	0.911	0.982				
		U_b	0.443	0.843	0.317	0.776	0.600	0.889	0.867	0.962				
		U_e	0.443	0.843	0.317	0.776	0.600	0.889	0.867	0.962				
-0.15	-0.15	B_EGEE	0.464	0.854	0.335	0.781	0.735	0.942	0.937	0.990				
		U_b	0.441	0.841	0.319	0.7809	0.616	0.893	0.880	0.970				

Allele Frequency q	Correlation Coefficient r	Analytical Method ^a	QTL Effects (h^2_1 & heterozygote GRR^b)									
			0 & 1.5	0 & 2.0	0.01 & 1.0	0.03 & 1.0	0.01 & 1.5	0.01 & 2.0	0.03 & 1.5	0.03 & 2.0		
		U_e	0.443	0.843	0.320	0.780	0.619	0.894	0.881	0.970		
	-0.45	B_EGEE	0.499	0.884	0.372	0.826	0.848	0.978	0.981	0.999		
		U_b	0.432	0.843	0.320	0.783	0.642	0.912	0.911	0.981		
		U_e	0.450	0.853	0.335	0.795	0.663	0.920	0.919	0.985		

Note —See note to Table 1 for definitions of a and b.

False-positive rate estimates of association tests for the mixture of the continuous and binary traits with the use of EGEE-based bivariate analyses and univariate analyses across various simulation scenarios under the significance level $\alpha = 0.05$.

Table 3

Sample Size <i>N</i>	Allele Frequency <i>q</i>	Correlation Coefficient <i>r</i>	Analytical Method ^a		
			B_EGEE	U_b	U_e
150	0.05	0.45	0.041	0.045	0.050
		0.15	0.041	0.047	0.048
		0	0.041	0.047	0.047
		-0.15	0.044	0.047	0.048
		-0.45	0.042	0.046	0.052
	0.15	0.45	0.047	0.047	0.053
		0.15	0.047	0.048	0.048
		0	0.051	0.050	0.050
		-0.15	0.048	0.050	0.050
		-0.45	0.050	0.051	0.053
300	0.05	0.45	0.044	0.050	0.055
		0.15	0.046	0.045	0.053
		0	0.044	0.052	0.052
		-0.15	0.046	0.046	0.053
		-0.45	0.046	0.045	0.057
	0.15	0.45	0.049	0.047	0.051
		0.15	0.049	0.050	0.052
		0	0.051	0.050	0.050
		-0.15	0.049	0.051	0.051
		-0.45	0.047	0.049	0.052

Note —See note to Table 1 for definition of a.

Table 4
The association signals for the 10 most significant SNPs in bivariate analyses on the empirical GWA data using **B_EGEE** method

SNP	Cytoband	Allele	MAF ^a	<i>p</i> _B -value ^b	<i>p</i> _U -value ^c
rs12613304	2p16.1	A/G	0.27	1.02×10 ⁻⁶	1.27×10 ⁻⁶
rs35267016	2p16.1	A/G	0.14	1.44×10 ⁻⁶	1.73×10 ⁻⁶
rs12613228	2p16.1	A/G	0.13	1.45×10 ⁻⁶	8.04×10 ⁻⁶
rs17049480	2p16.1	A/G	0.24	1.68×10 ⁻⁶	1.82×10 ⁻⁶
rs12468708	2p16.1	G/T	0.42	2.21×10 ⁻⁶	4.62×10 ⁻⁶
rs887956	2p16.1	A/C	0.26	2.40×10 ⁻⁶	1.25×10 ⁻⁵
rs17161597	7q21.13	A/G	0.07	2.75×10 ⁻⁶	2.02×10 ⁻⁵
rs12001727	9q21.32	C/T	0.39	3.91×10 ⁻⁶	1.29×10 ⁻⁵
rs4660472	1p34.2	A/G	0.35	4.55×10 ⁻⁶	7.54×10 ⁻⁶
rs7863577	9p13.2	G/T	0.10	5.58×10 ⁻⁶	8.20×10 ⁻⁶

^aNote.— MAF, Minor allele frequency
^b*p*_B-value, *p*-values obtained from **B_EGEE**
^c*p*_U-value, adjusted *p*-value for multiple testing for two phenotypes in univariate analyses by Bonferroni correction