

Bivariate Combined Linkage and Association Mapping of Quantitative Trait Loci

Jeesun Jung,¹ Ming Zhong,² Lian Liu,² and Ruzong Fan^{2*}

¹Department of Medical and Molecular Genetics, Indiana University, School of Medicine, Indianapolis, Indiana

²Department of Statistics, The Texas A&M University, Texas

In this paper, bivariate/multivariate variance component models are proposed for high-resolution combined linkage and association mapping of quantitative trait loci (QTL), based on combinations of pedigree and population data. Suppose that a quantitative trait locus is located in a chromosome region that exerts pleiotropic effects on multiple quantitative traits. In the region, multiple markers such as single nucleotide polymorphisms are typed. Two regression models, “genotype effect model” and “additive effect model”, are proposed to model the association between the markers and the trait locus. The linkage information, i.e., recombination fractions between the QTL and the markers, is modeled in the variance and covariance matrix. By analytical formulae, we show that the “genotype effect model” can be used to model the additive and dominant effects simultaneously; the “additive effect model” only takes care of additive effect. Based on the two models, *F*-test statistics are proposed to test association between the QTL and markers. By analytical power analysis, we show that bivariate models can be more powerful than univariate models. For moderate-sized samples, the proposed models lead to correct type I error rates; and so the models are reasonably robust. As a practical example, the method is applied to analyze the genetic inheritance of rheumatoid arthritis for the data of The North American Rheumatoid Arthritis Consortium, Problem 2, Genetic Analysis Workshop 15, which confirms the advantage of the proposed bivariate models. *Genet. Epidemiol.* 32:396–412, 2008. © 2008 Wiley-Liss, Inc.

Key words: multivariate analysis; linkage disequilibrium mapping; QTL

Contract grant sponsor: National Science; Contract grant number: DMS-0505025; Contract grant sponsor: GAW; Contract grant number: R01GM031575.

*Correspondence to: Ruzong Fan, Department of Statistics, The Texas A&M University, Texas 77843 3143. E-mail: rfan@stat.tamu.edu
Received 22 June 2007; Revised 20 November 2007; Accepted 2 January 2008

Published online 15 February 2008 in Wiley InterScience (www.interscience.wiley.com).

DOI: 10.1002/gepi.20313

INTRODUCTION

In almost all genetics studies, investigators collect data of multiple phenotypes. However, many studies only analyze one phenotype at a time. When several phenotypes are analyzed separately, potential problems arise. For instance, if two phenotypes are correlated to each other, it is not right to treat them as independent variables and analyze them separately [Allison et al., 1998; Amos et al., 2001; Evans, 2002; Kraft and de Andrade, 2003]. Moreover, separate analysis of several phenotypes can make the interpretation of results complicated. For related multiple traits, genetic analysis should be carried out by multivariate methods [Almasy et al., 1997; Blangero and Konigsberg, 1991; Lange, 2002]. In the human genome, gene clusters are sometimes found in a tightly linked chromosome region, like the leukocyte antigens and the beta-hemoglobins. It is important to know if a single gene or mutation has many different consequences in an individual, i.e., pleiotropy.

In this paper, we propose bivariate/multivariate variance component models for combined linkage and association mapping of quantitative trait loci/locus (QTL), which generalizes our previous univariate model [Fan and Xiong, 2003; Jung et al., 2005]. For the convenience

of presentation, a bivariate analysis framework is utilized to analyze two quantitative traits, which may be generalized to a multivariate analysis of multiple traits. In addition, we present the model using two markers, and the markers can be di-allelic such as single nucleotide polymorphisms (SNPs), or can be multi-allelic such as a micro-satellites. For each of the multiple phenotypes, the mean component and variance/covariance structure can be constructed as our previous work [Fan et al., 2005]. In addition, we construct the correlations among the multiple phenotypes using the framework presented in Chapter 8, Lange [2002]. In this way, the mean vectors and variance-covariance matrix of multiple phenotypes are fully characterized.

To construct the mean components of the model, two regression models (i.e., “genotype effect model” and “additive effect model”) are proposed to model the association between the markers and the trait locus. The linkage information, i.e., recombination fractions between the QTL and the markers, is modeled in the variance and covariance matrix. On building the models, theoretical analysis is presented to show that the models are valid in combined linkage and association mapping of QTL. By analytical formulae, we show that measures of linkage disequilibrium (LD) between the trait locus and the markers, i.e., the association between the traits and the

markers, are contained in the regression coefficients. Based on the two models, F -test statistics are proposed to test association between the QTL and markers. The non-centrality parameter approximations of F -test statistics are derived to make power calculation and comparison.

To show the robustness of the proposed models, type I error rates are calculated based on nuclear families and multi-generation pedigrees. The method is applied to analyze the genetic inheritance of rheumatoid arthritis for the data of The North American Rheumatoid Arthritis Consortium, Problem 2, Genetic Analysis Workshop (GAW) 15. A computer program in C++, named Combined Linkage and Association Mapping of QTL, is written to implement the proposed models, which is available on request from the authors.

METHODS

Consider a quantitative trait locus Q , which is located at an autosome. Suppose that there are two alleles Q_1 and Q_2 at the trait locus with frequencies q_1 and q_2 , respectively. Assume that the QTL exerts pleiotropic effects on multiple quantitative traits. For the convenience of presentation, we assume that the QTL affects two quantitative traits. In a region of the QTL Q , suppose that multiple markers are typed. For simplicity, we use two marker A and B in our analysis, but the models and methods can be easily generalized to use multiple markers. Suppose that the markers A and B are in Hardy-Weinberg equilibrium. Let us denote the alleles of marker A by A_1, \dots, A_m , where m is the number of alleles. Let the frequency of A_g be $P_{A_g}, g = 1, 2, \dots, m$. There are $J_A = m(m+1)/2$ possible genotypes, which can be listed as $A_1A_1, \dots, A_mA_m, A_1A_2, \dots, A_1A_m, \dots, A_{m-1}A_m$. Assume that the marker B has n alleles denoted by B_1, \dots, B_n . Let the frequency of allele B_k be $P_{B_k}, k = 1, 2, \dots, n$. There are $J_B = n(n+1)/2$ possible genotypes, which can be listed as $B_1B_1, \dots, B_nB_n, B_1B_2, \dots, B_1B_n, \dots, B_{n-1}B_n$.

POPULATION MODELS

Let y_1 and y_2 be two quantitative trait values of an individual with genotype G_A at marker A and genotype G_B at marker B . Following Fan et al. [2006], consider the following "genotype effect model" under normality:

$$y_i = w_i \gamma_i + \alpha_i + \sum_{g=1}^{m-1} x_A^{(g)} \alpha_{Aig} + \sum_{k=1}^{n-1} x_B^{(k)} \alpha_{Bik} + \sum_{1 \leq g < h \leq m} z_A^{(gh)} \delta_{Aigh} + \sum_{1 \leq k < l \leq n} z_B^{(kl)} \delta_{Bikl} + H_i + e_i, \quad (1)$$

where w_i is a row vector of co-variates such as sex and age for the i -trait, γ_i is a column vector of regression coefficients of w_i , H_i is polygenic effect, and e_i is error term. Assume that H_i is normal $N(0, \sigma_{G_i}^2)$, e_i is normal $N(0, \sigma_{e_i}^2)$, and H_i is independent of e_i . Here $\sigma_{G_i}^2$ is polygenic additive variance and $\sigma_{e_i}^2$ is error variance. The dummy

variables $x_A^{(g)}, z_A^{(gh)}, x_B^{(k)}$ and $z_B^{(kl)}$ are defined by

$$x_A^{(g)} = \begin{cases} 2 & \text{if } G_A = A_g A_g, \\ 1 & \text{if } G_A = A_g A_h, h \neq g, \\ 0 & \text{else,} \end{cases}$$

$$z_A^{(gh)} = \begin{cases} -P_{A_h}^2 & \text{if } G_A = A_g A_g, \\ P_{A_g} P_{A_h} & \text{if } G_A = A_g A_h, \\ -P_{A_g}^2 & \text{if } G_A = A_h A_h, \\ 0 & \text{else} \end{cases} \quad (2)$$

$$x_B^{(k)} = \begin{cases} 2 & \text{if } G_B = B_k B_k, \\ 1 & \text{if } G_B = B_k B_l, l \neq k, \\ 0 & \text{else,} \end{cases}$$

$$z_B^{(kl)} = \begin{cases} -P_{B_l}^2 & \text{if } G_B = B_k B_k, \\ P_{B_k} P_{B_l} & \text{if } G_B = B_k B_l, \\ -P_{B_k}^2 & \text{if } G_B = B_l B_l, \\ 0 & \text{else} \end{cases}$$

and $\alpha_i, \alpha_{Aig}, \alpha_{Bik}, \delta_{Aigh}, \delta_{Bikl}$ are regression coefficients of the dummy variables. In addition to the co-variate effects, there are $J_A + J_B - 1$ parameters $\alpha_i, \alpha_{Aig}, \alpha_{Bik}, \delta_{Aigh}, \delta_{Bikl}$ in model (1). Model (1) takes both additive and dominance effects into account [Fan et al., 2006]. If only the additive effect is modeled, model (1) can be modified to

$$y_i = w_i \gamma_i + \alpha_i + \sum_{g=1}^{m-1} x_A^{(g)} \alpha_{Aig} + \sum_{k=1}^{n-1} x_B^{(k)} \alpha_{Bik} + H_i + e_i. \quad (3)$$

In addition to the co-variate effects, there are $m+n-1$ parameters $\alpha_i, \alpha_{Aig}, g = 1, \dots, m-1, \alpha_{Bik}, k = 1, \dots, n-1$ in model (3). Model (3) only takes the additive effect into account; and we call it an "additive effect model" [Fan et al., 2006]. If both markers A and B are di-allelic such as SNPs, i.e., $m = n = 2$, models (1) and (3), respectively, simplify to

$$y_i = w_i \gamma_i + \alpha_i + x_A^{(1)} \alpha_{A1i} + x_B^{(1)} \alpha_{B1i} + z_A^{(12)} \delta_{A1i2} + z_B^{(12)} \delta_{B1i2} + H_i + e_i, \quad (4)$$

$$y_i = w_i \gamma_i + \alpha_i + x_A^{(1)} \alpha_{A1i} + x_B^{(1)} \alpha_{B1i} + H_i + e_i, \quad (5)$$

which extend the models of Fan and Xiong [2002]

Denote $X_A = (x_A^{(1)}, \dots, x_A^{(m-1)})^\tau, X_B = (x_B^{(1)}, \dots, x_B^{(n-1)})^\tau$, and $X_{AUB} = (X_A^\tau, X_B^\tau)^\tau$. Here the superscript τ denotes the transpose of a vector or a matrix. Let us denote the additive variance-covariance matrix of the indicator variables $x_A^{(g)}, x_B^{(k)}$ by $V_A = \text{Cov}(X_{AUB}, X_{AUB}) = E(X_{AUB} X_{AUB}^\tau) - E X_{AUB} (E X_{AUB}^\tau)$. Similarly, let $Z_A = (z_A^{(12)}, \dots, z_A^{(1m)}, z_A^{(23)}, \dots, z_A^{(2m)}, \dots, z_A^{(m-1)m})^\tau, Z_B = (z_B^{(12)}, \dots, z_B^{(1n)}, z_B^{(23)}, \dots, z_B^{(n-1)n})^\tau$, and $Z_{AUB} = (Z_A^\tau, Z_B^\tau)^\tau$. Let us denote the dominance variance-covariance matrix of the indicator variables $z_A^{(gh)}, z_B^{(kl)}$ by $V_D = \text{Cov}(Z_{AUB}, Z_{AUB})$. The elements of matrices V_A and V_D are provided in Appendix E, Fan et al. [2006]. For readers' convenience, we summarize V_A and V_D in Appendix A.

For $g = 1, 2, \dots, m$, let us denote $D_{A_g Q} = P(Q_1 A_g) - q_1 P_{A_g}$, which are measures of LD between

QTL Q and marker A . Here $P(Q_1A_g)$ is the frequency of haplotype Q_1A_g . For $k = 1, 2, \dots, n$, let us denote $D_{B_kQ} = P(Q_1B_k) - q_1P_{B_k}$, which are measures of LD between QTL Q and marker B . Here $P(Q_1B_k)$ is the frequency of haplotype Q_1B_k . For the i th trait, let $\mu_{ij}^{(i)}$ be the genotypic value of genotype $Q_jQ_l, i, j, l = 1, 2$ and so $\mu_{12} = \mu_{21}$. Denote $a_i = \mu_{11}^{(i)} - (\mu_{11}^{(i)} + \mu_{22}^{(i)})/2$ and $d_i = \mu_{12}^{(i)} - (\mu_{11}^{(i)} + \mu_{22}^{(i)})/2$. The average effect of gene substitution is $\alpha_{Qi} = a_i + (q_2 - q_1)d_i$, and dominance deviation is $\delta_{Qi} = 2d_i$ in view of traditional quantitative genetics [Falconer and Mackay, 1996]. Similar to Appendix E, Fan et al. [2006], we can show that the regression coefficients of models (1) and (3) are given by

$$\begin{pmatrix} \alpha_{Ai1} \\ \vdots \\ \alpha_{Ai(m-1)} \\ \alpha_{Bi1} \\ \vdots \\ \alpha_{Bi(n-1)} \end{pmatrix} = (V_A/2)^{-1} \begin{pmatrix} D_{A_1Q} \\ \vdots \\ D_{A_{m-1}Q} \\ D_{B_1Q} \\ \vdots \\ D_{B_{n-1}Q} \end{pmatrix} \alpha_{Qi},$$

$$\begin{pmatrix} \delta_{Ai12} \\ \vdots \\ \delta_{Ai(m-1)m} \\ \delta_{Bi12} \\ \vdots \\ \delta_{Bi(n-1)n} \end{pmatrix} = V_D^{-1} \begin{pmatrix} [P_{A_2}D_{A_1Q} - P_{A_1}D_{A_2Q}]^2 \\ \vdots \\ [P_{A_{m-1}}D_{A_mQ} - P_{A_m}D_{A_{m-1}Q}]^2 \\ [P_{B_2}D_{B_1Q} - P_{B_1}D_{B_2Q}]^2 \\ \vdots \\ [P_{B_{n-1}}D_{B_nQ} - P_{B_n}D_{B_{n-1}Q}]^2 \end{pmatrix} \delta_{Qi}. \quad (6)$$

If both markers A and B are di-allelic, equations (6) are given by

$$\begin{pmatrix} \alpha_{Ai1} \\ \alpha_{Bi1} \end{pmatrix} = \begin{pmatrix} P_{A_1}P_{A_2} & D_{A_1B_1} \\ D_{A_1B_1} & P_{B_1}P_{B_2} \end{pmatrix}^{-1} \begin{pmatrix} D_{A_1Q} \\ D_{B_1Q} \end{pmatrix} \alpha_{Qi},$$

$$\begin{pmatrix} \delta_{Ai12} \\ \delta_{Bi12} \end{pmatrix} = \begin{pmatrix} P_{A_1}^2P_{A_2}^2 & D_{A_1B_1}^2 \\ D_{A_1B_1}^2 & P_{B_1}^2P_{B_2}^2 \end{pmatrix}^{-1} \begin{pmatrix} D_{A_1Q}^2 \\ D_{B_1Q}^2 \end{pmatrix} \delta_{Qi}, \quad (7)$$

which are proved in Fan and Xiong [2002]. If multiple diallelic markers are used in the analysis, Jung et al. [2005] provide an extension of (7). Equations (6) and (7) show that the parameters of LD (i.e., D_{A_gQ} and D_{B_kQ}) and gene effect (i.e., α_{Qi} and δ_{Qi}) are contained in the regression coefficients. Models (1) and (3) simultaneously take care of the LD and the effects of the putative trait locus Q . The gene substitution effect α_{Qi} is contained only in α_{Aig} , α_{Bik} ; and the dominance effect δ_{Qi} is contained only in δ_{Aigh} , δ_{Bikl} . Therefore, V_A is called additive variance-covariance matrix; and V_D is called dominance variance-covariance matrix. Model (1) orthogonally decomposes genetic effect into summation of additive and dominance effects.

For the i th trait, it is well known that the additive variance $\sigma_{gai}^2 = 2q_1q_2\alpha_{Qi}^2$ and the dominance variance $\sigma_{gdi}^2 = (q_1q_2)^2\delta_{Qi}^2$. Let $\sigma_i^2 = \sigma_{gai}^2 + \sigma_{gdi}^2 + \sigma_{cai}^2 + \sigma_{ei}^2$ be the total variance of trait y_i . In addition, let σ_{ga12} and σ_{gd12} be additive and dominance cross covariances of y_1 and y_2 , respectively; and let σ_{ca12} and σ_{e12} be polygenic covariance and random error covariance of y_1 and y_2 , respectively. The covariance of y_1 and y_2 is $\sigma_{12} = \text{Cov}(y_1, y_2) = \sigma_{ga12} +$

$\sigma_{gd12} + \sigma_{ca12} + \sigma_{e12} = \sigma_{21}$, where $\sigma_{ga12} = 2q_1q_2\alpha_{Q1}\alpha_{Q2}$ and $\sigma_{gd12} = (q_1q_2)^2\delta_{Q1}\delta_{Q2}$.

FAMILIAL VARIANCE COMPONENT MODELS

Let $Y_1 = (y_{11}, \dots, y_{1t})^T$ and $Y_2 = (y_{21}, \dots, y_{2t})^T$ be their measured values on the t members of a non-inbred pedigree [Lange and Boehnke, 1983]. Here Y_1 is the column vector of trait values of the first trait of the t family members, and Y_2 is the column vector of trait values of the second trait. Under normality, regressions (1) and (3) can still be used to model the trait values. The notations need to be modified, accordingly, as follows. For instance, model (1) can be modified as:

$$y_{ij} = w_{ij}\gamma_i + \alpha_i + \sum_{g=1}^{m-1} x_{Aj}^{(g)}\alpha_{Aig} + \sum_{k=1}^{n-1} x_{Bj}^{(k)}\alpha_{Bik} + \sum_{1 \leq g < h \leq m} z_{Aj}^{(gh)}\delta_{Aigh} + \sum_{1 \leq k < l \leq n} z_{Bj}^{(kl)}\delta_{Bikl} + H_{ij} + e_{ij}, \quad (8)$$

where w_{ij} is a row vector of co-variables such as sex and age for the i -trait of the j th pedigree member, H_{ij} is polygenic effect, and e_{ij} is error term. Assume that H_{ij} is normal $N(0, \sigma_{Gi}^2), i = 1, 2$, e_{ij} is normal $N(0, \sigma_{ei}^2), i = 1, 2$, and H_{ij} is independent of e_{ij} . The dummy variables $x_{Aj}^{(g)}, z_{Aj}^{(gh)}, x_{Bj}^{(k)}$ and $z_{Bj}^{(kl)}$ are defined such as (2) according to the marker genotypes G_{Aj} at marker A and G_{Bj} at marker B of the j th pedigree member, e.g.,

$$x_{Aj}^{(g)} = \begin{cases} 2 & \text{if } G_{Aj} = A_gA_g \\ 1 & \text{if } G_{Aj} = A_gA_h, h \neq g. \\ 0 & \text{else} \end{cases}$$

If only the additive effect is modeled, model (8) can be modified to

$$y_{ij} = w_{ij}\gamma_i + \alpha_i + \sum_{g=1}^{m-1} x_{Aj}^{(g)}\alpha_{Aig} + \sum_{k=1}^{n-1} x_{Bj}^{(k)}\alpha_{Bik} + H_{ij} + e_{ij}. \quad (9)$$

The regression coefficients of models (8) and (9) are given by (6), which can be proved as Appendix E, Fan et al. [2006].

Let π_{Qjk} be the proportion of alleles shared identical by descent (IBD) at QTL Q by the j th and the k th individuals, and let Δ_{Qjk} be the probability that both alleles at QTL Q shared by the j th and the k th individuals are IBD. The trait variances and co-variances are given by

$$\begin{aligned} \text{Cov}(y_{1j}, y_{1k}) &= \pi_{Qjk}\sigma_{ga1}^2 + \Delta_{Qjk}\sigma_{gd1}^2 \\ &\quad + 2\Phi_{jk}\sigma_{ca1}^2 + 1_{(j=k)}\sigma_{e1}^2, \\ \text{Cov}(y_{2j}, y_{2k}) &= \pi_{Qjk}\sigma_{ga2}^2 + \Delta_{Qjk}\sigma_{gd2}^2 \\ &\quad + 2\Phi_{jk}\sigma_{ca2}^2 + 1_{(j=k)}\sigma_{e2}^2, \\ \text{Cov}(y_{1j}, y_{2k}) &= \pi_{Qjk}\sigma_{ga12} + \Delta_{Qjk}\sigma_{gd12} \\ &\quad + 2\Phi_{jk}\sigma_{ca12} + 1_{(j=k)}\sigma_{e12} \\ &= \text{Cov}(y_{2j}, y_{1k}), \end{aligned} \quad (10)$$

where Φ_{jk} is the kinship coefficient of individuals j and k [Lange, 2002]. Let $\Pi_Q = (\pi_{Qjk})_{t \times t}$ be a symmetric $t \times t$ matrix with π_{Qjk} as the entry in row j and column k ; let $\Delta_Q = (\Delta_{Qjk})_{t \times t}$ be a symmetric $t \times t$ matrix with Δ_{Qjk} as the entry in row j and column k ; let $\Phi = (\Phi_{jk})_{t \times t}$ be a symmetric $t \times t$ matrix with Φ_{jk} as the entry in row j and column k ; and let I_t be $t \times t$ identity matrix. In partitioned matrix and Kronecker product notation, these covariances can be collectively expressed as [Harville, 1997]

$$\begin{aligned} \text{Var} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} &= \sigma_{ga1}^2 \begin{pmatrix} \Pi_Q & 0 \\ 0 & 0 \end{pmatrix} + \sigma_{ga12} \begin{pmatrix} 0 & \Pi_Q \\ \Pi_Q & 0 \end{pmatrix} \\ &+ \sigma_{ga2}^2 \begin{pmatrix} 0 & 0 \\ 0 & \Pi_Q \end{pmatrix} + \sigma_{Ga1}^2 \begin{pmatrix} 2\Phi & 0 \\ 0 & 0 \end{pmatrix} \\ &+ \sigma_{Ga12} \begin{pmatrix} 0 & 2\Phi \\ 2\Phi & 0 \end{pmatrix} + \sigma_{Ga2}^2 \begin{pmatrix} 0 & 0 \\ 0 & 2\Phi \end{pmatrix} \\ &+ \sigma_{gd1}^2 \begin{pmatrix} \Delta_Q & 0 \\ 0 & 0 \end{pmatrix} + \sigma_{gd12} \begin{pmatrix} 0 & \Delta_Q \\ \Delta_Q & 0 \end{pmatrix} \\ &+ \sigma_{gd2}^2 \begin{pmatrix} 0 & 0 \\ 0 & \Delta_Q \end{pmatrix} + \sigma_{e1}^2 \begin{pmatrix} I_t & 0 \\ 0 & 0 \end{pmatrix} \\ &+ \sigma_{e12} \begin{pmatrix} 0 & I_t \\ I_t & 0 \end{pmatrix} + \sigma_{e2}^2 \begin{pmatrix} 0 & 0 \\ 0 & I_t \end{pmatrix} \\ &= \begin{pmatrix} \sigma_{ga1}^2 & \sigma_{ga12} \\ \sigma_{ga12} & \sigma_{ga2}^2 \end{pmatrix} \otimes \Pi_Q + \begin{pmatrix} \sigma_{gd1}^2 & \sigma_{gd12} \\ \sigma_{gd12} & \sigma_{gd2}^2 \end{pmatrix} \\ &\otimes \Delta_Q + \begin{pmatrix} \sigma_{Ga1}^2 & \sigma_{Ga12} \\ \sigma_{Ga12} & \sigma_{Ga2}^2 \end{pmatrix} \otimes \Phi \\ &+ \begin{pmatrix} \sigma_{e1}^2 & \sigma_{e12} \\ \sigma_{e12} & \sigma_{e2}^2 \end{pmatrix} \otimes I_t. \end{aligned} \tag{11}$$

The log-likelihood of Y_1 and Y_2 can be written as

$$\begin{aligned} L &= -t \log(2\pi) - \frac{1}{2} \log \det \left[\text{Var} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \right] \\ &- \frac{1}{2} \begin{pmatrix} Y_1 - EY_1 \\ Y_2 - EY_2 \end{pmatrix}^\tau \left[\text{Var} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \right]^{-1} \begin{pmatrix} Y_1 - EY_1 \\ Y_2 - EY_2 \end{pmatrix}, \end{aligned} \tag{12}$$

where $EY_1 = (Ey_{11}, \dots, Ey_{1t})^\tau$ and $EY_2 = (Ey_{21}, \dots, Ey_{2t})^\tau$ are mean vectors of trait values. For instance, $Ey_{ij} = w_{ij}\gamma_i + \alpha_i + \sum_{g=1}^{m-1} x_{Aj}^{(g)} \alpha_{Aig} + \sum_{k=1}^{n-1} x_{Bj}^{(k)} \alpha_{Bik} + \sum_{1 \leq g < h \leq m} z_{Aj}^{(gh)} \delta_{Aigh} + \sum_{1 \leq k < l \leq n} z_{Bj}^{(kl)} \delta_{Bikl}$ if model (8) is used in analysis.

PARAMETER ESTIMATION

Assume that the data are a combination of N unrelated individuals and I independent pedigrees. The I pedigrees can be multi-generation pedigrees of any sizes and any types of relatives, nuclear families, sibships or their combinations. For each pedigree, the log-likelihood can be written as (12). Assume that the pedigrees are unrelated, then the total log-likelihood is the summation of the individual log-likelihoods. Let $\Omega = (\sigma_{ga1}^2, \sigma_{ga12}, \sigma_{ga2}^2, \sigma_{gd1}^2, \sigma_{gd12}, \sigma_{gd2}^2, \sigma_{Ga1}^2, \sigma_{Ga12}, \sigma_{Ga2}^2, \sigma_{e1}^2, \sigma_{e12}, \sigma_{e2}^2)^\tau$ be the column vector of parameters of the variance-covariance matrix. Similarly, denote the regression coefficients as $\alpha_{i,AUB} = (\alpha_{Ai1}, \dots, \alpha_{Ai(m-1)}, \alpha_{Bi1}, \dots, \alpha_{Bi(n-1)})^\tau$ and

$\delta_{i,AUB} = (\delta_{Ai12}, \dots, \delta_{Ai(m-1)m}, \delta_{Bi12}, \dots, \delta_{Bi(n-1)n})^\tau$. Denote

$$Y_i = \begin{pmatrix} \gamma_i \\ \alpha_i \\ \alpha_{i,AUB} \\ \delta_{i,AUB} \end{pmatrix} \text{ and } Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}.$$

Standard methods, such as Newton-Raphson or Fisher scoring methods, can be used to estimate the parameters.

For the s th pedigree or individual, let Y_{1s} and Y_{2s} be the column vectors of the trait values of the two traits. In addition, let U_s be the model matrix of regression (1) or (8) for the pedigree or individual. Then U_s can be decomposed into $U_s = (W_s, X_{s,AUB}^\tau, Z_{s,AUB}^\tau)$, where W_s is the sub-matrix corresponding to regression coefficients γ_i and α_i , $X_{s,AUB}^\tau$ is the sub-matrix corresponding to regression coefficients $\alpha_{i,AUB}$, and $Z_{s,AUB}^\tau$ is the sub-matrix corresponding to regression coefficients $\delta_{i,AUB}$. Then regression (1) or (8) can be expressed by $\begin{pmatrix} Y_{1s} \\ Y_{2s} \end{pmatrix} = \begin{pmatrix} U_s & 0 \\ 0 & U_s \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$, where O is zero matrix, and ε_1 and ε_2 are column vectors of error terms.

Consider the overall log-likelihood $L = \sum_{s=1}^{N+I} L_s$, where $L_s, s = 1, \dots, N$ are the log-likelihood functions of N individuals, and $L_s, s = N + 1, \dots, N + I$ are log-likelihood functions of pedigrees as (12). Let Σ_s be the variance-covariance of Y_{1s} and Y_{2s} in the form (11) of s th pedigree or individual, and let $\hat{\Sigma}_s$ be the estimate of Σ_s . The regression coefficients Y can be estimated by

$$\begin{aligned} \hat{Y} &= \begin{pmatrix} \hat{Y}_1 \\ \hat{Y}_2 \end{pmatrix} = \left[\sum_{s=1}^{N+I} \begin{pmatrix} U_s^\tau & O \\ O & U_s^\tau \end{pmatrix} \hat{\Sigma}_s^{-1} \begin{pmatrix} U_s & O \\ O & U_s \end{pmatrix} \right]^{-1} \\ &\times \sum_{s=1}^{N+I} \begin{pmatrix} U_s^\tau & O \\ O & U_s^\tau \end{pmatrix} \hat{\Sigma}_s^{-1} \begin{pmatrix} Y_{1s} \\ Y_{2s} \end{pmatrix}. \end{aligned}$$

In practice, some of the parameters (e.g., variance parameters σ_{gd1}^2 and σ_{gd2}^2 , and covariance parameter σ_{gd12}) may not be estimable and identifiable due to the redundancy. One may need to specify the model carefully for specific types of data. Let us denote

$$\Psi_i = \begin{pmatrix} \gamma_i \\ \alpha_i \\ \alpha_{i,AUB} \end{pmatrix} \text{ and } \Psi = \begin{pmatrix} \Psi_1 \\ \Psi_2 \end{pmatrix}.$$

Let $V_s = (W_s, X_{s,AUB}^\tau)$ be the model matrix of regression (3) or (9). Then regression (3) or (9) can be expressed by $\begin{pmatrix} Y_{1s} \\ Y_{2s} \end{pmatrix} = \begin{pmatrix} V_s & 0 \\ 0 & V_s \end{pmatrix} \begin{pmatrix} \Psi_1 \\ \Psi_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$. The regression coefficients Ψ can be estimated by

$$\begin{aligned} \hat{\Psi} &= \begin{pmatrix} \hat{\Psi}_1 \\ \hat{\Psi}_2 \end{pmatrix} = \left[\sum_{s=1}^{N+I} \begin{pmatrix} V_s^\tau & O \\ O & V_s^\tau \end{pmatrix} \hat{\Sigma}_s^{-1} \begin{pmatrix} V_s & O \\ O & V_s \end{pmatrix} \right]^{-1} \\ &\times \sum_{s=1}^{N+I} \begin{pmatrix} V_s^\tau & O \\ O & V_s^\tau \end{pmatrix} \hat{\Sigma}_s^{-1} \begin{pmatrix} Y_{1s} \\ Y_{2s} \end{pmatrix}. \end{aligned}$$

ASSOCIATION STUDY AND NON-CENTRALITY PARAMETER APPROXIMATIONS OF F-TESTS

Evidence of association can be tested by the likelihood ratio test (LRT) procedure or F -test using the estimates $\hat{\Sigma}_s$ and \hat{Y} . For LRT procedure, let L_{ad} be the log-likelihood

under the alternative hypothesis of $H_{ABad,1}$ and L_0 be log-likelihood under the null hypothesis $H_{ABad,0}$. The LRT statistic $2[L_{ad} - L_0]$ is asymptotically distributed as χ^2 with the degrees of freedom $2(J_A + J_B - 2)$. As there are $m + n - 2$ measures of LD, $D_{A_1Q}, \dots, D_{A_{m-1}Q}, D_{B_1Q}, \dots, D_{B_{n-1}Q}$ due to $\sum_{g=1}^m D_{A_gQ} = 0$ and $\sum_{k=1}^n D_{B_kQ} = 0$, the number of coefficients $\alpha_{Aigr}, \alpha_{Bikr}, \delta_{Aighr}, \delta_{Biklr}, i = 1, 2$ having significant results should be less than $m + n - 2$.

For F -test procedure, the linear regression model theory is used to test genetic effect and LD coefficients [Graybill, 1976] and to evaluate the power of the F -test statistics. The approximations of the non-centrality parameters can be derived for power analysis. Let the total trait values as

$$Y = \begin{pmatrix} Y_{11} \\ Y_{21} \\ \vdots \\ Y_{1,N+I} \\ Y_{2,N+I} \end{pmatrix},$$

and let us denote the total model matrix

$$U = \begin{pmatrix} U_1 & O \\ O & U_1 \\ \vdots & \vdots \\ U_{N+I} & O \\ O & U_{N+I} \end{pmatrix}.$$

Then we can write the model as $Y = UY + \varepsilon$. Let total variance-covariance matrix be $\Sigma = \text{diag}(\Sigma_1, \dots, \Sigma_{N+I})$. In the s th pedigree, assume that there are n_s pedigree members. The total sample size is $2(N + \sum_{s=N+1}^{N+I} n_s)$.

Define a test matrix by $H = \begin{pmatrix} O & I_J & O & O \\ O & O & O & I_J \end{pmatrix}$, $J = J_A + J_B - 2$, where O are a zero matrices/vectors and I_J is $J \times J$ identity matrix. To test the null hypothesis $H_{ABad,0}$, the F -test statistic based on regression (1) or (8) is

$$F_{AB,ad} = \frac{(H\hat{Y})^\tau [H(U^\tau \hat{\Sigma}^{-1} U)^{-1} H^\tau]^{-1} (H\hat{Y})}{Y^\tau [\hat{\Sigma}^{-1} - \hat{\Sigma}^{-1} U(U^\tau \hat{\Sigma}^{-1} U)^{-1} U^\tau \hat{\Sigma}^{-1}] Y} \times \frac{2(N + \sum_{s=N+1}^{N+I} n_s) - 2(J_A + J_B - 2)}{2(J_A + J_B - 2)}, \quad (13)$$

with a non-central $F(2(J_A + J_B - 2), 2(N + \sum_{s=N+1}^{N+I} n_s) - 2(J_A + J_B - 2), \lambda_{ad})$ distribution under the alternative hypothesis. Here λ_{ad} is the non-centrality parameter given by

$$\lambda_{ad} = (HY)^\tau \left[H \left[\sum_{s=1}^{N+I} \begin{pmatrix} U_s^\tau & O \\ O & U_s^\tau \end{pmatrix} \Sigma_s^{-1} \begin{pmatrix} U_s & O \\ O & U_s \end{pmatrix} \right]^{-1} H^\tau \right]^{-1} \times (HY).$$

To test the additive genetic effect, the null hypothesis is $H_{ABa,0} : \alpha_{Ai1} = \dots = \alpha_{Ai(m-1)} = \alpha_{Bi1} = \dots = \alpha_{Bi(n-1)} = 0, i = 1, 2$. The test matrix is $H_1 = \begin{pmatrix} O & I_k & O & O \\ O & O & O & I_k \end{pmatrix}$, $k = m + n - 2$, where O are zero matrices/vectors and I_k is $k \times k$ identity matrix. In the same manner as $F_{AB,ad}$, one may construct an F -statistic $F_{AB,a}$ to test $H_{ABa,0}$ based on model (3) or (9), with degrees of freedom of $(2(m + n - 2), 2(N + \sum_{s=N+1}^{N+I} n_s) - 2(m + n - 1))$. The corresponding

non-centrality parameter of $F_{AB,a}$ is

$$\lambda_a = (H_1 \Psi)^\tau \left[H_1 \left[\sum_{s=1}^{N+I} \begin{pmatrix} V_s^\tau & O \\ O & V_s^\tau \end{pmatrix} \Sigma_s^{-1} \begin{pmatrix} V_s & O \\ O & V_s \end{pmatrix} \right]^{-1} H_1^\tau \right]^{-1} \times (H_1 \Psi).$$

Assume that there are no co-variates. In Appendix B, approximations of non-centrality parameters are provided for population data. For nuclear family data, Appendices C and D provide approximations of non-centrality parameters. In Appendix E, we consider a combination of N unrelated individuals, I_1 trio families (each has a single child and both parents), and I_2 nuclear families of size 4 (each has two siblings and both parents). Assume that N, I_1 , and I_2 are large enough that large sample theory applies. Appendix E provides approximations of non-centrality parameters of F -test statistics for the combination.

MAPPING STRATEGY

It is well known that association studies are prone to false positive due to population admixture or stratification, although their resolution can be high. Linkage analysis is robust to population structure, although its resolution can be low. The proposed models based on combinations of population and pedigree data can take advantage of robustness of linkage analysis and high resolution of an association study, and overcome the limits of each. In practice, linkage test based on pedigree data can be performed first to detect a broad region of a trait locus. An LRT can be constructed to test the null hypothesis of no linkage by considering a reduced variance component model with the i th trait of the j th pedigree member based on $y_{ij} = w_{ij}\gamma_i + \alpha_i + g_{ij} + H_{ij} + e_{ij}$, which does not model the association information [Amos, 1994]. Here g_{ij} is unobserved major gene effect, and the other notations are similar to above description. This prior linkage analysis can usually localize the trait locus in a chromosome region that ranges from a few cM to 20 cM.

With the prior linkage information, one may fit the models proposed in this article by using both pedigree and population data based on a dense genetic map for high resolution joint linkage study and LD mapping; i.e., fine disease gene mapping. Assume that significant evidence of linkage is found to be $\sigma_{ga1}^2 > 0, \sigma_{ga2}^2 > 0$, which implies $\alpha_{Q1} \neq 0, \alpha_{Q2} \neq 0$. Then the association test between two markers and the putative QTL can be carried out as following: $H_{ABa,0} : \alpha_{Ai1} = \dots = \alpha_{Ai(m-1)} = \alpha_{Bi1} = \dots = \alpha_{Bi(n-1)} = 0, i = 1, 2$ versus $H_{ABa,1}$: at least one of $\alpha_{Aigr}, \alpha_{Bikr}$'s is not equal to 0. This is equivalent to $H_{AB,a0} : D_{A_1Q} = \dots = D_{A_{m-1}Q} = D_{B_1Q} = \dots = D_{B_{n-1}Q} = 0$. Assume that the significant evidence of linkage is found to be $\sigma_{ga1}^2 > 0, \sigma_{ga2}^2 > 0, \sigma_{gd1}^2 > 0, \sigma_{gd2}^2 > 0$. The association test is $H_{ABad,0} : \alpha_{Ai1} = \dots = \alpha_{Ai(m-1)} = \alpha_{Bi1} = \dots = \alpha_{Bi(n-1)} = \delta_{Ai12} = \dots = \delta_{Ai1m} = \dots = \delta_{Ai(m-1)m} = \delta_{Bi12} = \dots = \delta_{Bi1n} = \dots = \delta_{Bi(n-1)n} = 0$ versus $H_{ABad,1}$: at least one of $\alpha_{Aigr}, \alpha_{Bikr}, \delta_{Aighr}, \delta_{Bijk}$ is not equal to 0. The significant variables, such as x_{Aij} and x_{Bij} , can be identified. Keeping the significant variables in the final model, one may calculate the LRT of the final model against a model which assumes neither linkage nor association between the trait values and markers, to test linkage and association simultaneously; or one may calculate the F -test or LRT of the final model

against a model which assumes linkage but no association, to test the association in the presence of prior linkage.

The proposed model explains linkage information in variance-covariance matrix and LD information in the mean coefficients. In the Appendices, we show that the non-centrality parameters of *F*-test statistics are functions of additive and dominance variances, and we may get high

TYPE I ERROR RATES

To investigate the robustness of the proposed models, we calculate type I error rates for four cases named *Null*, *Familiarity*, *Linkage* and *Composite*. The related parameter values are presented as follows ($\sigma_i^2 = \sigma_{gai}^2 + \sigma_{Gai}^2 + \sigma_{ei}^2, i = 1, 2$):

Location of genetic variant			Lod of univariate of anti-CCP	Lod of univariate of RF IgM	Lod of bivariate of anti-CCP and RF IgM
Chromosome	SNP	Position			
2	rs2685263	539873	3.73	3.65	7.48
4	rs1024461	106121591	3.27	4.50	7.31

power if linkage signal of QTL is large. Simultaneously using both linkage and LD information in our models has two advantages: (1) it is more likely to avoid spurious association of a separate LD mapping; (2) it increases the resolution of a separate linkage mapping. Using this strategy in study design, researchers may increase the power to localize the genetic location and to detect important genetic determinants of the traits of complex diseases.

RESULTS

AN EXAMPLE

The proposed method is applied to analyze two quantitative phenotypes, i.e., rheumatoid factor (RF) IgM and anti-cyclic citrullinated peptide (anti-CCP) for the data of The North American Rheumatoid Arthritis Consortium, which is provided by Genetic Analysis Workshop 15 (GAW 15), Problem 2. The data contain 642 Caucasian families [Amos et al., 2006]. As the two traits are unlikely to be normal, we transform the traits with an inverse Gaussian transformation of the ranks, which are approximately normal. By fitting additive effect models of univariate and bivariate, we obtain the following results:

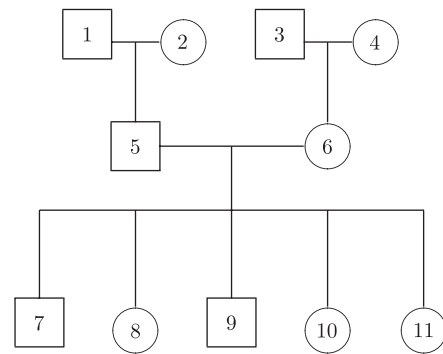


Fig. 1. Multi-generation pedigree used in type I error calculation, which are taken from Figure 1 of Abecasis et al. [2000] or Fan et al. [2005]. The number in the box or circle is individual ID.

For the case of *Null*, no genetic effect is assumed to affect the traits. For the case of *Familiarity*, polygenic effects are assumed, but no major locus effects are assumed. For

Test cases	$h_1^2 = \sigma_{ga1}^2 / \sigma_1^2$	$h_2^2 = \sigma_{ga2}^2 / \sigma_2^2$	$\sigma_{Gai1}^2 / \sigma_1^2$	$\sigma_{Gai2}^2 / \sigma_2^2$	σ_1^2	σ_2^2	θ_{AQ}	q_1	D_{A_gQ}
Null	0	0	0	0	15	10	Not applied	Not applied	Not applied
Familiarity	0	0	0.5	0.3	15	10	Not applied	Not applied	Not applied
Linkage	0.15	0.1	0	0	15	10	0.01	0.5	0
Composite	0.15	0.1	0.15	0.1	15	10	0.01	0.5	0

The above significant results for chromosomes 2 and 4 are found at the same marker for both RF and anti-CCP in univariate analysis. Hence, the two traits can be influenced by the genetic variants at the same locus. In addition, the two traits are strongly correlated to each other (*P* value <0.0001 of *t*-test 22.0 of testing slope = 0 of simple linear regression of the ranks of the two traits from running Statistical Analysis Software (SAS); the Pearson correlation coefficient = 0.501). This provides a motivation to pursue a bivariate analysis. It can be seen that the Lod scores of bivariate analysis are higher than those of separate univariate analysis. Hence, it is advantageous in performing bivariate analysis.

the case of *Linkage*, major locus effects are assumed, but no polygenic effects are assumed. For the case of *Composite*, both major locus effects and polygenic effects are assumed. For each test case, five types of pedigrees are simulated to calculate the type I error rates: the multi-generation pedigree in Figure 1, nuclear families with two to five kids. For each combination of test case and pedigree type, 10,000 datasets are simulated; and for each data set, 60 or 120 pedigrees are simulated. We evaluate a marker *A* which is di-allelic and tri-allelic, i.e., $m = 2, 3$. For di-allelic marker, equal allele frequencies are assumed, i.e., $P_{A_1} = P_{A_2} = 0.5$; for tri-allelic marker, the allele frequencies are given by $P_{A_1} = P_{A_2} = 0.3$ and $P_{A_3} = 0.4$. For each

of the 10,000 data sets, we fit the following model

$$y_{ij} = \alpha_i + \sum_{g=1}^{m-1} x_{Aj}^{(g)} \alpha_{Aig} + e_{ij},$$

under an assumption of $\sigma_{e12} = 0$. The vector of parameters of the variance-covariance matrix is $\Omega = (\sigma_{ga1}^2, \sigma_{ga12}, \sigma_{ga2}^2, \sigma_{e1}^2, \sigma_{e2}^2)^T$, and so there are five variance-covariance parameters and test the null hypothesis $H_{a0} : \alpha_{A11} = \dots = \alpha_{A(m-1)} = 0, i = 1, 2$. As the QTL Q is in linkage equilibrium with marker A or simply no major gene locus Q , an empirical test statistic which is larger than the

cutting point at a 0.05 or 0.01 significance level is treated as a false positive. Based on the LRT, type I error rates are calculated as the proportions of the 10,000 simulation data sets which give significant result at the 0.05 or 0.01 significant level. The results of type I error rates are presented in Table I. When each data set contains 60 pedigrees, the results of Table I show that the type I error rates are around the nominal level 0.05 or 0.01 for all cases, except for the case of two-kid nuclear family for a tri-allelic marker, $m = 3$. Hence, the model is reasonably robust. For the case of two-kid nuclear family for a tri-allelic marker, the type I error rates tend to be higher than the nominal

TABLE I. Type I error rates (%) at 0.01 and 0.05 significance levels based on likelihood ratio tests

No. of pedigrees	No. of alleles	Pedigree type	Test case	Error rates	
				$\alpha = 0.01$	$\alpha = 0.05$
60 pedigrees of each data set	Di-allele $m = 2$	Multi-generation pedigree Figure 1	Null	0.74	4.81
			Familiality	0.81	4.08
			Linkage	1.02	4.91
			Composite	1.01	4.93
		Five-kid nuclear family	Null	0.94	5.13
			Familiality	0.94	4.61
			Linkage	1.07	5.17
			Composite	0.93	4.71
		Four-kid nuclear family	Null	1.03	4.51
			Familiality	0.81	4.20
			Linkage	1.10	5.56
			Composite	1.08	5.37
		Three-kid nuclear family	Null	0.72	4.62
			Familiality	0.91	4.91
			Linkage	0.86	4.92
			Composite	0.97	5.41
		Two-kid nuclear family	Null	0.92	4.83
			Familiality	1.01	5.36
	Linkage		0.99	5.45	
	Composite		1.23	5.80	
	Tri-allele $m = 3$	Multi-generation pedigree in Figure 1	Null	0.86	4.69
			Familiality	0.81	4.61
			Linkage	0.95	5.22
			Composite	0.98	4.75
		Five-kid nuclear family	Null	0.77	4.62
			Familiality	0.85	4.40
			Linkage	0.89	5.46
			Composite	0.81	4.75
		Four-kid nuclear family	Null	0.80	4.80
			Familiality	1.02	4.52
			Linkage	1.07	5.47
			Composite	1.09	5.52
		Three-kid nuclear family	Null	0.84	5.05
			Familiality	0.93	4.94
			Linkage	0.87	5.20
			Composite	0.91	5.04
Two-kid nuclear family		Null	1.01	4.81	
		Familiality	1.18	5.42	
	Linkage	1.33	5.86		
	Composite	1.16	6.17		
120 pedigrees of each dataset	Tri-allele $m = 3$	Two-kid nuclear family	Null	0.89	4.75
			Familiality	0.97	5.00
			Linkage	1.08	5.25
			Composite	1.06	5.13

levels. This is most likely due to moderate sample size of 60 nuclear families of four people each, and relatively large number of parameters of $2 \times 3 + 5 = 11$, where there are 2×3 regression coefficients and five variance-covariance parameters. To further explore the issue, we increase the sample size by doubling the number of pedigrees to 120 two-kid nuclear families in the simulated data set. From the results on the bottom of Table I, we can see that the type I error rates are around the nominal level 0.05 or 0.01, when each data set contains 120 two-kid nuclear families.

POWER COMPARISON

To make power comparison, we consider the situation in which nuclear family and population data are combined.

The sizes of unrelated individuals, trio families and nuclear family of size 4 are 40, 30, and 20, respectively. In our previous work, the power comparison of one marker versus two markers, as well as di-allelic marker versus multi-allelic marker, was extensively performed [Fan et al., 2005, 2006; Jung et al., 2005]. In this paper, we focus on power comparison of two quantitative trait values versus one trait value. In addition, we consider two di-allelic markers *A* and *B* with equal allele frequencies for simplification. Two modes of inheritance are considered; a dominant mode of inheritance $a_1 = d_1 = 1.0$, $a_2 = d_2 = 1.0$ and a recessive mode of inheritance $a_1 = 1.0$, $d_1 = -0.5$, $a_2 = 1.0$, $d_2 = -0.5$. Moreover, let $h_1^2 = \sigma_{ga1}^2 / \sigma_1^2$, $h_2^2 = \sigma_{ga2}^2 / \sigma_2^2$ be the heritability of Y_1 and Y_2 , respectively.

Figure 2 shows power curves of test statistics $F_{2var,a}$ and $F_{2var,ad}$ using both traits Y_1 and Y_2 , and test statistics $F_{1var,a}$

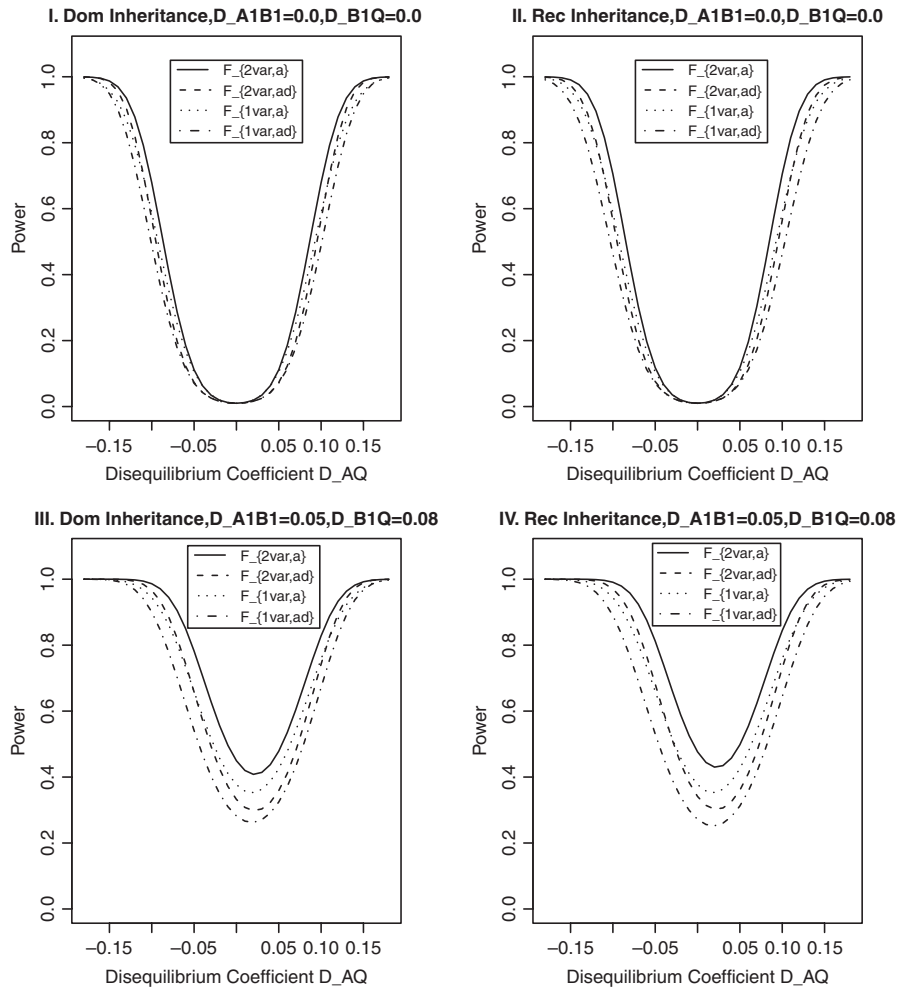


Fig. 2. Power curves against measure D_{A_1Q} of linkage disequilibrium at 0.01 level, where $q_1 = 0.5$, $\sigma_{Ga1}^2 = \sigma_{Ga2}^2 = \sigma_{G12} = \sigma_{e12} = 0$. For $F_{2var,a}$ and $F_{2var,ad}$, both traits Y_1 and Y_2 are used for a combined analysis. For $F_{1var,a}$ and $F_{1var,ad}$, only one trait Y_1 is used in analysis. $F_{2var,a}$ and $F_{1var,a}$ are based on the additive effect model, and $F_{1var,ad}$ and $F_{2var,ad}$ are based on the genotype effect model. For all graphs, a combination of nuclear families and population data are used (the number of population individuals = 60, the number of trio families = 30, and the number of nuclear families each with two kids = 20). Two di-allelic markers *A* and *B* with equal frequencies are used in the analysis. In addition, $h_1^2 = h_2^2 = 0.2$. In dominant (Dom) mode of inheritance of graphs I and III, $a_1 = 1, d_1 = 1, a_2 = 1, d_2 = 1$ for the two traits; in recessive (Rec) mode of inheritance of graphs II and IV, $a_1 = 1, d_1 = -0.5, a_2 = 1, d_2 = -0.5$ for the two traits. Here, we assume $\mu_{11}^{(i)} = a_i = -\mu_{22}^{(i)}$ and $\mu_{21}^{(i)} = \mu_{12}^{(i)} = d_i$.

and $F_{1var,ad}$ using the first trait Y_1 only. Here $F_{2var,a}$ and $F_{1var,a}$ are based on additive effect model, and $F_{1var,ad}$ and $F_{2var,ad}$ are based on genotype effect model. The powers are calculated across the measure $D_{AQ} = D_{A_1Q}$ of LD at 0.01 significance level. It can be seen that the power of test statistics using both traits is higher than that of test statistics using a single trait. Hence, it is advantageous to perform a unified multivariate analysis of multiple traits, instead of separate univariate analysis. In addition, additive effect models provides higher power than the genotype effect model, and this is consistent with the results of our previous univariate analysis (it is most likely due to that the dominance effect cannot compensate for the increasing degrees of freedom in the test statistics of genotype effect model). In Graphs I and II of Figure 2, we assume that there is no LD between markers A and B, and no LD between marker A and trait locus Q, i.e., $D_{A_1B_1} = D_{B_1Q} = 0$. Hence, the power curves reach minimum at $D_{A_1Q} = 0$, and the curves are symmetric in Graphs I and II. In Graphs III and IV of Figure 2, we assume a more

realistic situation of $D_{A_1B_1} = 0.05, D_{B_1Q} = 0.08$. Then, the curves are not symmetric any more, and the minimum powers are higher than 0.

Figure 3 shows power curves against heritability h_1^2 of the first trait: in Graphs I and II, the heritability of the second trait is fixed at $h_2^2 = 0.1$; in Graphs III and IV, $h_2^2 = 0.2$. Similar to Figure 2, we can easily see from the four graphs of Figure 3 that the power of bivariate test statistic $F_{2var,a}$ (or $F_{2var,ad}$) is higher than that of univariate test statistic $F_{1var,a}$ (or $F_{1var,ad}$). When heritability $h_2^2 = 0.1$, the test statistic $F_{2var,a}$ (or $F_{2var,ad}$) may have some power even when h_1^2 is small, while test statistic $F_{1var,a}$ (or $F_{1var,ad}$) may be powerful only when h_1^2 is large. As expected, the power increases as the heritability h_1^2 does.

Figure 4 shows power curves against trait allele frequency q_1 . Figure 5 shows power curves against marker allele frequency P_{A_1} . The powers across frequency of allele Q_1 in Figure 4 have a similar pattern as those across allele frequency P_{A_1} in Figure 5: the powers increase as q_1 increases in Figure 4, and the powers increase as P_{A_1}

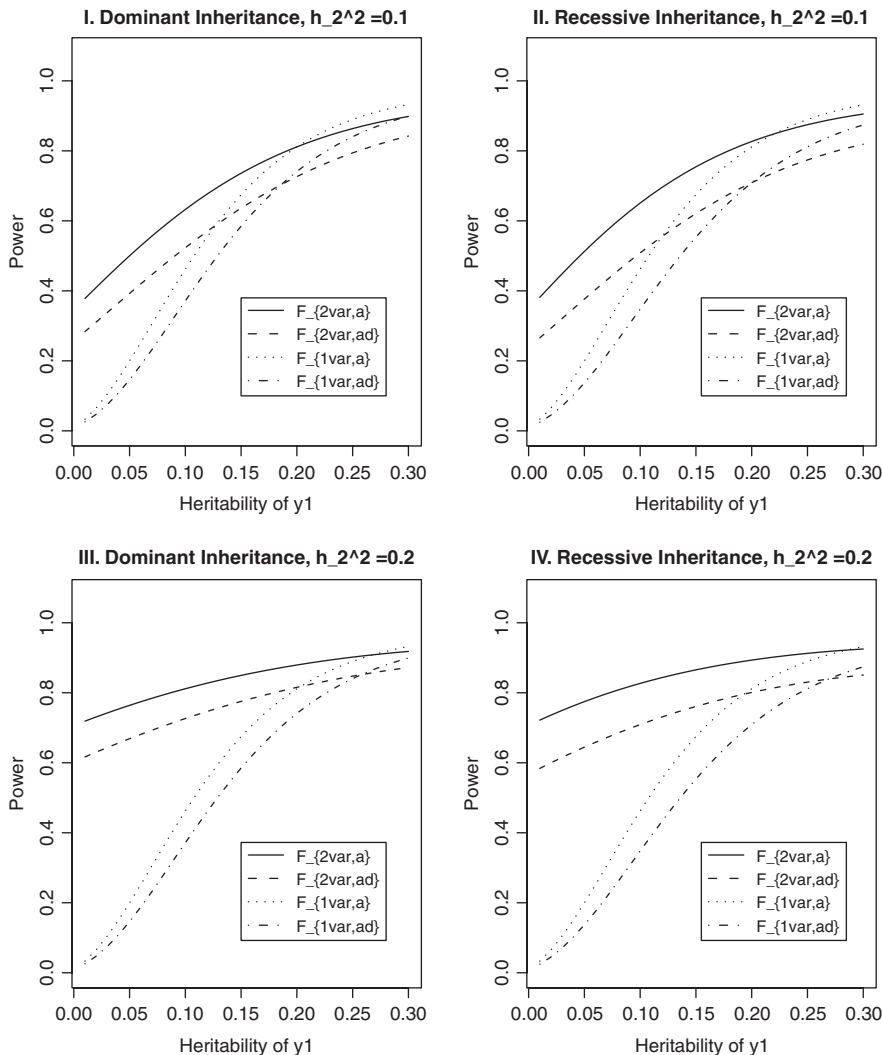


Fig. 3. Power curves against heritability h_1^2 of the first trait at 0.01 level. In the graphs, the measures of linkage disequilibrium are given by $D_{A_1B_1} = 0.08, D_{A_1Q} = D_{B_1Q} = 0.1$. The other parameters are the same as those of Figure 2.

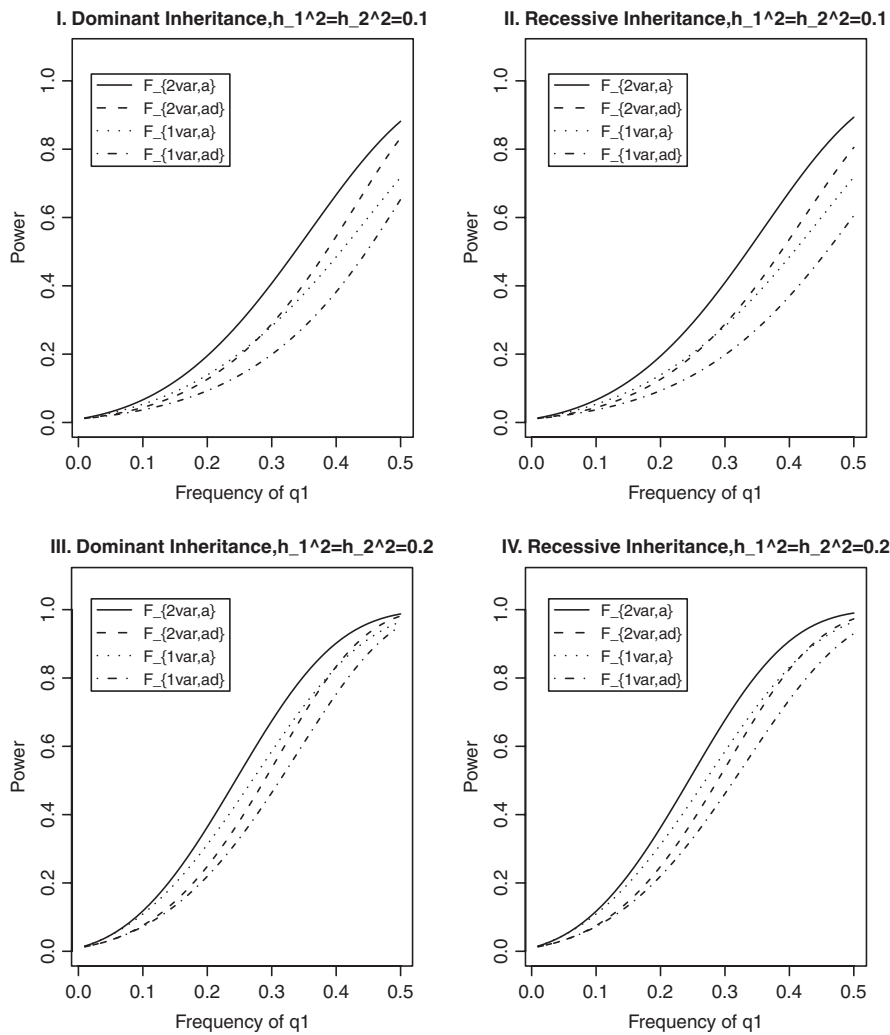


Fig. 4. Power curves against trait allele frequency q_1 at 0.01 level. The parameters are given by $D_{A_1B_1} = 0.08, D_{A_1Q} = (\min(P_{A_1}, q_1) - P_{A_1}q_1)/2, D_{B_1Q} = (\min(P_{B_1}, q_1) - P_{B_1}q_1)/2$. The other parameters are the same as those of Figure 2.

increases in Figure 5. In addition, two quantitative trait test statistics give higher power than a single quantitative trait test as that in Figures 2 and 3.

DISCUSSION

In this article, variance component models are developed for bivariate/multi-variate combined linkage and association mapping of QTL. By analytical power analysis and practical example, we show that bivariate models can be more powerful than univariate models. For moderate-sized samples, the proposed models lead to correct type I error rates; and so the models are reasonably robust. In the paper, we focus on bivariate models. Theoretically, it is straightforward to generalize the models for multivariate analysis. However, one may want to notice a potential problem in estimating the parameters: the number of parameters can increase rapidly if multiple traits are considered, and then it would be not easy to estimate the parameters accurately and the robustness of the models

can be problematic. Hence, trait selection can be important when a lot of traits are available, such as the expression phenotypes from microarray analysis, Problem 1 of GAW 15.

In the genetics literature, multivariate linkage models have been developed by various research groups [Amos et al., 1990; Arya et al., 2003; Jiang and Zeng, 1995]. To our knowledge, there has not been much research on multivariate combined linkage and association mapping of QTL. Thus, the proposed methods fill some of the gaps. In the paper, only one scenario is considered: one quantitative trait locus Q exerts pleiotropic effects on multiple quantitative traits. Other scenarios can be interesting for future research: (1) the different traits are influenced by separate tightly linked loci and (2) the different traits are influenced by unlinked loci which may be located on different chromosomes, but the different traits interact with each other. Each of these two scenarios represents an important research topic, and deserves more in-depth investigation. Actually, the proposed regression models can still be used to describe the traits if the

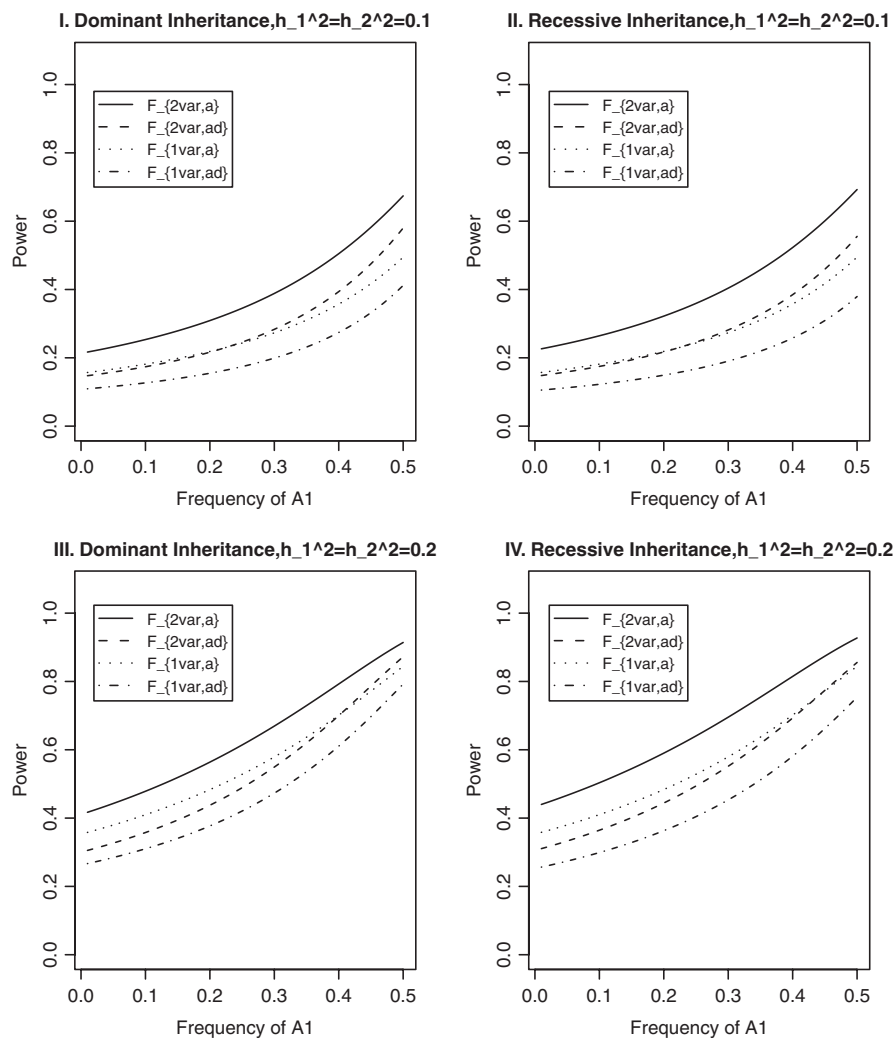


Fig. 5. Power curves against trait allele frequency P_{A_1} at 0.01 level. The parameters are given by $D_{AB} = (\min(P_{A_1}, P_{B_1}) - P_{A_1}P_{B_1})/2$, $D_{AQ_1} = (\min(P_{A_1}, q_1) - P_{A_1}q_1)/2$, $D_{BQ_1} = 0.08$. The other parameters are the same as those of Figure 2.

different traits are influenced by separate tightly linked loci, but the properties of the models need more investigation.

If genetic traits are influenced by multiple epistatic quantitative trait loci, it is interesting and important to detect gene-gene interactions. There have been a lot of research on gene-gene interactions of quantitative traits [Bensen et al., 2003; Cheverud, 2000; Cheverud and Routman, 1995, 1996; Kraft et al., 2003; Ulgen et al., 2003]. However, almost all the research on gene-gene interactions focus on population data. Moreover, no marker information is used in these studies. It would be exciting to develop models to use marker information to detect gene-gene interactions based on combinations of population data and pedigree data. Actually, variance component models can be developed to detect association between markers and the QTL, the gene-gene and gene-environment interactions. The research would take advantage of dense markers of the human genome research. Further investigations are necessary for the issues. As a final point, we assume that there is no missing genotype data in this paper. In the presence of genotyping error and

missing data, more research is needed to investigate their influence on the proposed models and methods [Ritchie et al., 2003].

ACKNOWLEDGMENTS

We thank the permission to use GAW15 data as an application of our method, and for the support to make our research possible. We thank one anonymous reviewer for very detailed and thoughtful critiques.

REFERENCES

- Allison DB, Thiel B, St. Jean P, Elston RC, Infante MC, Schork NJ. 1998. Multiple phenotype modeling in gene-mapping studies of quantitative traits: power advantages. *Am J Hum Genet* 63:1190–1201.
- Almasy L, Dyer TD, Blangero J. 1997. Bivariate quantitative trait linkage analysis: pleiotropy versus co-incident linkage. *Genet Epidemiol* 14:953–958.
- Amos CI. 1994. Robust variance-components approach for assessing linkage in pedigrees. *Am J Hum Genet* 54:534–543.

- Amos CI, Elston RC, Bonney GE, Keats BJB, 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.
- Amos CI, de Andrade M, Zhu DK. 2001. Comparison of multivariate tests for genetic linkage. *Hum Hered* 51:133–144.
- Amos CI, Chen WV, Lee A, Li W, Kern M, Lundsten R, Batliwalla F, Wener M, Remmers E, Kastner DA, Criswell LA, Seldin MA, Gregersen PK. 2006. High-density SNP analysis of 642 Caucasian families with rheumatoid arthritis identifies two new linkage regions on 11p12 and 2q33. *Genes Immun* 7:277–286.
- Arya R, Lehman D, Hunt KJ, Schneider J, Almasy L, Blangero J, Stern MP, Duggirala R. 2003. Evidence for bivariate linkage of obesity and HDL-C levels in the Framingham Heart Study. *BMC Genet* 4:S52.
- Bensen JT, Lange LA, Langefeld CD, Chang BL, Bleecker ER, Meyers DA, Xu JF. 2003. Exploring pleiotropy using principal components. *BMC Genet* 4:S53.
- Blangero J, Konigsberg LW. 1991. Multivariate segregation analysis using the mixed model. *Genet Epidemiol* 8:299–316.
- Cheverud JM. 2000. Detecting epistasis among quantitative trait loci. In: Wolf JB, Brodie III ED, Wade MJ, editors. *Epistasis and the Evolutionary Process*. Oxford: Oxford University Press. p 58–81.
- Cheverud JM, Routman EJ. 1995. Epistasis and its contribution to genetic variance components. *Genetics* 139:1455–1461.
- Cheverud JM, Routman EJ. 1996. Epistasis as a source of increased additive genetic variance at population bottlenecks. *Evolution* 50:1042–1051.
- Evans DM. 2002. The power of multivariate quantitative-trait locus linkage analysis is influenced by the correlation between the variables. *Am J Hum Genet* 70:1599–1602.
- Falconer DS, Mackay TFC. 1996. *Introduction to Quantitative Genetics*, 4th edition. London: Longman.
- Fan RZ, Xiong MM. 2002. High resolution mapping of quantitative trait loci by linkage disequilibrium analysis. *Eur J Human Genet* 10:607–615.
- Fan RZ, Xiong MM. 2003. Combined high resolution linkage and association mapping of quantitative trait loci. *Eur J Human Genet* 11:125–137.
- Fan RZ, Spinka C, Jin L, Jung JS. 2005. Pedigree linkage disequilibrium mapping of quantitative trait loci. *Eur J Hum Genet* 13:216–231.
- Fan RZ, Jung JS, Jin J. 2006. High resolution association mapping of quantitative trait loci, a population based approach. *Genetics* 172:663–686.
- Graybill FA. 1976. *Theory and Application of the Linear Model*. California: Pacific Grove.
- Harville DA. 1997. *Matrix Algebra From a Statistician’s Perspective*. New York: Springer.
- Jiang C, Zeng Z. 1995. Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* 140:1111–1127.
- Jung JS, Fan RZ, Jin L. 2005. Combined linkage and association mapping of quantitative trait loci by multiple markers. *Genetics* 170:881–898.
- Kraft P, de Andrade M. 2003. Group 6: pleiotropy and multivariate analysis. *Genet Epidemiol* 25:S50–S56.
- Kraft P, Bauman L, Yuan JY, Horvath S. 2003. Multivariate variance-components analysis of longitudinal blood pressure measurements from the Framingham Heart Study. *BMC Genet* 4:S55.
- Lange K. 2002. *Mathematical and Statistical Methods for Genetic Analysis*, 2nd edition. New York: Springer.
- Lange K, Boehnke M. 1983. Extensions to pedigree analysis. IV. Covariance component models for multivariate traits. *Am J Med Genet* 14:513–524.
- Ritchie MD, Hahn LW, Moore JH. 2003. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genet Epidemiol* 24:150–157.
- Ulgien A, Han Z, Li W. 2003. Correlation between quantitative traits and correlation between corresponding LOD scores: detection of pleiotropic effects. *BMC Genet* 4:S60.

APPENDIX A

For $g = 1, 2, \dots, m, k = 1, \dots, n$, let us denote $D_{A_g B_k} = P(A_g B_k) - P_{A_g} P_{B_k}$, which are measures of LD between two markers A and B . Here $P(A_g B_k)$ is frequency of haplotype $A_g B_k$. Denote

$$E_{AA} = \begin{pmatrix} 2P_{A_1}(P_{A_1} + 1) & \cdots & 2P_{A_1}P_{A_{a-1}} \\ \vdots & \cdots & \vdots \\ 2P_{A_1}P_{A_{a-1}} & \cdots & 2P_{A_{a-1}}(P_{A_{a-1}} + 1) \end{pmatrix},$$

$$E_{AB} = \begin{pmatrix} 2D_{A_1 B_1} + 4P_{A_1}P_{B_1} & \cdots & 2D_{A_1 B_{b-1}} + 4P_{A_1}P_{B_{b-1}} \\ \vdots & \cdots & \vdots \\ 2D_{A_{a-1} B_1} + 4P_{A_{a-1}}P_{B_1} & \cdots & 2D_{A_{a-1} B_{b-1}} + 4P_{A_{a-1}}P_{B_{b-1}} \end{pmatrix},$$

$$E_{BB} = \begin{pmatrix} 2P_{B_1}(P_{B_1} + 1) & \cdots & 2P_{B_1}P_{B_{b-1}} \\ \vdots & \cdots & \vdots \\ 2P_{B_{b-1}}P_{B_b} & \cdots & 2P_{B_{b-1}}(P_{B_{b-1}} + 1) \end{pmatrix}.$$

From the results of Appendix E, Fan et al. [2006], we may show $E(X_{AUB} X_{AUB}^T) = \begin{pmatrix} E_{AA} & E_{AB} \\ E_{AB}^T & E_{BB} \end{pmatrix}$. Notice $E x_{A_i} = 2P_{A_i}$ and $E x_{B_k} = 2P_{B_k}$. Therefore, we have

$$V_A = 2 \begin{pmatrix} P_{A_1}(1 - P_{A_1}) & -P_{A_1}P_{A_2} & \cdots & -P_{A_1}P_{A_{m-1}} & D_{A_1 B_1} & \cdots & D_{A_1 B_{n-1}} \\ -P_{A_1}P_{A_2} & P_{A_2}(1 - P_{A_2}) & \cdots & -P_{A_2}P_{A_{m-1}} & D_{A_2 B_1} & \cdots & D_{A_2 B_{n-1}} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ -P_{A_1}P_{A_{m-1}} & -P_{A_2}P_{A_{m-1}} & \cdots & P_{A_{m-1}}(1 - P_{A_{m-1}}) & D_{A_{m-1} B_1} & \cdots & D_{A_{m-1} B_{n-1}} \\ D_{A_1 B_1} & D_{A_2 B_1} & \cdots & D_{A_{m-1} B_1} & P_{B_1}(1 - P_{B_1}) & \cdots & -P_{B_1}P_{B_{n-1}} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ D_{A_1 B_{n-1}} & D_{A_2 B_{n-1}} & \cdots & D_{A_{m-1} B_{n-1}} & -P_{B_1}P_{B_{n-1}} & \cdots & P_{B_{n-1}}(1 - P_{B_{n-1}}) \end{pmatrix}.$$

For the dominance variance-covariance matrix V_D , it can be shown that $V_D = Cov(Z_{AUB}, Z_{AUB}) = E(Z_{AUB}Z_{AUB}^T)$, where

$$\begin{aligned} E[(z_A^{(gh)})^2] &= P_{A_g}^2 P_{A_h}^2 (P_{A_g} + P_{A_h})^2, \\ E[z_A^{(gh)} z_A^{(gh')}] &= [P_{A_g} P_{A_h} P_{A_h'}]^2, \\ E[z_A^{(gh)} z_A^{(g'h')}] &= 0, \\ E[(z_B^{(kl)})^2] &= P_{B_k}^2 P_{B_l}^2 (P_{B_k} + P_{B_l})^2, \\ E[z_B^{(kl)} z_A^{(kl')}] &= [P_{B_k} P_{B_l} P_{B_l'}]^2, \\ E[z_B^{(kl)} z_A^{(k'l')}] &= 0, \\ E[z_A^{(gh)} z_B^{(kl)}] &= [P_{A_h} (P_{B_l} D_{A_g B_k} - P_{B_k} D_{A_g B_l}) - P_{A_g} (P_{B_l} D_{A_h B_k} - P_{B_k} D_{A_h B_l})]^2. \end{aligned}$$

APPENDIX B

To get an approximation of non-centrality parameter λ_{ad} for population data, assume that there are N unrelated individuals. Here we assume that N is large enough that large sample theory applies. For each individual, the variance-covariance matrix is $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$, whose inverse is $\Sigma^{-1} = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$, where $a = \sigma_2^2 / (\sigma_1^2 \sigma_2^2 - \sigma_{12}^2)$, $b = -\sigma_{12} / (\sigma_1^2 \sigma_2^2 - \sigma_{12}^2)$, $c = \sigma_1^2 / (\sigma_1^2 \sigma_2^2 - \sigma_{12}^2)$. Therefore, the non-centrality parameter is

$$\begin{aligned} \lambda_{ad} &= (HY)^T \left[H \left[\sum_{s=1}^N \begin{pmatrix} U_s^T & O \\ O & U_s^T \end{pmatrix} \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} U_s & O \\ O & U_s \end{pmatrix} \right]^{-1} H^T \right]^{-1} (HY) \\ &= (HY)^T \left[H \left(\sum_{s=1}^N \begin{pmatrix} a U_s^T U_s & b U_s^T U_s \\ b U_s^T U_s & c U_s^T U_s \end{pmatrix} \right)^{-1} H^T \right]^{-1} (HY) \\ &= (HY)^T \left[H \left(\begin{pmatrix} a & b \\ b & c \end{pmatrix} \otimes \sum_{s=1}^N U_s^T U_s \right)^{-1} H^T \right]^{-1} (HY) \\ &= (HY)^T \left[H \left(\Sigma \otimes \left(\sum_{s=1}^N U_s^T U_s \right) \right)^{-1} H^T \right]^{-1} (HY). \end{aligned}$$

Notice that there are no co-variables. Hence, $U_s = (1, X_{s,AUB}^T, Z_{s,AUB}^T)$. By the results in Appendix A, large number law leads to the following approximation:

$$\left(\sum_{s=1}^N U_s^T U_s \right)^{-1} \approx N^{-1} \begin{pmatrix} 1 & EX_{AUB}^T & O \\ EX_{AUB} & E(X_{AUB} X_{AUB}^T) & O \\ O & O & V_D \end{pmatrix}^{-1} = N^{-1} \begin{pmatrix} 1 + EX_{AUB}^T V_A^{-1} EX_{AUB} & -EX_{AUB}^T V_A^{-1} & O \\ -V_A^{-1} EX_{AUB} & V_A^{-1} & O \\ O & O & V_D^{-1} \end{pmatrix},$$

which leads to

$$H \left(\begin{pmatrix} a & b \\ b & c \end{pmatrix} \otimes \sum_{s=1}^N U_s^T U_s \right)^{-1} H^T \approx N^{-1} \Sigma \otimes \begin{pmatrix} V_A^{-1} & O \\ O & V_D^{-1} \end{pmatrix}$$

Therefore, the non-centrality parameter can be approximated as $\lambda_{ad} \approx N(HY)^T \left(\Sigma^{-1} \otimes \begin{pmatrix} V_A & O \\ O & V_D \end{pmatrix} \right) (HY)$. Let us denote $D_{AQ} = (D_{A_1 Q}, \dots, D_{A_{m-1} Q})^T$ and $D_{BQ} = (D_{B_1 Q}, \dots, D_{B_{n-1} Q})^T$; $\Delta_{AQ} = ([P_{A_2} D_{A_1 Q} - P_{A_1} D_{A_2 Q}]^2, \dots, [P_{A_{m-1}} D_{A_{m-2} Q} - P_{A_{m-2}} D_{A_{m-1} Q}]^2)^T$ and $\Delta_{BQ} = ([P_{B_2} D_{B_1 Q} - P_{B_1} D_{B_2 Q}]^2, \dots, [P_{B_{n-1}} D_{B_{n-2} Q} - P_{B_{n-2}} D_{B_{n-1} Q}]^2)^T$. The relations (6) imply that

$$HY = \text{diag}(V_A^{-1}, V_D^{-1}, V_A^{-1}, V_D^{-1}) \times (2\alpha_{Q1}(D_{AQ}^T, D_{BQ}^T), \delta_{Q1}(\Delta_{AQ}^T, \Delta_{BQ}^T), 2\alpha_{Q2}(D_{AQ}^T, D_{BQ}^T), \delta_{Q2}(\Delta_{AQ}^T, \Delta_{BQ}^T))^T.$$

Finally, the non-centrality parameter λ_{ad} can be approximated by

$$\begin{aligned}\lambda_{ad} &\approx N[a(2\alpha_{Q1})^2 + 2b(2\alpha_{Q1})(2\alpha_{Q2}) + c(2\alpha_{Q2})^2](D_{AQ}^\tau, D_{BQ}^\tau)V_A^{-1} \begin{pmatrix} D_{AQ} \\ D_{BQ} \end{pmatrix} \\ &\quad + N[a\delta_{Q1}^2 + 2b\delta_{Q1}\delta_{Q2} + c\delta_{Q2}^2](\Delta_{AQ}^\tau, \Delta_{BQ}^\tau)V_D^{-1} \begin{pmatrix} \Delta_{AQ} \\ \Delta_{BQ} \end{pmatrix} \\ &= \frac{2N}{q_1q_2}[a\sigma_{ga1}^2 + 2b\sigma_{ga12} + c\sigma_{ga2}^2](D_{AQ}^\tau, D_{BQ}^\tau)V_A^{-1} \begin{pmatrix} D_{AQ} \\ D_{BQ} \end{pmatrix} \\ &\quad + \frac{N}{q_1^2q_2^2}[a\sigma_{gd1}^2 + 2b\delta_{Q1}\sigma_{gd12} + c\sigma_{gd2}^2](\Delta_{AQ}^\tau, \Delta_{BQ}^\tau)V_D^{-1} \begin{pmatrix} \Delta_{AQ} \\ \Delta_{BQ} \end{pmatrix}.\end{aligned}$$

Similarly, one may show the following approximation:

$$\begin{aligned}\lambda_a &\approx N[a(2\alpha_{Q1})^2 + 2b(2\alpha_{Q1})(2\alpha_{Q2}) + c(2\alpha_{Q2})^2](D_{AQ}^\tau, D_{BQ}^\tau)V_A^{-1} \begin{pmatrix} D_{AQ} \\ D_{BQ} \end{pmatrix} \\ &= \frac{2N}{q_1q_2}[a\sigma_{ga1}^2 + 2b\sigma_{ga12} + c\sigma_{ga2}^2](D_{AQ}^\tau, D_{BQ}^\tau)V_A^{-1} \begin{pmatrix} D_{AQ} \\ D_{BQ} \end{pmatrix}.\end{aligned}$$

APPENDIX C

Consider pedigree data with notations introduced in the Section of Method, and assume that there are no co-variables. For a relative pair (1,2) of individuals 1 and 2 who are non-inbred relatives, Table II gives the conditional probability $P(G_1, G_2|C)$ given their allele IBD sharing status. Here, G_j is genotype of individual j , and C is one event of $(IBD = k), k = 0, 1, 2$. For example, $P(A_g A_g, A_g A_g | IBD = 0) = P_{A_g}^4$, $P(A_g A_g, A_g A_h | IBD = 0) = 2P_{A_g}^3 P_{A_h}$ and $P(A_g A_g, A_h A_h | IBD = 0) = P_{A_g}^2 P_{A_h}^2$. Utilizing the conditional probabilities of Table II, the conditional covariances of variables $x_{Aj}^{(g)}, x_{Bj}^{(k)}, z_{Aj}^{(gh)}$ and $z_{Bj}^{(kl)}$ of a relative pair $j = 1, 2$ can be calculated and the results are listed in Table III. Given $(IBD = 0)$, the covariances are 0 as the two variables are independent and so unrelated (e.g., $\text{Cov}(x_{A1}^{(g)}, x_{A2}^{(g)} | IBD = 0) = 0$). Other entries of Table III can be calculated, accordingly.

Denote $X_{Aj} = (x_{Aj}^{(1)}, \dots, x_{Aj}^{(m-1)})^\tau$, $X_{Bj} = (x_{Bj}^{(1)}, \dots, x_{Bj}^{(n-1)})^\tau$, and $X_{AUB}^{(j)} = (X_{Aj}^\tau, X_{Bj}^\tau)^\tau, j = 1, 2$. Similarly, let $Z_{Aj} = (z_{Aj}^{(12)}, \dots, z_{Aj}^{(1m)}, z_{Aj}^{(23)}, \dots, z_{Aj}^{(2m)}, \dots, z_{Aj}^{(m-1,m)})^\tau$, $Z_{Bj} = (z_{Bj}^{(12)}, \dots, z_{Bj}^{(1n)}, z_{Bj}^{(23)}, \dots, z_{Bj}^{(2n)}, \dots, z_{Bj}^{(n-1,n)})^\tau$, and $Z_{AUB}^{(j)} = (Z_{Aj}^\tau, Z_{Bj}^\tau)^\tau$. Based on the results of Table III, it can be seen that $\text{Cov}(X_{AUB}^{(1)}, X_{AUB}^{(2)} | IBD = 0) = \text{Cov}(Z_{AUB}^{(1)}, Z_{AUB}^{(2)} | IBD = 0) = 0$ and $\text{Cov}(X_{AUB}^{(1)}, Z_{AUB}^{(2)} | IBD = k) = 0, k = 0, 1, 2$. In addition, we have $\text{Cov}(X_{AUB}^{(1)}, X_{AUB}^{(2)} | IBD = 1) = \frac{1}{2}\text{Cov}(X_{AUB}^{(1)}, X_{AUB}^{(2)} | IBD = 2) = V_A/2$, $\text{Cov}(Z_{AUB}^{(1)}, Z_{AUB}^{(2)} | IBD = 1) = 0$, and $\text{Cov}(Z_{AUB}^{(1)}, Z_{AUB}^{(2)} | IBD = 2) = V_D$. Recall that Φ_{12} is kinship coefficient of individuals 1 and 2, and let Δ_{712} be the probability that both alleles shared by the two individuals 1 and 2 are IBD at any locus [Lange, 2002]. Then it can be shown that the covariance matrix of variable vectors $X_{AUB}^{(1)}$ and $Z_{AUB}^{(2)}$ is a zero matrix, and

$$\begin{aligned}\text{Cov}(X_{AUB}^{(1)}, X_{AUB}^{(2)}) &= 2\Phi_{12}\text{Cov}(X_{AUB}^{(1)}, X_{AUB}^{(2)} | IBD = 2) = 2\Phi_{12}V_A, \\ \text{Cov}(Z_{AUB}^{(1)}, Z_{AUB}^{(2)}) &= \Delta_{712}\text{Cov}(Z_{AUB}^{(1)}, Z_{AUB}^{(2)} | IBD = 2) = \Delta_{712}V_D.\end{aligned}\tag{A1}$$

APPENDIX D

Again, assume that there are no co-variables. Consider I nuclear families, and assume that each nuclear family has both parents and K offspring. The total number of individuals is $I(K+2)$. Let us list the $K+2$ individuals of each family as $j = 1, 2, 3, \dots, K+2$, where individual 1 is the father and individual 2 is the mother, and the offspring are listed as $j = 3, \dots, K+2$. Suppose that variance-covariance matrices of the I families are the same, i.e., $\Sigma_1 = \dots = \Sigma_I = \Sigma$. Let us write variance-covariance matrix as $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^\tau & \Sigma_{22} \end{pmatrix}$, where $\Sigma_{11} = \text{Var}(Y_{1s})$ and $\Sigma_{22} = \text{Var}(Y_{2s})$ are symmetric matrices, and $\Sigma_{12} = \text{Cov}(Y_{1s}, Y_{2s})$ is covariance matrix of Y_{1s} and Y_{2s} . Then $\Sigma_s^{-1} = \begin{pmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} -\Sigma_{11}^{-1}\Sigma_{12} \\ I_{K+2} \end{pmatrix}(\Sigma_{22} - \Sigma_{12}^\tau \Sigma_{11}^{-1} \Sigma_{12})^{-1}(-\Sigma_{11}^{-1}\Sigma_{12})^\tau, I_{K+2}$ [Harville, 1997, p. 99]. Let Σ_s^{-1} be $\begin{pmatrix} A & B \\ B^\tau & C \end{pmatrix}$, where $A = \Sigma_{11}^{-1} - \Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{12}^\tau \Sigma_{11}^{-1} \Sigma_{12})^{-1}(\Sigma_{11}^{-1}\Sigma_{12})^\tau$, $B = \Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{12}^\tau \Sigma_{11}^{-1} \Sigma_{12})^{-1}$ and $C = (\Sigma_{22} - \Sigma_{12}^\tau \Sigma_{11}^{-1} \Sigma_{12})^{-1}$. The non-centrality

TABLE II. Conditional probability $P(G_1, G_2|C)$ of a relative pair (1,2) given their allele IBD sharing status

Conditional probability	Allele IBD sharing status C		
	IBD = 0	IBD = 1	IBD = 2
$P(A_g A_g, A_g A_g C)$	$P_{A_g}^4$	$P_{A_g}^3$	$P_{A_g}^2$
$P(A_g A_g, A_g A_h C)$	$2P_{A_g}^3 P_{A_h}$	$P_{A_h} P_{A_g}^2$	0
$P(A_g A_g, A_h A_h C)$	$P_{A_g}^2 P_{A_h}^2$	0	0
$P(A_g A_g, A_h A_{h'} C)$	$2P_{A_g}^2 P_{A_h} P_{A_{h'}}$	0	0
$P(A_g A_h, A_g A_h C)$	$4P_{A_g}^2 P_{A_h}^2$	$P_{A_g} P_{A_h}^2 + P_{A_g}^2 P_{A_h}$	$2P_{A_g} P_{A_h}$
$P(A_g A_h, A_g A_{h'} C)$	$4P_{A_g}^2 P_{A_h} P_{A_{h'}}$	$P_{A_g} P_{A_h} P_{A_{h'}}$	0
$P(A_g A_h, A_g A_{h'} C)$	$4P_{A_g} P_{A_h} P_{A_{h'}} P_{A_{h'}}$	0	0
$P(A_g A_g, B_k B_k C)$	$P_{A_g}^2 P_{B_k}^2$	$P_{A_g} P_{B_k} P(A_g B_k)$	$P(A_g B_k)^2$
$P(A_g A_g, B_k B_l C)$	$2P_{A_g}^2 P_{B_k} P_{B_l}$	$P_{A_g} P_{B_l} P(A_g B_k) + P_{A_g} P_{B_k} P(A_g B_l)$	$2P(A_g B_k)P(A_g B_l)$
$P(A_g A_h, B_k B_k C)$	$2P_{A_g} P_{A_h} P_{B_k}^2$	$P_{A_g} P_{B_k} P(A_h B_k) + P_{A_h} P_{B_k} P(A_g B_k)$	$2P(A_g B_k)P(A_h B_k)$
$P(A_g A_h, B_k B_l C)$	$4P_{A_g} P_{A_h} P_{B_k} P_{B_l}$	$P_{A_g} P_{B_k} P(A_h B_l) + P_{A_g} P_{B_l} P(A_h B_k) + P_{A_h} P_{B_k} P(A_g B_l) + P_{A_h} P_{B_l} P(A_g B_k)$	$2P(A_g B_k)P(A_h B_l)$

Here, G_j is genotype of individual j , and C is one event of $(IBD = k)$, $k = 0, 1, 2$. In the Table, we assume $g \neq h, g \neq g', g \neq h', h \neq g', h \neq h', g' \neq h', k \neq l$.

TABLE III. Conditional expectation of a relative pair (1,2) given their allele IBD sharing status

Conditional expectation	Allele IBD sharing status C		
	IBD = 0	IBD = 1	IBD = 2
$Cov(x_{A1}^{(g)}, x_{A2}^{(g)} C)$	0	$P_{A_g} [1 - P_{A_g}]$	$2P_{A_g} [1 - P_{A_g}]$
$Cov(x_{A1}^{(g)}, x_{A2}^{(h)} C)$	0	$-P_{A_g} P_{A_h}$	$-2P_{A_g} P_{A_h}$
$Cov(z_{A1}^{(g,h)}, z_{A2}^{(g,h)} C)$	0	0	$P_{A_g}^2 P_{A_h}^2 (P_{A_g} + P_{A_h})^2$
$Cov(z_{A1}^{(g,h)}, z_{A2}^{(g,h')} C)$	0	0	$[P_{A_g} P_{A_h} P_{A_{h'}}]^2$
$Cov(z_{A1}^{(g,h)}, z_{A2}^{(g,h')} C)$	0	0	0
$Cov(x_{A1}^{(g)}, z_{A2}^{(g,h)} C)$	0	0	0
$Cov(x_{A1}^{(g)}, z_{A2}^{(g,h')} C)$	0	0	0
$Cov(x_{A1}^{(g)}, x_{B1}^{(k)} C)$	0	$D_{A_g B_k}$	$2D_{A_g B_k}$
$Cov(x_{A1}^{(g)}, z_{B2}^{(k)} C)$	0	0	0
$Cov(z_{A1}^{(g,h)}, z_{B2}^{(k)} C)$	0	0	$E[z_{A1}^{(g,h)} z_{B2}^{(k)}] = E[z_{A2}^{(g,h)} z_{B2}^{(k)}]$

In the Table, we assume $g \neq h, g \neq g', g \neq h', h \neq g', h \neq h', g' \neq h', k \neq l$.

parameter λ_{ad} of F -test $F_{AB,ad}$ is given by

$$\begin{aligned} \lambda_{ad} &= (HY)^\tau \left[H \left[\sum_{s=1}^I \begin{pmatrix} U_s^\tau & O \\ O & U_s^\tau \end{pmatrix} \begin{pmatrix} A & B \\ B^\tau & C \end{pmatrix} \begin{pmatrix} U_s & O \\ O & U_s \end{pmatrix} \right]^{-1} H^\tau \right]^{-1} (HY) \\ &= (HY)^\tau \left[H \left(\sum_{s=1}^I \begin{pmatrix} U_s^\tau A U_s & U_s^\tau B U_s \\ U_s^\tau B^\tau U_s & U_s^\tau C U_s \end{pmatrix} \right)^{-1} H^\tau \right]^{-1} (HY), \end{aligned} \tag{A2}$$

where $HY = \begin{pmatrix} \alpha_{1,AUB} \\ \delta_{1,AUB} \\ \alpha_{2,AUB} \\ \delta_{2,AUB} \end{pmatrix}$.

Let $A = (A_{kl})_{(K+2) \times (K+2)}$, and let $D_A = (A_{13} + \dots + A_{1,K+2}) + (A_{23} + \dots + A_{2,K+2}) + \sum_{h=3}^{K+2} \sum_{j=h+1}^{K+2} A_{hj}$. Denote $S_A = \sum_{k=1}^{K+2} \sum_{l=1}^{K+2} A_{kl}$ and $tr(A) = \sum_{k=1}^{K+2} A_{kk}$. Notice that there is no co-variates. Large number law leads to

$$\sum_{s=1}^I U_s^\tau A U_s \approx I \begin{pmatrix} S_A & S_A [EX_{AUB}^{(1)}]^\tau & O \\ S_A EX_{AUB}^{(1)} & (tr(A) + D_A) V_A + (S_A - 2A_{12}) EX_{AUB}^{(1)} [EX_{AUB}^{(1)}]^\tau & O \\ O & O & tr(A) V_D + \sum_{k=3}^{K+2} \sum_{l=k+1}^{K+2} A_{kl} V_D / 2 \end{pmatrix}, \tag{A3}$$

where O are zero vectors or matrices, and $E(X_{AUB}^{(1)}) = (2P_{A_1}, \dots, 2P_{A_{m-1}}, 2P_{B_1}, \dots, 2P_{B_{n-1}})^\tau$. Similarly, let $B = (B_{kl})_{(K+2) \times (K+2)}$, and let $D_B = (B_{13} + \dots + B_{1,K+2}) + (B_{23} + \dots + B_{2,K+2}) + \sum_{h=3}^{K+2} \sum_{j=h+1}^{K+2} B_{hj}$. Denote $S_B = \sum_{k=1}^{K+2} \sum_{l=1}^{K+2} B_{kl}$ and $tr(B) = \sum_{k=1}^{K+2} B_{kk}$. Large number law leads to

$$\sum_{s=1}^I U_s^\tau B U_s \approx I \begin{pmatrix} S_B & S_B [EX_{AUB}^{(1)}]^\tau & O \\ S_B EX_{AUB}^{(1)} & (tr(B) + D_B) V_A + (S_B - 2B_{12}) EX_{AUB}^{(1)} [EX_{AUB}^{(1)}]^\tau & O \\ O & O & tr(B) V_D + \sum_{k=3}^{K+2} \sum_{l=k+1}^{K+2} B_{kl} V_D / 2 \end{pmatrix}. \quad (A4)$$

Let $C = (C_{kl})_{(K+2) \times (K+2)}$, and let $D_C = (C_{13} + \dots + C_{1,K+2}) + (C_{23} + \dots + C_{2,K+2}) + \sum_{h=3}^{K+2} \sum_{j=h+1}^{K+2} C_{hj}$. Denote $S_C = \sum_{k=1}^{K+2} \sum_{l=1}^{K+2} C_{kl}$ and $tr(C) = \sum_{k=1}^{K+2} C_{kk}$. Large number law leads to

$$\sum_{s=1}^I U_s^\tau C U_s \approx I \begin{pmatrix} S_C & S_C [EX_{AUB}^{(1)}]^\tau & O \\ S_C EX_{AUB}^{(1)} & (tr(C) + D_C) V_A + (S_C - 2C_{12}) EX_{AUB}^{(1)} [EX_{AUB}^{(1)}]^\tau & O \\ O & O & tr(C) V_D + \sum_{k=3}^{K+2} \sum_{l=k+1}^{K+2} C_{kl} V_D / 2 \end{pmatrix}. \quad (A5)$$

Assume that the number of families I is large enough that large sample theory applies. The non-centrality parameter λ_{ad} of statistic $F_{AB,ad}$ can be calculated by plugging approximations (A3)–(A5) into (A2).

Similarly, the non-centrality parameter λ_a of F -test $F_{AB,a}$ is given by

$$\begin{aligned} \lambda_a &= (H_1 \Psi)^\tau \left[H_1 \left[\sum_{s=1}^I \begin{pmatrix} V_s^\tau & O \\ O & V_s^\tau \end{pmatrix} \begin{pmatrix} A & B \\ B^\tau & C \end{pmatrix} \begin{pmatrix} V_s & O \\ O & V_s \end{pmatrix} \right]^{-1} H_1^\tau \right]^{-1} (H_1 \Psi) \\ &= (H_1 \Psi)^\tau \left[H_1 \left(\sum_{s=1}^I \begin{pmatrix} V_s^\tau A V_s & V_s^\tau B V_s \\ V_s^\tau B^\tau V_s & V_s^\tau C V_s \end{pmatrix} \right)^{-1} H_1^\tau \right]^{-1} (H_1 \Psi), \end{aligned}$$

where $H_1 \Psi = \begin{pmatrix} \alpha_{1,AUB} \\ \alpha_{2,AUB} \end{pmatrix}$. Again, large number law leads to the following approximations:

$$\begin{aligned} \sum_{s=1}^I V_s^\tau A V_s &\approx I \begin{pmatrix} S_A & S_A [EX_{AUB}^{(1)}]^\tau \\ S_A EX_{AUB}^{(1)} & (tr(A) + D_A) V_A + (S_A - 2A_{12}) EX_{AUB}^{(1)} [EX_{AUB}^{(1)}]^\tau \end{pmatrix}, \\ \sum_{s=1}^I V_s^\tau B V_s &\approx I \begin{pmatrix} S_B & S_B [EX_{AUB}^{(1)}]^\tau \\ S_B EX_{AUB}^{(1)} & (tr(B) + D_B) V_B + (S_B - 2B_{12}) EX_{AUB}^{(1)} [EX_{AUB}^{(1)}]^\tau \end{pmatrix}, \\ \sum_{s=1}^I V_s^\tau C V_s &\approx I \begin{pmatrix} S_C & S_C [EX_{AUB}^{(1)}]^\tau \\ S_C EX_{AUB}^{(1)} & (tr(C) + D_C) V_C + (S_C - 2C_{12}) EX_{AUB}^{(1)} [EX_{AUB}^{(1)}]^\tau \end{pmatrix}. \end{aligned}$$

APPENDIX E

Consider a combination of N unrelated individuals, I_1 trio families, and I_2 nuclear families of size 4. Let us assume that the variance-covariance matrices of the I_1 trio families are equal to $\Sigma_{N+1} = \dots = \Sigma_{N+I_1} = \begin{pmatrix} A_1 & B_1 \\ B_1^\tau & C_1 \end{pmatrix}^{-1}$ and assume that the variance-covariance matrices of the I_2 nuclear families of size 4 are equal to $\Sigma_{N+I_1+1} = \dots = \Sigma_{N+I_1+I_2} = \begin{pmatrix} A_2 & B_2 \\ B_2^\tau & C_2 \end{pmatrix}^{-1}$. Define

$$\begin{aligned} K_0 &= \sum_{s=1}^N \begin{pmatrix} U_s^\tau & O \\ O & U_s^\tau \end{pmatrix} \Sigma_s^{-1} \begin{pmatrix} U_s & O \\ O & U_s \end{pmatrix} \\ &= \sum_{s=1}^N \begin{pmatrix} U_s^\tau & O \\ O & U_s^\tau \end{pmatrix} \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}^{-1} \begin{pmatrix} U_s & O \\ O & U_s \end{pmatrix} \\ &\approx N \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}^{-1} \otimes \begin{pmatrix} 1 & EX_{AUB}^\tau & O \\ EX_{AUB} & E(X_{AUB} X_{AUB}^\tau) & O \\ O & O & V_D \end{pmatrix}, \end{aligned}$$

$$\begin{aligned}
K_1 &= \sum_{s=N+1}^{N+I_1} \begin{pmatrix} U_s^\tau & O \\ O & U_s^\tau \end{pmatrix} \Sigma_s^{-1} \begin{pmatrix} U_s & O \\ O & U_s \end{pmatrix} \\
&= \sum_{s=N+1}^{N+I_1} \begin{pmatrix} U_s^\tau & O \\ O & U_s^\tau \end{pmatrix} \begin{pmatrix} A_1 & B_1 \\ B_1^\tau & C_1 \end{pmatrix} \begin{pmatrix} U_s & O \\ O & U_s \end{pmatrix} \\
&= \sum_{s=N+1}^{N+I_1} \begin{pmatrix} U_s^\tau A_1 U_s & U_s^\tau B_1 U_s \\ U_s^\tau B_1^\tau U_s & U_s^\tau C_1 U_s \end{pmatrix}, \\
K_2 &= \sum_{s=N+I_1+1}^{N+I_1+I_2} \begin{pmatrix} U_s^\tau & O \\ O & U_s^\tau \end{pmatrix} \Sigma_s^{-1} \begin{pmatrix} U_s & O \\ O & U_s \end{pmatrix} \\
&= \sum_{s=N+I_1+1}^{N+I_1+I_2} \begin{pmatrix} U_s^\tau & O \\ O & U_s^\tau \end{pmatrix} \begin{pmatrix} A_2 & B_2 \\ B_2^\tau & C_2 \end{pmatrix} \begin{pmatrix} U_s & O \\ O & U_s \end{pmatrix}. \\
&= \sum_{s=N+I_1+1}^{N+I_1+I_2} \begin{pmatrix} U_s^\tau A_2 U_s & U_s^\tau B_2 U_s \\ U_s^\tau B_2^\tau U_s & U_s^\tau C_2 U_s \end{pmatrix}
\end{aligned}$$

Then the non-centrality parameter λ_{ad} can be calculated by

$$\begin{aligned}
\lambda_{ad} &= (HY)^\tau \left[H \left[\sum_{s=1}^{N+I_1+I_2} \begin{pmatrix} U_s^\tau & O \\ O & U_s^\tau \end{pmatrix} \Sigma_s^{-1} \begin{pmatrix} U_s & O \\ O & U_s \end{pmatrix} \right]^{-1} H^\tau \right]^{-1} (HY) \\
&= (HY)^\tau [H(K_0 + K_1 + K_2)^{-1} H^\tau]^{-1} (HY)
\end{aligned}$$

Notice that similar approximations as (A3)–(A5) can be utilized to calculate K_1 and K_2 . The non-centrality parameter λ_a of test statistic $F_{AB,a}$ can be calculated, accordingly.