Published in final edited form as:

Genet Epidemiol. 2015 July ; 39(5): 357–365. doi:10.1002/gepi.21900.

Fast eQTL Analysis for Twin Studies

Zhaoyu Yin¹, Kai Xia², Wonli Chung³, Patrick F. Sullivan⁴, and Fei Zou^{1,*}

¹Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA

²Department of Psychiatry, University of North Carolina, Chapel Hill, NC 27599, USA

³School of Public Health, Harvard, Boston, MA 02115, USA

⁴Department of Genetics, University of North Carolina, Chapel Hill, NC 27599, USA

Abstract

Twin data are commonly used for studying complex psychiatric disorders, and mixed effects models are one of the most popular tools for modeling dependence structures between twin pairs. However, for eQTL (expression quantitative trait loci) data where associations between thousands of transcripts and millions of single nucleotide polymorphisms need to be tested, mixed effects models are computationally inefficient and often impractical. In this paper, we propose a fast eQTL analysis approach for twin eQTL data where we randomly split twin pairs into two groups, so that within each group the samples are unrelated, and we then apply a multiple linear regression analysis separately to each group. A score statistic that automatically adjusts the (hidden) correlation between the two groups is constructed for combining the results from the two groups. The proposed method has well-controlled type I error. Compared to mixed effects models, the proposed method has similar power but drastically improved computational efficiency. We demonstrate the computational advantage of the proposed method via extensive simulations. The proposed method is also applied to a large twin eQTL data from the Netherlands Twin Register.

Keywords

Twin eQTL data; Score statistic; Mixed effects model; Matrix representation; High correlation

Introduction

Genome-wide association studies (GWAS) have been widely used over the last decade for identifying genetic variants associated with a diversity of complex human diseases, such as type 2 diabetes, breast cancer and psychiatric disorders [Garcia-Closas et al., 2013; Hanson et al., 2014; Winham et al., 2013]. These studies have identified a large number of disease associated SNPs (single nucleotide polymorphisms). However, the majority of the SNPs detected by GWAS individually explain a very small fraction of the total heritability associated with these traits, with no immediately clear functional or regulatory roles [McCarthy and Hirschhorn, 2008]. Gene expression, as an intermediate molecular

*Correspondence: Department of Biostatistics, University of North Carolina, McGavran-Greenberg Hall CB# 7420, Chapel Hill, NC 27599, USA. feizou@email.unc.edu. Phone:(919)843-4822.

phenotype, may provide additional insight into the regulatory roles of SNPs implicated by GWAS. With widely available high throughput technologies, gene expression and genetic variant data can be collected simultaneously on disease-relevant tissues from the same individuals, and expression quantitative trait loci (eQTL) analysis can be performed to assess which genomic regions and genetics variants lead to gene expression [Gilad et al., 2008]. eQTL analyses have been used to identify eQTL hotspots (genomic regions affecting multiple transcripts), construct causal networks, elucidate subclasses of clinical phenotypes, and determine lists of candidate genes for clinical trials [see the reviews of Kendziorski and Wang, 2006 and Wright et al., 2012]. Recent research has also shown that SNPs detected by GWAS are significantly more likely to be eQTL, which can be used to boost the discovery of trait-associated SNPs, and improve understanding of the biology underlying complex traits [Hsu et al., 2010; Nicolae et al., 2010; Schadt et al., 2008; Zhang et al., 2012; Zhong et al., 2010].

Various eQTL analytical approaches have been established where an association test is performed between one transcript and one SNP at a time by linear regression analysis, analysis of variance [Shabalin, 2012], generalized linear regression [Hernandez et al., 2012], or mixed effects models [Kang et al., 2008] depending on the type of eQTL data. More complicated analytical procedures, such as Bayesian regression [Bottolo et al., 2011; Chipman et al., 2011; Stegle et al., 2010] and partial least square regression [Chun et al., 2009] have also been applied to eQTL data. In addition, several methods have been proposed for detecting associations between a group of SNPs and the gene expression of each transcript [Hoggart et al., 2008; Michaelson et al., 2009; Zeng, 1994]. User friendly software such as Genevar [Yang et al., 2010] and eQTL viewer [Zou et al., 2007] have been developed for eQTL data analysis, output visualization and result interpretation. The high dimensionality of eQTL data makes modern eQTL analysis computationally intensive, given the fact that associations between several million SNPs and tens of thousands of transcripts need to be tested. To mitigate this heavy computational burden, analysis may be restricted to a small number of SNP-transcript pairs [Ghazalpour et al., 2008]. Alternatively, Shabalin [2012] has developed a fast eQTL analytical tool which is thousands of times faster than any existing QTL/eQTL software. Matrix eQTL is extremely computationally efficient because it expresses the association test between a SNP and the transcript pair as a function of the correlation between pair, which can be realized by a quick matrix operation.

For complex psychiatric disorders, such as schizophrenia and major depressive disorder, twin studies have received attention for establishing the general extent to which genes and environment are etiologically important [Boomsma et al., 2002; Chou et al., 2009; Neale and Cardon, 1992; Park et al., 2012; Silventoinen et al., 2003; Vaccarino et al., 2008]. Typical twin data include both monozygotic twins (MZ) and dizygotic twins (DZ), plus unpaired individual twins (singletons). MZ twins are assumed to be genetically identical, while DZ twins share 50% of their genes on average. Assuming that MZ and DZ twins share the same environment, a higher phenotypic similarity between MZ twins compared to DZ twins indicates that the phenotype is genetically controlled. Unlike data with independent samples, twin data require more careful statistical modeling since ignoring genetic relatedness and shared environment among twin pairs may lead to high false and/or low true positive findings. Several statistical approaches are available for twin data. One of the most common

approaches is structural equation modeling (SEM) Neale et al. [1989]. Several software programs to perform SEM are available, such as Mx [Neale et al., 1999], Mplus [Muthen and Muthen, 1998], LISREL [Jorsekog and Sorborn, 1986] and OpenMx [Boker et al., 2011, 2012]. Another popular alternative for twin and family data is the mixed-effects model where random effects are used to properly account for the correlations among subjects [Carlin et al., 2005; Kuna et al., 2012; Wang et al., 2011]. Mixed model have a well-established theory which is familiar to statisticians. Moreover, it is conveniently implemented in most statistical software and can flexibly adjust other non-genetic and genetic covariates [Feng et al., 2009; Rabe-Hesketh et al., 2008].

Though powerful, the mixed effects models are computationally intensive and impractical for modern twin eQTL analysis. Moreover, the ultra fast tool Matrix eQTL is not readily applicable to twin data since it does not model the dependence structure between twin pairs. To overcome these computational challenges, we propose a novel fast twin eQTL analysis approach. In this approach, we first randomly split the twin pairs into two groups such that within each group, the samples are unrelated. We then run a separate analysis for each group using any statistical procedure valid for independent data, such as multiple linear regression or analysis of variance. When combining the results from the two groups, we find traditional meta-analysis procedures, such as Fisher's test, is no longer applicable since the two sets of results are not independent due to the correlation between twin pairs. Naively combining the two sets of p-values without consideration of the dependence structure of the data would lead to inflated false positive findings. In this paper, we propose a novel score test which automatically adjusts the (hidden) correlation structures between twin pairs, and therefore controls the type I error accurately. To demonstrate the computational advantages and evaluate the performance of the proposed method, we conduct extensive simulations under various settings to mimic real world twin data.

Our simulation results establish that our proposed approach controls type I error rates well, with negligible power loss compared to the mixed effects model, which is the gold standard for analyzing twin data. Furthermore, the computational efficiency of the proposed method is dramatically improved. The proposed method is more than a thousand times faster than mixed effects models. The fast performance of the proposed method is achieved by computing the most computationally intensive part in the score test by matrix operations, similar to what has been done in matrix eQTL. The utility of the proposed method is further illustrated by analyzing a twin eQTL data where the twin samples (~ 4,000) are from the Netherlands twin registry.

Materials and Methods

The first step of the proposed approach is to randomly split each twin pair into two groups such that all samples within each group are unrelated. Thus a statistical procedure valid for independent data can be directly applied to each group for testing SNP-transcript pair associations. For simplicity and because of its popularity for eQTL data, multiple linear regression is considered. Specifically, for each group, given a transcript and a SNP, we fit the following linear regression model:

$$y_i = \beta g_i + \mathbf{x}_i \boldsymbol{\gamma} + \varepsilon_i, i=1, \dots, n,$$

where for the i th individual, y_i is the gene expression, g_i is the SNP genotype which is coded 0, 1 or 2 according to the number of minor alleles in the genotype, β is the fixed effect of the SNP; and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \dots, \gamma_q)'$ is a vector of parameters corresponding to the vector of non-genetic covariates plus the intercept $\mathbf{x}_i = (1, x_{i1}, \dots, x_{iq})$. Rewriting the above model in matrix form, we get

$$Y = \beta G + X\boldsymbol{\gamma} + \varepsilon,$$

where $Y = (y_1, y_2, \dots, y_n)'$, $G = (g_1, g_2, \dots, g_n)'$ and $X = (\mathbf{x}_1', \dots, \mathbf{x}_n')'$. The log-likelihood function is therefore

$$l(\theta) = \sum_{i=1}^n l_i(\theta)$$

where $\theta = (\beta, \boldsymbol{\gamma}, \sigma^2)$ and $l_i = -\frac{1}{2} \log \sigma^2 - \frac{(y_i - \beta g_i - \mathbf{x}_i \boldsymbol{\gamma})^2}{2\sigma^2}$. For the eQTL analysis, the null hypothesis is $H_0: \beta = 0$ where the vector of the nuisance parameters is $\boldsymbol{\eta} = (\boldsymbol{\gamma}, \sigma^2)$. Following Zou et al. [2004], we get the score function of the i th individual as

$$U_i = U_{\beta,i}(0, \boldsymbol{\eta}) - \Sigma_{\beta,\boldsymbol{\eta}}(0, \boldsymbol{\eta}) \Sigma_{\boldsymbol{\eta},\boldsymbol{\eta}}(0, \boldsymbol{\eta})^{-1} U_{\boldsymbol{\eta},i}(0, \boldsymbol{\eta}),$$

where $U_{\beta,i}(\beta, \boldsymbol{\eta})$ and $U_{\boldsymbol{\eta},i}(\beta, \boldsymbol{\eta})$ are defined as

$$U_{\beta,i}(\beta, \boldsymbol{\eta}) = \frac{\partial l_i(\beta, \boldsymbol{\eta})}{\partial \beta} = \frac{(y_i - \beta g_i - \mathbf{x}_i \boldsymbol{\gamma}) g_i}{\sigma^2},$$

$$U_{\boldsymbol{\eta},i}(\beta, \boldsymbol{\eta}) = \frac{\partial l_i(\beta, \boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \begin{pmatrix} \frac{(y_i - \beta g_i - \mathbf{x}_i \boldsymbol{\gamma}) \mathbf{x}_i'}{\sigma^2} \\ \frac{(y_i - \beta g_i - \mathbf{x}_i \boldsymbol{\gamma})^2}{2\sigma^4} - \frac{1}{2\sigma^2} \end{pmatrix}, \text{ respectively,}$$

and $\Sigma_{\beta,\boldsymbol{\eta}}(\beta, \boldsymbol{\eta})$ is the limit of $n^{-1} \sum_{i=1}^n \frac{\partial^2 l_i}{\partial \beta \partial \boldsymbol{\eta}}$, and $\Sigma_{\boldsymbol{\eta},\boldsymbol{\eta}}(\beta, \boldsymbol{\eta})$ is the limit of $n^{-1} \sum_{i=1}^n \frac{\partial^2 l_i}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}}$. The two Hessian matrices are

$$\frac{\partial^2 l(\beta, \boldsymbol{\eta})}{\partial \beta \partial \boldsymbol{\eta}} = \sum_{i=1}^n \left(-\frac{g_i z_i}{\sigma^2}, -\frac{(y_i - \beta g_i - \mathbf{x}_i \boldsymbol{\gamma}) g_i}{\sigma^4} \right) \text{ and}$$

$$\frac{\partial^2 l(\beta, \eta)}{\partial \eta \partial \eta} = n^{-1} \sum_{i=1}^n \begin{pmatrix} -\frac{\mathbf{x}_i' \mathbf{x}_i}{\sigma^2} & -\frac{(y_i - \beta g_i - \mathbf{x}_i \gamma) \mathbf{x}_i'}{\sigma^4} \\ -\frac{(y_i - \beta g_i - \mathbf{x}_i \gamma) \mathbf{x}_i'}{\sigma^4} & \frac{1}{2\sigma^4} - \frac{(y_i - \beta g_i - \mathbf{x}_i \gamma)^2}{\sigma^6} \end{pmatrix}.$$

Let the restricted MLE $\hat{\eta}$ be the solution of $\sum_{i=1} U_{\eta,i}(\beta, \eta) = 0$, where β is set to 0 in this equation. Specifically, we have

$$\hat{\gamma} = (X'X)^{-1}X'Y,$$

$$\hat{\sigma}^2 = \sum_{i=1}^n (y_i - \mathbf{x}_i \hat{\gamma})^2 / n = \|Y - X\hat{\gamma}\|^2 / n.$$

Note that $\hat{\eta}$ is estimated under H_0 and thus does not depend on the genotypes of any given SNP. Therefore it does not change from SNP to SNP, and needs only to be estimated once

for each transcript. Also note that all of the off-diagonal elements in $\left. \frac{\partial^2 l(\beta, \eta)}{\partial \eta \partial \eta} \right|_{\beta=0, \eta=\hat{\eta}}$ equal to zero. Replacing all unknown parameters by their sample estimators in the score function, we get

$$\begin{aligned} \hat{U}_i &= U_{\beta,i}(0, \hat{\eta}) - \Sigma_{\beta,\eta}(0, \hat{\eta}) \Sigma_{\eta,\eta}(0, \hat{\eta})^{-1} U_{\eta,i}(0, \hat{\eta}) \\ &= \frac{(y_i - \mathbf{x}_i \hat{\gamma}) g_i}{\hat{\sigma}^2} - \sum g_i \mathbf{x}_i \left(\sum \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \frac{(y_i - \mathbf{x}_i \hat{\gamma}) \mathbf{x}_i'}{\hat{\sigma}^2} + \frac{\sum (y_i - \mathbf{x}_i \hat{\gamma}) g_i}{n \hat{\sigma}^2} - \frac{\sum (y_i - \mathbf{x}_i \hat{\gamma}) g_i}{n \hat{\sigma}^4} (y_i - \mathbf{x}_i \hat{\gamma})^2. \end{aligned}$$

For computational efficiency, we express the score function $\hat{U}_i$ in terms of a matrix operation below. Denote

$$M_x = X(X'X)^{-1}X',$$

$$J_n = (1, \dots, 1)' \text{ and is of length } n,$$

$$a = \sum_{i=1}^n g_i \mathbf{x}_i = G'X,$$

$$A = \sum_{i=1}^{n_k} \mathbf{x}_i' \mathbf{x}_i = X'X,$$

$$c = \frac{\sum (y_i - \mathbf{x}_i \hat{\gamma}) g_i}{n \hat{\sigma}^2} = \frac{1}{n \hat{\sigma}^2} \left(G'Y - G'X(X'X)^{-1}X'Y \right) = \frac{1}{n \hat{\sigma}^2} G'(I - M_x)Y,$$

we have

$$\hat{U}_i = \frac{y_i - \mathbf{x}_i \hat{\gamma}}{\hat{\sigma}^2} [(g_i - aA^{-1}\mathbf{x}_i') - c(y_i - \mathbf{x}_i \hat{\gamma})] + c.$$

Let $\hat{U}$ be the vector of the score function of all individuals. That is,

$$\begin{aligned} \hat{U} = \begin{pmatrix} \hat{U}_1 \\ \hat{U}_2 \\ \vdots \\ \hat{U}_n \end{pmatrix} &= \begin{pmatrix} \frac{y_1 - \mathbf{x}_1 \hat{\gamma}}{\hat{\sigma}^2} \\ \frac{y_2 - \mathbf{x}_2 \hat{\gamma}}{\hat{\sigma}^2} \\ \vdots \\ \frac{y_n - \mathbf{x}_n \hat{\gamma}}{\hat{\sigma}^2} \end{pmatrix} \times \left[\begin{pmatrix} g_1 - \mathbf{x}_1 A'^{-1} a' \\ g_2 - \mathbf{x}_2 A'^{-1} a' \\ \vdots \\ g_n - \mathbf{x}_n A'^{-1} a' \end{pmatrix} - c \begin{pmatrix} y_1 - \mathbf{x}_1 \hat{\gamma} \\ y_2 - \mathbf{x}_2 \hat{\gamma} \\ \vdots \\ y_n - \mathbf{x}_n \hat{\gamma} \end{pmatrix} \right] + \begin{pmatrix} c \\ c \\ \vdots \\ c \end{pmatrix} \\ &= \frac{(I - M_x)Y}{\hat{\sigma}^2} \times [(G - XA'^{-1}a') - c(I - M_x)Y] + J_n c. \end{aligned}$$

Substituting a , A and c into the above equation, we get

$$\begin{aligned} \hat{U} &= \frac{(I - M_x)Y}{\hat{\sigma}^2} \times \left[(I - M_x)G - \frac{1}{n\hat{\sigma}^2} (I - M_x)YY'(I - M_x)G \right] + \frac{1}{n\hat{\sigma}^2} J_n Y'(I - M_x)G \\ &= \left\{ \frac{(I - M_x)Y}{\hat{\sigma}^2} \times \left[(I - M_x) - \frac{1}{n\hat{\sigma}^2} (I - M_x)YY'(I - M_x) \right] + \frac{1}{n\hat{\sigma}^2} J_n Y'(I - M_x) \right\} G. \quad (*) \end{aligned}$$

When the covariates X and G are independent of each other, the elements inside $\{\}$ can be simplified further to

$$\frac{(I - M_x)Y}{\hat{\sigma}^2} \times \left[\left(I - \frac{1}{n\hat{\sigma}^2} (I - M_x)YY' \right) + \frac{1}{n\hat{\sigma}^2} J_n Y' \right].$$

Note that the elements inside $\{\}$ only depend on Y and the covariates X , and thus are the same across all SNPs. This motivates us to derive the above matrix operation to calculate the score vectors across a large number of SNPs for a given transcript simultaneously.

Specifically, we replace the vector G in equation (*) by a matrix $H = (G_1, \dots, G_m)$, where G_j is the G vector corresponding to SNP j ($j = 1, \dots, m$). The score matrix $\hat{U}$ then is $n \times m$, where the j th column represents the score vector at SNP j . Remember that twin pairs are randomly split into two groups. Notationally let's add superscripts to $\hat{U}$ above and denote $\hat{U}^{(1)}$ and $\hat{U}^{(2)}$ as the score matrices of group 1 and group 2, respectively. Also let n_k be the number of individuals in the k th group ($k = 1, 2$) and the score function of the i th subject for the j th SNP in the k th group be $U_{i,j}^{(k)}$ ($i = 1, \dots, n_k$), or specifically we now have

$$\hat{U}^{(k)} = \begin{pmatrix} \hat{U}_{1,1}^{(k)} & \hat{U}_{1,2}^{(k)} & \cdots & \hat{U}_{1,m}^{(k)} \\ \hat{U}_{2,1}^{(k)} & \hat{U}_{2,2}^{(k)} & \cdots & \hat{U}_{2,m}^{(k)} \\ \vdots & \vdots & \cdots & \vdots \\ \hat{U}_{n_k,1}^{(k)} & \hat{U}_{n_k,2}^{(k)} & \cdots & \hat{U}_{n_k,m}^{(k)} \end{pmatrix}.$$

The linear regression results for the j th SNP from the two groups may be naively combined as

$$W_{\text{naive}}(j) = \frac{(\hat{\beta}(1) + \hat{\beta}(2))^2}{(SE(\hat{\beta}(1)))^2 + (SE(\hat{\beta}(2)))^2}$$

and assumed to follow χ_1^2 asymptotically under H_0 . Here, $\hat{\beta}(k)$ is the estimate of the SNP effect β with its standard error $SE(\hat{\beta}(k))$ from the linear regression analysis of the k th group ($k = 1, 2$). This test is only valid when the results from the two groups are independent of each other, which is not likely to be true for twin data. To account for the dependence structure of the two groups, we derive a new score statistic to automatically adjust the correlation between the two groups:

$$W_{\text{proposed}}(j) = \frac{(\sum_{i=1}^{n_1} \hat{U}_{i,j}^{(1)} + \sum_{i=1}^{n_2} \hat{U}_{i,j}^{(2)})^2}{\sum_{i=1}^{n_1} (\hat{U}_{i,j}^{(1)})^2 + \sum_{i=1}^{n_2} (\hat{U}_{i,j}^{(2)})^2 + 2 \sum_{i=1}^{n_{\text{twin}}} \hat{U}_{i,j}^{(1)} \times \hat{U}_{i,j}^{(2)}},$$

where n_{twin} is the total number of twin pairs, and groups 1 and 2 individuals are arranged in such way that the first n_{twin} samples in groups 1 and 2 are paired twins, and the remaining samples are singletons. The proposed test statistic $W_{\text{proposed}}(j)$ follows χ_1^2 asymptotically under H_0 .

Results

Simulation studies

The proposed method is applied to simulated twin data to evaluate its performance. Each dataset includes 900 or 1800 individuals consisting of MZ twins, DZ twins and singletons in the ratio 2 : 2 : 1. Two continuous covariates ($\mathbf{x}_1, \mathbf{x}_2$) with effects $\boldsymbol{\gamma} = (0.3, 0.1)$ are correlated to the response variable y , where $\mathbf{x}_1$ and $\mathbf{x}_2$ are generated from $N(3, 1)$ and $N(5, 2)$, respectively. The response variable y is generated from the following model,

$$y_i = \mu + g_i \beta + \mathbf{x}_i \boldsymbol{\gamma} + a_i + c_i + d_i + e_i, i = 1, \dots, n,$$

where y_i is the gene expression for the i th individual, μ is the grand mean, g_i is the SNP genotype, and $\mathbf{x}_i$ is a vector of non-genetic covariates. The SNP genotype g_i and the vector of covariates $\mathbf{x}_i$ are generated independently. The random terms a_i, d_i, c_i, e_i are the additive, dominant, common environment effects and random error, respectively, which are mutually

independent and normally distributed with mean zero and variance σ_a^2 , σ_c^2 , σ_d^2 and σ_e^2 , respectively. For subjects i and j who are a twin pair, we have $\text{cov}(a_i, a_j) = \sigma_a^2$ and $\text{cov}(d_i, d_j) = \sigma_d^2$ if they are *MZ* pair; $\text{cov}(a_i, a_j) = \sigma_a^2/2$ and $\text{cov}(d_i, d_j) = \sigma_d^2/4$ if they are *DZ* pair, while $\text{cov}(c_i, c_j) = \sigma_c^2$ for all twin pairs. According to Neale et al. [1989], the above model is referred to as the ACE or ACDE model depending $\sigma_d^2 = 0$ or not. For each simulation set up, 1000 datasets are generated. We set $\sigma_e^2 = 1$, $\sigma_a^2 = 0.75$ and $\sigma_c^2 = 0.75$, resulting in the additive heritability h_a^2 of 0.462 and the variance explained by the shared environmental c^2 , of 0.231 under the ACE model. For the ACDE model, we set σ_d^2 to $2\sigma_a^2$, leading to $h_a^2 = 0.316$ and $c^2 = 0.158$. Since there are 1000 association tests for each simulated dataset, 1000 datasets give a total of $1000 * 1000 = 1$ million tests, which are used for the type I error evaluation. The type I error of the proposed method is compared with three other methods: 1) a multiple linear regression model on the full twin data, where the dependence between the twin pairs is ignored; 2) the naive approach; and 3) the mixed effects model: $y_{ij} = g_{ij}\beta + x_{ij}\gamma + a_{ij} + d_{ij} + c_{ij} + \varepsilon_{ij}$, where i and j are family id and individual index respectively, x_{ij} is the vector of non-genetic covariates plus intercept, g_{ij} is the SNP genotype, and the definition of a_{ij} , c_{ij} , d_{ij} and their covariance structures are described in the above ACE and ACDE models.

Results from Tables 1 and 2 demonstrate that the type I error rates of both the proposed method and the mixed effects model are well controlled. In contrast, if the linear regression model (*lm*) is directly applied to the full twin data, the type I error rates have been dramatically inflated. For example, under the settings of $n = 1800$ and the targeted type I error rate $\alpha = 0.05$, the type I errors for the data from the ACE model and ACDE model are 0.098 and 0.102, respectively. The type I error inflation in the naive method is also clear.

For power comparisons, we generate 1000 datasets under the alternative hypothesis, where we set β to 0.32, 0.37, 0.45, and 0.50 in the ACE model and β to 0.40, 0.45, 0.50, and 0.55 in the ACDE model. For both of the models, the non-genetic effects γ and the variance components σ_a^2 , σ_e^2 , σ_c^2 and σ_d^2 are kept the same as in the above type I error investigation. The results in Tables 3 and 4 show that the proposed method has negligible power loss compared to the gold standard mixed effects models regardless of the sample size. All simulations and data analyses are conducted in R (a programming language, R Core Team [2014]).

Computational Efficiency

Our proposed method combines the advantages of multiple linear regression and the matrix operation to achieve fast computational performance. The proposed method took 361 seconds with one 2.93GHz Intel processor to perform 1 million association tests whereas the mixed effects model took 21,000 seconds using a cluster of 100 Intel processors to run the same number of tests. The R function that implements the proposed method can be found at www.bios.unc.edu/~feizou/software.

Netherlands Twin Registry (NTR) eQTL Study

A total of 2,752 individuals from the Netherlands Twin Registry (NTR) had their SNP genotypes and gene expression data measured [Wright et al., 2014] on Affymetrix 6.0 SNP arrays and U219 expression arrays. The goals of this project were to quantify the heritability of each transcript, to build a comprehensive list of eQTL in peripheral blood, and to assess the biomedical relevance of these eQTLs. After quality control [see Wright et al., 2014, for details], 686,895 SNPs and 47,585 transcripts on 2,494 individuals were used for the eQTL analysis. The 2,494 individuals include 638 MZ pairs, 529 DZ pairs and 160 singletons. A total of 96 covariates including the plate number, hybridization well position, age at blood sampling, sex, time intervals from sampling to RNA extraction and hybridization, and total white and red cell counts, 5 principal components (PCs) from the gene expression data, and 3 PCs derived from the SNP data are included as covariates in the eQTL analysis [see Wright et al., 2014].

In contrast to controlling the family-wise error, controlling the false discovery rate (FDR) is less stringent and handles the dependence between test statistics reasonably well [Benjamini and Yekutieli, 2001]. Instead of saving the p -values for all the SNP-transcript pairs, Shabalin [2012] only saves the p -values that pass a pre-specified significance threshold, on which the Benjamini and Hochberg procedure is applied. Following the procedure of Shabalin [2012], we identified 184,474 and 127,891 SNP-transcript pairs at $FDR = 0.01$ and $FDR = 0.001$, respectively, with the proposed method. In contrast, the naive method identified 225,419 and 136,509 significant SNP-transcript pairs at $FDR = 0.01$ and $FDR = 0.001$, which are significantly more than what have been detected by the proposed method. These additional findings are likely to be false positives, given the fact that the type I error inflation of the naive method. Based on the results from the proposed method, SNP-transcripts pairs with p -values passing certain thresholds are shown in the heat map (upper panel of Figure 1). Significant SNP-transcript pairs along the 45 degree diagonal line in the heat map are cis-eQTLs, whereas those off-diagonal are trans-eQTLs. The lower panel of Figure 1 clearly suggests that there are two master regulator regions on chromosomes 3 and 17 that affect a large number of transcripts, which deserve further investigation.

The top 20 SNP-transcript pairs detected by the proposed approach are summarized in Table 5. Note that the p -values of these pairs from the naive method and Matrix eQTL method are too small to be reported precisely and thus are reported as 0 in the software R.

Discussion

eQTL analysis usually involves thousands of transcripts and millions of SNPs assayed in thousands of individuals. The challenges for eQTL analysis for twin data are the heavy computational burden and the dependency of twins within pairs. The mixed effects model is the gold standard for analyzing twin data. However, it is extremely time-consuming for modern eQTL data. Although matrix eQTL is an ultra fast tool, it is not applicable to twin data for eQTL analysis, since the covariance matrix V is assumed to be known. This assumption is not true for twin data, where the covariance matrix can be decomposed into

several parts. For example, $V = \sigma_a^2(A + \frac{\sigma_e^2}{\sigma_a^2})$ where σ_a^2 and σ_e^2 are unknown. If the heritability

is much larger than random error, covariance for the additive genetic effect A can be used to approximate the covariance V , otherwise dropping the second term in the parentheses may result in an invalid inference. Alternatively, the heritability for each transcript needs to be estimated from the mixed effects model before the matrix eQTL is applied, and the computation advantage will be lost dramatically. In contrast to the mixed effects model, we randomly split the twin pairs into two groups and then applied multiple linear regression in each group to calculate the score vector for examining the association of each SNP-transcript pair. The simplicity of linear regression models mitigates the heavy computational burden dramatically. Simulation studies demonstrate the computational advantages of the proposed methods. Specifically, the proposed method is > 6000 times faster than the mixed effect model. The proposed method achieves the fast performance for two reasons: the simplicity of the linear regression model and large matrix operations (especially the matrix multiplication), expressing the most computationally intensive part in an efficient way.

To combine the analytical results from two groups, we propose a novel score statistic which automatically adjusts the correlation between the two groups in a natural way. Results from simulations show that the proposed method controls type I error rates much better at the target levels than do the linear regression model and the naive method. Moreover, this proposed method is competitive and does not lose much power compared to the mixed effects model but is much more computationally efficient. The slight power loss is likely due to the following two reasons: a) the fixed non-genetic covariate effects need to be estimated twice in our analysis while in the mixed effects model, they only need to be estimated once; and b) the shared environment effects are specifically modeled by the mixed effects model but our method lumps that effect into the random error term and inexplicitly models it. Score statistics have been well studied [Cox and Hinkley, 1974, Section 9.3(iii)] and widely used in practice. For example, the similar idea based on score statistic has been applied to meta-analysis of GWAS with overlapping samples, where a robust estimator for variance-covariance matrix was proposed [Lin and Sullivan, 2009]. Although we have focused our attention on twin studies, the proposed method could be applied and extended to multivariate phenotypes, where multiple correlated phenotypes are measured for each subject.

In simulation studies, we have also tried another test statistic,

$$T_2 = \begin{pmatrix} \sum_{i=1}^{n_1} \hat{U}_{i,j}^{(1)} \\ \sum_{i=1}^{n_2} \hat{U}_{i,j}^{(2)} \end{pmatrix}^T \begin{pmatrix} \sum_{i=1}^{n_1} (\hat{U}_{i,j}^{(1)})^2 & \sum_{i=1}^{n_{\text{twin}}} \hat{U}_{i,j}^{(1)} \times \hat{U}_{i,j}^{(2)} \\ \sum_{i=1}^{n_{\text{twin}}} \hat{U}_{i,j}^{(1)} \times \hat{U}_{i,j}^{(2)} & \sum_{i=1}^{n_2} (\hat{U}_{i,j}^{(2)})^2 \end{pmatrix} \begin{pmatrix} \sum_{i=1}^{n_1} \hat{U}_{i,j}^{(1)} \\ \sum_{i=1}^{n_2} \hat{U}_{i,j}^{(2)} \end{pmatrix} \sim \chi_2^2 \text{ under } H_0.$$

However, it is less efficient than the proposed one. This is probably due to the fact that our proposed score test follows χ_1^2 under H_0 while T_2 follows χ_2^2 under H_0 . In addition, we examined the robustness of the proposed method under two model misspecification scenarios where (1) if the IBD value for DZ pairs is not fixed at 0.5, but are randomly generated from $N(0.50, 0.04)$ to reflect the IBD variation among DZ samples that are commonly observed; (2) the random error e_i is non-normally distributed with a skewed distribution or a distribution with heavier tails (see Table 6). Expect these changes, the simulation set ups are kept the same as those in Table 1 with sample size $n = 900$. Clearly,

both the proposed method and the computationally intensive mixed effects model are very robust to the model misspecifications, further suggesting broad applications of the proposed method.

Finally, it is worth mentioning that the twin samples are randomly split, therefore the analysis result is not fixed and varies from one split to another split. As an example, we split 10,000 simulated data from the ACE model twice and the scatter plot of the proposed test statistics from the two splits is given in Figure 2. The two results are not identical but highly consistent with $R^2 > 0.99$. To reduce the splitting variation, we can repeat splitting of the samples multiple times and combine the analysis results from the multiple splits. But how to combine the highly correlated results deserves further investigation.

Acknowledgments

The authors are grateful for constructive comments and suggestions from the reviewers and the editor. Support was provided in part by National Institute of General Medical Sciences R01GM074175.

References

- Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*. 2001; 29(4):1165–1188.
- Boker SM, Neale MC, Maes HH, Wilde MJ, et al. OpenMx: An Open Source Extended Structural Equation Modeling Framework. *Psychometrika*. 2011; 76(2):306–317. [PubMed: 23258944]
- Boker SM, Neale MC, Maes HH, Wilde MJ, et al. OpenMx 1.2 User Guide. 2012
- Boomsma D, Busjahn A, Peltonen L. Classical twin studies and beyond. *Nature Reviews Genetics*. 2002; 3(11):872–882.
- Bottolo L, Petretto E, Blankenberg S, Cambien F, Cook SA, Tiret L, Richardson S. Bayesian detection of expression quantitative trait loci hot spots. *Genetics*. 2011; 189(4):1449–1459. [PubMed: 21926303]
- Carlin JB, Gurrin LC, Sterne JA, Morley R, Dwyer T. Regression models for twin studies: a critical review. *International Journal of Epidemiology*. 2005; 34(5):1089–1099. [PubMed: 16087687]
- Chipman KC, Singh AK. Bayesian detection of expression quantitative trait loci hot spots. *BMC Bioinformatics*. 2011; 12(7) 2011.
- Chou YY, Lepor N, Chiang MC, et al. Mapping genetic influences on ventricular structure in twins. *Neuroimage*. 2009; 44(4):1312–1323. [PubMed: 19041405]
- Chun H, Keles S. Expression quantitative trait loci mapping with multivariate sparse partial least squares regression. *Genetics*. 2009; 182(1):79–909. [PubMed: 19270271]
- Cox, DR.; Hinkley, DV. *Theoretical Statistics*. London: Chapman and Hall; 1974.
- Feng R, Zhou G, Zhang M, Zhang H. Analysis of twin data using SAS. *Biometrics*. 2009; 65(2):584–589. [PubMed: 18647295]
- Garcia-Closas M, Couch FJ, Lindstrom S, Michailidou K, Schmidt MK, et al. Genome-wide association studies identify four ER negative-specific breast cancer risk loci. *Nature Genetics*. 2013; 45(4):392–398. [PubMed: 23535733]
- Ghazalpour A, Doss S, Kang H, et al. High-resolution mapping of gene expression using association in an outbred mouse stock. *PLoS Genetics*. 2008; 4(8):e1000149. [PubMed: 18688273]
- Gilad Y, Rifkin SA, Pritchard JK. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends in Genetics*. 2008; 24(8):408–415. [PubMed: 18597885]
- Hanson RL, Muller YL, Kobes S, et al. A genome-wide association study in American Indians implicates DNER as a susceptibility locus for type 2 diabetes. *Diabetes*. 2014; 63(1):369–376. [PubMed: 24101674]

- Hernandez DG, Nalls MA, Moore M, et al. Integration of GWAS SNPs and tissue specific expression profiling reveal discrete eQTLs for human traits in blood and brain. *Neurobiology of Disease*. 2012; 47(1):20–28. [PubMed: 22433082]
- Hoggart CJ, Whittaker JC, DeIorio M, Balding DJ. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genetics*. 2008; 4(7):e1000130. [PubMed: 18654633]
- Hsu YH, Zillikens MC, Wilson SG, Farber CR, Demissie S, et al. An integration of genome-wide association study and gene expression profiling to prioritize the discovery of novel susceptibility Loci for osteoporosis-related traits. *PLoS Genetics*. 2010; 6(6):e1000977. [PubMed: 20548944]
- Jorsekog, KG.; Sorborn, D. Lisrel VI. Mooresville, Indiana: Scientific Software; 1986.
- Kang HM, Ye C, Eskin E. Accurate discovery of expression quantitative trait loci under con- founding from spurious and genuine regulatory hotspots. *Genetics*. 2008; 180(4):1909–1925. [PubMed: 18791227]
- Kendzierski C, Wang P. A review of statistical methods for expression quantitative trait loci mapping. *Mammalian Genome*. 2006; 17(6):509–517. [PubMed: 16783633]
- Kuna ST, Maislin G, Pack FM, Staley B, Hachadoorian R, Coccaro EF, Pack AI. Heritability of performance deficit accumulation during acute sleep deprivation in twins. *Sleep*. 2012; 35(9): 1223–1233. [PubMed: 22942500]
- Lin DY, Sullivan PF. Meta-analysis of genome-wide association studies with overlapping subjects. *The American Journal of Human Genetic*. 2009; 85(6):862–872.
- McCarthy MI, Hirschhorn JN. Genome-wide association studies: potential next steps on a genetic journey. *Human Molecular Genetics*. 2008; 17:R156–R165. (Review Issue 2). [PubMed: 18852205]
- Michaelson JJ, Loguercio S, Beyer A. Detection and interpretation of expression quantitative trait loci (eQTL). *Methods*. 2009; 48(3):265–276. [PubMed: 19303049]
- Muthen, LK.; Muthen, BO. Mplus User's Guide. Los Angeles, California: Muthen & Muthen; 1998.
- Neale, MC.; Cardon, LR. Methodology for genetic studies of twins and families. Dordrecht, the Netherlands: Kluwer Academic; 1992.
- Neale MC, Heath AC, Hewitt JK, Eaves LJ, Fulker DW. Fitting genetic models with LISREL: Hypothesis testing. *Behavior Genetics*. 1989; 19(1):37–49. [PubMed: 2712812]
- Neale, MC.; Boker, SM.; Xie, G.; Maes, HH. Mx:Statistical Modeling. Richmond, Virginia: Department of Psychiatry, Medical College of Virginia of Virginia Commonwealth University; 1999. 1999.
- Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genetics*. 2010; 6(4):e1000888. [PubMed: 20369019]
- Park JH, Song YM, Sung J, et al. The association between fat and lean mass and bone mineral density: the Healthy Twin Study. *Bone*. 2012; 50(4):1006–1011. [PubMed: 22306928]
- R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2014. URL <http://www.R-project.org>.
- Rabe-Hesketh S, Skrondal A, Gjessing HK. Biometrical modeling of twin and family data using standard mixed model software. *Biometrics*. 2008; 64(1):280–288. [PubMed: 17484777]
- Schadt EE, Molony C, Chudin E, Hao K, Yang X, et al. Mapping the Genetic Architecture of Gene Expression in Human Liver. *PLoS Biology*. 2008; 6(5):e107. [PubMed: 18462017]
- Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*. 2012; 28(10):1353–1358. [PubMed: 22492648]
- Silventoinen K, Sammalisto S, Perola M, et al. Heritability of adult body height: a comparative study of twin cohorts in eight countries. *Twin Research*. 2003; 6(5):399–408. [PubMed: 14624724]
- Stegle O, Parts L, Durbin R, Winn J. A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Computational Biology*. 2010; 6(5):e1000770. [PubMed: 20463871]
- Vaccarino V, Brennan ML, Miller AH, et al. Association of major depressive disorder with serum myeloperoxidase and other markers of inflammation: a twin study. *Biological Psychiatry*. 2008; 64(6):476–483. [PubMed: 18514165]

- Wang X, Guo X, He M, Zhang H. Statistical inference in mixed models and analysis of twin and family data. *Biometrics*. 2011; 67(3):987–995. [PubMed: 21306354]
- Winham SJ, Cuellar-Barboza AB, Oliveros A, McElroy SL, Crow S, Colby C, Choi D-S, Chauhan M, Frye M, Biernacka JM. Genome-wide association study of bipolar disorder accounting for effect of body mass index identifies a new risk allele in TCF7L2. *Molecular Psychiatry*. 2013
- Wright FA, Shabalin AA, Rusyn I. Computational tools for discovery and interpretation of expression quantitative trait loci. *Pharmacogenomics*. 2012; 13(3):343–352. [PubMed: 22304583]
- Wright FA, Sullivan PF, Brooks AI, Zou F, et al. Heritability and genomics of gene expression in peripheral blood. *Nature Genetics*. 2014; 46(5):430–437. [PubMed: 24728292]
- Yang TP, Beazley C, Montgomery SB, et al. Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. *Bioinformatics*. 2010; 26(19):2474–2476. [PubMed: 20702402]
- Zeng ZB. Precision mapping of quantitative trait loci. *Genetics*. 1994; 136(4):1457–1468. [PubMed: 8013918]
- Zhang M, Liang L, Morar N, Dixon AL, Lathrop GM, Ding J, Moffatt MF, Cookson WO, Kraft P, Qureshi AA, Han J. Integrating pathway analysis and genetics of gene expression for genome-wide association study of basal cell carcinoma. *Human Genetics*. 2012; 131(4):615–623. [PubMed: 22006220]
- Zhong H, Beaulaurier J, Lum PY, Molony C, Yang X, Macneil DJ, et al. Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. *PLoS Genetics*. 2010; 6:e1000932. [PubMed: 20463879]
- Zou F, Fine JP, Hu J, Lin DY. An efficient resampling method for assessing genome-wide statistical significance in mapping quantitative trait loci. *Genetics*. 2004; 168(4):2307–2316. [PubMed: 15611194]
- Zou W, Aylor DL, Zeng Z. eQTL viewer: visualizing how sequence variation affects genome-wide transcription. *BMC Bioinformatics*. 2007; 8(7)

Figure 1.

eQTL hotspot for NTR twin data

Upper panel: heat map of the SNP-transcript pairs with p -values $< 10^{-15}$. Green, blue, red and dark dots indicate the pairs whose $-\log_{10}(p)$ falls into $[15, 25)$, $[25, 50)$, $[50, 100)$ and $[100, \infty)$, respectively. The colored lines on the 4 edges are used to differentiate the 22 autosomes. X-axis: SNP locations; Y-axis: transcript locations.

Lower panel: the number of transcripts (y-axis) that are significantly associated with each SNP ($P < 10^{-15}$). X-axis: SNP locations.

Figure 2. Scatter plot of the proposed test statistics from 10,000 simulated data with each data randomly split twice. The data are generated from the ACE model and sample size is 1800.

Table 1

Type I error comparison for data from the ACE model

α	$n = 900$					$n = 1800$				
	0.05	0.01	0.001	0.0001	0.0001	0.05	0.01	0.001	0.0001	0.0001
proposed	0.0522	0.0108	0.0011	0.0001	0.0001	0.0507	0.0100	0.001	0.0001	0.0001
naïve	0.1000	0.0310	0.060	0.0012	0.0989	0.0301	0.0055	0.0010	0.0001	0.0001
mixed	0.0509	0.0103	0.0011	0.0001	0.0502	0.0100	0.0010	0.0001	0.0001	0.0001
lm	0.0988	0.0301	0.0057	0.0011	0.0984	0.0298	0.0053	0.0010	0.0001	0.0010

proposed: proposed score test statistic
naïve: naïve test statistic
mixed: mixed effects model
lm: multiple linear regression

Table 2

Type I error comparison for data from the ACDE model

α	$n = 900$					$n = 1800$				
	0.05	0.01	0.001	0.0001	0.05	0.01	0.001	0.0001	0.05	0.0001
proposed	0.0524	0.0107	0.0011	0.0001	0.0510	0.0104	0.0010	0.0001	0.0510	0.0001
naive	0.1027	0.0321	0.0064	0.0012	0.1025	0.0320	0.0062	0.0013	0.1025	0.0013
mixed	0.0509	0.0103	0.0010	0.0001	0.0502	0.0101	0.0010	0.0001	0.0502	0.0001
lm	0.1024	0.0317	0.0061	0.0012	0.1020	0.0317	0.0061	0.0012	0.1020	0.0012

proposed: proposed score test statistic
naive: naive test statistic
mixed: mixed effects model
lm: multiple linear regression

Table 3

Power comparison for data from the ACE model

		$\beta = 0.32$			$\beta = 0.37$			$\beta = 0.45$			$\beta = 0.50$		
Method	α	proposed	naive	mixed	pro	naive	mixed	pro	naive	mixed	pro	naive	mixed
$n = 900$	0.05	0.733	0.817	0.761	0.834	0.886	0.865	0.933	0.966	0.951	0.970	0.985	0.986
	0.01	0.503	0.662	0.557	0.664	0.790	0.710	0.830	0.903	0.882	0.899	0.951	0.933
	0.001	0.225	0.416	0.286	0.363	0.584	0.434	0.629	0.796	0.696	0.745	0.873	0.811
	0.0001	0.103	0.240	0.130	0.166	0.372	0.224	0.368	0.633	0.466	0.530	0.753	0.630
$n = 1800$	0.05	0.945	0.971	0.966	0.984	0.991	0.996	1.000	1.000	0.999	1.000	1.000	1.000
	0.01	0.831	0.919	0.889	0.940	0.973	0.963	0.989	0.998	0.999	0.999	1.000	0.999
	0.001	0.592	0.781	0.690	0.784	0.912	0.867	0.953	0.988	0.983	0.988	0.998	0.999
	0.0001	0.344	0.600	0.459	0.567	0.796	0.693	0.864	0.953	0.928	0.951	0.989	0.983

proposed: proposed score test statistic
naive: naive test statistic
mixed: mixed effects model
lm: multiple linear regression

Table 4

Power comparison for data from the ACDE model

		$\beta = 0.40$			$\beta = 0.45$			$\beta = 0.50$			$\beta = 0.55$		
Method	α	proposed	naive	mixed	pro	naive	mixed	pro	naive	mixed	pro	naive	mixed
$n = 900$	0.05	0.740	0.849	0.786	0.842	0.906	0.873	0.904	0.950	0.932	0.949	0.974	0.963
	0.01	0.501	0.670	0.561	0.633	0.792	0.688	0.755	0.870	0.804	0.844	0.926	0.890
	0.001	0.231	0.426	0.268	0.330	0.569	0.395	0.459	0.695	0.538	0.597	0.805	0.672
	0.0001	0.094	0.242	0.122	0.151	0.342	0.201	0.237	0.489	0.299	0.334	0.627	0.426
$n = 1800$	0.05	0.964	0.982	0.980	0.986	0.993	0.995	0.995	0.999	0.999	0.999	1.000	0.999
	0.01	0.859	0.942	0.913	0.948	0.980	0.971	0.981	0.991	0.990	0.994	0.999	0.998
	0.001	0.664	0.831	0.745	0.800	0.918	0.869	0.895	0.974	0.948	0.969	0.988	0.985
	0.0001	0.431	0.676	0.511	0.605	0.813	0.703	0.756	0.907	0.844	0.865	0.972	0.927

proposed: proposed score test statistic
naive: naive test statistic
mixed: mixed effects
lm: multiple linear regression

Table 5

Top 20 SNP-transcript pairs identified from the proposed method

	SNP	chr(SNP)	start(SNP)	end(SNP)	GENE	chr(GENE)	start(GENE)	end(GENE)	$\log_{10}(pval)$	q value
1	rs3909451	chr5	96295120	96295121	11745601_a_at	chr5	96215266	96225192	228.4077	1.278352e-218
2	rs27300	chr5	96363406	96363407	11745601_a_at	chr5	96215266	96225192	227.2055	8.544353e-218
3	rs12229020	chr12	10117690	10117691	11729479_a_at	chr12	10124007	10138056	227.1056	8.544353e-218
4	rs7313235	chr12	10132274	10132275	11729479_a_at	chr12	10124007	10138056	226.8456	1.166032e-217
5	rs2235918	chr1	17422605	17422606	11727597_at	chr1	17393255	17445948	226.6566	1.441486e-217
6	rs2076607	chr1	17422659	17422660	11727597_at	chr1	17393255	17445948	225.8381	7.908774e-217
7	rs9272346	chr6	32604371	32604372	11750528_x_at	chr6	32410960	32411702	225.5650	1.271430e-216
8	rs4808485	chr19	16439497	16439498	11764269_at	chr19	16438314	16438642	224.0849	3.360241e-215
9	rs27290	chr5	96350093	96350094	11745601_a_at	chr5	96215266	96225192	221.9954	3.670607e-213
10	rs4698634	chr4	17630191	17630192	11736024_at	chr4	17616279	17627249	221.6546	7.240198e-213
11	rs27300	chr5	96363406	96363407	11747137_x_at	chr5	96215265	96253609	221.4142	1.145032e-212
12	rs3909451	chr5	96295120	96295121	11747137_x_at	chr5	96215265	96253609	221.2025	1.708595e-212
13	rs2549794	chr5	96244540	96244541	11747137_x_at	chr5	96215265	96253609	221.0543	2.218743e-212
14	rs12229020	chr12	10117690	10117691	11753241_s_at	chr12	10124195	10137625	220.9615	2.550835e-212
15	rs7673500	chr4	17621378	17621379	11736024_at	chr4	17616279	17627249	220.8615	2.939277e-212
16	rs12229020	chr12	10117690	10117691	11762101_at	chr12	10124033	10136838	220.8420	2.939277e-212
17	rs2549794	chr5	96244540	96244541	11745601_a_at	chr5	96215266	96225192	220.7022	3.816818e-212
18	rs2076608	chr1	17422301	17422302	11727597_at	chr1	17393255	17445948	220.1213	1.373340e-211
19	rs22235914	chr1	17424844	17424845	11727597_at	chr1	17393255	17445948	220.0841	1.417414e-211
20	rs9902260	chr17	62399869	62399870	11757545_x_at	chr17	62399863	62400230	219.5656	4.443247e-211

Table 6

Type I error for data from misspecified ACE model with n=900

	$\rho_{dz} \sim N$		$E \sim \text{Gamma}^1$		$E \sim \text{Gamma}^2$		$E \sim t^1$		$E \sim t^2$	
	pro	mix	pro	mix	pro	mix	pro	mix	pro	mix
$\alpha = 0.05$	0.0522	0.0510	0.0524	0.0502	0.0524	0.0508	0.0525	0.0519	0.0545	0.0491
$\alpha = 0.01$	0.0108	0.0103	0.0110	0.0111	0.0108	0.0101	0.0110	0.0113	0.0116	0.0118
$\alpha = 0.001$	0.0011	0.0011	0.0011	0.0011	0.0011	0.0010	0.0012	0.0011	0.0013	0.0018
$\alpha = 0.0001$	0.0001	0.0001	0.0001	0.0004	0.0001	0.0001	0.0001	0.0002	0.0001	0.0001

pro: proposed score test statistic

mix: mixed effects

E: random error

ρ_{dz} : IBD value for DZ pairs

N: Normal(0.501,0.0389)

Gamma^1 : $\text{Gamma}(2, \sqrt{2})$ and Gamma^2 : $\text{Gamma}(6, \sqrt{6})$

t^1 : $tdf=4$ and t^2 $tdf=3$