Family-Based Bivariate Association Tests for Quantitative Traits

Lei Zhang[1,2], Aaron J. Bonham[2], Jian Li[2], Yu-Fang Pei[1,2], Jie Chen[3], Christopher J. Papasian[2], Hong-Wen Deng[1,2,4,5]*

1 Key Laboratory of Biomedical Information Engineering, Ministry of Education and Institute of Molecular Genetics, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi, People's Republic of China, **2** School of Medicine, University of Missouri-Kansas City, Kansas City, Missouri, United States of America, **3** Department of Mathematics and Statistics, University of Missouri-Kansas City, Kansas City, Missouri, United States of America, **4** Center of System Biomedical Sciences, Shanghai University of Science and Technology, Shanghai, People's Republic of China, **5** College of Life Sciences and Engineering, Beijing Jiao Tong University, Beijing, People's Republic of China

Abstract

The availability of a large number of dense SNPs, high-throughput genotyping and computation methods promotes the application of family-based association tests. While most of the current family-based analyses focus only on individual traits, joint analyses of correlated traits can extract more information and potentially improve the statistical power. However, current TDT-based methods are low-powered. Here, we develop a method for tests of association for bivariate quantitative traits in families. In particular, we correct for population stratification by the use of an integration of principal component analysis and TDT. A score test statistic in the variance-components model is proposed. Extensive simulation studies indicate that the proposed method not only outperforms approaches limited to individual traits when pleiotropic effect is present, but also surpasses the power of two popular bivariate association tests termed FBAT-GEE and FBAT-PC, respectively, while correcting for population stratification. When applied to the GAW16 datasets, the proposed method successfully identifies at the genome-wide level the two SNPs that present pleiotropic effects to HDL and TG traits.

Citation: Zhang L, Bonham AJ, Li J, Pei Y-F, Chen J, et al. (2009) Family-Based Bivariate Association Tests for Quantitative Traits. PLoS ONE 4(12): e8133. doi:10.1371/journal.pone.0008133

Editor: Philip Awadalla, University of Montreal, Canada

Received August 9, 2009; **Accepted** October 6, 2009; **Published** December 2, 2009

Funding: The study was partially supported by grants from the National Institutes of Health (R01AR050496, R21 AG027110, R01 AG026564, P50 AR055081 and R21 AA015973). The investigators of this work also benefited from grants from the National Science Foundation of China, Huo Ying Dong Education Foundation, Shanghai Leading Academic Discipline Project (S30501), HuNan Province, Xi'an Jiaotong University, and the Ministry of Education of China. Computing support was partially provided by the High Performance Computing Cluster Center at Xi'an Jiaotong University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: dengh@umkc.edu

Introduction

Recent technological advances in genotyping along with the capacity to detect increasingly large numbers of single nucleotide polymorphisms (SNPs) have created great demand for developing new strategies to identify genes that underlie phenotypic variation. The availability of high-throughput SNP genotype data is prompting the development of genetic association analyses, including family based association tests (FBAT).

For family data sets, such as the Framingham heart study [1], multiple phenotypes are usually recorded. While most of the current analyses focus only on traits individually, explicitly modeling genetic and environmental correlations among traits can theoretically extract more information and consequently provide a greater power of test. In linkage studies, it has been shown that joint analyses of two correlated traits may substantially improve power for localizing genes that jointly influence complex traits, and for evaluating their effects [2–7]. In association studies, however, only a limited few methods are available [8–10]. Therein, Lange et al. [10] proposed a multivariate generalized estimating equations (GEEs) based method, termed FBAT-GEE. The method FBAT-GEE makes no assumptions on phenotypic distributions and thus can be applied to phenotypes with arbitrary distributions. For quantitative traits, Lange et al. [9] also proposed a generalized principal component analysis (PCA), termed FBAT-PC, which is more powerful than FBAT-GEE.

Both the methods FBAT-GEE and FBAT-PC possess the property of protection against population stratification by a transmission disequilibrium test (TDT)-like strategy. Despite its potential merit, this property comes at the cost of a substantial loss of statistical power by the fact that genotypes at each single marker need to be decomposed in order to correct for population stratification and test association simultaneously. The loss of power may become problematic in the context of genomewide association studies (GWAS) where it is critical to achieve a genomewide significance level in order to judge a positive finding.

Alternatively, the issue of population stratification can be handled at the population level by studying population structures from genotype data of multiple markers [11–17]. Among these approaches, Principal component analysis based methods [12,14,16,17] summarize individual genetic background information. PCA-based methods are proven to be both computationally fast and statistically effective. The extensions of PCA to familial data are also proposed by Zhu et al. [14] and by us previously [18]. Thus, with the availability of large numbers of genotyped markers, a more flexible scheme that would enhance the feasibility of applying FBAT would be to correct for population stratification from multiple markers rather than from any single marker.

In this study, under the framework of the variance-components (VCs) model [19,20], we develop a method for tests of association by joint analysis of two correlated quantitative traits in families. Specifically, Individual genotype scores and phenotypes are adjusted by the use of the principal component analysis to guide against potential population stratification. A score test is proposed with the residual of genotypes and of phenotypes. Statistical properties of the proposed method are investigated through extensive simulations under a variety of conditions, and its performance is compared with existing both univariate and bivariate methods.

Methods

Multivariate Variance-Components Pedigree Model

We describe the problem in the variance-components (VCs) [8,19,20] framework which is a powerful tool for modeling phenotypic variation for continuous traits in families. We describe the model in the context of multivariate phenotypes, although we consider only bivariate situation in our analysis.

Assume that there are N nuclear families with n_i individuals in the i-th family ($i = 1, \ldots, N$). Let $\mathbf{y}_{ij} = (y_{ij1}, \ldots, y_{ijm})'$ be the vector of m ($m = 2$ for bivariate) phenotypes for individual j ($j = 1, \ldots, n_i$) in family i. Further, let $\mathbf{Y}_i = (\mathbf{y}_{i1}', \ldots, \mathbf{y}_{in_i}')'$ be the vector of phenotypes for all members in family i. Under the additive mode of inheritance, the genotype score g_{ij} is defined as 0, 1 and 2 for genotypes "11", "12" and "22", respectively. In the variance-components model, genetic components contributing to phenotypes are decomposed into the fixed effects, e.g., the effects at the specified locus, and the random effects, e.g., the effects of unknown polygenes. Specifically, the phenotype vector $\mathbf{y}_{ij}$ can be described by the following multivariate variance-components model

$$\mathbf{y}_{ij} = \boldsymbol{\mu} + \mathbf{x}_{ij}\boldsymbol{\beta}_{\mathbf{x}} + \mathbf{g}_{ij}\boldsymbol{\beta}_{\mathbf{g}} + \boldsymbol{\alpha}_{ij} + \boldsymbol{\varepsilon}_{ij} \quad \text{for } i = 1, \ldots N, j = 1, \ldots, n_i, \quad (1)$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)'$ denotes the vector of grand means for the m phenotypes; $\mathbf{x}_{ij}$ is a $m \times t$ design matrix for t covariates, e.g., age, sex, and known environment factors, to the m phenotypes, and $\boldsymbol{\beta}_{\mathbf{x}}$ is the vector of corresponding covariate effects; $\mathbf{g}_{ij}$ is a $m \times m$ design matrix for genotype scores with the m principal diagonal elements being g_{ij} and the other elements being 0, and $\boldsymbol{\beta}_{\mathbf{g}}$ the corresponding additive major gene effects. At last, $\boldsymbol{\alpha}_{ij}$ and $\boldsymbol{\varepsilon}_{ij}$ are vectors of length m denoting, respectively, the additive polygenic effects and the residual effects.

Here, the covariate effects $\mathbf{x}_{ij}\boldsymbol{\beta}_{\mathbf{x}}$ and the major gene effects $\mathbf{g}_{ij}\boldsymbol{\beta}_{\mathbf{g}}$ are modeled as fixed, whereas the polygene effects $\boldsymbol{\alpha}_{ij}$ and the residuals $\boldsymbol{\varepsilon}_{ij}$ are modeled as random. Let $\boldsymbol{\alpha}_{ij}$ and $\boldsymbol{\varepsilon}_{ij}$ follow multivariate normal distributions

$$\boldsymbol{\alpha}_{ij} \sim N(\mathbf{0}, \mathbf{A}) \text{ and } \boldsymbol{\varepsilon}_{ij} \sim N(\mathbf{0}, \mathbf{E}),$$

where $\mathbf{A}$ and $\mathbf{E}$ are the $m \times m$ variance-covariance matrices accounting for polygenic ($\boldsymbol{\alpha}_{ij}$) and environmental ($\boldsymbol{\varepsilon}_{ij}$) variation among the traits, respectively, so that

$$\mathrm{E}\left(\mathbf{y}_{ij}\right) = \boldsymbol{\mu} + \mathbf{x}_{ij}\boldsymbol{\beta}_{\mathbf{x}} + \mathbf{g}_{ij}\boldsymbol{\beta}_{\mathbf{g}}, \quad (2)$$

The covariance matrix of $\mathbf{y}_{ij}$, $\sum_{ij}$, has elements

$$\sum_{ij} = \mathbf{G} + \mathbf{A} + \mathbf{E},$$

where $\mathbf{G}$, $\mathbf{A}$, and $\mathbf{E}$ are the $m \times m$ variance-covariance matrices accounting for major gene ($\mathbf{g}_{ij}$), polygenic ($\boldsymbol{\alpha}_{ij}$) and environmental ($\boldsymbol{\varepsilon}_{ij}$) variations, respectively.

The phenotype vector for the family i, $\mathbf{Y}_i$, will then follow a multivariate normal distribution with the mean vector

$$\mathrm{E}(\mathbf{Y}_i) = \boldsymbol{\mu}_i + \mathbf{x}_i\boldsymbol{\beta}_{\mathbf{x}} + \mathbf{g}_i\boldsymbol{\beta}_{\mathbf{g}}, \quad (3)$$

and the covariance matrix

$$\sum_i = \Pi_i \otimes \mathbf{G} + 2\varphi_{\mathbf{i}} \otimes \mathbf{A} + \mathbf{I} \otimes \mathbf{E}, \quad (4)$$

where $\boldsymbol{\mu}_i = \left(\boldsymbol{\mu}', \ldots, \boldsymbol{\mu}'\right)'$ is the mean vector with length mn_i for the family i; $\mathbf{X}_i = \begin{pmatrix} \mathbf{X}_{i1} \\ \ldots, \\ \mathbf{X}_{ini} \end{pmatrix}_{mn_i\mathrm{x}t}$ is the design matrix for covariates, and $\mathbf{g}_i = \begin{pmatrix} \mathbf{g}_{i1} \\ \ldots, \\ \mathbf{g}_{ini} \end{pmatrix}_{mn_i\mathrm{x}m}$ is the design matrix for genotypes at the tested locus; Π_i is the $n_i \times n_i$ identity-by-descent (IBD) matrix at the tested locus (estimated from the genotype data) and $\varphi_{\mathbf{i}}$ is the $n_i \times n_i$ kinship coefficient matrix (estimated from the relationships among individuals), both of which can be calculated from pedigree structures and available genotype data. Finally, $\mathbf{I}$ is the $n_i \times n_i$ identity matrix and $\otimes$ is the Kronecker-product operator for matrices.

Correcting for Population Stratification

When the issue of population stratification exists, the model described above may not provide a valid test. We previously proposed to extend the principal component analysis to include familial data [18]. The method is briefly outlined as follows: founders in each family are selected to form an unrelated sample on which principal component analysis is performed with available genotype data. The calculated principal components are used to estimate these founders' genetic background information and to adjust their genotype scores and phenotypes, as described by Price et al. [12]. Principal components for non-founders in each family are inferred according to those for their founder relatives through a TDT strategy. The inferred principal components are then used to adjust non-founders' genotypes and phenotypes. The approach is also extended to the scenarios where parental information is missing. Denote the adjusted genotypes and phenotypes with an asterisk (*), and we rewrite the equation (3) as

$$\mathrm{E}\left(\mathbf{Y}_i^*\right) = \boldsymbol{\mu}_i + \mathbf{x}_i\boldsymbol{\beta}_{\mathbf{x}} + \mathbf{g}_i^*\boldsymbol{\beta}_{\mathbf{g}}. \quad (5)$$

Tests of Association

With the assumptions of independent families and of within-family multivariate normality distributed phenotypes, the likelihood function of the adjusted genotype and phenotype data is written as

$$L = \Pi_i^N (2\pi)^{-n_i/2} \left|\sum_i\right|^{-1/2} \exp\left\{ -\frac{1}{2}\left(\mathbf{Y}_i^* - \mathrm{E}\left(\mathbf{Y}_i^*\right)\right)' \sum_i^{-1}\left(\mathbf{Y}_i^* - \mathrm{E}\left(\mathbf{Y}_i^*\right)\right) \right\}.$$

Evidence of association is assessed by a statistical hypothesis testing of the null hypothesis H_0: $\boldsymbol{\beta}_{\mathbf{g}} = \mathbf{0}$ (no association) versus the alternative hypothesis H_1: $\boldsymbol{\beta}_\mathbf{g} \neq \mathbf{0}$ (evidence of association). Generally, the hypothesis can be tested by a likelihood ratio test (LRT) where for each marker the maximal likelihoods under the null and alternative hypotheses are estimated respectively. However, the LRT is rather computationally intensive when large numbers of markers are involved, making it prohibitive for large-scale scans. Here we propose a multivariate score test as the extension of that proposed by Chen & Abecasis [21]. The set of parameter in the likelihood function is $\boldsymbol{\theta} = [\boldsymbol{\mu}, \boldsymbol{\beta}_\mathbf{x}, \boldsymbol{\beta}_\mathbf{g}, \mathbf{G}, \mathbf{A}, \mathbf{E}]$. We first fit the model under the null hypothesis (without $\boldsymbol{\beta}_\mathbf{g}$), from which we obtain the maximal likelihood estimates (MLE) of $\sum_i$ and $\mathrm{E}\left(\mathbf{Y}_i^*\right)$, denoted by $\hat{\boldsymbol{\Sigma}}_i^{(\mathrm{base})}$ and $\mathrm{E}\left(\mathbf{Y}_i^*\right)^{(\mathrm{base})}$, respectively, for family i. Under the null hypothesis, the score vector with respect to $\boldsymbol{\beta}_\mathbf{g}$ is defined as

$$\mathbf{S} = \sum\nolimits_{i=1}^{N} \left(\mathbf{g}_i^* - \bar{\mathbf{g}}^*\right)' \hat{\boldsymbol{\Sigma}}_i^{(\mathrm{base})^{-1}} \left(\mathbf{Y}_i^* - \mathrm{E}\left(\mathbf{Y}_i^*\right)^{(\mathrm{base})}\right),$$

and the corresponding variance-covariance matrix is

$$\mathrm{V}(\mathbf{S}) = \sum\nolimits_{i=1}^{N} \left(\mathbf{g}_i^* - \bar{\mathbf{g}}^*\right)' \hat{\boldsymbol{\Sigma}}_i^{(\mathrm{base})^{-1}} \left(\mathbf{g}_i^* - \bar{\mathbf{g}}^*\right).$$

The score test statistic is then defined as

$$T = \mathbf{S}' \mathrm{V}(\mathbf{S})^{-1} \mathbf{S},$$

which asymptotically follows a χ^2 distribution with degree of freedoms (df) being the rank of $\mathrm{V}(\mathbf{S})$, which is equal to m unless there are linear dependences between the phenotypes. The statistic T is valid regardless of the presence of population stratification.

Data Simulations

Statistical properties and performances of the proposed method were investigated via extensive simulation studies. For genotype data, we simulated 998 SNP markers, with the allele frequency for each marker being drawn from the Uniform distribution U(0.1, 0.9). We also simulated two additional SNPs, both with MAF 0.3, as the causal and the test SNPs, respectively. Two hundred nuclear families were simulated by sampling parental genotypes according to allele frequencies, and then by randomly selecting two parents to produce children. Unless otherwise specified, the number of children per family was drawn from a Poisson distribution with mean 2.

Two quantitative traits were simulated by the equations (3) and (4). To each trait, the causal SNP was assumed to explain a specific proportion of phenotypic variation, which was set to 2% by default, and the background polygenic effects were assumed to explain additional 40% of phenotypic variation. The polygenic (ρ_a) and environmental (ρ_e) correlations between the traits were set to 0.4 unless otherwise specified.

When needed, population stratification was generated by mixing equal numbers of families from two discrete populations A and B. Marker allele frequencies in the two populations were generated using the Balding-Nichols model [22]. Specifically, for each SNP, an ancestry allele frequency p was drawn from the Uniform distribution U(0.1, 0.9). The allele frequencies for populations A and B were then drawn from a Beta distribution with parameters $p(1-F_{ST})/F_{ST}$ and $(1-p)(1-F_{ST})/F_{ST}$, where F_{ST} is the measure of genetic distance between the two populations. We set F_{ST} to 0.05 to simulate moderate population stratification. The two populations were generated separately with different phenotypic means (μ_A and μ_B) and different causal and test SNP MAFs ($p_A = 0.2$ and $p_B = 0.4$ for both causal and test SNPs). The values of μ_A and μ_B were selected such that 20% of the phenotypic variance of each trait in the combined population was explained by stratification.

We evaluated the statistical properties, including type I error rates and power, of the proposed method. In all the simulations, the null hypothesis was produced by setting the LD measure r^2 between the causal and the test sites to 0.0, whereas the alternative hypothesis was produced by setting a certain value of r^2 between the two sites. Unless otherwise specified, the value of r^2 under the alternative was set to 1.0 to produce a perfect association between the two sites.

We also studied the effects of various influential factors, including locus effect, correlation coefficient between traits, the level of LD, and family structure, on the performance of the proposed method. For comparison, we also included extensive popular univariate and bivariate methods into analysis. For univariate method, we selected the commonly used method QTDT proposed by Abecasis et al. [23], and the univariate score test proposed by us previously [18]. For bivariate analyses, we selected two popular methods: FBAT-GEE [10] and FBAT-PC [9], which are implemented in the programs FBAT [24] and PBAT [25], respectively. We denote the proposed test and the other methods as T, UT, QTDT, FBAT and PBAT, respectively, throughout the study.

GAW16 Datasets

As an application, we analyzed the Genetic Analysis Workshop 16 (GAW16) Problem 3 simulated data sets with the proposed method. The access and analyses of the GAW16 simulated data sets have been approved by the Institutional Review Board (IRB) of the University of Missouri-Kansas City (UMKC). The GAW16 data sets consist of 6476 subjects from the Framingham Heart Study (FHS), where each subject has real genotypes at approximately 550,000 (549,872) SNPs and simulated phenotypes.

Subjects are distributed among three generations and singletons. After dividing large families into smaller nuclear families and applying some quality controls to the data (for example, as the proposed test cannot analyze half-sibs, we deleted half-sibs from the data), we finally identified 5456 family members from a total of 1815 nuclear families.

A total of four correlated quantitative traits, termed HDL, LDL, TG, CHOL, respectively, are simulated in order to mimic the lipid pathway underlying the development of cardiovascular disease [26]. We focused on the traits HDL and TG. Genetic components underlying each of both traits consist of several SNPs with major effects and 1,000 SNPs with polygenic effects. Two major SNPs (*rs3200218* and *rs8192719*) present pleiotropic effects to both traits in the simulation. Phenotype data are simulated at three pseudo-visits with 10 years apart to mimic the context of longitudinal study, and at each visit, 200 simulated data sets are replicated. The dataset from the first replicate of the first visit was analyzed as suggested by the workshop. Both phenotypes were adjusted by age and sex.

Results

Type I Error Rates

We first estimate type I error rates under a variety of polygenic (ρ_a) and environmental (ρ_e) correlations in homogeneous population setting, as listed in Table 1. Two modes of linkage are considered: 1) the marker is tightly linked to but not associated with the QTL (Linkage); and 2) the marker is neither linked to nor associated with the QTL (No linkage). It is obvious that all

Table 1. Type I error rates at various levels of residual correlations under homogeneous population.

	ρ_a																
	−0.8			**−0.4**			**0.0**			**0.4**			**0.8**				
ρ_e	**T**	**FBAT**	**PBAT**	**T**	**FBAT**	**PBAT**	**T**	**FBAT**	**PBAT**	**T**	**FBAT**	**PBAT**	**T**	**FBAT**	**PBAT**	**QTDT**	**UT**
Linkage[a]																	
−0.8	4.9	4.6	5.1	3.8	4.5	5.3	5.0	5.8	5.1	4.7	5.7	5.3	4.1	4.6	3.8	4.9	5.5
−0.4	7.4	5.8	5.2	5.2	4.6	5.2	5.1	5.7	4.4	5.0	4.5	4.9	5.2	6.0	4.8		
0.0	5.3	4.5	5.3	4.7	5.8	6.8	5.3	3.9	4.3	4.7	6.6	4.0	5.8	5.7	5.2		
0.4	5.3	4.8	4.2	5.3	4.5	5.1	5.1	3.7	4.6	5.0	5.2	5.8	4.3	4.6	5.5		
0.8	4.7	4.6	5.3	4.3	4.2	4.9	4.8	4.8	5.4	4.8	5.7	5.3	3.9	4.9	3.9		
No Linkage[b]																	
−0.8	4.7	4.5	3.8	5.7	4.9	5.4	4.6	4.6	4.2	5.8	4.9	4.2	5.9	4.3	4.9	5.2	4.0
−0.4	5.1	3.9	4.5	5.7	5.1	5.4	5.4	5.5	5.2	4.1	6.5	4.1	4.3	4.8	5.3		
0.0	5.5	4.8	3.9	4.4	4.2	5.1	3.8	6.5	5.8	4.0	6.0	5.1	4.3	5.2	4.1		
0.4	5.2	5.4	3.7	4.6	4.9	4.2	4.9	6.1	3.8	4.1	4.6	4.9	5.6	5.4	4.6		
0.8	5.2	6.2	5.2	6.4	4.1	4.4	4.4	5.0	4.8	4.3	5.4	6.0	4.2	4.5	4.6		

Two hundred nuclear families were simulated, with the number of children per family being drawn from a Beta distribution with mean 2. Type I error rate was estimated at nominal level 5% on 1,000 replicates, with various levels of polygenic correlation (ρ_a) and environmental correlation (ρ_e).

[a] the test site was linked to but not associated with the causal site.

[b] the test site was neither linked to nor associated with the causal site.

Abbreviations: T, the proposed bivariate method; FBAT, the method FBAT-GEE [10] implemented in the software FBAT [24]; PBAT, the method FBAT-PC [9] implemented in the software PBAT [25]; QTDT, the method proposed by Abecasis et al. [23] and implemented in the software QTDT; UT, the univariate test in our previous study [18].

doi:10.1371/journal.pone.0008133.t001

methods have correct type I error rates that are close to the target levels, regardless of the existence of linkage.

Table 2 lists the type I error rates when families from two populations are admixed. All methods again have correct error rates, implying their ability to protect against population stratification.

We also estimate the type I error rates when parents in each family are missing, as presented in table 3. In the table, the number of children per family varies from 2 to 4, with the total number of children being fixed at 480. Again, all investigated methods have correct type I error rates regardless of the presence of linkage or population stratification, indicating the proposed method as well as the others is applicable to studies with missing parental information.

Power Estimates

Powers of various methods affected by ρ_a and ρ_e are listed in table 4. For all the three bivariate approaches, power increases as

Table 2. Type I error rates at various levels of residual correlations under admixed population.

	ρ_a																
	−0.8			**−0.4**			**0.0**			**0.4**			**0.8**				
ρ_e	**T**	**FBAT**	**PBAT**	**T**	**FBAT**	**PBAT**	**T**	**FBAT**	**PBAT**	**T**	**FBAT**	**PBAT**	**T**	**FBAT**	**PBAT**	**QTDT**	**UT**
Linkage[a]																	
−0.8	2.0	4.5	4.5	3.5	5.7	4.0	3.2	4.0	4.3	3.8	5.4	5.5	4.3	5.7	5.4	4.6	4.4
−0.4	3.5	5.4	5.7	3.3	4.1	5.1	4.1	5.2	5.8	2.8	3.9	5.7	3.8	4.1	4.0		
0.0	4.3	4.2	3.7	2.9	4.5	4.3	3.8	4.2	5.1	5.5	4.7	5.1	4.0	5.0	5.1		
0.4	2.8	4.0	4.9	2.7	4.4	4.5	3.0	5.3	3.8	2.2	3.9	4.3	3.2	4.9	3.7		
0.8	2.9	5.8	5.8	3.2	4.9	3.2	2.8	4.9	4.8	2.6	4.0	5.8	3.6	6.0	4.8		
No Linkage[b]																	
−0.8	1.9	5.2	4.2	2.6	4.9	5.7	3.6	5.1	4.9	2.9	4.8	4.0	4.1	4.6	6.0	5.0	4.5
−0.4	3.7	4.9	5.0	4.2	4.4	5.2	3.0	6.0	5.8	3.4	5.4	5.4	4.6	5.3	5.3		
0.0	2.8	4.6	4.4	3.1	5.4	5.9	3.4	5.8	5.0	3.8	6.0	5.5	4.5	4.8	5.0		
0.4	3.9	5.4	5.3	2.8	4.8	4.6	4.1	5.4	5.5	4.0	4.9	4.8	4.7	4.5	4.7		
0.8	3.7	5.1	5.9	4.4	5.1	5.8	4.1	4.8	5.3	3.9	6.5	5.7	3.9	4.2	4.7		

Two hundred nuclear families were simulated by admixing 100 from two populations A and B. See Table 1 legend for simulation and abbreviation details.

doi:10.1371/journal.pone.0008133.t002

Table 3. Type I error rates when parents are missing.

	No. of children per family														
	2					3					4				
	T	FBAT	PBAT	QTDT	UT	T	FBAT	PBAT	QTDT	UT	T	FBAT	PBAT	QTDT	UT
Homogeneous															
Linkage	6.3	4.7	5.9	4.5	4.7	3.3	5.1	3.6	5.2	4.9	5.3	4.7	5.0	5.2	4.2
No Linkage	5.0	3.3	4.3	4.7	3.6	4.9	4.1	4.6	4.9	4.7	5.9	4.7	5.0	4.8	5.3
Admixture															
Linkage	3.9	4.7	5.2	5.2	3.5	4.6	5.3	5.0	4.8	3.9	4.2	4.3	5.1	4.1	4.5
No Linkage	4.0	4.5	4.6	4.7	3.8	4.7	5.9	6.1	5.8	4.7	5.3	4.5	5.2	5.1	4.8

The number of children per family varied from 2 to 4, and the number of families varied accordingly with the constraint that the total number of children was fixed at 480. Both polygenic and environmental correlations were set to 0.4. See Table 1 legend for simulation and abbreviation details.

doi:10.1371/journal.pone.0008133.t003

residual correlations (ρ_a and/or ρ_e) decrease from +1.0 to −1.0. For example, under homogenous population, the power of the proposed method is 87.6% when both ρ_a and ρ_e are +0.8, and increases to 100.0% when both correlations decrease to −0.8. In additional simulations where the major gene correlation between the traits is constrained to −1.0 rather than +1.0, we observe an opposite trend in power change; power increases as ρ_a and/or ρ_e increase from −1.0 to +1.0 (data not shown). Therefore, our simulation results indicate that the power of the bivariate approaches increases when the effects of the major-gene and those of the residuals (polygenic and environmental) act in increasingly dissimilar directions, which coincides with previous studies in the literature of linkage analyses [3].

Among the three bivariate methods, the proposed one has the highest power under all the parameter settings. The power improvement is remarkably large. For example, when both ρ_a and ρ_e are +0.8 under homogeneous population, the power for the proposed method is 90.5%, whereas it is only 48.8% for FBAT and 58.0% for PBAT. For the other two methods, PBAT has a higher power than FBAT.

When comparing the bivariate and univariate approaches, the proposed method has higher power than UT under most conditions, and than QTDT under all the conditions. UT has higher power than QTDT, consistent with our previous study [18]. FBAT and PBAT have higher power than QTDT unless the traits are highly positively correlated.

Power when parental information is missing is presented in table 5. The trends in power changes are similar with those when parental information is available. The proposed method again has the highest power, and the bivariate tests have higher power over univariate tests in most situations.

We also study the effects of two factors, including the level of LD and family structure, on power estimations. As presented in table 6, all the methods have increased power with increased r^2. As the number of children per family increases, the power of the proposed method and UT decreases a little, whereas that of the

Table 4. Power estimates at various levels of residual correlations.

	ρ_a																
	−0.8			−0.4			0.0			0.4			0.8				
ρ_e	T	FBAT	PBAT	T	FBAT	PBAT	T	FBAT	PBAT	T	FBAT	PBAT	T	FBAT	PBAT	QTDT	UT
Homogeneous																	
−0.8	100.0	100.0	100.0	100.0	99.3	99.7	100.0	95.9	98.3	100.0	87.5	94.5	100.0	80.0	88.7	61.3	93.7
−0.4	100.0	97.3	99.5	100.0	90.9	96.1	100.0	83.8	92.0	99.8	74.9	85.7	99.3	71.5	82.4		
0.0	100.0	89.2	93.2	100.0	80.5	89.6	99.4	75.2	84.6	99.2	65.4	77.1	98.4	59.7	70.5		
0.4	100.0	79.0	88.4	99.4	70.6	79.7	98.0	62.8	76.5	97.4	60.3	71.8	96.1	53.7	65.8		
0.8	99.7	66.6	76.9	98.2	61.8	73.6	96.0	58.2	65.5	95.6	55.1	64.1	90.5	48.8	58.0		
Admixture																	
−0.8	100.0	85.8	94.5	99.7	76.0	83.1	99.9	71.2	80.3	100.0	64.5	74.6	99.8	57.7	66.8	55.1	91.4
−0.4	99.8	75.1	82.4	100.0	69.8	80.3	99.3	61.0	74.8	99.3	56.3	63.1	98.4	49.7	57.2		
0.0	99.9	64.0	78.3	99.3	57.9	64.7	98.0	53.3	57.6	97.5	50.9	54.2	94.9	44.7	50.3		
0.4	98.4	55.7	64.7	97.6	50.4	58.6	96.3	49.1	53.4	93.4	44.2	47.7	92.0	40.4	44.5		
0.8	95.6	48.8	55.5	94.0	45.5	52.2	92.3	40.4	48.0	91.3	35.7	42.5	87.6	37.4	39.6		

Power was estimated at the nominal level of 5% based on 1,000 replicates. The value of r^2 between test and causal sites was set to 1.0, and the recombination rate between the sites was set to 0.01. See Table 1 and 2 legends for simulation and abbreviation details.

doi:10.1371/journal.pone.0008133.t004

Table 5. Power estimates at various levels of residual correlations when parents are missing.

	ρ_a																
	−0.8			**−0.4**			**0.0**			**0.4**			**0.8**				
ρ_e	**T**	**FBAT**	**PBAT**	**T**	**FBAT**	**PBAT**	**T**	**FBAT**	**PBAT**	**T**	**FBAT**	**PBAT**	**T**	**FBAT**	**PBAT**	**QTDT**	**UT**
Homogeneous																	
−0.8	99.0	100.0	100.0	99.2	96.7	99.2	99.3	90.1	97.5	98.4	85.5	94.4	97.5	77.5	88.9	43.4	72.8
−0.4	99.7	85.8	95.7	99.2	79.1	92.8	96.8	71.2	85.8	94.7	66.9	83.1	90.3	60.5	76.9		
0.0	99.0	67.7	86.1	95.8	60.2	80.7	92.3	57.2	76.6	85.6	51.4	69.9	81.6	46.4	66.0		
0.4	94.3	54.3	74.5	89.0	46.1	70.2	84.8	42.6	61.9	79.7	40.7	56.8	76.4	39.1	60.2		
0.8	88.2	42.7	66.7	81.4	37.8	59.5	77.7	37.2	58.2	76.3	35.7	57.7	68.9	31.4	48.8		
Admixture																	
−0.8	99.9	99.3	99.7	99.4	94.7	97.3	98.7	87.9	94.6	94.6	77.5	88.2	93.1	70.1	78.6	40.3	63.2
−0.4	98.2	83.4	96.2	94.9	70.2	90.7	90.8	64.1	83.1	86.8	58.9	78.4	81.0	54.7	74.3		
0.0	92.0	61.2	87.7	87.6	57.0	78.9	83.3	52.6	75.5	78.9	47.3	72.3	74.5	42.9	65.8		
0.4	82.7	45.8	72.1	78.8	42.6	69.9	73.2	40.8	63.4	68.7	39.0	59.6	64.1	35.0	57.2		
0.8	76.9	41.3	65.2	69.8	36.4	61.7	63.1	34.7	55.4	62.3	32.5	54.1	57.8	29.5	39.5		

Parents were deleted from the simulation. See Table 1 and 2 legends for simulation and abbreviation details.

doi:10.1371/journal.pone.0008133.t005

other methods increases instead. Again, the proposed method has the highest power in all situations.

Analysis of GAW16 Datasets

As an application, we analyze the GAW16 simulated datasets as described in the preceding section. On a desktop computer with a single 2.8GHz Intel Xeon CPU, the computation from the proposed method for the scan with 550K SNPs takes a total time of approximately 20 hours. Figure 1A presents the quantile-quantile (QQ) plot (left) and log-QQ plot (right), and Figure 1B plots raw p-values over 22 autosomes. Obviously, the overall p-values distribute uniformly between 0 and 1. The most significant signal from the proposed method is observed at SNP *rs10820738*, with a p-value 9.55e-17. UT has a p-value 4.42e-17 at this SNP. The second most significant signal corresponds to one of its flanking SNPs *rs2297398* (7.7kb apart), with a p-value 1.88e-15 from the proposed method. The SNP *rs10820738* explains the most phenotypic variation (1.0%) for HDL trait in the GAW16 simulation, but none for TG trait. The third most significant SNP is *rs3200218* with a p-value 3.47e-12. This SNP is exactly one of the two SNPs that present pleiotropic effects to both traits. The p-value at the other major pleiotropic SNP *rs8192719* is 3.07e-6, which does not achieve the genomewide significance level. However, one of its flanking SNPs *rs7249735* (12.8kb apart) presents a p-value 2.77e-8 that is significant at the genomewide level. Generally speaking, the proposed method successfully identifies both pleiotropic SNPs at the genomewide level. We also list p-values at these two SNPs from other methods in Table 7. Obviously, most of them do not achieve genomewide significant level, further demonstrating the advantage of the proposed method.

Table 6. The effects of LD level and family structures on power estimates.

	No. of children per family														
	1					**2**					**3**				
r^2	**T**	**FBAT**	**PBAT**	**QTDT**	**UT**	**T**	**FBAT**	**PBAT**	**QTDT**	**UT**	**T**	**FBAT**	**PBAT**	**QTDT**	**UT**
Homogeneous															
0.25	46.2	14.7	17.3	14.5	41.1	42.1	15.7	20.9	18.1	43.6	41.9	18.9	23.4	20.8	40.8
0.50	72.7	18.2	26.4	20.5	70.3	74.6	26.1	37.9	30.4	66.2	72.2	34.2	41.1	36.3	67.9
0.75	89.4	28.8	39.0	31.5	86.0	91.0	41.3	53.3	42.0	85.1	89.5	49.5	59.7	50.8	83.7
1.00	97.1	36.9	49.5	39.6	92.8	96.0	52.2	63.7	54.0	92.4	95.7	59.9	70.5	62.5	92.5
Admixture															
0.25	32.5	11.4	15.4	13.2	31.6	33.0	13.7	18.9	16.5	29.0	29.2	12.8	19.5	16.4	33.1
0.50	60.8	13.9	18.2	17.7	58.2	61.0	22.9	34.8	28.8	58.7	60.6	25.1	35.4	30.7	57.5
0.75	80.9	22.2	36.3	26.7	77.4	81.6	32.2	48.6	37.6	75.7	79.5	33.4	55.9	43.7	74.2
1.00	92.3	25.4	41.6	34.1	87.1	92.4	42.0	57.1	49.3	86.3	91.9	45.7	66.8	55.6	89.5

The number of children per family varied from 1 to 3, with the total number of individuals being constrained at 800. See Table 1 and 2 legends for simulation and abbreviation details.

doi:10.1371/journal.pone.0008133.t006

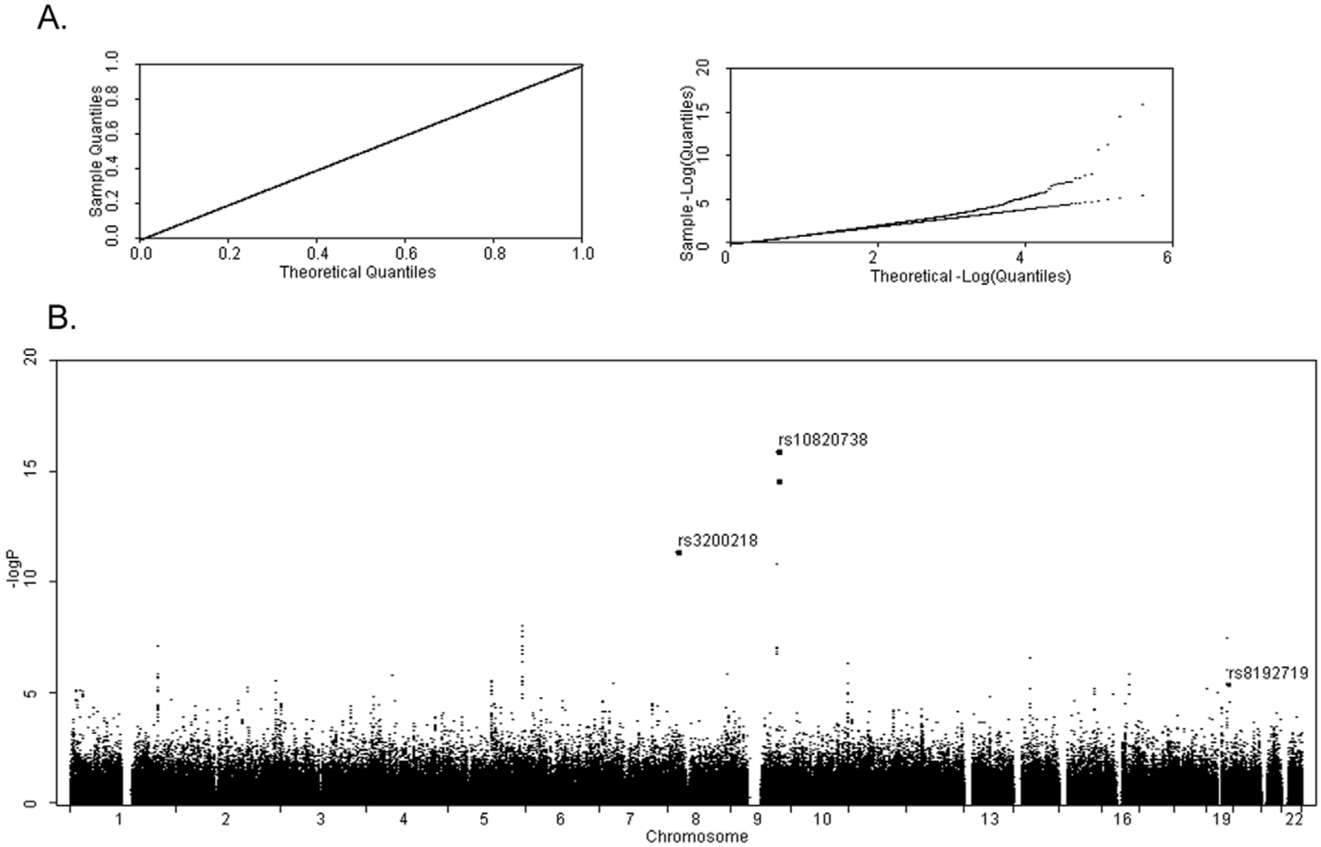

Figure 1. Application of the proposed method to the GAW16 simulated datasets. The GAW16 simulated HDL and TG traits were analyzed. Figure 1A, the quantile-quantile (QQ) plot (left), and log-QQ plot (right); Figure 1B, raw p-values of the genome-wide scan.
doi:10.1371/journal.pone.0008133.g001

Discussion

In this study, we have presented a bivariate test of association for quantitative traits in families, by the use of the multivariate variance-components model. In particular, the proposed method uses principal component analysis to correct for population stratification. Simulation studies have shown that the proposed method not only outperforms the analysis focused on individual traits when pleiotropic effect is present, but also has increased power compared with the existing bivariate methods, while correcting for population stratification.

The strength of bivariate analyses is influenced by correlations between traits. Our simulation results show that bivariate analyses are more powerful when the major genes and the residual factors act in more dissimilar ways. For example, by constraining the major-gene correlation to +1.0, bivariate approaches are most powerful when both ρ_a and ρ_e are equal to -0.8, corresponding to an approximate phenotypic correlation of -0.76. When pleiotropic effect is present, bivariate analysis is more powerful than univariate analysis unless the residual correlation is high and in the same direction to the correlation of major gene effect, coinciding with the findings in linkage studies [27]. When pleiotropic effect is not present and there is weak or no correlation between traits, on the other hand, bivariate analysis is less powerful than univariate analysis [28]. In our analyses of the SNP *rs10820738* that has no pleiotropic effect in the GAW16 simulation, the p-value from the proposed bivariate method, 9.55e-17, is slightly higher than that from the univariate method UT, 4.42e-17. In practical applications in which the existence of pleiotropic effect is unknown *in prior*, bivariate analysis is not necessarily more powerful than univariate analysis, even when the traits are strongly correlated. Bivariate analysis should thus be processed with caution, and combing the results of bivariate and univariate analyses is warranted.

Thus, our simulations provide a statement in demonstrating that bivariate approaches are more powerful than univariate analyses unless major-gene effects and residual effects are very highly correlated in the same direction, which coincides with the conclusion of Amos et al. [27].

Our method that is developed by extending the variance-components model offers a way to powerfully/robustly perform bivariate association analysis in the presence of linkage in general pedigrees. The variance-components model is advantageous in detecting QTL in the following two aspects: first, it combines the

Table 7. P-values at pleiotropic SNPs.

	p-value				
SNP	**T**	**FBAT**	**PBAT**	**QTDT**	**UT**
rs3200218	3.47e-12	0.03	5.20E-04	1.91E-04	2.72e-08
rs8192719	3.07e-6	0.94	0.04	0.04	2.38e-06

For univariate tests, e.g., QTDT and UT, the uniform bivariate p-value was obtained by adjusting the minimum of the two univairate p-values by multiple testing correction (multiplying by 2). See Table 1 legend for abbreviation details.
doi:10.1371/journal.pone.0008133.t007

analysis of linkage and association and would increase the power of detecting QTL when the marker, itself is not the QTL, is associated with the QTL. Second, the prior evidence on linkage can be incorporated to indicate LD strength between the QTL and the marker [29].

Another strengthen of our method is decomposing individual genotype scores by principal component analysis rather than by TDT-like strategy for controlling population stratification. The resulting test statistic provides largely improved power over existing TDT-based methods, where the latter may be prohibitive for application to genome scans due to their poor powers. For example, under the moderate setting where both polygenic and environmental correlation coefficients were set to 0.4 and the locus effect were set to 2%, we observed 164 and 24 significant results over 1,000 replicates for the proposed method and UT, respectively, but only 6, 8, and 4 for FBAT-GEE, FBAT-PC, and QTDT, respectively, at the genome-wide level 1.0e-6.

An interesting observation from our simulations is that family structures influence the power of the investigated methods in different patterns. For FBAT-GEE, FBAT-PC and QTDT that control population stratification through TDT, small numbers of large families provide more power than large numbers of small families. This is not surprising, since with these methods the parental information is used to control the stratification, and consequently, only the information of children contributes to test statistics. Contrary, the power decreases slightly as the number of children per family increases for the proposed method. This appears to be caused by the fact that a large number of small independent families provides more information on allele distributions than a small number of larger families can provide.

In this manuscript, we focus our attention on data with nuclear families. However, the proposed method is applicable to extended pedigrees as well. The described multivariate variance-components model can be directly applied to extended pedigrees. As for correcting for population stratification, the extension of principal component analysis coupled with TDT strategy to extended pedigrees is also straightforward. For example, all founders can be collected to form an unrelated sample. In cases where there is no founder, one sib can be randomly selected into the unrelated sample, as we described in reference [18]. PCA will then be performed on the unrelated sample, and both the genotype and the phenotype for each subject in the unrelated sample are adjusted accordingly. For subjects not in the unrelated sample, their principal components can be calculated by that of their parents or sib that is in the unrelated sample. The process will carry on recursively until all non-founders are adjusted. Some specialized algorithms, e.g., the one described in [30], can be adopted in a relatively simple manner. However, the performance when applying to extended pedigrees is unknown deserves further endeavor.

In summary, we develop a novel method for family based bivariate association test. Our method is more powerful than currently available bivariate methods. The proposed method is computationally effective and can complete a typical GWAS scan within hours. The program implementing our proposed method is available upon request to us.

Acknowledgments

We acknowledge the Genetic Analysis Workshop 16 (GAW16) and the Framingham Heart Study (FHS) for permission of use of the data. We also acknowledge Yanfang Guo and Feng Zhang for helpful discussions. Computing support was partially provided by the High Performance Computing Cluster Center at Xi'an Jiaotong University.

Author Contributions

Conceived and designed the experiments: LZ HWD. Performed the experiments: LZ. Analyzed the data: LZ AJB YFP. Contributed reagents/materials/analysis tools: AJB JL YFP. Wrote the paper: LZ AJB JL YFP JC CJP HWD.

References

1. Cupples LA, Arruda HT, Benjamin EJ, D'Agostino RB, Sr., Demissie S, et al. (2007) The Framingham Heart Study 100K SNP genome-wide association study resource: overview of 17 phenotype working group reports. BMC Med Genet 8 Suppl 1: S1.
2. Almasy L, Dyer TD, Blangero J (1997) Bivariate quantitative trait linkage analysis: pleiotropy versus co-incident linkages. Genet Epidemiol 14: 953–958.
3. Amos CI, Elston RC, Bonney GE, Keats BJ, Berenson GS (1990) A multivariate method for detecting genetic linkage, with application to a pedigree with an adverse lipoprotein phenotype. Am J Hum Genet 47: 247–254.
4. de Andrade M, Mendell NR (2005) Summary of contributions to GAW Group 12: multivariate methods. Genet Epidemiol 29 Suppl 1: S91–95.
5. Jiang C, Zeng ZB (1995) Multiple trait analysis of genetic mapping for quantitative trait loci. Genetics 140: 1111–1127.
6. Li Y, Sung WK, Liu JJ (2007) Association mapping via regularized regression analysis of single-nucleotide-polymorphism haplotypes in variable-sized sliding windows. Am J Hum Genet 80: 705–715.
7. Williams JT, Van Eerdewegh P, Almasy L, Blangero J (1999) Joint multipoint linkage analysis of multivariate qualitative and quantitative traits. I. Likelihood formulation and simulation results. Am J Hum Genet 65: 1134–1147.
8. Jung J, Zhong M, Liu L, Fan R (2008) Bivariate combined linkage and association mapping of quantitative trait loci. Genet Epidemiol 32: 396–412.
9. Lange C, van Steen K, Andrew T, Lyon H, DeMeo DL, et al. (2004) A family-based association test for repeatedly measured quantitative traits adjusting for unknown environmental and/or polygenic effects. Stat Appl Genet Mol Biol 3: Article17.
10. Lange C, Silverman EK, Xu X, Weiss ST, Laird NM (2003) A multivariate family-based association test using generalized estimating equations: FBAT-GEE. Biostatistics 4: 195–206.
11. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55: 997–1004.
12. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904–909.
13. Wall JD, Pritchard JK (2003) Assessing the performance of the haplotype block model of linkage disequilibrium. Am J Hum Genet 73: 502–515.
14. Zhu X, Li S, Cooper RS, Elston RC (2008) A unified association analysis approach for family and unrelated samples correcting for stratification. Am J Hum Genet 82: 352–365.
15. Chen HS, Zhu X, Zhao H, Zhang S (2003) Qualitative semi-parametric test for genetic associations in case-control designs under structured populations. Ann Hum Genet 67: 250–264.
16. Zhang S, Zhu X, Zhao H (2003) On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. Genet Epidemiol 24: 44–56.
17. Zhu X, Zhang S, Zhao H, Cooper RS (2002) Association mapping, using a mixture model for complex traits. Genet Epidemiol 23: 181–196.
18. Zhang L, Li J, Pei Y-F, Liu Y, Deng H-W (2009) Tests of association for quantitative traits in nuclear families using principal components to correct for population stratification. Annals of Human Genetics In Press.
19. Schaid DJ (1996) General score tests for associations of genetic markers with disease using cases and their parents. Genet Epidemiol 13: 423–449.
20. Amos CI (1994) Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet 54: 535–543.
21. Chen WM, Abecasis GR (2007) Family-based association tests for genomewide association scans. Am J Hum Genet 81: 913–926.
22. Balding DJ, Nichols RA (1995) A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica 96: 3–12.
23. Abecasis GR, Cardon LR, Cookson WO (2000) A general test of association for quantitative traits in nuclear families. Am J Hum Genet 66: 279–292.
24. Laird NM, Horvath S, Xu X (2000) Implementing a unified approach to family-based tests of association. Genet Epidemiol 19 Suppl 1: S36–42.
25. Lange C, DeMeo D, Silverman EK, Weiss ST, Laird NM (2004) PBAT: tools for family-based association studies. Am J Hum Genet 74: 367–369.
26. Kannel WB, Dawber TR, Kagan A, Revotskie N, Stokes J 3rd (1961) Factors of risk in the development of coronary heart disease–six year follow-up experience. The Framingham Study. Ann Intern Med 55: 33–50.
27. Amos CI, Laing AE (1993) A comparison of univariate and multivariate tests for genetic linkage. Genet Epidemiol 10: 671–676.

28. Zhang L, Pei YF, Li J, Papasian CJ, Deng HW (2009) Univariate/multivariate genome-wide association scans using data from families and unrelated samples. PLoS One 4: e6502.

29. Fulker DW, Cherny SS, Sham PC, Hewitt JK (1999) Combined linkage and association sib-pair analysis for quantitative traits. Am J Hum Genet 64: 259–267.

30. Abecasis GR, Cookson WO, Cardon LR (2000) Pedigree tests of transmission disequilibrium. Eur J Hum Genet 8: 545–551.