

Combined Linkage and Association Mapping of Quantitative Trait Loci with Missing Completely at Random Genotype Data

Ruzong Fan · Lian Liu · Jeesun Jung ·
Ming Zhong

Received: 26 June 2006 / Accepted: 31 January 2008 / Published online: 27 February 2008
© Springer Science+Business Media, LLC 2008

Abstract In genetics study, the genotypes or phenotypes can be missing due to various reasons. In this paper, the impact of missing genotypes is investigated for high resolution combined linkage and association mapping of quantitative trait loci (QTL). We assume that the genotype data are missing completely at random (MCAR). Two regression models, “genotype effect model” and “additive effect model”, are proposed to model the association between the markers and the trait locus. If the marker genotype is not missing, the model is exactly the same as those of our previous study, i.e., the number of genotype or allele is used as weight to model the effect of the genotype or allele in single marker case. If the marker genotype is missing, the expected number of genotype or allele is used as weight to model the effect of the genotype or allele. By analytical formulae, we show that the “genotype effect model” can be used to model the additive and dominance effects simultaneously, and the “additive effect model” can only be used to model the additive effect. 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 calculate power and to compare power, which show that

the power of the F -tests is reduced due to the missingness. By simulation study, we show that the two models have reasonable type I error rates for a dataset of moderate sample size. However, the type I error rates can be very slightly inflated if all individuals with missing genotypes are removed from analysis. Hence, the proposed method can help to get correct type I error rates although it does not improve power. As a practical example, the method is applied to analyze the angiotensin-1 converting enzyme (ACE) data.

Keywords Missing genotype · Linkage disequilibrium mapping · QTL

Introduction

In disease gene mapping, linkage analysis and linkage disequilibrium (LD) mapping or association study can be carried out. Linkage analysis is based on pedigree data, and association study can be based on either population data or pedigree data or combinations of population and pedigree data. Linkage analysis is robust to population structure, and is appropriate for low resolution genetic mapping to localize trait loci to broad chromosome regions within a few cM. In contrast to linkage analysis, association study for genetic traits is useful in high resolution genetic mapping, i.e., fine disease gene mapping; however, association study is prone to population structure and the rate of false positives can be high. In recent years, there has been great interest in carrying out combined linkage and association mapping of complex genetic traits (Li et al. 2005; Xiong and Jin 2000). The combined analysis of linkage and association can take the advantage of the robustness of

Edited by Pak Sham.

R. Fan (✉) · L. Liu · M. Zhong
Department of Statistics, Texas A&M University, 447 Blocker
Building, 3143 TAMUS, College Station, TX 77843, USA
e-mail: rfan@stat.tamu.edu

J. Jung
Department of Medical and Molecular Genetics, Indiana
University, School of Medicine, 975 West Walnut Street, IB 130,
Indianapolis, IN 46202, USA

linkage analysis, and the high resolution of association study. In addition, it may minimize the limits of each.

For many complex traits, such as diabetes, depression, alcoholism and hypertension, quantitative phenotypes can be very informative. Hence, it is of importance to develop statistical methods for mapping of quantitative trait loci/locus (QTL). There has been a long history in the research of linkage mapping of QTL (Almasy and Blangero 1998; Feingold 2002; Fulker et al. 1995; Goldgar 1990; Haseman and Elston 1972; Pratt et al. 2000). Moreover, variance component models have been proposed for combined linkage and association mapping of QTL (Abecasis et al. 2000a, 2000b; Allison 2001; Almasy et al. 1999; Boerwinkle et al. 1986; George et al. 1999; Fulker et al. 1999; Sham et al. 2000).

Based on combinations of population and pedigree data, we have developed variance component models for combined linkage and association mapping of QTL for complex diseases (Fan and Jung 2003; Fan et al. 2006, 2005; Fan and Xiong 2003; Jung et al. 2005). However, there is limited research to investigate the impact of missing data on our models. In genetics study, the genotypes or phenotypes can be missing due to various reasons. It is important to develop models which account for missing data. In this article, we are going to develop models which account for missing data, and to investigate the impact of missing genotypes on combined linkage and association mapping of QTL. Two regression models, “genotype effect model” and “additive effect model”, are proposed to model the association between the markers and the trait locus when there are missing genotypes. Based on the two models, *F*-test statistics or likelihood ratio test statistics can be used to test association between the QTL and markers. We will investigate the impact of missing data on the models, under an assumption that the genotype data are missing completely at random (MCAR). Simulation study will be performed to evaluate the robustness of the proposed models, and to make comparison with models which exclude the individuals with missing genotypes from analysis. In addition, the method will be applied to analyze the angiotensin-1 converting enzyme data (Farrall et al. 1999; Keavney et al. 1998).

Method

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. In a region of the QTL Q , suppose that one marker or multiple markers are typed for a sample; and the sample may include multi-generation pedigrees of any sizes and any types of relatives, nuclear families, sib-ships and unrelated

individuals. However, the marker information may be missing for some individuals of the sample at some markers. That is to say, some genotype information may not be available for some individuals. In multiple marker case, the genotypes of an individual may be missing at some markers and may be available at the other markers. In this paper, we assume that the genotype data are MCAR (Little and Rubin 2002), i.e., the missingness does not depend on the genotype and phenotype data. In the following, we first present the models by one marker, and extend to use two/multiple markers in analysis.

Log-likelihood and mapping strategy

Log-likelihood

Suppose that the data are composed of a combination of N unrelated individuals and I independent families. The I families can be multi-generation pedigrees, nuclear families, sib-ships, or their combinations. In the following, we first define log-likelihood of the data. Let us list the log-likelihoods of the N individuals by L_1, \dots, L_N , and the log-likelihoods of the I families by L_{N+1}, \dots, L_{N+I} . The overall log-likelihood is $L = \sum_{i=1}^{N+I} L_i$. In the i -th family, let t_i be the total number of individuals who are listed as $j = 1, 2, \dots, t_i$; each individual j is preceded by all his/her ancestors. Let us denote the quantitative traits of i -th family by a vector $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{it_i})^\tau$. Here, the superscript τ denotes vector/matrix transpose. In addition, assume that marker genotypes are either available or missing for a family member. The log-likelihood is defined by $L_i = -\frac{t_i}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_i| - \frac{1}{2} (\mathbf{y}_i - X_i \phi)^\tau \Sigma_i^{-1} (\mathbf{y}_i - X_i \phi)$, under the assumption of multi-variate normality. In the log-likelihood, Σ_i is the variance–covariance matrix which is defined in the paragraph below; X_i is a model matrix defined in Subsections below, and ϕ is a column vector of regression coefficients related to the model matrix.

Σ_i is a $t_i \times t_i$ matrix defined as

$$\Sigma_i = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1t_i} \\ \rho_{12} & 1 & \cdots & \rho_{2t_i} \\ \vdots & \vdots & \cdots & \vdots \\ \rho_{1t_i} & \rho_{2t_i} & \cdots & 1 \end{pmatrix} \sigma^2,$$

where $\sigma^2 = \sigma_g^2 + \sigma_{Ga}^2 + \sigma_e^2$, σ_g^2 is variance explained by the putative QTL Q , σ_{Ga}^2 is polygenic additive variance, and σ_e^2 is error variance. The genetic variance $\sigma_g^2 = \sigma_{ga}^2 + \sigma_{gd}^2$ is decomposed into additive and dominance components. As in the traditional quantitative genetics, let a be the effect of genotype Q_1Q_1 , d be the effect of genotype Q_1Q_2 , and $-a$ be the effect of genotype Q_2Q_2 (Falconer and Mackay 1996). Let $\alpha_Q = a + (q_2 - q_1)d$ be the average effect of gene substitution, and $\delta_Q = 2d$ be the dominance

deviation. In addition, let $\mu = a(q_1 - q_2) + 2dq_1q_2$ be the aggregate effect of the QTL on the trait mean in the population. It is well known that the additive variance $\sigma_{ga}^2 = 2q_1q_2\alpha_Q^2$ and the dominance variance $\sigma_{gd}^2 = (q_1q_2)^2\delta_Q^2$. $\rho_{jk} = (\pi_{jkQ}\sigma_{ga}^2 + \Delta_{jkQ}\sigma_{gd}^2 + 2\Phi_{jk}\sigma_{Ga}^2)/\sigma^2$ is correlation between the j -th individual and the k -th individual of the family, where π_{jkQ} is the proportion of alleles shared identically by descent (IBD) at QTL Q by the j -th and the k -th individuals, Δ_{jkQ} is the probability that both alleles at QTL Q shared by the j -th and the k -th individuals are IBD, and Φ_{jk} is the kinship coefficient of individuals j and k . π_{jkQ} and Δ_{jkQ} are usually estimated by marker information (Amos 1994; Amos and Elston 1989). The recombination fractions between the genotyped markers and the unobserved QTL are contained in the estimations of π_{jkQ} and Δ_{jkQ} . Hence, linkage information is modeled in variance–covariance matrix.

For the N unrelated individuals, the log-likelihoods are $L_i = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_{i1} - X_i\phi)^T(y_{i1} - X_i\phi)$, $i = 1, \dots, N$. Here, y_{i1} is the trait value of the i -th individual. It can be seen that no linkage information is contained in the log-likelihoods of the N unrelated individuals. The linkage is modeled solely in variance–covariance matrices of the I families. Therefore, family data can be used for linkage analysis. In subsections below, we will show that LD information is contained in the regression coefficients ϕ . Thus, LD information is contained in both population data and family data.

Mapping strategy

The linkage analysis can usually locate the trait locus in a broad chromosome region within a few cM or even around 15 cM. Linkage analysis is less sensitive to population structures of subdivisions and admixtures, although its resolution can be low. In contrast, the LD analysis has an advantage for high resolution mapping of trait locus, but can be prone to false positives. In practice, linkage analysis can be performed as the first step of analysis to obtain suggestive linkage information. With evidence of suggestive linkage from linkage study, population data and family data can be combined together for LD analysis for fine mapping of QTL. Using this strategy to map the trait locus, one may take advantage of both linkage analysis and LD mapping, and be more likely to avoid the spurious association.

Mixed effect models by one marker

In a region of the QTL Q , suppose that one marker A is typed, which may be di-allelic or multi-allelic. Let us

denote the alleles of marker A by A_1, \dots, A_m , where m is the number of alleles. Suppose that the marker A is in Hardy–Weinberg equilibrium (HWE). Let the frequency of A_g be P_{A_g} , $g = 1, 2, \dots, m$. Consider the j -th pedigree member of the i -th family with trait value y_{ij} and genotype G_{Aij} . If the genotype G_{Aij} is not missing, there are $J_A = m(m + 1)/2$ possibilities for G_{Aij} , which can be listed as $A_1A_1, \dots, A_mA_m, A_1A_2, \dots, A_1A_m, \dots, A_{m-1}A_m$. In practice, the genotype G_{Aij} can be missing. Therefore, the genotype G_{Aij} can be one of the J_A genotypes if it is not missing and can be missing. If G_{Aij} is missing, we denote it by $G_{Aij} = ?$; and if G_{Aij} is not missing, we denote it by $G_{Aij} \neq ?$, i.e., the complementary set of $G_{Aij} = ?$. Let us denote the probability that the genotype G_{Aij} is missing by ε_A , i.e., $P(G_{Aij} = ?) = \varepsilon_A$. Notice that $P(G_{Aij} \neq ?) = 1 - \varepsilon_A$. In addition, let $P(G_{Aij} = A_gA_h|G_{Aij} = ?)$ or $P(G_{Aij} = A_gA_h|G_{Aij} \neq ?)$ be the conditional probability of genotype A_gA_h given $G_{Aij} = ?$ or $G_{Aij} \neq ?$. Since the missing mechanism is missing completely at random, the probability is

$$P(G_{Aij} = A_gA_h|G_{Aij} = ?) = P(G_{Aij} = A_gA_h|G_{Aij} \neq ?) = P(A_gA_h) = \begin{cases} P_{A_g}^2 & \text{if } g = h \\ 2P_{A_g}P_{A_h} & \text{if } g \neq h \end{cases}$$

Genotype effect model

For the listed J_A genotypes, let $\beta_{11}, \dots, \beta_{mm}$, $\beta_{12}, \dots, \beta_{1m}, \dots, \beta_{m-1,m}$ be the corresponding effects on quantitative trait. The “genotype effect model” can be written as

$$y_{ij} = w_{ij}\gamma + \sum_{1 \leq g \leq h \leq m} [1_{(G_{Aij}=A_gA_h)} + P(G_{Aij} = A_gA_h|G_{Aij} = ?)1_{(G_{Aij}=?)}] \beta_{gh} + H_{ij} + e_{ij} = w_{ij}\gamma + \sum_{1 \leq g \leq h \leq m} [1_{(G_{Aij}=A_gA_h)} + P(A_gA_h)1_{(G_{Aij}=?)}] \beta_{gh} + H_{ij} + e_{ij}, \tag{1}$$

where

$$1_{(G_{Aij}=A_gA_h)} = \begin{cases} 1 & \text{if } G_{Aij} = A_gA_h \\ 0 & \text{else} \end{cases}$$

and

$$1_{(G_{Aij}=?) } = \begin{cases} 1 & \text{if } G_{Aij} = ? \\ 0 & \text{else} \end{cases}$$

are indicator functions of genotypes A_gA_h and $(G_{Aij} = ?)$, w_{ij} is a row vector of co-variates such as sex and age, γ is a column vector of regression coefficients of w_{ij} , H_{ij} is polygenic additive effect, and e_{ij} is the error term. Assume that H_{ij} is random normal $N(0, \sigma_{Ga}^2)$, and e_{ij} is normal $N(0, \sigma_e^2)$. In addition, γ and β_{gh} are fixed effects. Hence, model (1) is a mixed effect model (Pinheiro and

Bates 2000). The contribution of polygenic additive effect to the variance–covariance matrix Σ_i is from the terms which contain $\sigma_{G_a}^2$.

Now let us show that model (1) extends the “genotype effect model” in Fan et al. (2006). If the genotype is not missing and $G_{Aij} = A_g A_h$, the model (1) becomes $y_{ij} = w_{ij}\gamma + \beta_{gh} + H_{ij} + e_{ij}$, which is similar to “genotype effect model” (1), Fan et al. (2006). Note that the polygenic additive effect H_{ij} is not modeled in Fan et al. (2006). Since only population data are used in Fan et al. (2006), polygenic effect is not modeled to avoid redundancy and the models therein are fixed effect models. In the current paper, the polygenic effect is modeled as a random effect and so the models are mixed effect models. Since we use both population data and family data, the polygenic additive effect is assumed to be estimable.

If $G_{Aij} = ?$ is missing, the model (1) is $y_{ij} = w_{ij}\gamma + \sum_{1 \leq g \leq h \leq m} P(A_g A_h) \beta_{gh} + H_{ij} + e_{ij}$, which uses the conditional probability $P(G_{Aij} = A_g A_h | G_{Aij} = ?) = P(A_g A_h)$ as the weight to model the effect β_{gh} of genotype $A_g A_h$. Let us denote

$$x_{Aij}^{(gh)} = 1_{(G_{Aij}=A_g A_h)} + P(A_g A_h) 1_{(G_{Aij}=?)}, \tag{2}$$

which can be thought as the expected number of genotype $A_g A_h$ given observed genotype G_{Aij} at marker A. The “genotype effect model” (1) can be re-written as $y_{ij} = w_{ij}\gamma + \sum_{1 \leq g \leq h \leq m} x_{Aij}^{(gh)} \beta_{gh} + H_{ij} + e_{ij}$. Here, we add the polygenic effect to the model proposed in Fan et al. (2006). Based on “genotype effect model” (1), we may get the model matrix X_i and regression coefficient vector ϕ as follows: $\phi = (\gamma^\tau, \beta_{11}, \dots, \beta_{mm}, \beta_{12}, \dots, \beta_{1m}, \dots, \beta_{m-1,m})^\tau$ and $X_i = (X_{Ail}, \dots, X_{Ait_i})^\tau$, where $X_{Aij} = (w_{ij}, x_{Aij}^{(11)}, \dots, x_{Aij}^{(mm)}, x_{Aij}^{(12)}, \dots, x_{Aij}^{(1m)}, \dots, x_{Aij}^{(m-1,m)})^\tau, j = 1, 2, \dots, t_i$.

Additive effect model

Assume that the genetic effect is additive, i.e., $\beta_{gh} = \alpha_g + \alpha_h$, where α_g is effect of allele A_g . Then model (1) becomes an “additive effect model”, which can be written as

$$y_{ij} = w_{ij}\gamma + \sum_{1 \leq g \leq h \leq m} [1_{(G_{Aij}=A_g A_h)} + P(A_g A_h) 1_{(G_{Aij}=?)}] (\alpha_g + \alpha_h) + H_{ij} + e_{ij}. \tag{3}$$

If the genotype is not missing and $G_{Aij} = A_g A_h$, the model (3) becomes $y_{ij} = w_{ij}\gamma + \alpha_g + \alpha_h + H_{ij} + e_{ij}$, which is similar to “additive effect model”, Fan et al. (2006). Therefore, model (3) extends the “additive effect model”, Fan et al. (2006). If $G_{Aij} = ?$ is missing, the model (3) is

$$\begin{aligned} y_{ij} &= w_{ij}\gamma + \sum_{1 \leq g \leq h \leq m} P(A_g A_h) (\alpha_g + \alpha_h) + H_{ij} + e_{ij} \\ &= w_{ij}\gamma + \sum_{g=1}^m P_{A_g} \alpha_g + \sum_{h=1}^m P_{A_h} \alpha_h + H_{ij} + e_{ij} \\ &= w_{ij}\gamma + \sum_{g=1}^m 2P_{A_g} \alpha_g + H_{ij} + e_{ij}. \end{aligned}$$

Note that $2 P_{A_g} = 2P(G_{Aij} = A_g A_g | G_{Aij} = ?) + \sum_{h \neq g} P(G_{Aij} = A_g A_h | G_{Aij} = ?)$ is the expected number of alleles A_g given $G_{Aij} = ?$, which is the weight to model the effect α_g of allele A_g . Let us denote

$$x_{Aij}^{(g)} = 2 \cdot 1_{(G_{Aij}=A_g A_g)} + \sum_{h \neq g} 1_{(G_{Aij}=A_g A_h)} + 2P_{A_g} 1_{(G_{Aij}=?)}, \tag{4}$$

which is the expected number of alleles A_g given observed genotype G_{Aij} at marker A. From the discussion above, we may re-write the “additive effect model” (3) as $y_{ij} = w_{ij}\gamma + \sum_{g=1}^m x_{Aij}^{(g)} \alpha_g + H_{ij} + e_{ij}$. Again, we add the polygenic effect to the model proposed in Fan et al. (2006). Based on the “additive effect model” (3), we may get the model matrix X_i and regression coefficient vector ϕ as follows: $\phi = (\gamma^\tau, \alpha_1, \dots, \alpha_m)^\tau$ and $X_i = (Z_{Ail}, \dots, Z_{Ait_i})^\tau$, where $Z_{Aij} = (w_{ij}, x_{Aij}^{(1)}, \dots, x_{Aij}^{(m)})^\tau, j = 1, 2, \dots, t_i$.

Property of model coefficients

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, $P(Q_1 A_g)$ is the frequency of haplotype $Q_1 A_g$. In Appendix A, the regression coefficients of “genotype effect model” (1) are calculated as

$$\begin{aligned} \beta_{gh} &= \mu + \alpha_Q [D_{A_g Q} / P_{A_g} + D_{A_h Q} / P_{A_h}] \\ &\quad - \delta_Q D_{A_g Q} D_{A_h Q} / [P_{A_g} P_{A_h}]. \end{aligned} \tag{5}$$

In Appendix B, we show that the regression coefficients of “additive effect model” (3) are given by

$$\alpha_g = \mu/2 + \alpha_Q D_{A_g Q} / P_{A_g}. \tag{6}$$

Notice that relations (5) and (6) are exactly the same as those of Fan et al. (2006). Assume that the additive effect is significantly present, but the dominance effect is not significantly present, i.e., $\alpha_Q \neq 0$ but $\delta_Q = 0$. To test association between the marker A and the QTL Q, one may test hypotheses $H_{a0} : \alpha_1 = \dots = \alpha_m$ versus H_{a1} : at least two α_g 's are not equal. On the other hand, assume that both additive and dominance effects are significantly present at the putative QTL Q, i.e., $\alpha_Q \neq 0$ and $\delta_Q \neq 0$. To test association between the marker A and the QTL Q, one may test hypotheses $H_{ad0} : \beta_{11} = \dots = \beta_{mm} = \beta_{12} = \dots = \beta_{1m} = \dots = \beta_{m-1,m}$ versus H_{ad1} : at least two β_{gh} are not equal. One may test H_{a0} and H_{ad0} by likelihood ratio test

χ^2 -statistics. In the following, we will introduce F -test statistics and the related properties. By large sample theory, the F -test statistics and the likelihood ratio test statistics are close to each other (Graybill, 1976, pp. 187–188).

F-tests and non-centrality parameter approximations

Assume that there are no covariates. Let us denote $X = (X_1^\tau, \dots, X_N^\tau, X_{N+1}^\tau, \dots, X_{N+I}^\tau)^\tau$, $Y = (y_{11}, \dots, y_{N1}, y_{N+1,1}, \dots, y_{N+I,1})^\tau$, $H = (H_{11}, \dots, H_{N1}, H_{N+1,1}, \dots, H_{N+I,1})^\tau$, and $e = (e_{11}, \dots, e_{N1}, e_{N+1,1}, \dots, e_{N+I,1})^\tau$. Here, $H_i = (H_{i1}, \dots, H_{ii})^\tau$ and $e_i = (e_{i1}, \dots, e_{ii})^\tau$, $i = N + 1, \dots, N + I$. Then “genotype effect model” (1) or “additive effect model” (3) can be expressed as $Y = X\phi + H + e$. Let $\hat{\Sigma}_i$ and $\hat{\phi}$ be the maximum likelihood estimations of Σ_i and ϕ . By standard regression theory, the coefficients can be estimated by $\hat{\phi} = [\sum_{i=1}^{N+I} X_i^\tau \hat{\Sigma}_i^{-1} X_i]^{-1} \sum_{i=1}^{N+I} X_i^\tau \hat{\Sigma}_i^{-1} Y_i$.

For the “genotype effect model” (1), denote regression coefficient vector $\eta = (\beta_{11}, \dots, \beta_{mm}, \beta_{12}, \dots, \beta_{1m}, \dots, \beta_{m-1,m})^\tau$. Let us define a $(J_A - 1) \times J_A$ matrix by

$$T = \begin{pmatrix} 1 & -1 & 0 & 0 & \dots & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & \dots & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \dots & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 & \dots & 0 & 0 & -1 \end{pmatrix}_{(J_A-1) \times J_A}$$

Then, $(T\eta)^\tau = (\beta_{11} - \beta_{22}, \dots, \beta_{11} - \beta_{mm}, \beta_{11} - \beta_{12}, \dots, \beta_{11} - \beta_{1m}, \dots, \beta_{11} - \beta_{m-1,m})^\tau$. Hence, the hypothesis H_{ad0} is equivalent to $T\eta = (0, \dots, 0)^\tau$. By Graybill (1976), Chapter 6, the test statistic of a hypothesis H_{ad0} is non-central $F(J_A - 1, \sum_{i=1}^{N+I} t_i - J_A)$ defined by

$$F_{m,ad} = \frac{(T\hat{\eta})^\tau [T(X^\tau \hat{\Sigma}^{-1} X)^{-1} T^\tau]^{-1} (T\hat{\eta})}{Y^\tau [\hat{\Sigma}^{-1} - \hat{\Sigma}^{-1} X (X^\tau \hat{\Sigma}^{-1} X)^{-1} X^\tau \hat{\Sigma}^{-1}] Y} \frac{\sum_{i=1}^{N+I} t_i - J_A}{J_A - 1}$$

where $\Sigma = \text{diag}(\Sigma_1, \dots, \Sigma_{N+I})$ is the overall variance–covariance matrix with matrices Σ_i on the diagonal, and $\hat{\Sigma}$ is its maximum likelihood estimation. The non-centrality parameter of above F -statistic is $\lambda_{m,ad} = (T\eta)^\tau [T(X^\tau \Sigma^{-1} X)^{-1} T^\tau]^{-1} (T\eta)$.

Assume that the dataset is a population sample, i.e., $I = 0$. Under the assumption of large sample size N , we show in Appendix C the following approximation

$$\lambda_{m,ad} \approx \frac{N(1 - \varepsilon_A)}{\sigma^2} \left[\sigma_{ga}^2 R_{AQ}^2 + \sigma_{gd}^2 R_{AQ}^4 \right], \tag{7}$$

where R_{AQ}^2 is a general measure of the degree of LD between marker A and the QTL Q defined by $R_{AQ}^2 = \sum_{g=1}^m \sum_{s=1}^2 \frac{[P(Q_s A_g) - P_{A_g} q_s]^2}{[P_{A_g} q_s]}$ (Hedrick 1987; Sham et al. 2000). Notice that R_{AQ}^2 is the χ^2 statistic of the

$m \times 2$ table of haplotype frequencies of the marker A and trait locus Q . Approximation (7) shows that the non-centrality parameter $\lambda_{m,ad}$ is reduced by a factor of $1 - \varepsilon_A$. If there is no missing genotype data, i.e., $\varepsilon_A = 0$, approximation (7) is exactly the same as that of the “genotype effect model” in Fan et al. (2006). In the presence of missing data, the model developed extends our previous work. In addition, $\lambda_{m,ad}$ is reduced by a factor of R_{AQ}^2 for additive variance σ_{ga}^2 and a factor of R_{AQ}^4 for dominance variance σ_{gd}^2 .

For the “additive effect model” (3), denote $\psi = (\alpha_1, \dots, \alpha_m)^\tau$. Let K be a $(m-1) \times m$ matrix defined by

$$K = \begin{pmatrix} 1 & -1 & 0 & 0 & \dots & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & \dots & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \dots & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 & \dots & 0 & 0 & -1 \end{pmatrix}_{(m-1) \times m}$$

Then, $(K\psi)^\tau = (\alpha_1 - \alpha_2, \dots, \alpha_1 - \alpha_m)$. Hence, the hypothesis H_{a0} is equivalent to $K\psi = (0, \dots, 0)^\tau$. By Graybill (1976), Chapter 6, the test statistic of the hypothesis H_{a0} is non-central $F(m - 1, \sum_{i=1}^{N+I} t_i - m)$ defined by

$$F_{m,a} = \frac{(K\hat{\psi})^\tau [K(X^\tau \hat{\Sigma}^{-1} X)^{-1} K^\tau]^{-1} (K\hat{\psi})}{Y^\tau [\hat{\Sigma}^{-1} - \hat{\Sigma}^{-1} X (X^\tau \hat{\Sigma}^{-1} X)^{-1} X^\tau \hat{\Sigma}^{-1}] Y} \frac{\sum_{i=1}^{N+I} t_i - m}{m - 1}$$

Here, the model matrix X is built from the “additive effect model” (3). The non-centrality parameter of above F -statistic is $\lambda_{m,a} = (K\psi)^\tau [K(X^\tau \Sigma^{-1} X)^{-1} K^\tau]^{-1} (K\psi)$. Assume that the dataset is a population sample, i.e., $I = 0$. Under an assumption of large sample size N , we show in Appendix D the following approximation

$$\lambda_{m,a} = \frac{1}{\sigma^2} (K\psi)^\tau [K(X^\tau X)^{-1} K^\tau]^{-1} (K\psi) \approx \frac{N(1 - \varepsilon_A) \sigma_{ga}^2}{\sigma^2} R_{AQ}^2. \tag{8}$$

Again, the approximation (8) shows that the non-centrality parameter $\lambda_{m,a}$ is reduced by a factor of $1 - \varepsilon_A$. If there is no missing genotype data, i.e., $\varepsilon_A = 0$, approximation (8) is exactly the same as that of the “additive effect model” in Fan et al. (2006). Besides, $\lambda_{m,a}$ is reduced by a factor of R_{AQ}^2 for additive variance. The dominance variance is not present in $\lambda_{m,a}$.

Mixed effect models by two markers

In addition to marker A , assume that a second marker B is typed, which has n alleles denoted by B_1, \dots, B_n . Suppose that the marker B is also in HWE. 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 $B_1B_1, \dots, B_nB_n, B_1B_2, \dots, B_1B_n, \dots, B_{n-1}B_n$. Let y_{ij} be the trait value of the j -th pedigree member of the i -th family with genotype G_{Aij} at marker A and genotype G_{Bij} at marker B . Such as G_{Aij} discussed above, G_{Bij} can be missing. If G_{Bij} is missing, we denote it as $G_{Bij} = ?$; and if G_{Bij} is not missing, we denote it by $G_{Bij} \neq ?$. Let us denote the probability that the genotype G_{Bij} is missing by ε_B , i.e., $P(G_{Bij} = ?) = \varepsilon_B$. Notice that $P(G_{Bij} \neq ?) = 1 - \varepsilon_B$. In addition, let $P(G_{Bij} = B_kB_l | G_{Bij} = ?)$ or $P(G_{Bij} = B_kB_l | G_{Bij} \neq ?)$ be the conditional probability of genotype B_kB_l given $G_{Bij} = ?$ or $G_{Bij} \neq ?$. Since the missing mechanism is MCAR, the probability

$$P(G_{Bij} = B_kB_l | G_{Bij} = ?) = P(G_{Bij} = B_kB_l | G_{Bij} \neq ?) = P(B_kB_l) = \begin{cases} P_{B_k}^2 & \text{if } k = l \\ 2P_{B_k}P_{B_l} & \text{if } k \neq l \end{cases}$$

Such as relations (4) to define $x_{Aij}^{(g)}$, let us denote the expected number of alleles B_k given the observed genotype G_{Bij} at marker B

$$x_{Bij}^{(k)} = 2 \cdot 1_{(G_{Bij}=B_kB_k)} + \sum_{l \neq k} 1_{(G_{Bij}=B_kB_l)} + 2P_{B_k} 1_{(G_{Bij}=?)}. \quad (9)$$

The ‘‘additive effect model’’ (13) of Fan et al. (2006) can be extended to

$$y_{ij} = w_{ij}\gamma + \alpha + \sum_{g=1}^{m-1} x_{Aij}^{(g)}\alpha_{A_g} + \sum_{k=1}^{n-1} x_{Bij}^{(k)}\alpha_{B_k} + H_{ij} + e_{ij}, \quad (10)$$

where w_{ij} and γ are the same as those in model (1), and α, α_{A_g} , and α_{B_k} are regression coefficients. To understand that model (10) extends model (13) of Fan et al. (2006), consider four possible cases as follows:

Case 1: both the genotypes G_{Aij} and G_{Bij} are not missing, the model (10) is similar to (13) of Fan et al. (2006). In the model (10), we model the polygenic effect, which is not modeled in Fan et al. (2006).

Case 2: both the genotypes $G_{Aij} = ?$ and $G_{Bij} = ?$ are missing, the model (10) becomes

$$y_{ij} = w_{ij}\gamma + \alpha + 2 \sum_{g=1}^{m-1} P_{A_g}\alpha_{A_g} + 2 \sum_{k=1}^{n-1} P_{B_k}\alpha_{B_k} + H_{ij} + e_{ij}.$$

Case 3: the genotype G_{Aij} is not missing and the genotype G_{Bij} is missing, the model (10) becomes

$$y_{ij} = w_{ij}\gamma + \alpha + \sum_{g=1}^{m-1} \left[2 \cdot 1_{(G_{Aij}=A_gA_g)} + \sum_{h \neq g} 1_{(G_{Aij}=A_gA_h)} \right] \alpha_{A_g} + 2 \sum_{k=1}^{n-1} P_{B_k}\alpha_{B_k} + H_{ij} + e_{ij}.$$

Case 4: the genotype G_{Aij} is missing and the genotype G_{Bij} is not missing, the model (10) becomes

$$y_{ij} = w_{ij}\gamma + \alpha + 2 \sum_{g=1}^{m-1} P_{A_g}\alpha_{A_g} + \sum_{k=1}^{n-1} \left[2 \cdot 1_{(G_{Bij}=B_kB_k)} + \sum_{l \neq k} 1_{(G_{Bij}=B_kB_l)} \right] \alpha_{B_k} + H_{ij} + e_{ij}.$$

To extend the ‘‘genotype effect model’’ (14) of Fan et al. (2006), let us denote

$$\begin{aligned} z_{Aij}^{(gh)} &= -P_{A_h}^2 1_{(G_{Aij}=A_gA_g)} + P_{A_g}P_{A_h} 1_{(G_{Aij}=A_gA_h)} - P_{A_g}^2 1_{(G_{Aij}=A_hA_h)}, \\ z_{Bij}^{(kl)} &= -P_{B_l}^2 1_{(G_{Bij}=B_kB_k)} + P_{B_k}P_{B_l} 1_{(G_{Bij}=B_kB_l)} - P_{B_k}^2 1_{(G_{Bij}=B_lB_l)}. \end{aligned} \quad (11)$$

If the genotypes G_{Aij} and G_{Bij} are not missing, the variables $x_{Aij}^{(g)}, x_{Bij}^{(k)}, z_{Aij}^{(gh)}$ and $z_{Bij}^{(kl)}$ are the same as those defined in Fan et al. (2006). If the genotype G_{Aij} or G_{Bij} is missing, $x_{Aij}^{(g)}$ or $x_{Bij}^{(k)}$ is simply the expected number $2P_{A_g}$ or $2P_{B_k}$ of alleles A_g or B_k , and $z_{Aij}^{(gh)}$ or $z_{Bij}^{(kl)}$ is 0. The reason that $z_{Aij}^{(gh)}$ is 0 on $G_{Aij} = ?$ is as follows: $-P_{A_h}^2 P(G_{Aij} = A_gA_g | G_{Aij} = ?) + P_{A_g}P_{A_h} P(G_{Aij} = A_gA_h | G_{Aij} = ?) - P_{A_g}^2 P(G_{Aij} = A_hA_h | G_{Aij} = ?) = 0$; the same reasoning applies to $z_{Bij}^{(kl)}$. The ‘‘genotype effect model’’ (14) of Fan et al. (2006) can be extended to

$$y_{ij} = w_{ij}\gamma + \alpha + \sum_{g=1}^{m-1} x_{Aij}^{(g)}\alpha_{A_g} + \sum_{k=1}^{n-1} x_{Bij}^{(k)}\alpha_{B_k} + \sum_{1 \leq g < h \leq m} z_{Aij}^{(gh)}\delta_{Agh} + \sum_{1 \leq k < l \leq n} z_{Bij}^{(kl)}\delta_{Bkl} + H_{ij} + e_{ij}, \quad (12)$$

where δ_{Agh} and δ_{Bkl} are regression coefficients of variables $z_{Aij}^{(gh)}$ and $z_{Bij}^{(kl)}$, respectively; other terms are the same as those of model (10). It can be shown that model (12) extends model (14) of Fan et al. (2006).

Denote $X_{Aij} = (x_{Aij}^{(1)}, \dots, x_{Aij}^{(m-1)})^\tau$, $X_{Bij} = (x_{Bij}^{(1)}, \dots, x_{Bij}^{(n-1)})^\tau$, and $X_{A \cup B}^{(ij)} = (X_{Aij}^\tau, X_{Bij}^\tau)^\tau$. Let us denote the additive variance–covariance matrix of the indicator variables $x_{Aij}^{(g)}, x_{Bij}^{(k)}$ by $V_A = \text{Cov}(X_{A \cup B}^{(ij)}, X_{A \cup B}^{(ij)}) = E(X_{A \cup B}^{(ij)}(X_{A \cup B}^{(ij)})^\tau) - E X_{A \cup B}^{(ij)} E(X_{A \cup B}^{(ij)})^\tau$. Similarly, let $Z_{Aij} = (z_{Aij}^{(12)}, \dots, z_{Aij}^{(1m)}, z_{Aij}^{(23)}, \dots, z_{Aij}^{(2m)}, \dots, z_{Aij}^{(m-1,m)})^\tau$, $Z_{Bij} = (z_{Bij}^{(12)}, \dots, z_{Bij}^{(1n)}, z_{Bij}^{(23)}, \dots, z_{Bij}^{(2n)}, \dots, z_{Bij}^{(n-1,n)})^\tau$, and $Z_{A \cup B}^{(ij)} = (Z_{Aij}^\tau, Z_{Bij}^\tau)^\tau$. Let us denote the dominance variance–covariance matrix of the indicator variables $z_{Aij}^{(gh)}, z_{Bij}^{(kl)}$ by $V_D = \text{Cov}(Z_{A \cup B}^{(ij)}, Z_{A \cup B}^{(ij)})$. The elements of matrices V_A and V_D are provided in Appendix E.

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 . In Appendix E, we show that the regression coefficients of models (10) and (12) are

$$\begin{pmatrix} \alpha_{A1} \\ \vdots \\ \alpha_{A(m-1)} \\ \alpha_{B1} \\ \vdots \\ \alpha_{B(n-1)} \end{pmatrix} = (V_A/2)^{-1} \begin{pmatrix} D_{A_1Q}(1-\varepsilon_A) \\ \vdots \\ D_{A_{m-1}Q}(1-\varepsilon_A) \\ D_{B_1Q}(1-\varepsilon_B) \\ \vdots \\ D_{B_{n-1}Q}(1-\varepsilon_B) \end{pmatrix} \alpha_Q,$$

$$\begin{pmatrix} \delta_{A12} \\ \vdots \\ \delta_{A(m-1)m} \\ \delta_{B12} \\ \vdots \\ \delta_{B(n-1)n} \end{pmatrix} = V_D^{-1} \begin{pmatrix} [P_{A_2}D_{A_1Q} - P_{A_1}D_{A_2Q}]^2(1-\varepsilon_A) \\ \vdots \\ [P_{A_{m-1}}D_{A_mQ} - P_{A_m}D_{A_{m-1}Q}]^2(1-\varepsilon_A) \\ [P_{B_2}D_{B_1Q} - P_{B_1}D_{B_2Q}]^2(1-\varepsilon_B) \\ \vdots \\ [P_{B_{n-1}}D_{B_nQ} - P_{B_n}D_{B_{n-1}Q}]^2(1-\varepsilon_B) \end{pmatrix} \delta_Q. \tag{13}$$

Equations (13) show that the parameters of LD (i.e., D_{A_gQ} and D_{B_kQ}) and gene effects (i.e., α_Q and δ_Q) are contained in the regression coefficients. Models (10) and (12) simultaneously take care of the LD and the effects of the putative trait locus Q . The gene substitution effect α_Q is contained only in α_{Ag} , α_{Bk} ; and the dominance effect δ_Q is contained only in δ_{Agh} , δ_{Bkl} . Therefore, V_A is called an additive variance–covariance matrix; and V_D is called a dominance variance–covariance matrix. The model (12) orthogonally decomposes genetic effect into summation of additive and dominance effects.

Based on equations (13), we may use models (10) and (12) to test the association between the trait locus Q and the two markers A and B . Assume that the additive genetic effect is significantly present, but the dominance genetic effect is not significantly present, i.e., $\alpha_Q \neq 0$ but $\delta_Q = 0$. To test association between the markers A & B and the QTL Q , one may test hypotheses $H_{ABa0} : \alpha_{A1} = \dots = \alpha_{A(m-1)} = \alpha_{B1} = \dots = \alpha_{B(n-1)} = 0$ versus $H_{ABa1} : \text{at least one } \alpha_{Ag}, \alpha_{Bk} \text{'s is not equal to 0}$. On the other hand, assume that both additive and dominance genetic effects are significantly present at the putative QTL Q , i.e., $\alpha_Q \neq 0$ and $\delta_Q \neq 0$. To test association between the markers A & B and the QTL Q , one may test hypotheses $H_{ABad0} : \alpha_{A1} = \dots = \alpha_{A(m-1)} = \alpha_{B1} = \dots = \alpha_{B(n-1)} = \delta_{A12} = \dots = \delta_{A1m} = \dots = \delta_{A(m-1)m} = \delta_{B12} = \dots = \delta_{B1n} = \dots = \delta_{B(n-1)n} = 0$ versus $H_{ABad1} : \text{at least one } \alpha_{Ag}, \alpha_{Bk}, \delta_{Agh}, \delta_{Bkl} \text{ is not equal to 0}$. Again, one may test H_{ABa0} and H_{ABad0} by likelihood ratio test χ^2 -statistics. We will introduce F -tests in the following.

Regression models and F-tests

Based on regression (12), one may construct an F -test statistic $F_{AB, ad}$ to test the null hypothesis H_{ABad0} in the

same way to construct $F_{m,ad}$ or $F_{m,a}$ (Graybill 1976, Chapter 6). Under the null hypothesis of H_{ABad0} , $F_{AB,ad}$ is central $F(J_A + J_B - 2, \sum_{i=1}^{N+I} t_i - J_A - J_B + 1)$. Similarly, one may construct an F -test statistic $F_{AB,a}$ to test the null hypothesis H_{ABa0} based on the ‘‘additive effect model’’ (10). Under the null hypothesis of H_{ABa0} , $F_{AB,a}$ is central $F(m + n - 2, \sum_{i=1}^{N+I} t_i - m - n + 1)$.

Population sample and non-centrality parameter approximations

Assume that there are no covariates, and the dataset is a population sample, i.e., $I = 0$. Suppose the sample size N is large enough that the large sample theory applies. Denote $D_{AQ} = (D_{A_1Q}, \dots, D_{A_{m-1}Q})^\tau$ and $D_{BQ} = (D_{B_1Q}, \dots, D_{B_{n-1}Q})^\tau$; $\Delta_{AQ} = ([P_{A_2}D_{A_1Q} - P_{A_1}D_{A_2Q}]^2, \dots, [P_{A_{m-1}}D_{A_mQ} - P_{A_m}D_{A_{m-1}Q}]^2)^\tau$ and $\Delta_{BQ} = ([P_{B_2}D_{B_1Q} - P_{B_1}D_{B_2Q}]^2, \dots, [P_{B_{n-1}}D_{B_nQ} - P_{B_n}D_{B_{n-1}Q}]^2)^\tau$. Under the alternative hypothesis of H_{ABad1} , $F_{AB,ad}$ is non-central $F(J_A + J_B - 2, N - J_A - J_B + 1)$, and it can be shown that the corresponding non-centrality parameter is approximated by

$$\begin{aligned} \lambda_{ABad} \approx & \frac{N}{\sigma^2} \left[(D_{AQ}^\tau(1-\varepsilon_A), D_{BQ}^\tau(1-\varepsilon_B)) \right. \\ & \times (V_A/2)^{-1} \begin{pmatrix} D_{AQ}(1-\varepsilon_A) \\ D_{BQ}(1-\varepsilon_B) \end{pmatrix} \sigma_{ga}^2 / (q_1q_2) \\ & + (\Delta_{AQ}^\tau(1-\varepsilon_A), \Delta_{BQ}^\tau(1-\varepsilon_B)) V_D^{-1} \\ & \left. \times \begin{pmatrix} \Delta_{AQ}(1-\varepsilon_A) \\ \Delta_{BQ}(1-\varepsilon_B) \end{pmatrix} \sigma_{gd}^2 / (q_1^2q_2^2) \right]. \end{aligned}$$

Under the null hypothesis of H_{ABa0} , $F_{AB,a}$ is central $F(m + n - 2, N - n - m + 1)$. Under the alternative hypothesis of H_{ABa1} , $F_{AB,a}$ is non-central $F(m + n - 2, N - m - n + 1)$, and it can be shown that the corresponding non-centrality parameter is approximated by

$$\begin{aligned} \lambda_{ABa} \approx & \frac{N}{\sigma^2} (D_{AQ}^\tau(1-\varepsilon_A), D_{BQ}^\tau(1-\varepsilon_B)) (V_A/2)^{-1} \\ & \times \begin{pmatrix} D_{AQ}(1-\varepsilon_A) \\ D_{BQ}(1-\varepsilon_B) \end{pmatrix} \sigma_{ga}^2 / (q_1q_2). \end{aligned}$$

Pedigree sample and non-centrality parameter approximations

Consider pedigree data, and assume that there are no covariates. For a relative pair (1,2) of individuals 1 and 2 who are non-inbred relatives, Table 1 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, $P(A_gA_g, A_gA_g | IBD = 0) = P_{A_g}^4$, $P(A_gA_g, A_gA_h | IBD = 0) = 2P_{A_g}^3 P_{A_h}$ and

Table 1 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
$P(A_g A_g, A_g A_g C)$	$P_{A_g}^4$	$P_{A_g}^3$	$P_{A_g}^2$
$P(A_g A_g, A_g A_h C)$	$2P_{A_g} P_{A_h}^3$	$P_{A_h} P_{A_g}^2$	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
$P(A_g A_h, A_g' A_{h'} C)$	$4P_{A_g} P_{A_h} P_{A_{g'}} P_{A_{h'}}$	0	0
$P(A_g A_g, B_k B_k C)$	$P_{A_g}^2 P_{B_k}^2$	$P_{A_g} P_{B_k} P(A_g B_k)$	$P(A_g B_k)^2$
$P(A_g A_g, B_k B_l C)$	$2P_{A_g}^2 P_{B_k} P_{B_l}$	$P_{A_g} P_{B_l} P(A_g B_k) + P_{A_g} P_{B_k} P(A_g B_l)$	$2P(A_g B_k)P(A_g B_l)$
$P(A_g A_h, B_k B_k C)$	$2P_{A_g} P_{A_h} P_{B_k}^2$	$P_{A_g} P_{B_k} P(A_h B_k) + P_{A_h} P_{B_k} P(A_g B_k)$	$2P(A_g B_k)P(A_h B_k)$
$P(A_g A_h, B_k B_l C)$	$4P_{A_g} P_{A_h} P_{B_k} P_{B_l}$	$P_{A_g} P_{B_k} P(A_h B_l) + P_{A_g} P_{B_l} P(A_h B_k) + P_{A_h} P_{B_k} P(A_g B_l) + P_{A_h} P_{B_l} P(A_g B_k)$	$2P(A_g B_k)P(A_h B_l) + 2P(A_g B_l)P(A_h B_k)$

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 2 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
$Cov(x_{Ai1}^{(g)}, x_{Ai2}^{(g)} C)$	0	$P_{A_g} [1 - P_{A_g}] (1 - \epsilon_A)^2$	$2P_{A_g} [1 - P_{A_g}] (1 - \epsilon_A)^2$
$Cov(x_{Ai1}^{(g)}, x_{Ai2}^{(h)} C)$	0	$-P_{A_g} P_{A_h} (1 - \epsilon_A)^2$	$-2P_{A_g} P_{A_h} (1 - \epsilon_A)^2$
$Cov(z_{Ai1}^{(gh)}, z_{Ai2}^{(gh)} C)$	0	0	$P_{A_g}^2 P_{A_h}^2 (P_{A_g} + P_{A_h})^2 (1 - \epsilon_A)^2$
$Cov(z_{Ai1}^{(gh)}, z_{Ai2}^{(g'h')} C)$	0	0	$[P_{A_g} P_{A_h} P_{A_{h'}} (1 - \epsilon_A)]^2$
$Cov(z_{Ai1}^{(gh)}, z_{Ai2}^{(g'h)} C)$	0	0	0
$Cov(x_{Ai1}^{(g)}, z_{Ai2}^{(gh)} C)$	0	0	0
$Cov(x_{Ai1}^{(g)}, z_{Ai2}^{(g'h')} C)$	0	0	0
$Cov(x_{Ai1}^{(g)}, x_{Bi2}^{(k)} C)$	0	$(1 - \epsilon_A)(1 - \epsilon_B) D_{A_g B_k}$	$2(1 - \epsilon_A)(1 - \epsilon_B) D_{A_g B_k}$
$Cov(x_{Ai1}^{(g)}, z_{Bi2}^{(kl)} C)$	0	0	0
$Cov(z_{Ai1}^{(gh)}, z_{Bi2}^{(kl)} C)$	0	0	$E[z_{Ai1}^{(gh)} z_{Bi1}^{(kl)}] = E[z_{Ai2}^{(gh)} z_{Bi2}^{(kl)}]$

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

$P(A_g A_g, A_h A_h|IBD = 0) = P_{A_g}^2 P_{A_h}^2$. Utilizing the conditional probabilities of Table 1, the conditional covariances of variables $x_{Aij}^{(g)}, x_{Bij}^{(k)}, z_{Aij}^{(gh)}$ and $z_{Bij}^{(kl)}$ of a relative pair (1,2) can be calculated and the results are listed in Table 2. Given $(IBD=0)$, the covariances are 0 since the two variables are independent and so unrelated (for instance, $Cov(x_{Ai1}^{(g)}, x_{Ai2}^{(g)}|IBD = 0) = 0$). Other entries of Table 2 can be calculated, accordingly. Based on Table 2, it can be seen that $Cov(X_{AUB}^{(i1)}, X_{AUB}^{(i2)}|IBD = 0) = Cov(Z_{AUB}^{(i1)}, Z_{AUB}^{(i2)}|IBD = 0) = 0$ and $Cov(X_{AUB}^{(i1)}, Z_{AUB}^{(i2)}|IBD = k) = 0, k = 0, 1, 2$. In addition, we have

$$Cov(X_{AUB}^{(i1)}, X_{AUB}^{(i2)}|IBD = 1) = \frac{1}{2} Cov(X_{AUB}^{(i1)}, X_{AUB}^{(i2)}|IBD = 2),$$

$$Cov(Z_{AUB}^{(i1)}, Z_{AUB}^{(i2)}|IBD = 1) = 0.$$

Let Φ_{12} be their kinship coefficient of individuals 1 and 2, and 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 $X_{AUB}^{(i1)}$ and $Z_{AUB}^{(i2)}$ is a zero matrix, and

$$Cov(X_{AUB}^{(i1)}, X_{AUB}^{(i2)}) = 2\Phi_{12} Cov(X_{AUB}^{(i1)}, X_{AUB}^{(i2)}|IBD = 2) = 2\Phi_{12} V_{A2},$$

$$Cov(Z_{AUB}^{(i1)}, Z_{AUB}^{(i2)}) = \Delta_{712} Cov(Z_{AUB}^{(i1)}, Z_{AUB}^{(i2)}|IBD = 2) = \Delta_{712} V_{D2}, \tag{14}$$

where the elements of $V_{A2} = Cov(X_{AUB}^{(i1)}, X_{AUB}^{(i2)}|IBD = 2)$ and $V_{D2} = Cov(Z_{AUB}^{(i1)}, Z_{AUB}^{(i2)}|IBD = 2)$ are given by the entries of the last column of Table 2.

Nuclear family data. Consider I families each has both parents and s offspring. The total number of individuals is $I(s + 2)$. Let us list the $s + 2$ individuals of each family as $j = 1, 2, 3, \dots, s + 2$, where individual 1 is the father and individual 2 is the mother, and the offspring are listed as $j = 3, \dots, s + 2$. Suppose that variance–covariance matrices

of the I families are the same, i.e., $\Sigma_1 = \dots = \Sigma_I$. Denote $\Sigma_i^{-1} = \frac{1}{\sigma^2} (\gamma_{hj})_{(s+2) \times (s+2)}$, and let $b = (\gamma_{13} + \dots + \gamma_{1,s+2}) + (\gamma_{23} + \dots + \gamma_{2,s+2}) + \sum_{h=3}^{s+2} \sum_{j=h+1}^{s+2} \gamma_{hj}$. If the number of families I is large enough, we show in Appendix F that the non-centrality parameter of statistic $F_{AB,ad}$ is approximated by

$$\begin{aligned} \lambda_{ABad} &\approx \frac{2I\sigma_{ga}^2}{q_1q_2\sigma^2} \left(D_{AQ}^\tau(1-\varepsilon_A), D_{BQ}^\tau(1-\varepsilon_B) \right) \\ &\times V_A^{-1} \left(\sum_{k=1}^{s+2} \gamma_{kk} V_A + b V_{A2} \right) V_A^{-1} \begin{pmatrix} D_{AQ}(1-\varepsilon_A) \\ D_{BQ}(1-\varepsilon_B) \end{pmatrix} \\ &+ \frac{I\sigma_{gd}^2}{(q_1q_2)^2\sigma^2} \left(\Delta_{AQ}^\tau(1-\varepsilon_A), \Delta_{BQ}^\tau(1-\varepsilon_B) \right) \\ &\times V_D^{-1} \left(\sum_{k=1}^{s+2} \gamma_{kk} V_D + \sum_{k=3}^{s+2} \sum_{l=k+1}^{s+2} \gamma_{kl} V_{D2}/2 \right) \\ &\times V_D^{-1} \begin{pmatrix} \Delta_{AQ}(1-\varepsilon_A) \\ \Delta_{BQ}(1-\varepsilon_B) \end{pmatrix}. \end{aligned} \quad (15)$$

Similarly, the non-centrality parameter of statistic $F_{AB,a}$ is approximated by

$$\begin{aligned} \lambda_{ABa} &\approx \frac{2I\sigma_{ga}^2}{q_1q_2\sigma^2} \left(D_{AQ}^\tau(1-\varepsilon_A), D_{BQ}^\tau(1-\varepsilon_B) \right) \\ &\times V_A^{-1} \left(\sum_{k=1}^{s+2} \gamma_{kk} V_A + b V_{A2} \right) V_A^{-1} \begin{pmatrix} D_{AQ}(1-\varepsilon_A) \\ D_{BQ}(1-\varepsilon_B) \end{pmatrix}. \end{aligned}$$

Multi-generation pedigree data. Consider I families given in graph A or graph B of Fig. 1 (Fig. 1 in Abecasis et al. 2000B; Fan et al. 2005). For each individual in Fig. 1, an ID is assigned. For the grand parents of graph B, both phenotypes and genotypes are unavailable and so no IDs are assigned. The total number of individuals is tI , where $t = 11$ for graph A and $t = 18$ for graph B of Fig. 1, respectively. Again, assume that variance–covariance matrices of the I families are the same, i.e., $\Sigma_1 = \dots = \Sigma_I$. Denote $\Sigma_i^{-1} = \frac{1}{\sigma^2} (\gamma_{hj})_{t \times t}$. If the number of families I is large enough, we can show in the same way as Appendix F that the non-centrality parameter of statistic $F_{AB,ad}$ is approximated by

$$\begin{aligned} \lambda_{ABad} &\approx \frac{2I\sigma_{ga}^2}{q_1q_2\sigma^2} \left(D_{AQ}^\tau(1-\varepsilon_A), D_{BQ}^\tau(1-\varepsilon_B) \right) \\ &\times V_A^{-1} \left(\sum_{k=1}^t \gamma_{kk} V_A + b_1 V_{A2} \right) V_A^{-1} \begin{pmatrix} D_{AQ}(1-\varepsilon_A) \\ D_{BQ}(1-\varepsilon_B) \end{pmatrix} \\ &+ \frac{I\sigma_{gd}^2}{(q_1q_2)^2\sigma^2} \left(\Delta_{AQ}^\tau(1-\varepsilon_A), \Delta_{BQ}^\tau(1-\varepsilon_B) \right) \\ &\times V_D^{-1} \left(\sum_{k=1}^t \gamma_{kk} V_D + b_2 V_{D2}/2 \right) V_D^{-1} \begin{pmatrix} \Delta_{AQ}(1-\varepsilon_A) \\ \Delta_{BQ}(1-\varepsilon_B) \end{pmatrix}, \end{aligned} \quad (16)$$

where b_1 and b_2 are provided in Appendix G. The non-centrality parameter of $F_{AB,a}$ is approximated by

$$\begin{aligned} \lambda_{ABa} &\approx \frac{2I\sigma_{ga}^2}{q_1q_2\sigma^2} \left(D_{AQ}^\tau(1-\varepsilon_A), D_{BQ}^\tau(1-\varepsilon_B) \right) \\ &\times V_A^{-1} \left(\sum_{k=1}^t \gamma_{kk} V_A + b_1 V_{A2} \right) V_A^{-1} \begin{pmatrix} D_{AQ}(1-\varepsilon_A) \\ D_{BQ}(1-\varepsilon_B) \end{pmatrix}. \end{aligned}$$

Results

Type I error rates

Simulation studies are performed to evaluate the robustness of the proposed models. We evaluate a marker A which is di-allelic, tri-allelic and quadri-allelic, i.e., $m = 2, 3$ and 4 . 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 $P_{A_3} = 0.4$; and for quadri-allelic marker, equal allele frequencies are assumed, i.e., $P_{A_1} = \dots = P_{A_4} = 0.25$. Five test cases are considered in type I error rate calculation. Table 3 presents parameters of four test cases. Trait values are constructed by normal distribution with mean 100 and total variance $\sigma^2 = 1$ except for test case of *Admixture*. Here $\sigma^2 = \sigma_{ga}^2 + \sigma_{Ga}^2 + \sigma_e^2$ is the summation of the additive major gene effect σ_{ga}^2 , the variance of polygenic effect σ_{Ga}^2 , and the error variance σ_e^2 . In the test cases of *Null*, *Familiarity*, and *Admixture*, no major gene effect is assumed, i.e., $\sigma_{ga}^2 = 0$. In the test cases of *Linkage* and *Composite*, major gene effect is assumed, and recombination fraction $\theta_{AQ} = 0$; in the meantime, linkage equilibrium is assumed between QTL Q and the marker A . In the test case of *Admixture*, population admixture is generated by mixing families equally drawn from one of the two sub-populations C and D. In both sub-populations C and D, no major gene effect or familial effect is assumed, i.e., $\sigma_{ga}^2 = \sigma_{Ga}^2 = 0$. However, the trait mean of sub-population C is fixed as 1 and the variance is fixed as 1. The trait mean of sub-population D is fixed as 0 and the variance is fixed as 1. Therefore, the total variance in the mixing population is $\sigma^2 = 1.25$. The admixture contributed to $(1-0)^2/[4 \times 1.25] = 0.20$ of the total variance.

To calculate the type I error rates of Table 4, 1,000 datasets are simulated for each test case. Each dataset contains 50 pedigrees of either graph A or graph B of Fig. 1, respectively. Using the datasets, we fit the model (3) and test the null hypothesis $H_{a0}: \alpha_1 = \dots = \alpha_m$. Since the QTL Q is in linkage equilibrium with marker A , an empirical test statistic which is larger than the cutting point at a 0.05 significance level is treated as a false positive. Based on likelihood ratio test, type I error rates are calculated as the proportions of the 1,000 simulation datasets which give significant result at the 0.05 significant level. The results of type error rates are presented in Table 4. The results of Table 4 show that the type I error rates are

Fig. 1 Multi-generation pedigrees used in power calculations and comparison, which are taken from Fig. 1 of Abecasis et al. (2000B) or Fan et al. (2005). The number in the box or circle is individual ID

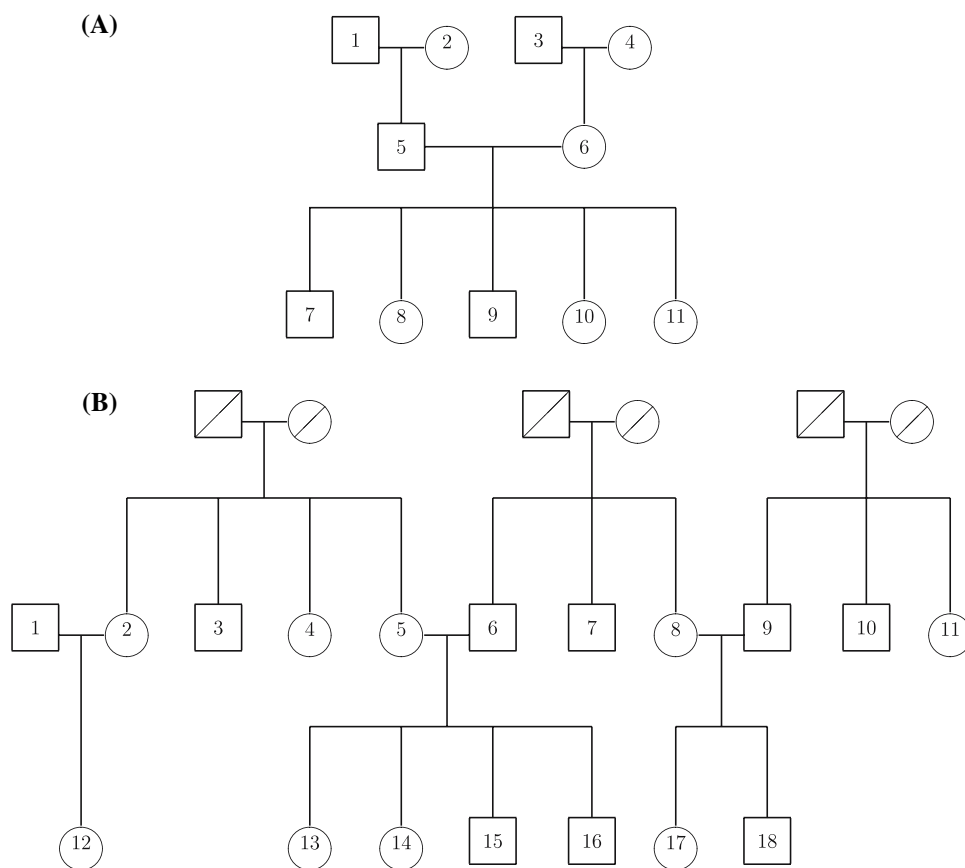


Table 3 The parameters of the simulated genetic cases

Test cases	σ_{Ga}^2	σ_{Ga}^2	σ_e^2	σ^2	$\theta_{A Q}$	q_1	$D_{A_g Q}, g = 1, \dots, m$
Null	0	0	1.0	1.0	0.5	Not Applied	Not Applied
Familiarity	0	0.5	0.5	1.0	0.5	Not Applied	Not Applied
Linkage	0.5	0	0.5	1.0	0	0.5	0
Composite	0.2	0.3	0.5	1.0	0	0.5	0

The total variance is fixed as $\sigma^2 = \sigma_{ga}^2 + \sigma_{Ga}^2 + \sigma_e^2 = 1$, and $\sigma_{gd}^2 = 0$. *Admixture*: no major gene effect or familial effect $\sigma_g^2 = \sigma_{Ga}^2 = 0$, but with population admixture (see text for explanation)

around the nominal level 0.05. Hence, the model is reasonably robust. In all the four missing rate cases ($\epsilon_A = 0.05, 0.10, 0.15, 0.20$), the type I error rates are reasonable. Hence, the missingness does not affect the robustness of the model.

In the Table 5, we show the type I error rates using tri-nuclear families. Each tri-nuclear family contains 3 people, parents and an offspring. Again, 1,000 datasets are simulated for each test case. Each dataset contains 50 tri-nuclear pedigrees. Two types of calculation are performed: (1) imputing genotypes which are missing by the proposed method, and keeping every individual in the analysis; (2) removing all individuals from analysis whose genotypes are missing. It can be seen that the proposed method can help to get correct type I error rates, by imputing genotypes which are missing, since the type I error rates are around

the nominal level 0.05. In the previous approach, an individual is deleted from analysis once his/her genotype is missing; it may inflate type I error rates by the results of Table 5, since the type I error rates are around 0.06. In the column of $\epsilon_A = 0$ of Table 5, no individual is removed since there is no missing genotypes; hence, the type I error rates are the same for the two types of calculation.

Power comparison

Let us denote the heritability by h^2 , which is defined as $h^2 = \sigma_{ga}^2 / \sigma^2$ (Falconer and Mackay 1996). To compare power, we plot power curves of two cases using the related non-centrality parameter approximation: a dominant mode of inheritance, $a = 1, d = 1$, and a recessive mode of

Table 4 Type I error rates (%) at a 0.05 significance level based on likelihood ratio tests.

No. of alleles	Pedigree type	Test case	Error rates			
			$\epsilon_A = 0.05$	$\epsilon_A = 0.10$	$\epsilon_A = 0.15$	$\epsilon_A = 0.20$
Di-allele $m = 2$	Small 3-generation pedigree (A)	Null	4.2	4.9	4.9	4.6
		Familiality	3.8	4.3	5.2	4.2
		Admixture	4.2	4.5	4.4	4.5
		Linkage	5.7	4.8	5.0	5.3
		Composite	5.2	4.6	4.9	5.0
	Large 3-generation pedigree (B)	Null	4.1	4.3	4.0	4.0
		Familiality	5.4	5.6	5.0	4.7
		Admixture	5.4	4.8	4.6	3.5
		Linkage	4.5	4.4	4.7	4.4
		Composite	5.3	5.1	4.8	5.0
Tri-allele $m = 3$	Small 3-generation pedigree (A)	Null	4.8	3.9	4.5	3.7
		Familiality	5.1	5.1	5.4	5.4
		Admixture	4.0	3.7	4.2	4.4
		Linkage	5.6	3.9	5.0	5.8
		Composite	4.5	4.6	5.0	4.9
	Large 3-generation pedigree (B)	Null	4.8	5.0	5.8	5.0
		Familiality	5.0	4.8	5.7	5.6
		Admixture	3.8	5.4	5.5	5.8
		Linkage	5.4	4.9	5.2	5.1
		Composite	5.9	5.6	5.6	5.1
Quadri-allele $m = 4$	Small 3-generation pedigree (A)	Null	4.9	4.8	5.4	4.6
		Familiality	4.6	5.2	4.6	5.4
		Admixture	4.4	3.7	4.4	4.4
		Linkage	5.4	5.3	5.7	5.2
		Composite	5.4	5.4	5.4	4.8
	Large 3-generation pedigree (B)	Null	4.3	4.6	5.6	5.2
		Familiality	4.3	3.9	4.5	4.6
		Admixture	4.8	5.2	4.3	5.4
		Linkage	5.6	5.8	4.9	4.8
		Composite	5.1	5.2	5.4	4.3

Table 5 Type I error rates (%) at a 0.05 significance level of 50 tri-nuclear families based on likelihood ratio tests

No. of alleles	Method of treating missing genotypes	Test case	Error rates				
			$\epsilon_A = 0$	$\epsilon_A = 0.05$	$\epsilon_A = 0.10$	$\epsilon_A = 0.15$	$\epsilon_A = 0.20$
Quadri-allele $m = 4$	Imputing all genotypes which are missing, and keeping everyone in analysis	Null	4.8	4.4	5.3	4.1	4.9
		Familiality	5.6	5.0	5.6	5.3	5.3
		Admixture	5.6	5.6	5.2	5.1	5.1
		Linkage	5.3	5.7	5.2	5.1	5.2
		Composite	5.3	5.0	5.5	5.5	4.9
	Removing all individuals whose genotypes are missing	Null	4.8	6.1	6.4	5.6	6.0
		Familiality	5.6	6.2	5.7	5.9	5.8
		Admixture	5.6	6.2	5.7	5.8	6.0
		Linkage	5.3	5.7	6.0	6.0	6.2
		Composite	5.3	5.3	5.7	6.0	5.8

inheritance, $a = 1, d = -0.5$. Figures 2 and 3 show power curves of population samples at 0.01: Fig. 2 is based on models (1) and (3) using one marker A; Fig. 3 is based on

models (10) and (12) using two markers A and B. The power curves of Fig. 2 are plotted against the heritability h^2 ; for graphs I and II, the marker A has three equal

frequency alleles; for graphs III and IV, the marker *A* has four frequency alleles; and the related parameters are given in the legend of the figure. Two features can be noted from Fig. 2: (1) the power based on “genotype effect model” (1) is generally lower than that of the “additive effect model” (3); (2) the power is reasonably high when the heritability h^2 is larger than 0.15. The power curves of Fig. 3 are plotted against the LD measure D_{A_1Q} ; for graphs I and II, there are no missing genotypes, i.e., $\varepsilon_A = \varepsilon_B = 0.0$; for graphs III and IV, there are missing genotypes, and $\varepsilon_A = \varepsilon_B = 0.25$. It is obvious that missing genotypes lead to power decreasing, since the non-centrality parameter approximations are reduced.

The power curves of Figs. 4–6 are based on pedigree data: Fig. 4 is based on nuclear families in which each family consists of two parents and two children; Fig. 5 is

based on 45 small 3-generation pedigrees (Graph I, Fig. 1); and Fig. 6 is based on 30 large 3-generation pedigrees (Graph II, Fig. 1). Such as the population data, the three figures show that missing genotypes lead to power decreasing. In addition, the power based on the “genotype effect model” (12) is generally lower than that of the “additive effect model” (10).

Example

The proposed method is applied to analyze angiotensin-1 converting enzyme data (Farrall et al. 1999; Keavney et al. 1998). The data consist of 83 extended families with between 4 and 18 members. Circulating ACE levels were measured for 405 individuals. Ten bi-allelic

Fig. 2 Power curves of population sample at 0.01 level based on models (1) and (3), where $N = 250, \varepsilon_A = 0.1, q_1 = 0.5, \sigma_{Ga}^2 = 0.10$. For graphs I and II, the marker *A* has three alleles and $P_{A_i} = 1/3, i = 1, 2, 3, D_{A_1Q} = 0.12, D_{A_2Q} = D_{A_3Q} = -0.06$; for graphs III and IV, the marker *A* has four alleles, $P_{A_i} = 0.25, i = 1, \dots, 4, D_{A_1Q} = -D_{A_2Q} = D_{A_3Q} = -D_{A_4Q} = 0.08$; for dominant mode of inheritance of graphs I and III, $a = 1, d = 1$; for recessive mode of inheritance of graphs II and IV, $a = 1, d = -0.5$

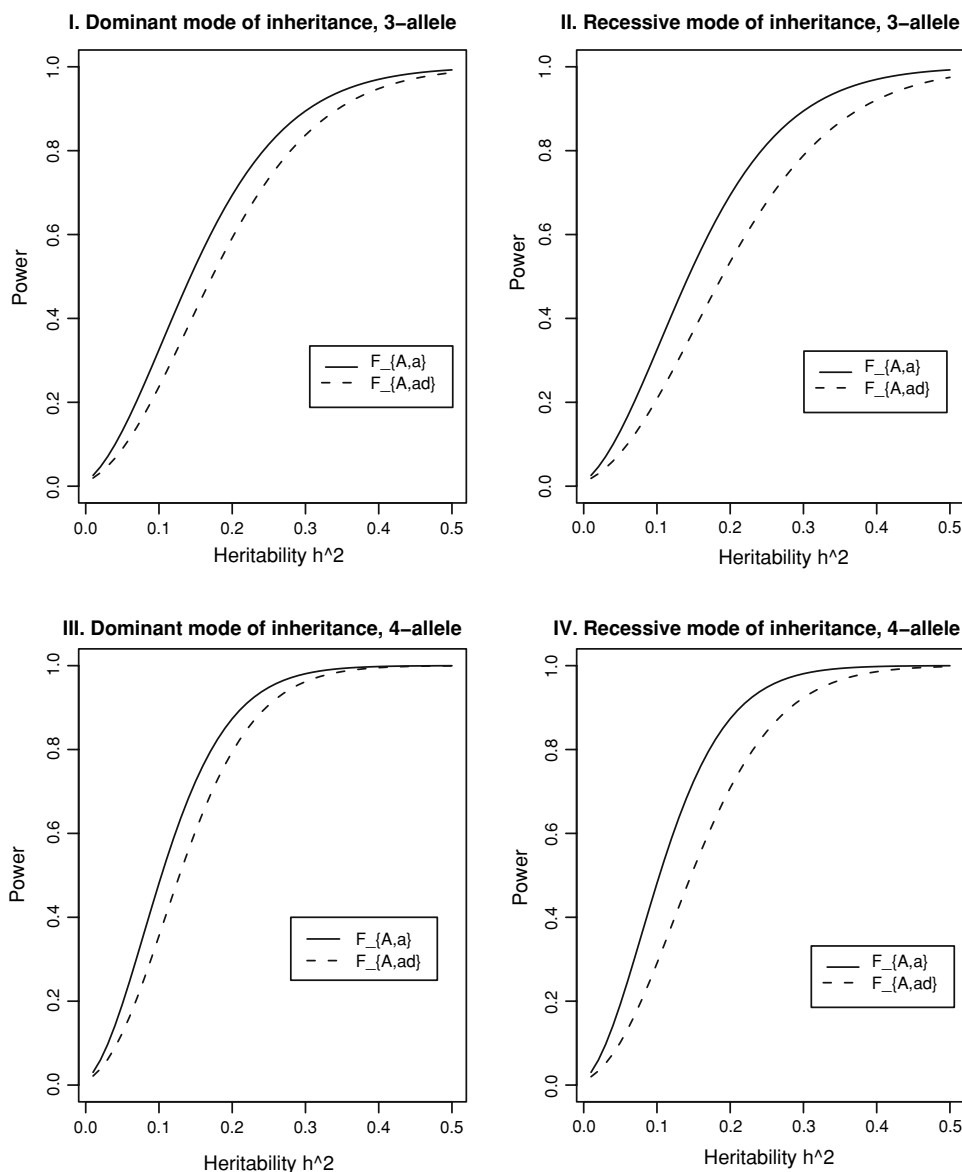
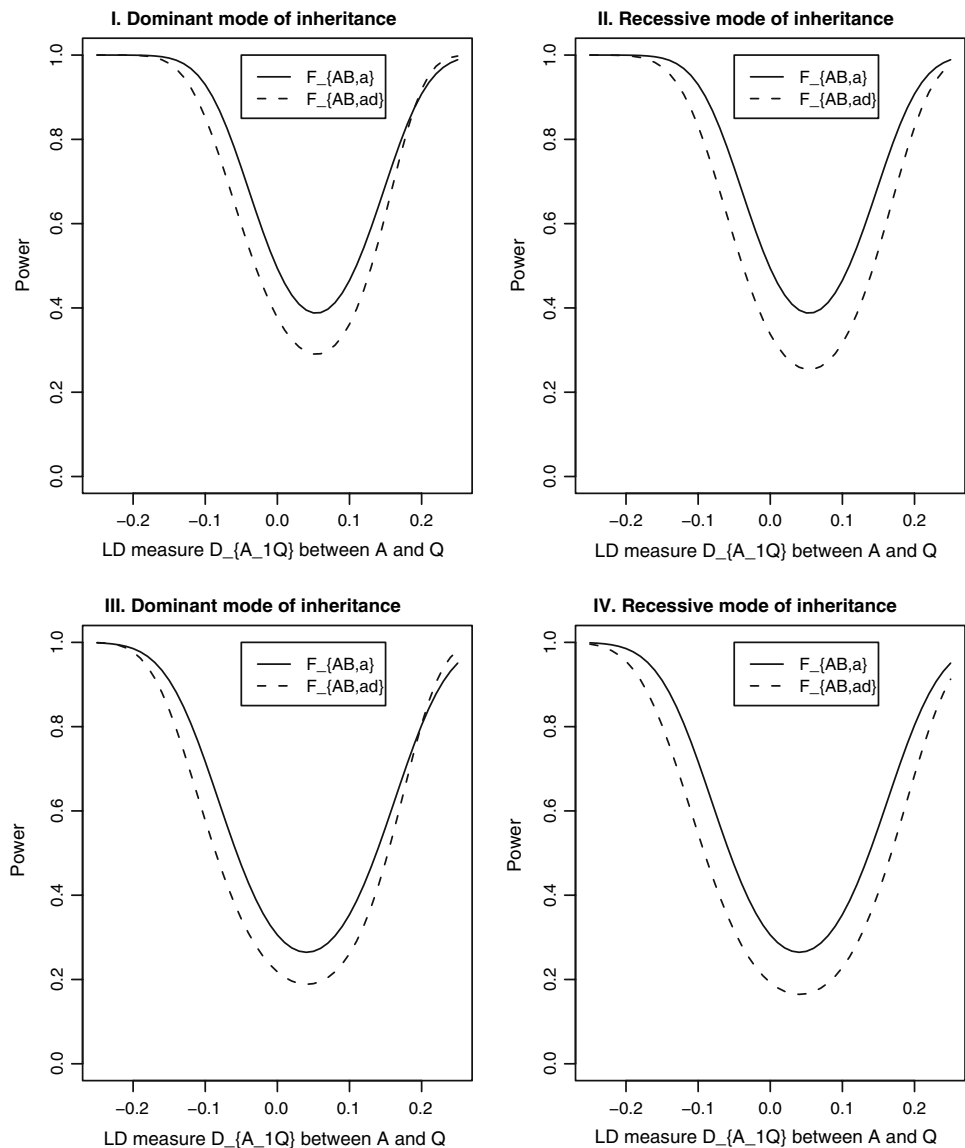


Fig. 3 Power curves of population sample at 0.01 level based on models (10) and (12), where $N = 200, m = 2, n = 3, q_1 = q_2 = P_{A_1} = P_{A_2} = 0.5, P_{B_i} = 1/3, i = 1, 2, 3, D_{B_1Q} = D_{B_2Q} = 0.06, D_{A_1B_1} = D_{A_1B_2} = 0.05, \sigma_{Ga}^2 = 0.10, h^2 = 0.15$. For graphs I and II, $\varepsilon_A = \varepsilon_B = 0.0$; for graphs III and IV, $\varepsilon_A = \varepsilon_B = 0.25$; for dominant mode of inheritance of graphs I and III, $a = 1, d = 1$; for recessive mode of inheritance of graphs II and IV, $a = 1, d = -0.5$

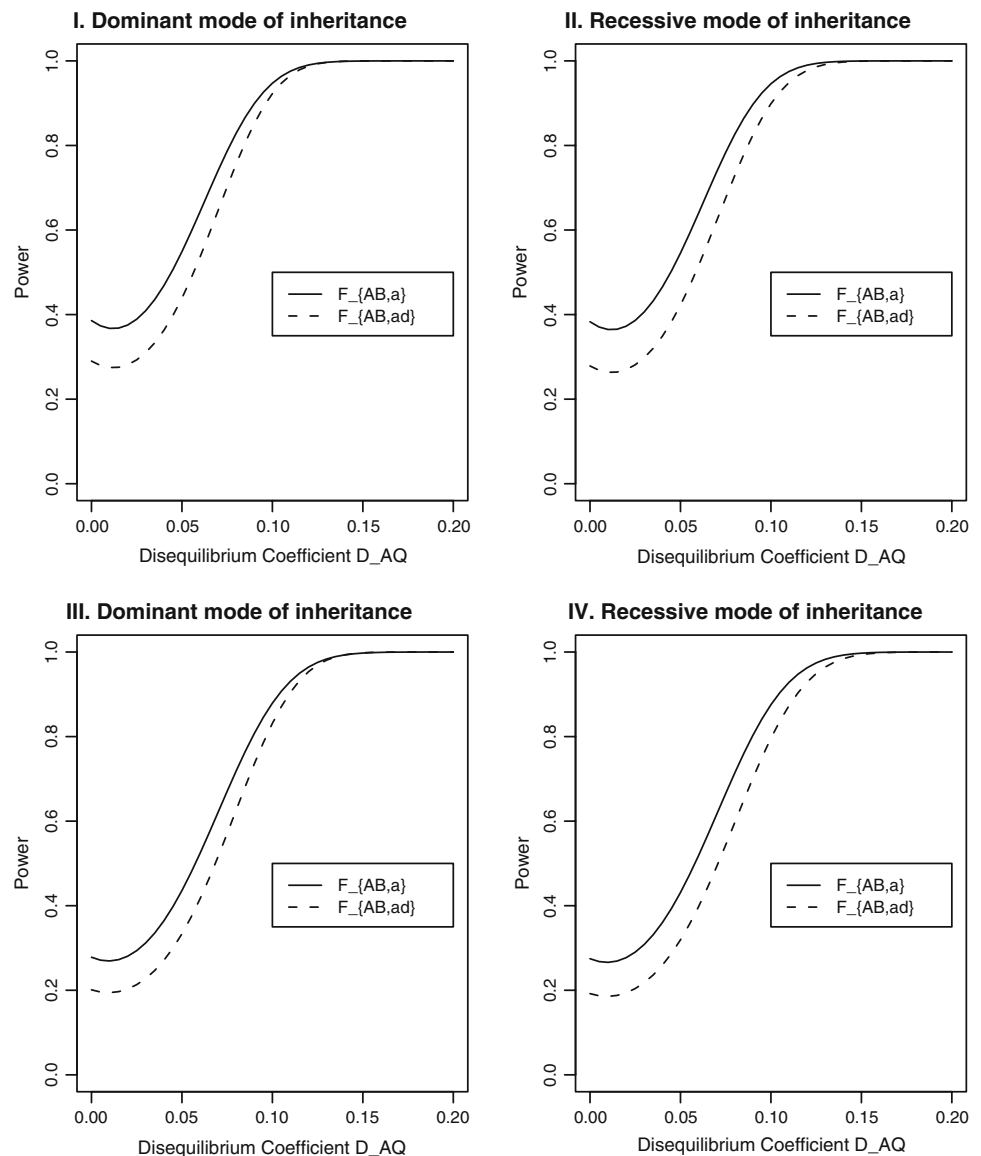


polymorphisms in the ACE gene were genotyped. There is missing genotype information at markers. Although we can not rigorously show that the missingness is missing completely at random, it is roughly correct since there is no systematic pattern in the missingness; actually, either founder's or non-founder's genotypes or both can be missing in a pedigree. In addition, the missingness is different from marker to marker, i.e., genotypes of an individual at some markers are missing, and are not missing at other markers. In our previous study, all individuals with missing genotypes are deleted, and so the total number of individuals is different from marker to marker (Fan et al. 2005; refer to column 2 of Table 6). For instance, there are 4 individuals whose genotype information is missing at marker I/D and the total number $N = 401$ in our previous study; at marker G2350A, on the other hand, there are more missing genotype data, and $N = 368$.

In this paper, all 405 individuals are used in the analysis using the proposed variance component models for each marker.

Before fitting the models, multi-point IBD at each marker are calculated by Merlin (Abecasis et al. 2002). Variance component linkage analysis shows that additive variance is significantly larger than 0, but dominant variance is not significantly larger than 0 (Abecasis et al. 2000B). Hence, dominant effects can be excluded from regression equation, and the total variance is modeled as $\sigma^2 = \sigma_{ga}^2 + \sigma_{Ga}^2 + \sigma_e^2$. Table 6 shows LD analysis of the ACE gene by individual marker. After fitting the proposed models in this article, lod is calculated by $LRT/(2 \ln 10)$, where $LRT = 2(L_1 - L_0)$, L_1 is the log-likelihood under $y_{ij} = \alpha + x_{Aij}^{(1)}\alpha_A + H_{ij} + e_{ij}$ which is equivalent to model (3), and L_0 is the log-likelihood under $y_{ij} = \alpha + H_{ij} + e_{ij}$.

Fig. 4 Power curves of 200 nuclear families at 0.01 level based on models (10) and (12), where $m = n = 2$, $q_1 = P_{A_1} = P_{B_1} = 0.5$, $D_{A_1B_1} = 0.05$, $D_{B_1Q} = 0.06$, $\sigma_{G_a}^2 = 0.10$, $h^2 = 0.15$. Here, each nuclear family has two children. For graphs I and II, $\varepsilon_A = \varepsilon_B = 0$; for graphs III and IV, $\varepsilon_A = \varepsilon_B = 0.25$; for dominant mode of inheritance of graphs I and III, $a = 1$, $d = 1$; for recessive mode of inheritance of graphs II and IV, $a = 1$, $d = -0.5$



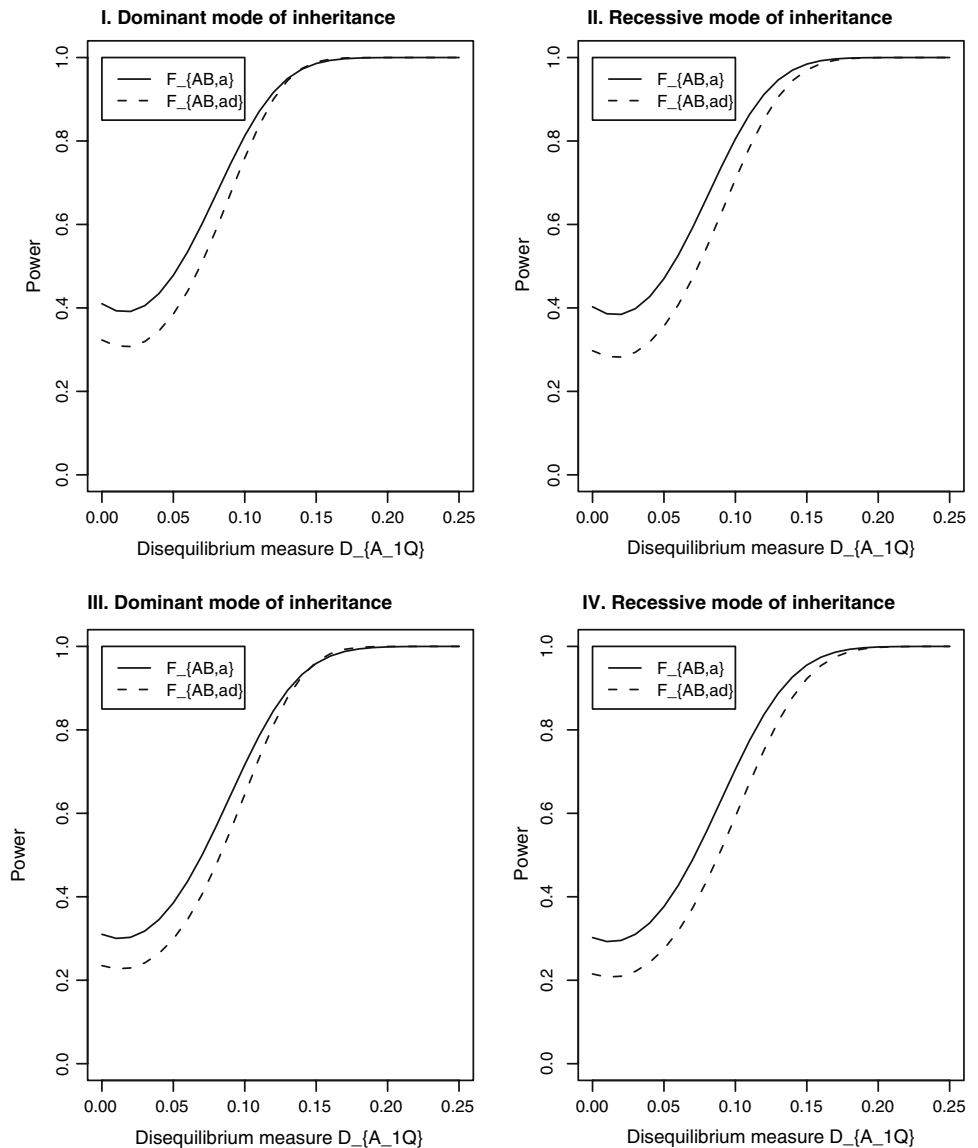
The lod scores calculated by the proposed method in this article are similar to those in our previous study for most markers (column 4 and column 5, Table 6). Hence, whether the individuals with missing genotypes are removed from the analysis or not does not influence the conclusion we reached. The results of Table 6 confirm the previous finding that the association is strongest around the G2215A, I/D, G2350A and 4656(CT)3/2 polymorphisms (Abecasis et al. 2000b). Therefore, these markers are likely in complete LD with the trait alleles. The Lod scores calculated by the proposed method at three markers, T-1237C, G2215A and G2350A, show big decreases compared with those of our previous study. This is most likely due to the fluctuations from the missingness. In our previous study, we found that allele *I* at marker I/D is almost always present with allele *A* at marker G2350A, and allele *D* at marker I/D is almost always present with allele *G* at marker

G2350A (Fan et al. 2005). The two markers are almost in complete LD with each other. However, the lod scores of these two markers are different from each other, which is most likely because there are more missing genotypes at marker G2350A.

Discussion

In searching for common genes of complex traits, large samples are needed that are likely to come only from combining family and population based data. In addition, sophisticated methods are needed to analyze these combinations. The statistical and mathematical methods and models must account for missing data, and must account for environmental covariates which are certain to play a role in complex diseases. In this article, we have extended

Fig. 5 Power curves of 45 small 3-generation pedigrees (Graph A, Fig. 1) at 0.01 level based on models (10) and (12), where $m = n = 2$, $q_1 = P_{A_1} = P_{B_1} = 0.5$, $D_{B_1Q} = 0.08$, $D_{A_1B_1} = 0.05$, $\sigma_{Ga}^2 = 0.10$, $h^2 = 0.15$. For graphs I and II, $\varepsilon_A = \varepsilon_B = 0.0$; for graphs III and IV, $\varepsilon_A = \varepsilon_B = 0.25$; for dominant mode of inheritance of graphs I and III, $a = 1$, $d = 1$; for recessive mode of inheritance of graphs II and IV, $a = 1$, $d = -0.5$



our previous variance component models for combined linkage and association mapping of QTL in the presence of missing genotypes. Under an assumption of MCAR, two regression models, “genotype effect model” and “additive effect model”, are proposed to model the association between the markers and the trait locus. 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, which show that the power of the F -tests is reduced due to the missingness. In the Table 4, Fan et al. (2005), we showed that the non-centrality parameter approximations compare well to what is produced via simulation, under a circumstance of no missing data and di-allelic markers. Under the assumption that the genotype data are MCAR, we perform some simulations to confirm the accuracy of the non-centrality

parameter approximations (data not shown). Moreover, simulation studies are performed to calculate the type I error rates to evaluate the robustness of the proposed models. It is found the type I error rates are reasonable. The method is applied to analyze the European angiotensin-1 converting enzyme data.

In this paper, the genotypes are assumed to be MCAR. This assumption is roughly true for some study, such as the angiotensin-1 converting enzyme data. For certain studies, the missingness can be systematic, e.g., some or all of the founder genotypes can be missing (Wang and Elston 2005). It is unclear how this will affect the proposed models, if the assumption of MCAR is not valid. In many situations, the assumption of MCAR is unlikely to be met. Hence, the utility of the proposed method in a practical situation can be limited although it may improve robustness. The issue of how to properly account for missing data to improve

Fig. 6 Power curves of 30 large 3-generation pedigrees (Graph B, Fig. 1) at 0.01 level based on models (10) and (12), where $m = n = 2, q_1 = P_{A_1} = P_{B_1} = 0.5, D_{B_1Q} = 0.06, D_{A_1B_1} = 0.05, \sigma_{\epsilon_a}^2 = 0.10, h^2 = 0.15$. For graphs I and II, $\epsilon_A = \epsilon_B = 0.0$; for graphs III and IV, $\epsilon_A = \epsilon_B = 0.25$; for dominant mode of inheritance of graphs I and III, $a = 1, d = 1$; for recessive mode of inheritance of graphs II and IV, $a = 1, d = -0.5$

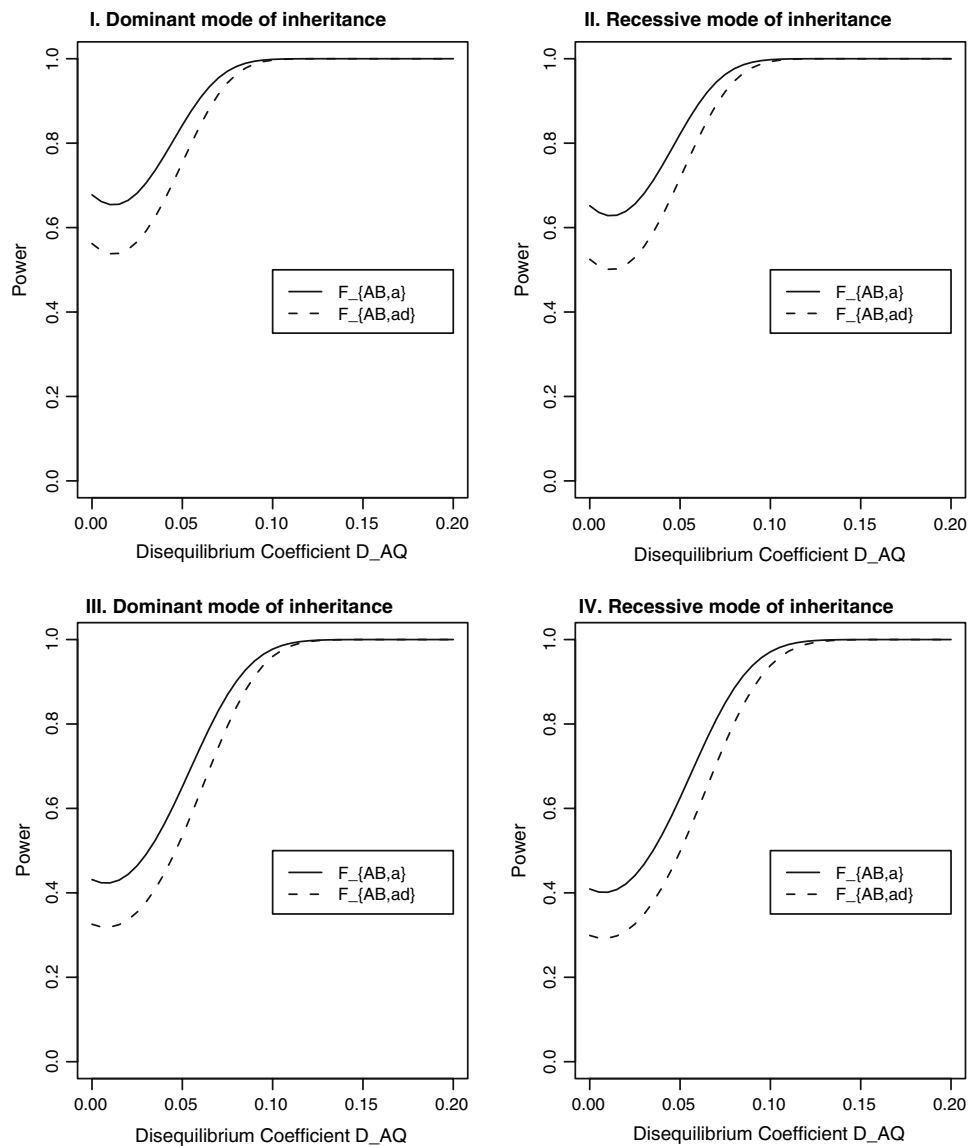


Table 6 Linkage disequilibrium analysis of the European ACE data by individual marker

The *AbAw's lod* is taken from Table 4 of Abecasis et al. (2000B). The *lod without missing* is taken from Table 5, column 4, Fan et al. (2005), which is calculated by deleting all individuals when their genotypes are missing. The *lod with missing* is calculated based on the model developed in this paper. * There is error in Table 5, column 4, Fan et al. (2005)

Marker A	Previous results			Lod with missing # of individuals=405
	# Of individuals	AbAw's lod	Lod without missing	
T-5491C	395*	9.86	13.91	13.34
A-5466C	392	9.04	14.06	14.25
T-3892C	400	12.49	18.27	17.58
A-240T	401	10.81	13.35	13.66
T-93C	392	10.93	13.00	13.13
T-1237C	377	11.52	20.59	17.94
G2215A	372	14.91	27.01	24.78
I/D	401	15.76	27.37	27.59
G2350A	368*	14.40	28.01	26.13
4656(CT)3/2	390	14.22	27.93	27.16

power while maintaining correct type I error rates in genetic studies is paramount. For instance, consider a trinuclear family. Assume that the two parents’ genotypes are 12 and 11, and the genotype of the offspring is missing. It is easy to see that the genotype of the offspring can be 11 or 21 with a probability 0.5 for each genotype. Hence, the right imputing method is to assign a weight of 0.5 for each of these two genotypes. Besides, other genotypes such as 22 should be assigned a weight of 0. In short, information of genotypes of family members can be used to infer the missing genotype of another family member. In this way, it is very likely that the power can be improved. However, it won’t be easy to get neat non-centrality parameter approximations as the ones we have in the paper under an assumption of MCAR. Instead, simulation study is a possible method for the investigation. Due to the length of this paper, we leave this issue to be investigated in the future study.

The proposed models can be used to analyze either single marker or multiple genotype data. However, the models can only be used to analyze one phenotype. As the ability to generate more genetics data, both for phenotypes and for genotypes, increases, statistical methods are required to evaluate multiple phenotypes and multiple genotypes simultaneously. That is, more research is necessary to extend existing theoretical methods to analyze multivariate phenotypes of complex diseases.

Acknowledgments The research was supported by the National Science Foundation Grant DMS-0505025. We thank two anonymous reviewers for very detailed and thoughtful critiques, which make the paper better.

Appendix A

Multiplying both sides of the “genotype effect model” (1) by $1_{(G_{Aij}=A_gA_h)}$ and taking expectation lead to

$$\begin{aligned}
 & E(y_{ij}1_{(G_{Aij}=A_gA_h)}) \\
 &= w_{ij}\gamma E[1_{(G_{Aij}=A_gA_h)}] + E[1_{(G_{Aij}=A_gA_h)}]\beta_{gh} \\
 &= \begin{cases} (1 - \varepsilon_A)[w_{ij}\gamma + \beta_{gg}]P_{A_g}^2 & \text{if } g = h \\ (1 - \varepsilon_A)[w_{ij}\gamma + \beta_{gh}] \cdot 2P_{A_g}P_{A_h} & \text{if } g \neq h \end{cases} \quad (17)
 \end{aligned}$$

Let G_{Qij} be genotype of the j -th individual of the i -th family at the trait locus Q . A true random effect model describing the trait value is $y_{ij} = w_{ij}\gamma + g_{ij} + H_{ij} + e_{ij}$, where

$$g_{ij} = \begin{cases} a & G_{Qij} = Q_1Q_1 \\ d & G_{Qij} = Q_1Q_2 \\ -a & G_{Qij} = Q_2Q_2 \end{cases}$$

Since the missing mechanism is missing completely at random, we have

$$\begin{aligned}
 & P(G_{Qij} = Q_1Q_1, G_{Aij} = A_gA_g | G_{Aij} \neq ?) = [P(Q_1A_g)]^2, \\
 & P(G_{Qij} = Q_1Q_2, G_{Aij} = A_gA_g | G_{Aij} \neq ?) = 2P(Q_1A_g)P(Q_2A_g), \\
 & P(G_{Qij} = Q_2Q_2, G_{Aij} = A_gA_g | G_{Aij} \neq ?) = [P(Q_2A_g)]^2.
 \end{aligned}$$

Utilizing relations $P(Q_1A_g) = -D_{A_gQ} + P_{A_g}q_1$ and $P(Q_2A_g) = -D_{A_gQ} + P_{A_g}q_2$, we have

$$\begin{aligned}
 & E(y_{ij}1_{(G_{Aij}=A_gA_g)}) \\
 &= w_{ij}\gamma E[1_{(G_{Aij}=A_gA_g)}] + E[g_{ij}1_{(G_{Aij}=A_gA_g)}] \\
 &= w_{ij}\gamma P(A_gA_g | G_{Aij} \neq ?) P(G_{Aij} \neq ?) \\
 &\quad + E[g_{ij}1_{(G_{Aij}=A_gA_g)} | G_{Aij} \neq ?] P(G_{Aij} \neq ?) \\
 &= (1 - \varepsilon_A) [w_{ij}\gamma P_{A_g}^2 + a[P(Q_1A_g)]^2 \\
 &\quad + d \cdot 2P(Q_1A_g)P(Q_2A_g) - a[P(Q_2A_g)]^2] \\
 &= (1 - \varepsilon_A) [w_{ij}\gamma P_{A_g}^2 + \mu P_{A_g}^2 + 2D_{A_gQ}\alpha_Q P_{A_g} - \delta_Q D_{A_gQ}^2]. \quad (18)
 \end{aligned}$$

Equating Eqs. 17 and 18, we show the Eq. 5 when $g = h$. Now assume that $g \neq h$. Since the missing mechanism is missing completely at random, we have

$$\begin{aligned}
 & P(G_{Qij} = Q_1Q_1, G_{Aij} = A_gA_h | G_{Aij} \neq ?) = 2P(Q_1A_g)P(Q_1A_h), \\
 & P(G_{Qij} = Q_1Q_2, G_{Aij} = A_gA_h | G_{Aij} \neq ?) = 2P(Q_1A_g)P(Q_2A_h) \\
 &\quad + 2P(Q_1A_h)P(Q_2A_g), \\
 & P(G_{Qij} = Q_2Q_2, G_{Aij} = A_gA_h | G_{Aij} \neq ?) = 2P(Q_2A_g)P(Q_2A_h).
 \end{aligned}$$

Utilizing relations $P(Q_1A_g) = D_{A_gQ} + P_{A_g}q_1$, $P(Q_2A_g) = -D_{A_gQ} + P_{A_g}q_2$, $P(Q_1A_h) = D_{A_hQ} + P_{A_h}q_1$, $P(Q_2A_h) = -D_{A_hQ} + P_{A_h}q_2$, we have

$$\begin{aligned}
 & E(y_{ij}1_{(G_{Aij}=A_gA_h)}) \\
 &= w_{ij}\gamma E[1_{(G_{Aij}=A_gA_h)}] + E[g_{ij}1_{(G_{Aij}=A_gA_h)}] \\
 &= w_{ij}\gamma P(A_gA_h | G_{Aij} \neq ?) P(G_{Aij} \neq ?) \\
 &\quad + E[g_{ij}1_{(G_{Aij}=A_gA_h)} | G_{Aij} \neq ?] P(G_{Aij} \neq ?) \\
 &= (1 - \varepsilon_A) [w_{ij}\gamma \cdot 2P_{A_g}P_{A_h} + 2a(P(Q_1A_g)P(Q_1A_h) \\
 &\quad - P(Q_2A_g)P(Q_2A_h)) + d(2P(Q_1A_g)P(Q_2A_h) \\
 &\quad + 2P(Q_2A_g)P(Q_1A_h))] \\
 &= (1 - \varepsilon_A) [2P_{A_g}P_{A_h}w_{ij}\gamma + 2P_{A_g}P_{A_h}\mu \\
 &\quad + 2\alpha_Q(D_{A_gQ}P_{A_h} + D_{A_hQ}P_{A_g}) - 2\delta_Q D_{A_gQ}D_{A_hQ}]. \quad (19)
 \end{aligned}$$

Equating Eqs. 17 and 18, we show the Eq. 5 when $g \neq h$.

Appendix B

In relations (17), replacing β_{gh} by $\alpha_g + \alpha_h$ and taking summation lead to

$$\begin{aligned} E(y_{ij}1_{(G_{Aij} \neq ?)}) &= \sum_{1 \leq g \leq h \leq m} E(y_{ij}1_{(G_{Aij}=A_gA_h)}) \\ &= (1 - \varepsilon_A) \sum_{g=1}^m \sum_{h=1}^m (w_{ij}\gamma + \alpha_g + \alpha_h) P_{A_g} P_{A_h} \\ &= (1 - \varepsilon_A) \left(w_{ij}\gamma + 2 \sum_{g=1}^m \alpha_g P_{A_g} \right). \end{aligned}$$

Since the missing mechanism is missing completely at random, one has $E(y_{ij}1_{(G_{Aij} \neq ?)}) = E(y_{ij}|G_{Aij} \neq ?)(1 - \varepsilon_A) = (1 - \varepsilon_A)E y_{ij} = (1 - \varepsilon_A)(w_{ij}\gamma + \mu)$. Thus, $\sum_{g=1}^m \alpha_g P_{A_g} = \mu/2$.

Again, replacing β_{gh} by $\alpha_g + \alpha_h$ in relations (17) and taking summation with respect to h lead to

$$\begin{aligned} E \left[y_{ij}1_{(G_{Aij}=A_gA_g)} + \frac{1}{2} \sum_{h \neq g} y_{ij}1_{(G_{Aij}=A_gA_h)} \right] \\ &= (1 - \varepsilon_A) \sum_{h=1}^m (w_{ij}\gamma + \alpha_g + \alpha_h) P_{A_g} P_{A_h} \\ &= (1 - \varepsilon_A) P_{A_g} \left(w_{ij}\gamma + \alpha_g + \sum_{h=1}^m \alpha_h P_{A_h} \right) \\ &= (1 - \varepsilon_A) P_{A_g} (w_{ij}\gamma + \alpha_g + \mu/2). \end{aligned} \tag{20}$$

Notice $\sum_{g=1}^m D_{A_g} Q = 0$. Taking summation of relations (18) and (19) leads to

$$\begin{aligned} E \left[y_{ij}1_{(G_{Aij}=A_gA_g)} + \frac{1}{2} \sum_{h \neq g} y_{ij}1_{(G_{Aij}=A_gA_h)} \right] \\ &= (1 - \varepsilon_A) P_{A_g} [w_{ij}\gamma + \mu + D_{A_g} Q / P_{A_g}]. \end{aligned} \tag{21}$$

Equating the right-hand terms of relations (20) and (21) leads to (6).

Appendix C

Assume that there are no covariates, and the dataset is a population sample. Then the model matrix of “genotype effect model” (1) is $X_i = X_{Ail}^\tau = (x_{Ail}^{(11)}, \dots, x_{Ail}^{(nm)}, x_{Ail}^{(12)}, \dots, x_{Ail}^{(1m)}, \dots, x_{Ail}^{(m-1,m)})$, $i = 1, \dots, N$. To show non-centrality parameter approximation (7), we first notice the following relation

$$E[X_1^\tau X_1] = (1 - \varepsilon_A) \text{diag}(P_{A_1}^2, v^\tau) + \varepsilon_A \begin{pmatrix} P_{A_1}^2 \\ v \end{pmatrix} (P_{A_1}^2, v^\tau), \tag{22}$$

where v is a column vector given by $v^\tau = (P_{A_2}^2, \dots, P_{A_m}^2, 2P_{A_1}P_{A_2}, \dots, 2P_{A_1}P_{A_m}, \dots, 2P_{A_{m-1}}P_{A_m})$. In addition, $\text{diag}(P_{A_1}^2, v^\tau)$ is a diagonal matrix, whose elements on the diagonal are given by the elements of $(P_{A_1}^2, v^\tau)$. We may verify (22) by $E[(x_{A11}^{(gh)})^2] = E1_{(G_{A11}=A_gA_h)} + P(A_gA_h)^2 E1_{(G_{A11}=?)}$ and for

$$(g, h) \neq (k, l), E[x_{A11}^{(gh)} x_{A11}^{(kl)}] = P(A_gA_h)P(A_kA_l)E1_{(G_{A11}=?)} = P(A_gA_h)P(A_kA_l)\varepsilon_A.$$

Let us denote $u = (P_{A_2}^{-2}, \dots, P_{A_m}^{-2}, [2P_{A_1}P_{A_2}]^{-2}, \dots, [2P_{A_1}P_{A_m}]^{-2}, \dots, [2P_{A_{m-1}}P_{A_m}]^{-2})$. Applying the large number law and a fact of inverse matrix $(M + ab^\tau)^{-1} = M^{-1} - (M^{-1}a)(b^\tau M^{-1}) / (1 + b^\tau M^{-1}a)$, we can calculate the following approximation

$$\begin{aligned} T(X^\tau X)^{-1} T^\tau &\approx T[NE(X_1^\tau X_1)]^{-1} T^\tau \\ &= N^{-1} \cdot T \left[(1 - \varepsilon_A) \text{diag}(P_{A_1}^2, v^\tau) \right. \\ &\quad \left. + \varepsilon_A \begin{pmatrix} P_{A_1}^2 \\ v \end{pmatrix} (P_{A_1}^2, v^\tau) \right]^{-1} T^\tau \\ &= [(1 - \varepsilon_A)N]^{-1} \\ &\quad \cdot T \left[\text{diag}(P_{A_1}^{-2}, u^\tau) - \varepsilon_A \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} (1, 1, \dots, 1) \right] T^\tau \\ &= [(1 - \varepsilon_A)N]^{-1} \cdot T \text{diag}(P_{A_1}^{-2}, u^\tau) T^\tau. \end{aligned}$$

Utilizing above relation, we may show non-centrality parameter approximation (7) in the same way as Appendix III, Fan et al. (2006).

Appendix D

Assume that there are no covariates, and the dataset is a population sample. Then the model matrix of “additive effect model” (3) is $X_i = Z_{Ail}^\tau = (x_{Ail}^{(1)}, \dots, x_{Ail}^{(m)})$, $i = 1, \dots, N$. To show non-centrality parameter approximation (8), we first notice the following relation

$$\begin{aligned} E[Z_{A11} Z_{A11}^\tau] \\ &= 2(1 - \varepsilon_A) \left[\text{diag}(P_{A_1}, \dots, P_{A_m}) + \begin{pmatrix} P_{A_1} \\ \vdots \\ P_{A_m} \end{pmatrix} (P_{A_1}, \dots, P_{A_m}) \right] \\ &\quad + 4\varepsilon_A \begin{pmatrix} P_{A_1} \\ \vdots \\ P_{A_m} \end{pmatrix} (P_{A_1}, \dots, P_{A_m}), \end{aligned}$$

which can be verified by $E[(x_{A11}^{(g)})^2] = 4E1_{(G_{A11}=A_gA_g)} + \sum_{h \neq g} E1_{(G_{A11}=A_gA_h)} + 4P_{A_g}^2 E1_{(G_{A11}=?)}$ and for $h \neq g$, $E[x_{A11}^{(g)} x_{A11}^{(h)}] = (1 - \varepsilon_A) \cdot 2P_{A_g}P_{A_h} + 4P_{A_g}P_{A_h}\varepsilon_A$. Let $X = (Z_{A11}, \dots, Z_{AN1})^\tau$. Applying

the large number law and a fact of inverse matrix $(M + ab^\tau)^{-1} = M^{-1} - (M^{-1}a)(b^\tau M^{-1}) / (1 + b^\tau M^{-1}a)$, we can calculate the following approximation

$$\begin{aligned} &K(X^\tau X)^{-1}K^\tau \\ &\approx K[NE(Z_{A11}Z_{A11}^\tau)]^{-1}K^\tau \\ &= N^{-1} \cdot K \left[2(1 - \varepsilon_A)\text{diag}(P_{A_1}, \dots, P_{A_m}) + 2(1 + \varepsilon_A) \right. \\ &\quad \times \left. \begin{pmatrix} P_{A_1} \\ \vdots \\ P_{A_m} \end{pmatrix} (P_{A_1}, \dots, P_{A_m}) \right]^{-1} K^\tau \\ &= [2(1 - \varepsilon_A)N]^{-1} \\ &\quad \cdot K \left[\text{diag}(P_{A_1}^{-1}, \dots, P_{A_m}^{-1}) - (1 + \varepsilon_A) \right. \\ &\quad \times \left. \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} (1, 1, \dots, 1)/2 \right] K^\tau \\ &= [2(1 - \varepsilon_A)N]^{-1} \cdot K \text{diag}(P_{A_1}^{-1}, \dots, P_{A_m}^{-1}) K^\tau. \end{aligned}$$

Utilizing above relation, we may show non-centrality parameter approximation (8) in the same way as Appendix IV, Fan et al. (2006).

Appendix E

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 markers A and B . Here, $P(A_g B_k)$ is frequency of haplotype $A_g B_k$. It can be shown that for $g \neq h$, $k \neq l$, $h \neq h'$, $l \neq l'$, $(g, h) \neq (g', h')$, $(k, l) \neq (k', l')$

The quantities in (23) imply that the elements of V_A are given by

$$\begin{aligned} \text{Cov}(x_{Aij}^{(g)}, x_{Aij}^{(h)}) &= -2P_{A_g} P_{A_h} (1 - \varepsilon_A), \\ \text{Var}(x_{Aij}^{(g)}) &= 2P_{A_g} (1 - P_{A_g}) (1 - \varepsilon_A), \\ \text{Cov}(x_{Aij}^{(g)}, x_{Bij}^{(k)}) &= 2D_{A_g B_k} (1 - \varepsilon_A) (1 - \varepsilon_B), \\ \text{Cov}(x_{Bij}^{(k)}, x_{Bij}^{(l)}) &= -2P_{B_k} P_{B_l} (1 - \varepsilon_B), \\ \text{Var}(x_{Bij}^{(k)}) &= 2P_{B_k} (1 - P_{B_k}) (1 - \varepsilon_B). \end{aligned}$$

Since $E Z_{A \cup B}^{(ij)}$ is a vector of 0s by the quantities in (23), it can be shown that $V_D = \text{Cov}(Z_{A \cup B}^{(ij)}, Z_{A \cup B}^{(ij)}) = E(Z_{A \cup B}^{(ij)} (Z_{A \cup B}^{(ij)})^\tau)$. Moreover, the quantities in (23) imply that the covariance matrix $\text{Cov}(X_{A \cup B}^{(ij)}, Z_{A \cup B}^{(ij)})$ is a 0 matrix. In addition, the covariances between the trait value y_{ij} and variables $x_{Aij}^{(g)}, x_{Bij}^{(k)}, z_{Aij}^{(gh)}$ and $z_{Bij}^{(kl)}$ are

$$\begin{aligned} \text{Cov}(y_{ij}, x_{Aij}^{(g)}) &= 2\alpha_Q (1 - \varepsilon_A) D_{A_g Q}, \\ \text{Cov}(y_{ij}, x_{Bij}^{(k)}) &= 2\alpha_Q (1 - \varepsilon_B) D_{B_k Q}, \\ \text{Cov}(y_{ij}, z_{Aij}^{(gh)}) &= E[y_{ij} z_{Aij}^{(gh)}], \text{Cov}(y_{ij}, z_{Bij}^{(kl)}) = E[y_{ij} z_{Bij}^{(kl)}]. \end{aligned}$$

Taking variance–covariance between y_{ij} and $x_{Aij}^{(g)}, x_{Bij}^{(k)}, z_{Aij}^{(gh)}, z_{Bij}^{(kl)}$ based on relation (12), we may get the regression coefficients (13) of models (10) and (12).

$$\begin{aligned} E x_{Aij}^{(g)} &= 2P_{A_g}, E(x_{Aij}^{(g)})^2 = (1 - \varepsilon_A)(2P_{A_g}^2 + 2P_{A_g}) + 4P_{A_g}^2 \varepsilon_A, E[x_{Aij}^{(g)} x_{Aij}^{(h)}] = 2P_{A_g} P_{A_h} (1 - \varepsilon_A) + 4P_{A_g} P_{A_h} \varepsilon_A, \\ E x_{Bij}^{(k)} &= 2P_{B_k}, E(x_{Bij}^{(k)})^2 = (1 - \varepsilon_B)(2P_{B_k}^2 + 2P_{B_k}) + 4P_{B_k}^2 \varepsilon_B, E[x_{Bij}^{(k)} x_{Bij}^{(l)}] = 2P_{B_k} P_{B_l} (1 - \varepsilon_B) + 4P_{B_k} P_{B_l} \varepsilon_B, \\ E z_{Aij}^{(gh)} &= 0, E(z_{Aij}^{(gh)})^2 = (1 - \varepsilon_A)P_{A_g}^2 P_{A_h}^2 [P_{A_g} + P_{A_h}]^2, E z_{Bij}^{(kl)} = 0, E(z_{Bij}^{(kl)})^2 = (1 - \varepsilon_B)P_{B_k}^2 P_{B_l}^2 [P_{B_k} + P_{B_l}]^2, \\ E[x_{Aij}^{(g)} z_{Aij}^{(gh)}] &= E[x_{Aij}^{(g)} z_{Aij}^{(hh')}] = E[x_{Bij}^{(k)} z_{Bij}^{(kl)}] = E[x_{Bij}^{(k)} z_{Bij}^{(ll')}] = E[x_{Aij}^{(g)} z_{Bij}^{(kl)}] = E[x_{Bij}^{(k)} z_{Aij}^{(gh)}] = 0, \\ E[x_{Aij}^{(g)} x_{Bij}^{(k)}] &= 2D_{A_g B_k} (1 - \varepsilon_A) (1 - \varepsilon_B) + 4P_{A_g} P_{B_k} \varepsilon_B, E[z_{Aij}^{(gh)} z_{Aij}^{(gh')}] = (P_{A_g} P_{A_h} P_{A_h}')^2 (1 - \varepsilon_A), \\ E[z_{Aij}^{(gh)} z_{Aij}^{(g'h')}] &= 0, E[z_{Bij}^{(kl)} z_{Bij}^{(k'l')}] = (P_{B_k} P_{B_l} P_{B_l}')^2 (1 - \varepsilon_B), E[z_{Bij}^{(kl)} z_{Bij}^{(k'l')}] = 0, \\ E[z_{Aij}^{(gh)} z_{Bij}^{(kl)}] &= [P_{A_h} (P_{B_l} D_{A_g B_k} - P_{B_k} D_{A_g B_l}) - P_{A_g} (P_{B_l} D_{A_h B_k} - P_{B_k} D_{A_h B_l})]^2 (1 - \varepsilon_A) (1 - \varepsilon_B), \\ E[y_{ij} x_{Aij}^{(g)}] &= 2P_{A_g} (w_{ij} \gamma + \mu) + 2\alpha_Q D_{A_g Q} (1 - \varepsilon_A), E[y_{ij} x_{Bij}^{(k)}] = 2P_{B_k} (w_{ij} \gamma + \mu) + 2\alpha_Q D_{B_k Q} (1 - \varepsilon_B), \\ E[y_{ij} z_{Aij}^{(gh)}] &= \delta_Q [P_{A_g} D_{A_h Q} - P_{A_h} D_{A_g Q}]^2 (1 - \varepsilon_A), E[y_{ij} z_{Bij}^{(kl)}] = \delta_Q [P_{B_k} D_{B_l Q} - P_{B_l} D_{B_k Q}]^2 (1 - \varepsilon_B). \end{aligned} \tag{23}$$

Appendix F

Notice $\Sigma_i^{-1} = \frac{1}{\sigma^2} (\gamma_{hj})_{(s+2) \times (s+2)}$. Let X_i be the model matrix of family $i = 1, 2, \dots, I$. Then

$$X_i = \begin{pmatrix} 1 & x_{Ai1}^{(1)} & \dots & x_{Ai1}^{(m-1)} & x_{Bi1}^{(1)} & \dots & x_{Bi1}^{(n-1)} & z_{Ai1}^{(12)} & \dots & z_{Ai1}^{(m-1,m)} & z_{Bi1}^{(12)} & \dots & z_{Bi1}^{(n-1,n)} \\ 1 & x_{Ai2}^{(1)} & \dots & x_{Ai2}^{(m-1)} & x_{Bi2}^{(1)} & \dots & x_{Bi2}^{(n-1)} & z_{Ai2}^{(12)} & \dots & z_{Ai2}^{(m-1,m)} & z_{Bi2}^{(12)} & \dots & z_{Bi2}^{(n-1,n)} \\ \vdots & \vdots & \dots & \vdots & \vdots & \dots & \vdots & \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{Ai,s+2}^{(1)} & \dots & x_{Ai,s+2}^{(m-1)} & x_{Bi,s+2}^{(1)} & \dots & x_{Bi,s+2}^{(n-1)} & z_{Ai,s+2}^{(12)} & \dots & z_{Ai,s+2}^{(m-1,m)} & z_{Bi,s+2}^{(12)} & \dots & z_{Bi,s+2}^{(n-1,n)} \end{pmatrix}.$$

Denote $\gamma = \sum_{k=1}^{s+2} \sum_{l=1}^{s+2} \gamma_{kl}$. Applying large number law leads to an approximation as

$$\sum_{i=1}^I X_i^T \Sigma_i^{-1} X_i / I \approx \frac{1}{\sigma^2} \begin{pmatrix} \gamma & \gamma [E(X_{AUB}^{(11)})]^T & O_1 \\ \gamma E(X_{AUB}^{(11)}) & \sum_{k=1}^{s+2} \gamma_{kk} V_A + bV_{A2} + \gamma E(X_{AUB}^{(11)}) [E(X_{AUB}^{(11)})]^T & O_2 \\ O_3 & O_4 & \sum_{k=1}^{s+2} \gamma_{kk} V_D + \sum_{k=3}^{s+2} \sum_{l=k+1}^{s+2} \gamma_{kl} V_{D2} / 2 \end{pmatrix}, \tag{24}$$

where $O_i, i = 1,2,3,4$ are zero vectors or matrices, and $E(X_{AUB}^{(11)}) = (2P_{A1}, \dots, 2P_{A_{m-1}}, 2P_{B1}, \dots, 2P_{B_{n-1}})^T$. Let

$$S = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}$$

be the test matrix corresponding to hypothesis H_{ABad0} , and $\phi = (\alpha, \alpha_{A1}, \dots, \alpha_{A(m-1)}, \alpha_{B1}, \dots, \alpha_{B(m-1)}, \delta_{A12}, \dots, \delta_{A(m-1)m}, \delta_{B12}, \dots, \delta_{B(n-1)n})^T$ be the column vector of regression coefficient of “genotype effect model” (12). Utilizing regression coefficients (13), we may show (15) by plugging approximation (24) into $\lambda_{ABad} = (S\phi)^T [S(\sum_{i=1}^I X_i^T \Sigma_i^{-1} X_i)^{-1} S^T]^{-1} (S\phi)$. One may want to notice that we may use Theorem 8.5.11, Harville

(1997), to calculate the inverse of the right-hand matrix of (24).

Appendix G

For pedigrees in graph A of Fig. 1, the constants b_1 and b_2 of $\lambda_{AB,ad}$ in (16) are given by

$$b_1 = [\gamma_{15} + (\gamma_{17} + \dots + \gamma_{1,11})/2] + [\gamma_{25} + (\gamma_{27} + \dots + \gamma_{2,11})/2] + [\gamma_{36} + (\gamma_{37} + \dots + \gamma_{3,11})/2] + [\gamma_{46} + (\gamma_{47} + \dots + \gamma_{4,11})/2] + (\gamma_{57} + \dots + \gamma_{5,11}) + (\gamma_{67} + \dots + \gamma_{6,11}) + \sum_{k=7}^{11} \sum_{l=k+1}^{11} \gamma_{kl},$$

$$b_2 = \sum_{k=7}^{11} \sum_{l=k+1}^{11} \gamma_{kl} / 2.$$

For pedigrees in graph B of Fig. 1, constants b_1 and b_2 are given by

$$\begin{aligned}
b_1 = & \gamma_{1,12} + [\gamma_{2,12} + (\gamma_{2,13} + \dots + \gamma_{2,16})/2] \\
& + [\gamma_{3,12} + \dots + \gamma_{3,16}]/2 \\
& + [\gamma_{4,12} + \dots + \gamma_{4,16}]/2 \\
& + [\gamma_{5,12}/2 + (\gamma_{5,13} + \dots + \gamma_{5,16})] \\
& + [(\gamma_{6,13} + \dots + \gamma_{6,16}) + (\gamma_{6,17} + \gamma_{6,18})/2] \\
& + [\gamma_{7,13} + \dots + \gamma_{7,18}]/2 \\
& + [(\gamma_{8,13} + \dots + \gamma_{8,16})/2 + (\gamma_{8,17} + \gamma_{8,18})] \\
& + (\gamma_{9,17} + \gamma_{9,18}) + (\gamma_{10,17} + \gamma_{10,18})/2 \\
& + (\gamma_{11,17} + \gamma_{11,18})/2 + (\gamma_{12,13} + \dots + \gamma_{12,16})/4 \\
& + (\gamma_{13,14} + \gamma_{13,15} + \gamma_{13,16}) \\
& + (\gamma_{14,15} + \gamma_{14,16}) + \gamma_{15,16} + [\gamma_{13,17} + \dots + \gamma_{16,17}]/4 \\
& + [\gamma_{13,18} + \dots + \gamma_{16,18}]/4 + \gamma_{17,18}, \\
b_2 = & [(\gamma_{13,14} + \gamma_{13,15} + \gamma_{13,16}) + (\gamma_{14,15} + \gamma_{14,16}) \\
& + \gamma_{15,16}]/2 + \gamma_{17,18}/2.
\end{aligned}$$

References

- Abecasis GR, Cardon LR, Cookson WOC (2000a). A general test of association for quantitative traits in nuclear families. *Am J Hum Genet* 66:279–292
- Abecasis GR, Cherny SS, Cookson WOC, Cardon LR (2002) Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 30:97–101
- Abecasis GR, Cookson WOC, Cardon LR (2000b). Pedigree tests of linkage disequilibrium. *Eur J Hum Genet* 8:545–551
- Allison DB (2001) Joint tests of linkage and association for quantitative traits. *Theor Popul Biol* 60:239–251
- Almasy L, Blangero J (1998) Multipoint quantitative trait linkage analysis in general pedigrees. *Am J Hum Genet* 62:1198–1211
- Almasy L, Williams JT, Dyer TD, Blangero J (1999) Quantitative trait locus detection using combined linkage/disequilibrium analysis. *Genet Epidemiol* 17(Suppl 1):S31–S36
- Amos CI (1994) Robust variance-components approach for assessing linkage in pedigrees. *Am J Hum Genet* 54:534–543
- Amos CI, Elston RC (1989) Robust methods for the detection of genetic linkage for quantitative data from pedigrees. *Genet Epidemiol* 6:349–360
- Boerwinkle E, Chakraborty E, Sing CF (1986) The use of measured genotype information in the analysis of quantitative phenotype in man. I. models and analytical methods. *Ann Hum Genet* 50:181–194
- Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th edn. Longman, London
- Fan RZ, Jung JS (2003) High resolution joint linkage disequilibrium and linkage mapping of quantitative trait loci based on sibship data. *Hum Hered* 56:166–187
- Fan RZ, Jung JS, Jin J (2006) High resolution association mapping of quantitative trait loci, a population based approach. *Genetics* 172:663–686
- 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, Xiong MM (2003) Combined high resolution linkage and association mapping of quantitative trait loci. *Eur J Hum Genet* 11:125–137
- Farrall M, Keavney B, MckKenzie CA, Delèpine M, Matsuda F, Lathrop GM (1999) Fine mapping of an ancestral recombination break-point in DCPI. *Nat Genet* 23:270–271
- Feingold E (2002) Invited editorial: regression-based quantitative-trait-locus mapping in the 21st century. *Am J Hum Genet* 71:217–222
- Fulker DW, Cherny SS, Cardon LR (1995) Multiple interval mapping of quantitative trait loci, using sib-pairs. *Am J Hum Genet* 56:1224–1233
- 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
- George V, Tiwari HK, Zhu XF, Elston RC (1999) A test of transmission/disequilibrium for quantitative traits in pedigree data, by multiple regression. *Am J Hum Genet* 65:236–245
- Goldgar DE (1990) Multipoint analysis of human quantitative genetic variation. *Am J Hum Genet* 47:957–967
- Graybill FA (1976) Theory and application of the linear model. Pacific Grove, California
- Harville DA (1997) Matrix algebra from a statistician's perspective. Springer
- Haseman JK, Elston RC (1972) The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* 2:3–19
- Hedrick PW (1987) Gametic disequilibrium measures: proceed with caution. *Genetics* 117:331–341
- Jung JS, Fan RZ, Jin L (2005) Combined linkage and association mapping of quantitative trait loci by multiple markers. *Genetics* 170:881–898
- Keavney B, MckKenzie CA, Connell JM, Julier C, Ratcliffe PJ, Sobel E, Lathrop M, Farrall M (1998) Measured haplotype analysis of the angiotensin-I converting enzyme gene. *Hum Mol Genet* 7:1745–1751
- Lange K (2002) Mathematical and Statistical methods for genetic analysis, 2nd edn. Springer
- Li M, Boehnke M, Abecasis GR (2005) Joint modeling of linkage and association: identifying SNPs responsible for a linkage signal. *Am J Hum Genet* 76:934–49
- Little RJA, Rubin DB (2002) Statistical analysis with missing data, 2nd edn. Wiley Inter-Science, Wiley, Inc., Publication
- Pinheiro JC, Bates DM (2000) Mixed-effects models in S and S-plus. Springer
- Pratt SC, Daly M, Kruglyak L (2000) Exact multipoint quantitative-trait linkage analysis in pedigrees by variance components. *Am J Hum Genet* 66:1153–1157
- Sham PC, Cherny SS, Purcell S, Hewitt JK (2000) Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *Am J Hum Genet* 66:1616–1630
- Wang T, Elston RC (2005) The bias introduced by population stratification in IBD based linkage analysis. *Hum Hered* 60:134–142
- Xiong MM, Jin L (2000) Combined linkage and linkage disequilibrium mapping for genome screens. *Genet Epidemiol* 19:211–234