

# Association Studies of QTL for Multi-Allele Markers by Mixed Models

Ruzong Fan<sup>a,b</sup> Jeusun Jung<sup>a</sup><sup>a</sup>Department of Statistics, Texas A&M University, College Station, Tex., USA; <sup>b</sup>Institute of Medical Biometry, Informatics and Epidemiology, University of Bonn, Bonn, Germany

---

## Key Words

Linkage analysis · Association study · Mixed models · Quantitative trait locus

---

## Abstract

In this paper, we extend association study methods of both Fan et al. [Hum Hered 2002;53:130–145], in which a quantitative trait locus (QTL) and a multi-allele marker are considered for trio families, and Fan and Xiong [Biostatistics 2003, in press], in which a QTL and a bi-allelic marker are considered for nuclear families. The objective is to build mixed models for association study between a QTL and a multi-allelic marker for nuclear families with any number of offspring. Two types of nuclear family data are considered: the first is genetic data of offspring from at least one heterozygous parents, and the second is genetic data of offspring of nuclear family. (1) For the data of offspring from at least one heterozygous parents, we assume that at least one parent is heterozygous at the marker locus, and we may infer clearly the transmission of parental marker alleles to the offspring. We show that it can be used in association study in the presence of linkage. The theoretical basis is the difference between the conditional mean of trait value given an allele is transmitted and the conditional mean of trait value given the allele is not transmitted from a heterozygous parent. To build valid models, we calculate the variance covariance structure of trait values of offspring. Besides, the reduc-

tion of the number of parameters is discussed under an assumption of tight linkage between the trait locus and the marker. (2) For the data of offspring of nuclear family, we show that it can be used in general association study. In this case, the theoretical basis is the difference between the conditional mean of trait values given an allele is transmitted from a parent and the population mean. Then, we calculate variance-covariance structure of trait values of offspring. (3) Based on the theoretical analysis, mixed models are built for each type of the data, and related test statistics are proposed for association study. By power calculation and comparison, we show that, in some instances, the proposed test statistics have higher power than that by collapsing alleles to be new ones. The proposed models are used to analyze chromosomes 4 and chromosome 16 data of the Oxford asthma data, Genetic Analysis Workshop 12.

Copyright © 2002 S. Karger AG, Basel

## 1. Introduction

In recent years, there has been a growing interest in linkage analysis and association study or linkage disequilibrium (LD) mapping between a quantitative trait locus (QTL) and a marker locus. Allison [1997] and Xiong et al. [1998] proposed transmission disequilibrium tests (TDT) for mapping QTL. Rabinowitz [1997] performed simulation study of QTL. Based on variance component models,

---

## KARGER

Fax +41 61 306 12 34  
E-Mail [karger@karger.ch](mailto:karger@karger.ch)  
[www.karger.com](http://www.karger.com)© 2002 S. Karger AG, Basel  
0001-5652/02/0543-0132\$18.50/0Accessible online at:  
[www.karger.com/hhe](http://www.karger.com/hhe)Dr. Ruzong Fan  
Department of Statistics, Texas A&M University  
447 Blocker Building  
College Station, TX 77843-3143 (USA)  
Tel. +1 979 845 3156, Fax +1 979 845 3144, E-Mail [rfan@stat.tamu.edu](mailto:rfan@stat.tamu.edu)

Abecasis et al. [2000], Fulker et al. [1999] and Sham et al. [2000] explored linkage and association study between a QTL and a bi-allelic marker. George et al. [1999] proposed linear regression models for TDT analysis of quantitative traits for general pedigrees. Conditional on the parental traits, Zhu and Elston [2000, 2001] extended the method of George et al. [1999], and proposed better test statistics in detecting linkage and association. Using a bi-allelic marker, Fan and Xiong [2003] proposed mixed models for linkage and association studies of QTL for nuclear families with any number of offspring. However, almost all these investigations used bi-allele markers in the analysis. In certain circumstances, one may have to deal with multiple allele markers such as micro-satellites. One may collapse a multiple allele marker to be a bi-allelic marker in the study, but this may not be a good method since much information may be lost by collapsing different alleles to be new alleles. Moreover, different ways to collapse a multiple allele marker can lead to different results which make the interpretation of results a complicated task. There is a need to extend the bi-allelic marker method to fit multi-allele markers, and to perform composite tests for the association study.

Fan et al. [2003] investigated models and composite tests to detect association and linkage between a QTL and a multi-allele marker locus for trio families. Here trio families mean that the families have two parents and one single offspring. The methods of Fan et al. [2002] are not valid for general nuclear families with more than one offspring, since the methods do not consider the correlation of offspring's trait values. For a nuclear family with more than one offspring, the trait values of the offspring are not independent. To build valid test statistics and models, one needs to consider the variance-covariance structure of the trait values of offspring in nuclear families, in addition to the mean structure under the normal assumption.

In this paper, mixed models are proposed for two types of nuclear family data in the association study between a QTL and a multiple allele marker. The first is genetic data of offspring from at least one heterozygous parent, and the second is the genetic data of offspring of nuclear families. In a similar manner as those in Fan et al. [2002], Fan and Xiong [2003], we calculate the conditional mean and conditional variance-covariance matrix of trait values of the offspring for each type of the data. For multiple allele marker, the number of parameters can be too large in using the data of offspring from at least one heterozygous parent [Fan et al. 2002; Sham and Curtis 1995]. One needs to consider appropriate method to reduce the number of parameters, and build valid models to fit data.

Under the assumption of tight linkage between the trait locus and the marker, we show that the number of parameters can be significantly reduced by approximations. Based on the two types of the data and the related conditional mean and conditional variance-covariance structures, mixed models and test statistics are introduced for association study. The non-centrality parameters of the test statistics are calculated based on the theory of linear statistical analysis of mixed models. By power calculation and comparison, the merit of the proposed methods and test statistics is presented by graphical methods. In addition, chromosome 4 and chromosome 16 data of the Oxford asthma data, Genetic Analysis Workshop 12, are analyzed by using the models and methods proposed in the paper [Cookson and Abecasis 2001; Daniel et al. 1996].

## 2. Mean and Variance-Covariance Structures: Heterozygous Parent Data

Consider one QTL  $Q$  with 2 alleles  $Q_1$  and  $Q_2$  occurring with frequencies  $q_1$  and  $q_2$ , respectively. Assume that the expected phenotypic trait value of a person with genotype  $Q_r Q_s$  is  $v + \mu_{rs}$ ,  $r, s = 1, 2$ , respectively, where  $v$  is overall mean and  $\mu_{rs}$  is the effect of genotype  $Q_r Q_s$ . Obviously,  $\mu_{12} = \mu_{21}$ . Suppose that a marker locus  $M$  is linked to the trait locus  $Q$ . Denote the recombination fraction between the marker locus  $M$  and the trait locus  $Q$  by  $\theta$ . Assume that  $m$  alleles  $M_1, \dots, M_m$  are typed at the marker locus  $M$  occurring with frequencies  $p_1, \dots, p_m$ . The haplotype frequency is denoted by  $h_{ri}$  for haplotype  $Q_r M_i$ ,  $r = 1, 2, i = 1, \dots, m$ . If  $h_{ri} = q_r p_i$  for all  $r$  and  $i$ , then the trait locus  $Q$  and the marker  $M$  are in linkage equilibrium. Otherwise, the trait locus  $Q$  and the marker  $M$  are in LD or association. The measure of LD between the trait allele  $Q_1$  and the marker allele  $M_i$  is defined by  $\delta_i = h_{1i} - q_1 p_i$ ,  $i = 1, \dots, m$ . Since  $\sum_{i=1}^m \delta_i = 0$ , one of  $\delta_1, \dots, \delta_m$  can be expressed by others, e.g.,  $\delta_m = -\sum_{i=1}^{m-1} \delta_i$ . Let  $Y$  be the phenotypic trait variable, which is decomposed into  $Y = v + g + G + e$ , where  $v$  is overall mean,  $g$  is random major gene effect,  $G$  is polygenic effect which is distributed as normal  $N(0, \sigma_G^2)$ , and  $e$  is sampling error which is distributed as normal  $N(0, \sigma_e^2)$ .  $g, G$  and  $e$  are independent of each other. If an individual has genotype  $Q_s Q_r$  at the trait locus, then  $E(g|Q_s Q_r) = \mu_{rs}$ . Let  $TQ$  denote the abbreviation of 'transmitted quantitative trait allele'. Then we have the conditional mean  $E[Y|TQ = Q_r] = v + \sum_{s=1}^2 \mu_{rs} q_s = v + \mu_r$ . Let  $P(M_i, M_j)$  be the probability of an offspring who receives marker allele  $M_i$  from his/her heterozygous parent but not allele  $M_j$ . Then

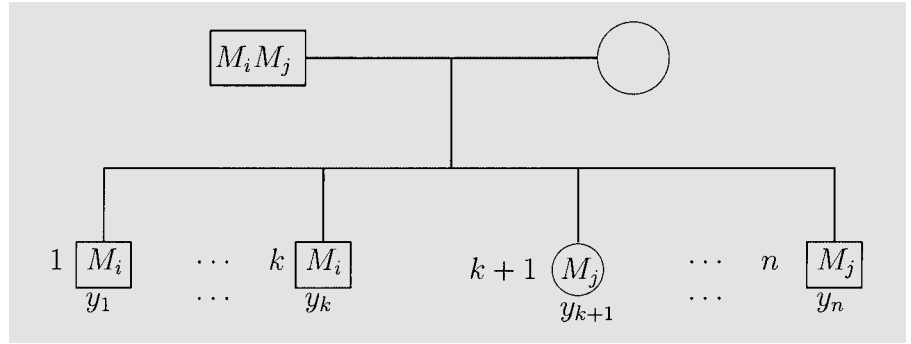


Fig. 1. A nuclear family with  $n$  offspring. Assume that the genotype of the father at the marker locus is heterozygous  $M_i M_j$ ,  $i \neq j$ . Moreover, the father transmits allele  $M$  to kids 1, ...,  $k$ , and transmits allele  $M_j$  to kids  $k + 1, \dots, n$ .

$P(M_i, M_j) = P(M_j, M_i) = p_i p_j$ . Let  $P(Q_r M_i, M_j)$  be the probability of a child who receives haplotype  $Q_r M_i$  from his/her heterozygous parent but not allele  $M_j$ . Then we have a relation  $P(Q_r M_i, M_j) = (1 - \theta) h_{ri} p_j + \theta h_{rj} p_i$ .

### 2.1. Mean and Variance-Covariance Structures

For a family with two parents and at least one offspring, we assume that at least one parent is heterozygous at the marker locus  $M$ . Moreover, assume we may infer clearly the transmission of parental marker alleles to the offspring. If both parents and an offspring have the same genotype  $M_i M_j$ ,  $i \neq j$ , it is impossible to tell which parent transmits which allele to the offspring, and hence the data can not be used in analysis. Actually, this is the only type of data which needs to be excluded. For a bi-allelic marker, one needs to exclude the heterozygous offspring of a mating heterozygous  $\times$  heterozygous [Fan and Xiong 2003; George et al. 1999; Zhu and Elston 2000, 2001]. For a multi-allelic marker, any offspring from a mating  $M_i M_i \times M_j M_k$ ,  $j \neq k$  or  $M_i M_j \times M_l M_k$ ,  $i \neq j$ ,  $i \neq k$ ,  $j \neq k$  or a mating  $M_i M_j \times M_l M_k$ ,  $i \neq j$ ,  $i \neq l$ ,  $i \neq k$ ,  $j \neq l$ ,  $j \neq k$ ,  $l \neq k$  can be included in analysis since one can infer clearly the transmission of parental marker alleles to the offspring. Hence, a heterozygous offspring of a mating heterozygous  $\times$  heterozygous may not necessarily be excluded in multi-allelic marker case unless both parents and offspring have exactly the same heterozygous genotype.

Let us look at a pedigree depicted in figure 1. Assume that the genotype of the father at the marker locus is heterozygous  $M_i M_j$ ,  $i \neq j$ . Moreover, the father transmits allele  $M_i$  to children 1, ...,  $k$ , and transmits allele  $M_j$  to children  $k + 1, \dots, n$ . For child  $i$ , the quantitative trait value is denoted by  $y_i$ ,  $i = 1, \dots, n$ . Let  $TM$  denote the abbreviation of 'transmitted marker allele', and  $NM$  of 'non-transmitted marker allele'. Given that marker allele  $M_i$  is transmitted and allele  $M_j$  is not transmitted from the heterozygous father for children 1, ...,  $k$ , the conditional

expected mean can be calculated in the same way as equation (1) or (2) of Fan and Xiong [2003],

$$\begin{aligned} \alpha_{i,j} &= E[Y | TM = M_i, NM = M_j] \\ &= v + \sum_{r=1}^2 \mu_r [(1 - \theta) h_{ri} p_j + \theta h_{rj} p_i] / [p_i p_j]. \end{aligned} \quad (1)$$

Similarly, the conditional expected mean of the children  $k + 1, \dots, n$  in figure 1 is

$$\begin{aligned} \alpha_{j,i} &= E[Y | TM = M_j, NM = M_i] \\ &= v + \sum_{r=1}^2 \mu_r [(1 - \theta) h_{rj} p_i + \theta h_{ri} p_j] / [p_i p_j]. \end{aligned} \quad (2)$$

Since  $h_{2i} p_j - h_{2j} p_i = (p_i - h_{1i}) p_j - (p_j - h_{1j}) p_i = -h_{1i} p_j + h_{1j} p_i$ , equations (1) and (2) lead to a difference between  $\alpha_{i,j}$  and  $\alpha_{j,i}$

$$\begin{aligned} \alpha_{i,j} - \alpha_{j,i} &= (1 - 2\theta) \sum_{r=1}^2 \mu_r (h_{ri} p_j - h_{rj} p_i) / (p_i p_j) \\ &= (1 - 2\theta) (\mu_1 - \mu_2) (h_{1i} p_j - h_{1j} p_i) / (p_i p_j) \\ &= (1 - 2\theta) (\mu_1 - \mu_2) (\delta_i p_j - \delta_j p_i) / (p_i p_j). \end{aligned} \quad (3)$$

Assume that the trait locus  $Q$  is linked to the marker locus  $M$ , i.e.,  $0 \leq \theta < 1/2$ . Then, one may construct statistics and models to test  $\delta_i p_j - \delta_j p_i \neq 0$ . Interestingly,  $\delta_i p_j - \delta_j p_i \neq 0$  implies that at least one of  $\delta_i$  and  $\delta_j$  is not equal to 0. That is to say that the marker  $M$  is in LD with trait locus  $Q$ . Hence, one may construct statistics and models to test association in the presence of linkage between the marker  $M$  and the trait locus  $Q$  based on the difference (3).

To build valid test statistics and models, we need to calculate the variance-covariances of the trait values of offspring in nuclear families. In a similar manner as Appendix A of Fan and Xiong [2003], we may show that the conditional variance of trait value of the children 1, ...,  $k$  is  $\sigma_{i,j}^2 = \sigma_e^2 + \sigma_G^2 + \Sigma_{i,j}^2$ , where  $\Sigma_{i,j}^2 = \Sigma_{r=1}^2 \Sigma_{s=1}^2 (v + \mu_{rs} - \alpha_{i,j})^2 q_s P(Q_r M_i, M_j) / P(M_i, M_j)$ . Similarly, the conditional variance of trait values of the children  $k + 1, \dots, n$  is

$\sigma_{j,i}^2 = \sigma_e^2 + \sigma_G^2 + \Sigma_{j,i}^2$ , where  $\Sigma_{j,i}^2 = \Sigma_{r=1}^2 \Sigma_{s=1}^2 (v + \mu_{rs} - \alpha_{j,i})^2 q_s P(Q_r M_j, M_i) / P(M_j, M_i)$ . For the conditional covariances, let us denote the expected conditional covariance between  $y_l (l = 1, \dots, k)$  and  $y_t (t \neq l, t = 1, \dots, k)$  by  $\Sigma_{ij,ij}$ , the expected conditional covariance between  $y_l (l = 1, \dots, k)$  and  $y_t (t = k + 1, \dots, n)$  as  $\Sigma_{ij,ji} = \Sigma_{ji,ij}$ , and the expected conditional covariance between  $y_l (l = k + 1, \dots, n)$  and  $y_t (t \neq l, t = k + 1, \dots, n)$  as  $\Sigma_{ji,ji}$ . Such as in Appendix B in Fan and Xiong [2003], we may calculate  $\Sigma_{ij,ij}$  and  $\Sigma_{ij,ji}$  which are given in Appendix A.

On the other hand, we need to build a model under the null hypothesis of no association in the presence of linkage. To do this, we need to calculate the mean and variance-covariance parameters. Under the assumption of linkage equilibrium between the marker locus and the trait locus, we show in Appendix B that  $\alpha_{i,j} = \Sigma_{r=1}^2 (v + \mu_r) q_r = v + \mu = \alpha$ ,  $\sigma_{i,j}^2 = \sigma^2$ ,  $\Sigma_{ij,ij} = \Sigma_{ts}$  and  $\Sigma_{ij,ji} = \Sigma_{td}$ , which do not depend on subscripts  $i$  and  $j$  ( $\sigma^2$ ,  $\Sigma_{ts}$  and  $\Sigma_{td}$  are given in Appendix B).

In figure 1, we depict the transmission of alleles from the heterozygous father to the offspring. Based on the transmission, we calculate the conditional expectations  $\alpha_{i,j}$ ,  $\alpha_{j,i}$ , their difference and the related conditional variance-covariances. For the mother, we can perform similar analysis. If the mother is homozygous  $M_i M_i$ , every offspring receives an allele  $M_i$  from her and so she does not provide useful information [Spielman et al. 1993]. If the mother is heterozygous, one should examine if an allele is transmitted to an offspring by the mother. Keeping all offspring with whom one may infer clearly the transmission of allele from the mother, we may repeat the same process for the mother-offspring pairs as the above for the father-offspring pairs.

## 2.2. Parameter Reductions

In Subsection 2.1, we work out the mean and variance-covariance structures of siblings for a nuclear family. Although the structure is valid theoretically, the number of parameters can be very large for a multi-allele marker  $M$ . The number of mean parameters  $\alpha_{i,j}$  is  $m(m - 1)$ , and the number of variance-covariances  $\sigma_{i,j}^2$ ,  $\Sigma_{ij,ij}$ ,  $\Sigma_{ij,ji}$  is  $5[m(m - 1)/2]$  for a marker  $M$  with  $m$  alleles. Hence, the total number of the parameters is  $7m(m - 1)/2$ . For a marker with 3 alleles, the number of parameters is 21; for a marker with 4 alleles, the number of parameters is 42. One needs to reduce the number of parameters to build valid models and obtain robust test statistics.

In a population, the presence of LD is usually the result of tight linkage between a trait locus and a marker locus [Falconer and Mackay 1996; Fan et al. 2002; Sham and

Curtis 1995]. Assume that the recombination fraction  $\theta \approx 0$ , i.e. there is tight linkage between the trait locus and the marker. In Appendix C, we show that approximately  $\alpha_{i,j} \approx \alpha_i$ ,  $\sigma_{i,j}^2 \approx \sigma_i^2$  and  $\Sigma_{ij,ij} \approx \Sigma_{i,i}$  only depend on subscript  $i$ , and the covariance  $\Sigma_{ij,ji} \approx \Sigma_{i,j} = \Sigma_{j,i}$  depends on both  $i$  and  $j$ . Therefore, the expected conditional variance-covariance matrix of  $y_l, l = 1, \dots, n$ , in figure 1 can be expressed as

$$\begin{pmatrix} \sigma_{i,j}^2 & \Sigma_{ij,ij} & \dots & \Sigma_{ij,ij} & \Sigma_{ij,ji} & \dots & \Sigma_{ij,ji} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \Sigma_{ij,ij} & \Sigma_{ij,ij} & \dots & \sigma_{i,j}^2 & \Sigma_{ij,ji} & \dots & \Sigma_{ij,ji} \\ \Sigma_{ji,ij} & \Sigma_{ji,ij} & \dots & \Sigma_{ji,ij} & \sigma_{j,i}^2 & \dots & \Sigma_{ji,ji} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \Sigma_{ji,ij} & \Sigma_{ji,ij} & \dots & \Sigma_{ji,ij} & \Sigma_{ji,ji} & \dots & \sigma_{j,i}^2 \end{pmatrix} \approx \begin{pmatrix} \sigma_i^2 & \Sigma_{i,i} & \dots & \Sigma_{i,i} & \Sigma_{i,j} & \dots & \Sigma_{i,j} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \Sigma_{i,i} & \Sigma_{i,i} & \dots & \sigma_i^2 & \Sigma_{i,j} & \dots & \Sigma_{i,j} \\ \Sigma_{j,i} & \Sigma_{j,i} & \dots & \Sigma_{j,i} & \sigma_j^2 & \dots & \Sigma_{j,j} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \Sigma_{j,i} & \Sigma_{j,i} & \dots & \Sigma_{j,i} & \Sigma_{j,j} & \dots & \sigma_j^2 \end{pmatrix}.$$

With these parameter reductions, the number of mean parameters  $\alpha_i$  is  $m$ , and the number of variance-covariance parameters  $\sigma_i^2$ ,  $\Sigma_{i,i}$ ,  $\Sigma_{i,j}$  is  $2m + m(m - 1)/2$ . Hence, the total number of the parameters is  $3m + m(m - 1)/2$ . Such as in Fan and Xiong [2003], the number of parameters for a bi-allele marker is 7. For a marker with 3 alleles, the number of parameters is 12, and for a marker with 4 alleles, the number of parameters is 18. Therefore, the number of parameters can be significantly reduced under the assumption of tight linkage between the trait locus and the marker.

## 3. Mean and Variance-Covariance Structures: General Nuclear Family Data

Consider a sample of nuclear families. For each family, suppose there are two parents and at least one offspring. For each parent-offspring pair, one first determines which allele is transmitted from the parent to the offspring. Here we use a different strategy from that in Section 2. For instance, we simply assume that an allele  $M_i$  is transmitted from a homozygous parent  $M_i M_i$  to any of his/her offspring, and ignore which one it is. If both parents and an offspring have the same genotype  $M_i M_j, i \neq j$ , we assume

that one parent transmits  $M_i$  to the offspring and the other parent transmits  $M_j$  to the offspring. In this way for each parent-offspring pair, we may define an transmission of allele from the parent to the offspring. Putting all data together, we may arrange the trait values of offspring in a way as table 1 in Fan et al. [2002]. Hence, all data from a nuclear family can be used in analysis. Based on which marker allele is transmitted from a parent, one may calculate conditional mean  $\beta_i = E(Y|TM = M_i)$ . In Appendix D, we show  $(\beta_i - \alpha)/(1 - \theta) = (\mu_1 - \mu_2)\delta_i/p_i$ . Hence, the absence of association between trait locus  $Q$  and marker  $M$ , i.e.,  $\delta_i = 0$ , is equivalent to  $\beta_i = \alpha$ . This constitutes the basis to build models and to construct appropriate statistics to test the association between trait locus  $Q$  and marker  $M$  by comparing the estimates of parameters  $\beta_i$  and  $\alpha$ . To do this, we need variance covariance structures of the trait values of offspring. In Appendix D, we calculate conditional variance  $\sigma_{ir}^2 = \text{Var}(Y|TM = M_i)$ . For two offspring of a nuclear family, let  $TM_1$  be the abbreviation of ‘transmitted marker allele for child 1’, and let  $TM_2$  be the abbreviation of ‘transmitted marker allele for child 2’. For  $i \neq j$ , the conditional covariance  $\Sigma_{i,jr} = \text{Cov}(Y_1, Y_2|TM_1 = M_i, TM_2 = M_j) = \Sigma_{ij,ji}$ . The conditional covariance  $\Sigma_{i,ir} = \text{Cov}(Y_1, Y_2|TM_1 = M_i, TM_2 = M_i)$  is calculated in Appendix D.

#### 4. Models

##### 4.1. Heterozygous Parent Data

Suppose that the data consist of nuclear families. For each family, suppose that both parents are typed at the marker locus  $M$  and at least one of the parents is heterozygous. From the analysis in Section 2, one may utilize linear mixed model to analyze the data. Suppose that there are  $I$  heterozygous parents, each of them has at least one offspring. For the offspring of each heterozygous parent, assume that one may clearly determine which allele at the marker locus  $M$  are transmitted from the heterozygous parent. For each child, a quantitative trait value is observed.

For the  $l$ th heterozygous parent, assume that the genotype at the marker locus is  $M_iM_j$ ,  $i \neq j$ . Moreover, he/she has  $n_l$  offspring, and the offspring’s trait values are listed as  $y_{l1}, \dots, y_{ln_l}$ . Assume that the offspring consist of two parts: (1)  $k_l$  offspring correspond to that allele  $M_i$  is transmitted and allele  $M_j$  is not transmitted from their heterozygous parent, and their trait values are listed as  $y_{l1}, \dots, y_{lk_l}$ ; (2) the rest of the offspring correspond to that allele  $M_i$  is not transmitted and allele  $M_j$  is transmitted from their

heterozygous parent, and their trait values are listed as  $y_{l,k_l+1}, \dots, y_{ln_l}$ . Under the null hypothesis of no association in the presence of linkage between the trait locus  $Q$  and the marker locus  $M$ , one may use a multivariate linear model

$$y_{lu} = v + g_{lu} + G_{lu} + e_{lu}, u = 1, 2, \dots, n_l, \text{ reduced model}, \quad (4)$$

where  $y_{lu}$  are normal variables with mean  $\alpha$  and  $n_l \times n_l$  variance-covariance matrix

$$V_l = \begin{pmatrix} \sigma^2 & \Sigma_{ts} & \dots & \Sigma_{ts} & \Sigma_{td} & \dots & \Sigma_{td} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \Sigma_{ts} & \Sigma_{ts} & \dots & \sigma^2 & \Sigma_{td} & \dots & \Sigma_{td} \\ \Sigma_{td} & \Sigma_{td} & \dots & \Sigma_{td} & \sigma^2 & \dots & \Sigma_{ts} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \Sigma_{td} & \Sigma_{td} & \dots & \Sigma_{td} & \Sigma_{ts} & \dots & \sigma^2 \end{pmatrix}.$$

Under the alternative hypothesis of association in the presence of linkage, one may use a full model

$$y_{lu} = v + g_{lu|(TM = M_i, NM = M_j)} + G_{lu} + e_{lu}, u = 1, 2, \dots, k_l, \\ y_{lu} = v + g_{lu|(TM = M_j, NM = M_i)} + G_{lu} + e_{lu}, u = k_l + 1, \dots, n_l. \quad (5)$$

$y_{lu}$  are normal variables with mean  $\alpha_i$  for  $u = 1, \dots, k_l$  and mean  $\alpha_j$  for  $u = k_l + 1, \dots, n_l$ , and a variance-covariance matrix

$$\Gamma_l = \begin{pmatrix} \sigma_i^2 & \Sigma_{i,i} & \dots & \Sigma_{i,i} & \Sigma_{i,j} & \dots & \Sigma_{i,j} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \Sigma_{i,i} & \Sigma_{i,i} & \dots & \sigma_i^2 & \Sigma_{i,j} & \dots & \Sigma_{i,j} \\ \Sigma_{j,i} & \Sigma_{j,i} & \dots & \Sigma_{j,i} & \sigma_j^2 & \dots & \Sigma_{j,j} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \Sigma_{j,i} & \Sigma_{j,i} & \dots & \Sigma_{j,i} & \Sigma_{j,j} & \dots & \sigma_j^2 \end{pmatrix}.$$

Putting all data together, we may perform association studies based on models (4) and (5). Denote  $n = \sum_{l=1}^I n_l$ ,  $\vec{y}_l = (y_{l1}, \dots, y_{ln_l})^T$ ,  $\vec{y} = (\vec{y}_1^T, \dots, \vec{y}_I^T)^T$ ,  $V = \text{diag}(V_1, V_2, \dots, V_I)$  and  $\Gamma = \text{diag}(\Gamma_1, \Gamma_2, \dots, \Gamma_I)$ . Let  $I_n$  be the identity  $n \times n$  matrix. Under the reduced model,  $\vec{y}$  is normal with mean  $\alpha I_n$  and variance-covariance matrix  $V$ . Similarly,  $\vec{y}$  is normal with mean  $X(\alpha_1, \dots, \alpha_m)^T$  and variance-covariance matrix  $\Gamma$  under the full model, where  $X$  is an  $n \times m$  model matrix based on model (5).

##### 4.2. General Nuclear Family Data

In this Subsection, we are going to build models and construct statistics to test association between the trait locus  $Q$  and marker  $M$  by using general nuclear family data. For a sample of nuclear families, assume that there is at least one offspring for each family. For a homozygous parent with genotype  $M_iM_i$  at the marker  $M$  and  $n_l$  off-

spring, let the trait values of the offspring be  $y_1, \dots, y_{n_l}$ . One may use a multivariate linear model for data analysis

$$y_u = v + g_{u|(TM=M)} + G_u + e_u, u = 1, 2, \dots, n_l, \quad (6)$$

where  $y_u$  are normal variables with mean  $\beta_i$  and  $n_l \times n_l$  variance-covariance matrix

$$\begin{pmatrix} \sigma_{ir}^2 & \Sigma_{i,ir} & \dots & \Sigma_{i,ir} \\ \Sigma_{i,ir} & \sigma_{ir}^2 & \dots & \Sigma_{i,ir} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{i,ir} & \Sigma_{i,ir} & \dots & \sigma_{ir}^2 \end{pmatrix}.$$

For a heterozygous parent with genotype  $M_i M_j$ ,  $i \neq j$  at the marker  $M$  and  $n_l$  offspring, let the trait values of the offspring be  $y_1, \dots, y_{n_l}$ . Suppose that: (1)  $k_l$  offspring correspond to that allele  $M_i$  is transmitted and allele  $M_j$  is not transmitted from their heterozygous parent, and their trait values are listed as  $y_1, \dots, y_{k_l}$ ; (2) the rest of the offspring correspond to that allele  $M_i$  is not transmitted and allele  $M_j$  is transmitted from their heterozygous parent, and their trait values are listed as  $y_{k_l+1}, \dots, y_{n_l}$ . One may use a model

$$y_u = v + g_{u|(TM=M)} + G_u + e_u, u = 1, 2, \dots, k_l, \quad (7)$$

$$y_u = v + g_{u|(TM=M)} + G_u + e_u, u = k_l + 1, \dots, n_l.$$

$y_u$  are normal variables with mean  $\beta_i$  for  $u = 1, \dots, k_l$  and mean  $\beta_j$  for  $u = k_l + 1, \dots, n_l$ , and an  $n_l \times n_l$  variance-covariance matrix

$$\begin{pmatrix} \sigma_{ir}^2 & \Sigma_{i,ir} & \dots & \Sigma_{i,ir} & \Sigma_{i,jr} & \dots & \Sigma_{i,jr} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \Sigma_{i,ir} & \Sigma_{i,ir} & \dots & \sigma_{ir}^2 & \Sigma_{i,jr} & \dots & \Sigma_{i,jr} \\ \Sigma_{j,ir} & \Sigma_{j,ir} & \dots & \Sigma_{j,ir} & \sigma_{jr}^2 & \dots & \Sigma_{j,jr} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \Sigma_{j,ir} & \Sigma_{j,ir} & \dots & \Sigma_{j,ir} & \Sigma_{j,jr} & \dots & \sigma_{jr}^2 \end{pmatrix}.$$

Such as in Subsection 4.1, we may combine all data together and perform analysis.

## 5. Test Statistics and Non-Centrality Parameters

### 5.1. Heterozygous Parent Data

Let  $\hat{\alpha}_i, \hat{\sigma}_i^2, \hat{\Sigma}_{i,i}, \hat{\Sigma}_{i,j}$  be the maximum likelihood estimators of parameters  $\alpha_i, \sigma_i^2, \Sigma_{i,i}, \Sigma_{i,j}$  of the full model (5). Then the estimate of  $\gamma = (\alpha_1, \dots, \alpha_m)^\tau$  is  $\hat{\gamma} = (\hat{\alpha}_1, \dots, \hat{\alpha}_m)^\tau = [X^\tau \hat{\Gamma}^{-1} X]^{-1} X^\tau \hat{\Gamma}^{-1} \vec{y}$ . Assume that the sample size is large. In Appendix E, we show that the test statistic of the null hypothesis  $H_0: \alpha_1 = \dots = \alpha_m$ , is non-central  $F(m-1, n-m)$  defined by (details are given in Appendix E)

$$F_{het} = \frac{(H\hat{\gamma})^\tau [H(X^\tau \hat{\Gamma}^{-1} X)^{-1} H^\tau]^{-1} H\hat{\gamma} / (m-1)}{\vec{y}^\tau [\hat{\Gamma}^{-1} - \hat{\Gamma}^{-1} X (X^\tau \hat{\Gamma}^{-1} X)^{-1} X^\tau \hat{\Gamma}^{-1}] \vec{y} / (n-m)},$$

$$H = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 \\ 1 & 0 & -1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & 0 & 0 & \dots & -1 \end{pmatrix}.$$

Here  $H$  is a  $(m-1) \times m$  testing matrix. The non-centrality parameter of the test statistic  $F$  can be calculated by  $\lambda_{het} \approx (H\hat{\gamma})^\tau [H(X^\tau \hat{\Gamma}^{-1} X)^{-1} H^\tau]^{-1} H\hat{\gamma}$ . If  $n_i = 1$  for each family, then there is only one single child in each family and the above formula can be simplified. Let  $k_i, i = 1, 2, \dots, m$  be the number of offspring who receive allele  $M_i$  from their heterozygous parents. In Appendix F, we show that the non-centrality parameter of the singleton test statistic  $F_{het, singleton}$  is

$$\lambda_{het, singleton} \approx \sum_{i=2}^m (\alpha_1 - \alpha_i)^2 k_i \sigma_i^2 - \frac{\left[ \sum_{i=2}^m (\alpha_1 - \alpha_i) k_i \sigma_i^2 \right]^2}{\sum_{i=1}^m k_i \sigma_i^2}.$$

When  $m = 2$ , i.e., the marker  $M$  is bi-allelic, one may simplify  $\lambda_{het, singleton} \approx (\alpha_1 - \alpha_2)^2 / [\sigma_1^2 / k_1 + \sigma_2^2 / k_2]$ , which is the same as that of Fan and Xiong (2003).

Assume that the data consist of both singleton families and sib-pair families. Suppose there are  $k_i$  singleton offspring who receive allele  $M_i$  from their heterozygous parents,  $k_{ii}$  ( $i = 1, 2, \dots, m$ ) sib pairs in each of them both sibs receive allele  $M_i$  from their heterozygous parents, and  $k_{ij} = k_{ji}$ ,  $i \neq j$  sib pairs in each of them one sib receives allele  $M_i$  from his/her heterozygous parent and the other receives allele  $M_j$  from the same heterozygous parent. In Appendix G, we obtain the matrix

$$X^\tau \Gamma^{-1} X = \text{diag} \left( \frac{k_1}{\sigma_1^2} + \frac{2k_{11}}{\sigma_1^2 + \Sigma_{1,1}}, \dots, \frac{k_m}{\sigma_m^2} + \frac{2k_{mm}}{\sigma_m^2 + \Sigma_{m,m}} \right) + X_3^\tau \Gamma_3^{-1} X_3,$$

where matrix  $X_3$ , sub-variance-covariance matrix  $\Gamma_3$ , and  $X^\tau \Gamma_3^{-1} X$  are given in Appendix G. Inserting the above matrix to the formula  $\lambda_{het}$ , one may calculate the non-centrality parameter  $\lambda_{het, singleton, sibs}$  of a test statistic  $F_{het, singleton, sibs}$ . For a bi-allele marker  $M$ , it is the same as that in Fan and Xiong [2003].

### 5.2. General Nuclear Family Data

For the model introduced in Subsection 4.2, we may calculate the non-centrality parameter of statistic  $F_{Gen\_Nuc}$  to test the null hypothesis  $H_0: \beta_1 = \dots = \beta_m$  in a similar manner. First, assume that each family has only one child. Let  $k_i, i = 1, 2, \dots, m$  be the number of offspring who receive allele  $M_i$  from their parents. We can show that the

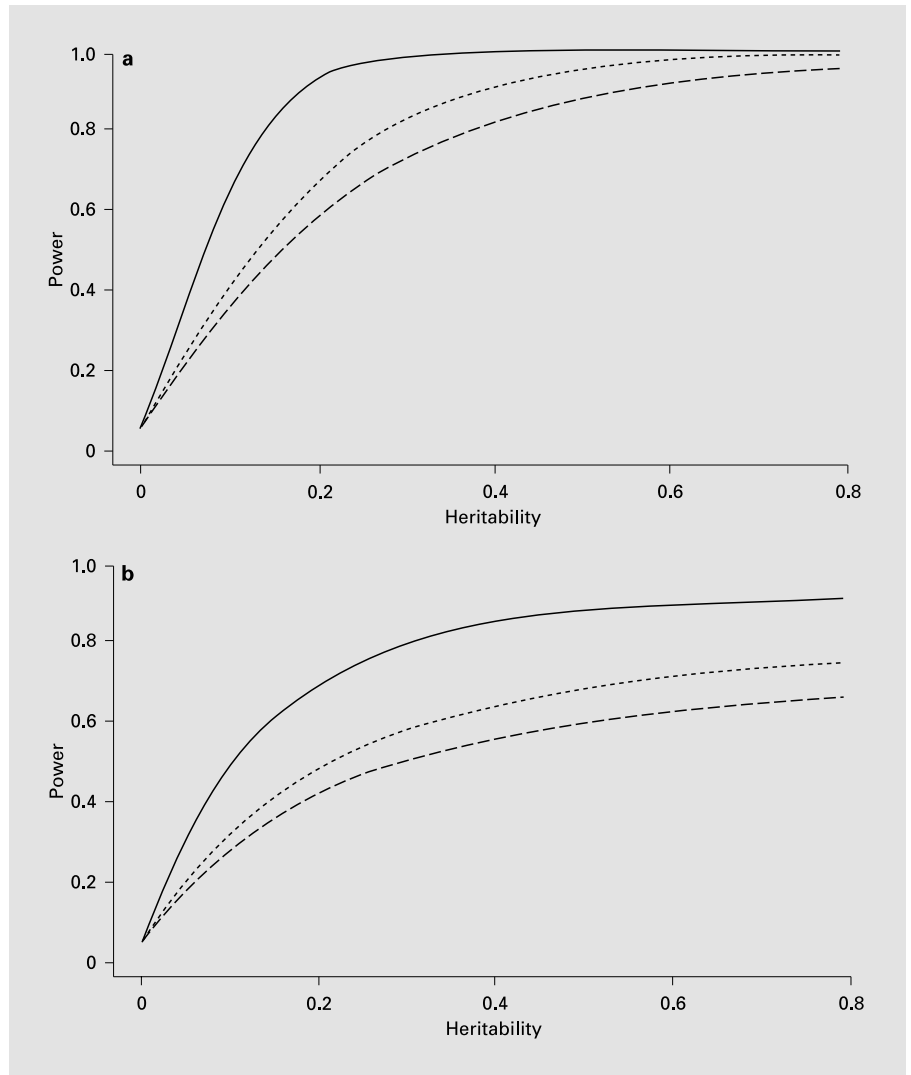


Fig. 2. Power curves of  $F_{het, singleton}$  for 2 (---), 3 (-----) and 4 (—) allele markers against the heritability at 0.05 significant level, when  $q_1 = 0.25$ ,  $\sigma_G^2 = 0.75$ ,  $A = 20$ ,  $\theta = 0.005$  for a dominant trait  $a = d = 1.0$  (a); and a recessive trait  $a = 1.0$  and  $d = -0.5$  (b). For a 2 allele marker,  $p_1 = 0.50$ ,  $k_1 = k_2 = 100$ ; for a 3 allele marker,  $p_1 = 0.4$ ,  $p_2 = 0.3$ ,  $k_1 = 100$ ,  $k_2 = k_3 = 50$ ; for a 4 allele marker,  $p_1 = 0.25$ ,  $k_i = 50$ ,  $i = 1, \dots, 4$ .

corresponding non-centrality parameter of a singleton test statistic  $F_{Gen\_Nuc, singleton}$  is  $\lambda_{Gen\_Nuc, singleton}$

$$\approx \sum_{i=2}^m (\beta_1 - \beta_i)^2 k_{i/} \sigma_{ir}^2 - \frac{\left[ \sum_{i=2}^m (\beta_1 - \beta_i) k_{i/} \sigma_{ir}^2 \right]^2}{\sum_{i=1}^m k_{i/} \sigma_{ir}^2}.$$

Second, the data consist of both singleton families and sib-pair families. Suppose there are  $k_i$  singleton offspring who receive allele  $M_i$  from their parents,  $k_{ij}$  ( $i = 1, 2, \dots, m$ ) sib pairs in each of them both sibs receive allele  $M_i$  from their parents, and  $k_{ij} = k_{ji}$ ,  $i \neq j$  sib pairs in each of them one sib receives allele  $M_i$  from his/her heterozygous parent and the other receives allele  $M_j$  from the same hetero-

zygous parent. We may calculate the corresponding non-centrality parameter

$$\lambda_{Gen\_Nuc, singleton, sibs} \approx (H\beta)^T [H\Pi^{-1}H^T]^{-1} H\beta$$

of a statistic  $F_{Gen\_Nuc, singleton, sibs}$ , where

$$\Pi = \text{diag} \left( \frac{k_1}{\sigma_{1r}^2} + \frac{2k_{11}}{\sigma_{1r}^2 + \Sigma_{1,1r}}, \dots, \frac{k_m}{\sigma_{mr}^2} + \frac{2k_{mm}}{\sigma_{mr}^2 + \Sigma_{m,mr}} \right) + \Pi_3, \text{ and } \Pi_3 =$$

$$\begin{pmatrix} \sum_{i \neq 1} \frac{k_{1i} \sigma_{ir}^2}{\sigma_{1r}^2 \sigma_{ir}^2 - \Sigma_{1,ir}^2} & -\frac{k_{12} \Sigma_{1,2r}}{\sigma_{1r}^2 \sigma_{2r}^2 - \Sigma_{1,2r}^2} & \cdots & -\frac{k_{1m} \Sigma_{1,mr}}{\sigma_{1r}^2 \sigma_{mr}^2 - \Sigma_{1,mr}^2} \\ -\frac{k_{12} \Sigma_{1,2r}}{\sigma_{1r}^2 \sigma_{2r}^2 - \Sigma_{1,2r}^2} & \sum_{i \neq 2} \frac{k_{2i} \sigma_{ir}^2}{\sigma_{2r}^2 \sigma_{ir}^2 - \Sigma_{2,ir}^2} & \cdots & -\frac{k_{2m} \Sigma_{2,mr}}{\sigma_{2r}^2 \sigma_{mr}^2 - \Sigma_{2,mr}^2} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{k_{1m} \Sigma_{1,mr}}{\sigma_{1r}^2 \sigma_{mr}^2 - \Sigma_{1,mr}^2} & -\frac{k_{2m} \Sigma_{2,mr}}{\sigma_{2r}^2 \sigma_{mr}^2 - \Sigma_{2,mr}^2} & \cdots & \sum_{i \neq m} \frac{k_{mi} \sigma_{ir}^2}{\sigma_{mr}^2 \sigma_{ir}^2 - \Sigma_{m,ir}^2} \end{pmatrix}.$$

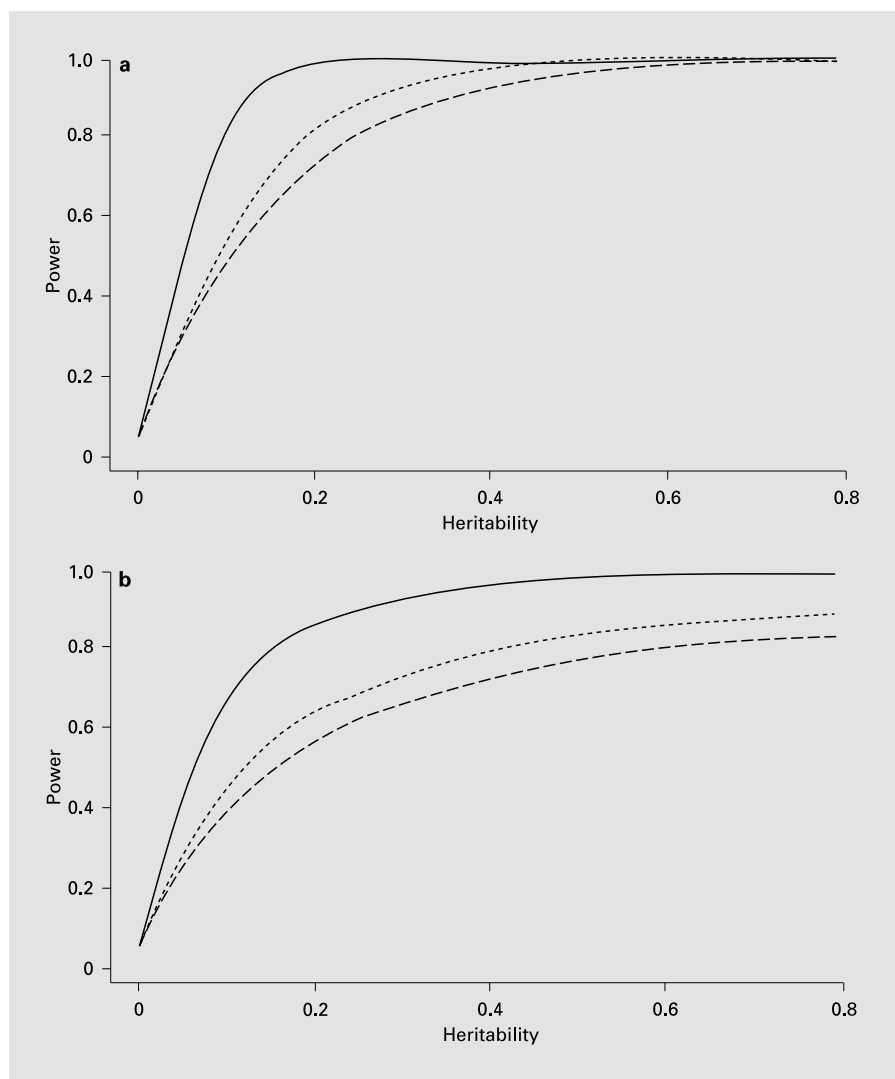


Fig. 3. Power curves of  $F_{het, singleton, sibs}$  for 2 (---), 3 (- - - -) and 4 (—) allele markers against the heritability at 0.05 significant level, when  $q_1 = 0.25$ ,  $\sigma_G^2 = 0.75$ ,  $A = 20$ ,  $\theta = 0.005$  for a dominant trait  $a = d = 1.0$  (a); and a recessive trait  $a = 1.0$  and  $d = -0.5$  (b). For a 2 allele marker,  $p_1 = 0.50$ ,  $k_i = 60$ ,  $k_{ij} = 30$ ,  $i, j = 1, 2$ ; for a 3 allele marker,  $p_1 = 0.4$ ,  $p_2 = 0.3$ ,  $k_1 = 60$ ,  $k_2 = k_3 = 30$ ,  $k_{ij} = 15$ ,  $i, j = 1, 2, 3$ ; for a 4 allele marker,  $p_i = 0.25$ ,  $k_i = 30$ ,  $k_{ij} = 9$ ,  $i, j = 1, \dots, 4$ .

## 6. Power Calculation and Power Comparison

Such as the standard theory of quantitative genetics [Falconer and Mackay 1996], assume that  $v = 0$ ,  $\mu_{11} = a$ ,  $\mu_{12} = \mu_{21} = d$ ,  $\mu_{22} = -a$ . Let the additive variance be  $\sigma_a^2 = 2q_1q_2(a + d(q_2 - q_1))^2$ , and the dominant variance be  $\sigma_d^2 = (2q_1q_2d)^2$ . Let the heritability be denoted by  $h^2$ , which is defined by  $\sigma_a^2 / (\sigma_a^2 + \sigma_d^2 + \sigma_e^2)$ . In the history of a population, the disease genes are usually due to a mutation. Because of the evolutionary process, the haplotype frequencies  $h_{ri}$  change from generation to generation. The expected haplotype frequency can be calculated by  $E[h_{ri}] = h_{ri}(0)e^{-\theta A} + p_iq_r(1 - e^{-\theta A})$ , where  $A$  is the age of the most recent mutation at the trait locus,  $h_{ri}(0)$  is the initial haplotype frequency of haplotype  $Q_rM_i$  at the generation of occur-

rence of the mutation at the trait locus. If there is only a single mutation in the population, one may assume that  $h_{11}(0) = q_1$ ,  $h_{1i}(0) = 0$ , and  $h_{21}(0) = p_1 - q_1 \geq 0$ ,  $h_{2i}(0) = p_i$ ,  $i = 2, \dots, m$ . Replacing  $h_{ri}$  in  $P(Q_rM_i, M_j)$  by  $E[h_{ri}]$ , we may calculate the approximations of the non-centrality parameters using the non-centrality parameters given in Section 5. To calculate the non-centrality parameters, we need parameter values such as the marker allele frequencies  $p_1$  and  $p_2$ , trait allele frequencies  $q_1$  and  $q_2$ , heritability  $h^2$ , mutation age  $A$ , haplotype frequencies  $h_{ri}$ , recombination fraction  $\theta$ , additive effect  $a$ , dominant effect  $d$ , polygenic variance  $\sigma_G^2$ , and error variance  $\sigma_e^2$ .

Assume that the frequencies of marker alleles are evenly distributed. Figures 2 and 3 plot the power curves of  $F_{het, singleton}$  and  $F_{het, singleton, sibs}$  against the heritability at

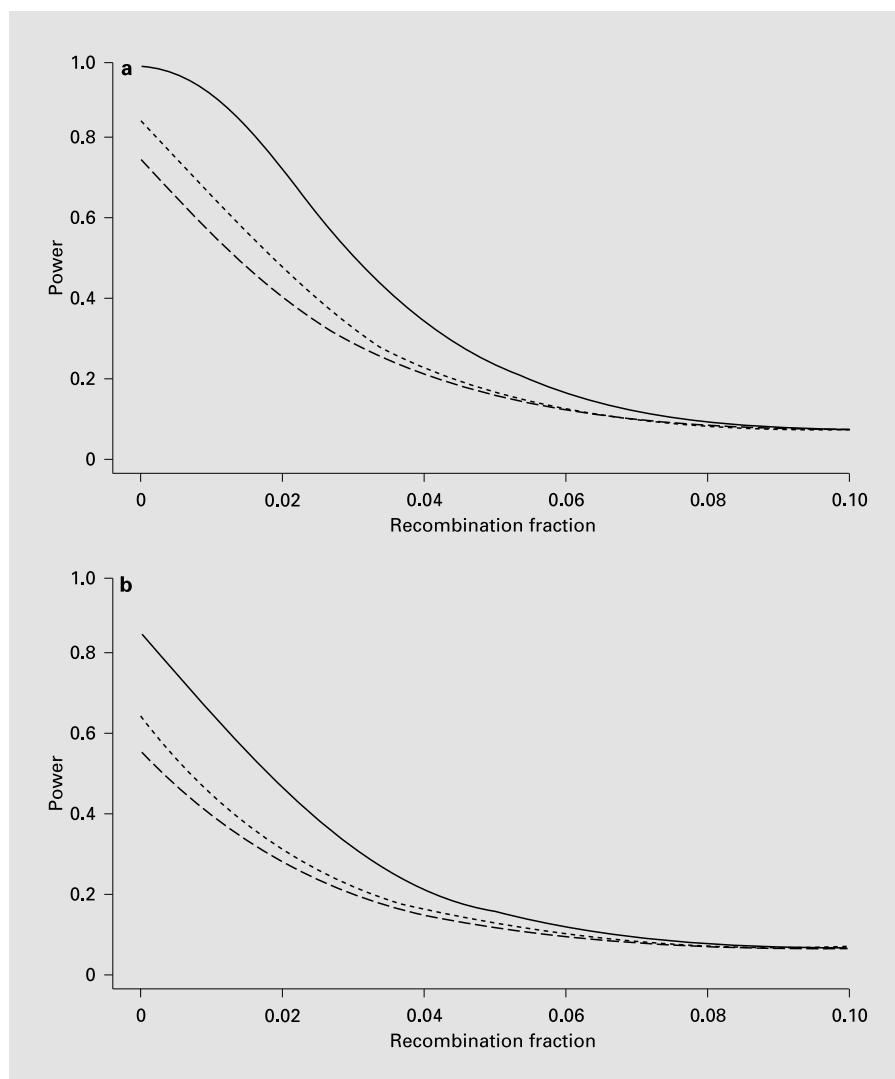


Fig. 4. Power curves of  $F_{Gen\_Nuc\_singleton}$  for 2 (---), 3 (-----) and 4 (—) allele markers against the recombination fraction at 0.05 significant level, when  $q_1 = 0.25$ ,  $\sigma_G^2 = 0.75$ ,  $A = 20$ ,  $h^2 = 0.25$  for a dominant trait  $a = d = 1.0$  (a); and a recessive trait  $a = 1.0$  and  $d = -0.5$  (b). For a 2 allele marker,  $p_1 = 0.50$ ,  $k_1 = k_2 = 100$ ; for a 3 allele marker,  $p_1 = 0.4$ ,  $p_2 = 0.3$ ,  $k_1 = 100$ ,  $k_2 = k_3 = 50$ ; for a 4 allele marker,  $p_i = 0.25$ ,  $k_i = 50$ ,  $i = 1, \dots, 4$ .

0.05 significant level, for dominant and recessive traits for 2, 3 and 4 allele markers, respectively. In each graph of the two figures, the total numbers of offspring for 2, 3 and 4 allele markers are the same. Hence, the comparison of the power is meaningful. It is clear from the 4 graphs of the two figures 2 and 3 that the power of the test statistic using 4 allele marker is higher than that of the test statistic using 3 allele marker, which in turn is higher than that of the test statistic using 2 allele marker. Figures 4 and 5 plot the power curves of  $F_{Gen\_Nuc\_singleton}$  and  $F_{Gen\_Nuc\_singleton,sibs}$  against the recombination fraction at 0.05 significant level, for dominant and recessive traits for 2, 3 and 4 allele markers, respectively. The four graphs in the two figures 4 and 5 show that the power of the test statistic using 4 allele marker is higher than that of the test statistic using 3 allele

marker, which in turn is higher than that of the test statistic using 2 allele marker. In addition, the power is high when the trait locus is tightly linked to the marker ( $\theta < 0.01$ ); otherwise, the power decreases very rapidly once the trait locus is getting far away from the marker ( $\theta > 0.02$ ).

Assume that the frequencies of marker alleles are not evenly distributed. Figure 6 plots the power curves of  $F_{het, singleton, sibs}$  against the heritability at 0.05 significant level, for dominant and recessive traits for 2, 3 and 4 allele markers, respectively. In each of two graphs in the figure, the power of the test statistic using 3 allele marker is higher than that of the test statistic using 4 allele marker, which in turn is higher than that of the test statistic using 2 allele marker in general.

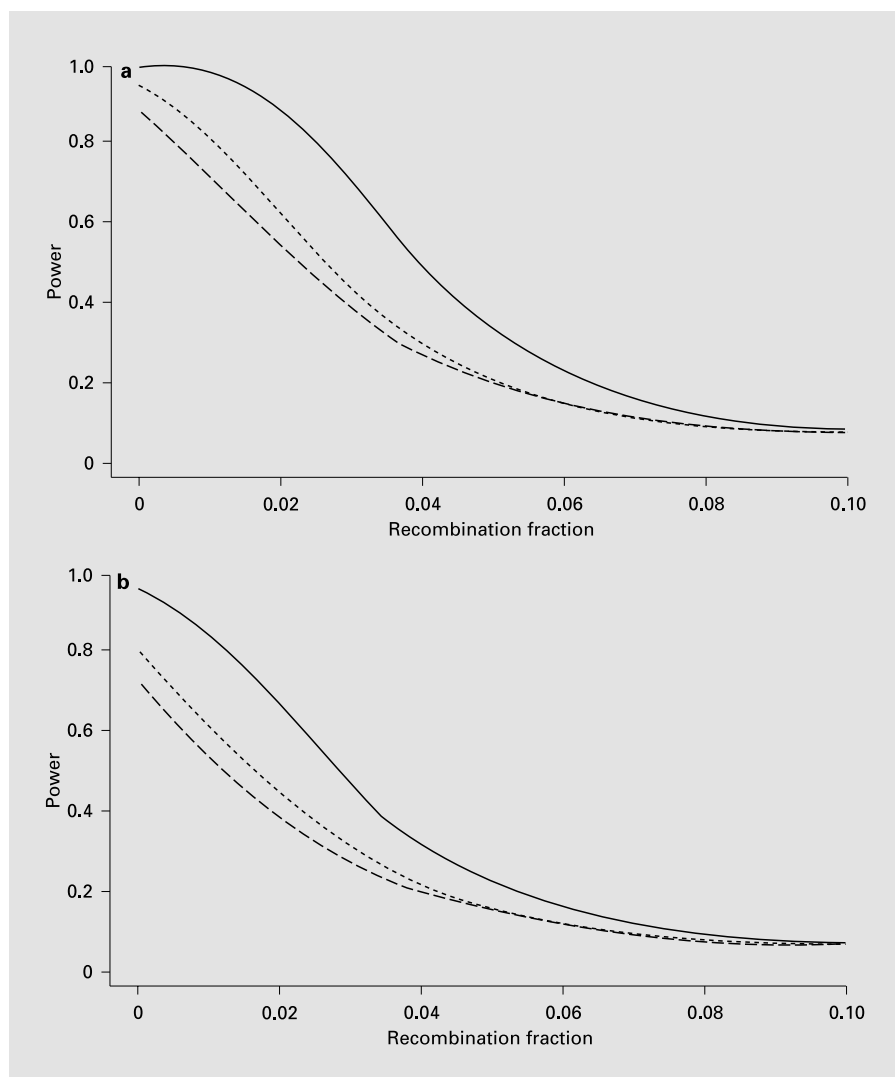


Fig. 5. Power curves of  $F_{Gen\_Nuc, singleton, sibs}$  for 2 (---), 3 (- - - -) and 4 (—) allele markers against the recombination fraction at 0.05 significant level, when  $q_1 = 0.25$ ,  $\sigma_G^2 = 0.75$ ,  $A = 20$ ,  $h^2 = 0.25$  for a dominant trait  $a = d = 1.0$  (a); and a recessive trait  $a = 1.0$  and  $d = -0.5$  (b). For a 2 allele marker,  $p_1 = 50$ ,  $k_i = 60$ ,  $k_{ij} = 30$ ,  $i, j = 1, 2$ ; for a 3 allele marker,  $p_1 = 0.4$ ,  $p_2 = 0.3$ ,  $k_1 = 60$ ,  $k_2 = k_3 = 30$ ,  $k_{ij} = 15$ ,  $i, j = 1, 2, 3$ ; for a 4 allele marker,  $p_i = 0.25$ ,  $k_i = 30$ ,  $k_{ij} = 9$ ,  $i, j = 1, \dots, 4$ .

## 7. An Example

The methods and models are applied to analyze the chromosomes 4 and 16 data of the Oxford asthma data, Genetic Analysis Workshop 12 [Cookson and Abecasis 2001]. The data consist of 80 nuclear families with a total of 203 offspring. In these 80 families, 43 have two offspring, 31 have three offspring, and 6 have four offspring. On chromosome 4, 18 markers are typed and each marker has 4 alleles. On chromosome 16, 22 markers are typed and each marker has 4 alleles. In Daniel et al. [1996], linkage to bronchial responsiveness to methacholine (slope) and other quantitative traits was tested by the Haseman-Elston sib-pair technique [Haseman and Elston 1972]. Two regions of potential linkage to autosomal markers

were detected with  $\log_e$ slope on chromosomes 4 and 16 [Daniel et al. 1996].

In the four alleles typed, the frequency of one allele is too low (around 3%). When we use the four alleles in data analysis, the convergence is problematic and the results are not stable. This may be due to the large number of parameters for the data set. To reduce the number of parameters and to make the results stable, we collapse each of the 4 allele markers to be 3 allele marker. Table 1 shows the results of test statistics  $F_{het}$  and  $F_{Gen\_Nuc}$ , the results from Fan and Xiong [2003], and Daniel et al. [1996]. Three markers, D4S1450, D16S515 and D16S289 show association with the asthma phenotypic trait  $\log_e$ slope at significant levels 0.05. The results confirms the findings in Fan and Xiong [2003] and Daniel et al. [1996].

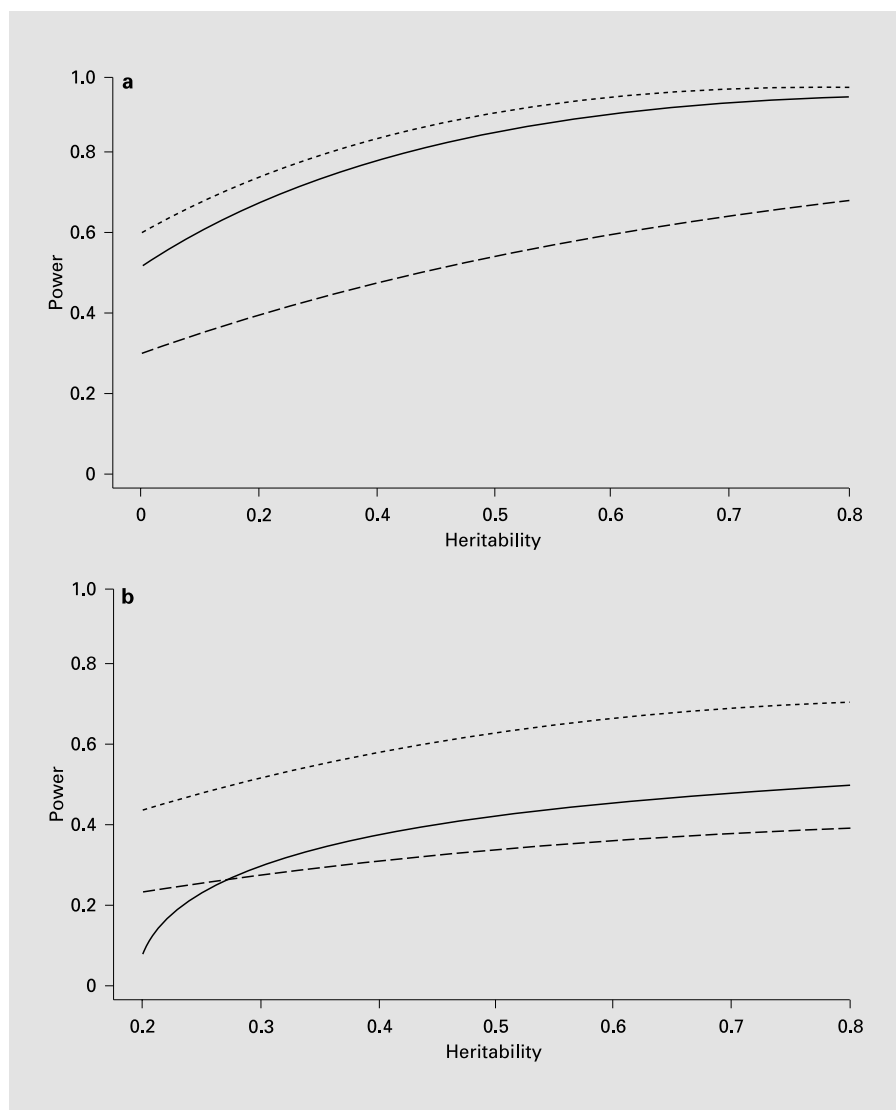


Fig. 6. Power curves of  $F_{het, singleton, sibs}$  for 2 (---), 3 (-----) and 4 (—) allele markers against the heritability at 0.05 significant level. For a 2 allele marker,  $p_1 = 0.90$ ,  $p_2 = 0.10$ ; for a 3 allele marker,  $p_1 = 0.5$ ,  $p_2 = 0.45$ ,  $p_3 = 0.05$ ; for a 4 allele marker,  $p_1 = 0.45$ ,  $p_2 = p_3 = 0.25$ ,  $p_4 = 0.05$ . All other parameters are the same as those in figure 3.

Table 1. Results of test statistics of asthma data

Marker locus	p values of $F_{Het}$	p values of $F_{Gen\_Nuc}$	p values of Fan and Xiong (2003)	p values of Daniel et al. (1996)
D4S1540	0.03	0.003	0.02	<0.05
D16S515	<0.0001	<0.0001	<0.04	<0.05
D16S289	0.001	<0.0001	<0.0001	<0.05

## 8. Discussion

This paper explores mixed models and test statistics for association study between a QTL and a multiple allele marker. For the data of offspring from at least one heterozygous parent with clear transmission of parental alleles, we show that it can be used in association study in the presence of linkage. The theoretical basis is the difference between the conditional mean of trait value given an allele is transmitted and the conditional mean of trait value given the allele is not transmitted from a heterozygous parent. For the data of offspring of nuclear family, we show that it can be used in general association study. In this case, the theoretical basis is the difference between the conditional mean of trait value given an allele is transmitted from a parent and the population mean. To build valid models, we calculate the variance-covariance structure of trait values of offspring. Besides, the reduction of the number of parameters is discussed under an assumption of tight linkage between the trait locus and the marker for the data of offspring from at least one heterozygous parent. Based on mixed models, two test statistics are proposed for association study. By power calculation and comparison, we show that the proposed test statistics have higher power than that by collapsing alleles to be new ones if the marker allele frequencies are evenly distributed. Hence, it is more advantageous to use a multiple allele marker in association study than that to collapse it to be a bi-allelic marker. Using the data of offspring of nuclear family, we show that the power is high when the trait locus is tightly linked to the marker ( $\theta < 0.01$ ); otherwise, the power decreases very rapidly once the trait locus is getting far away from the marker ( $\theta > 0.02$ ). The proposed models are used to analyze chromosomes 4 and chromosome 16 data of the Oxford asthma data, Genetic Analysis Workshop 12.

In mapping a disease gene locus, one may carry out both linkage analysis and association study. Linkage analysis is based on family data, and is useful in localizing a genetic trait locus in a broad chromosome region of a few centimorgans. Hence, linkage analysis can provide suggestive linkage between a trait locus and a marker locus based on a sparse marker map. Besides, linkage analysis is robust to the population stratification which heavily affects the results of population-based association study. Association study, on the other hand, is useful in fine gene mapping of genetic trait locus since the allelic association due to LD usually operates over very short genetic distance. Hence, association study can provide high resolution in genetic trait mapping. In practice, the first step in

mapping disease genes can be linkage analysis to get suggestive linkage. Then, an association study can be used in disease gene fine mapping. Due to the suggestive linkage from the linkage analysis, the results of the follow-up association study are less likely the outcomes of population structures.

Using a bi-allelic marker, Fan and Xiong [2003] proposed mixed models for both linkage analysis in the presence of association and association study in the presence of linkage. For a multi-allelic marker, it is not clear how to extend Fan and Xiong [2003] for linkage analysis in the presence of association since the way to reduce the number of parameters is unclear. In this paper, we assume that data are available for all members in a nuclear family. This may not be true for late-on-set genetic traits such as Alzheimer's disease, heart disease, many forms of cancer, non-insulin-dependent diabetes mellitus (NIDDM), and osteoporosis, for which the parental data are hard to obtain. It would be interesting if the methods and models in this paper can be extended to study the data when the parental genetic data are not available [Cardon 2000]. One method is permutation tests such as that for qualitative traits [Spielman and Ewens 1996]. In our models and analysis, we do not include any covariates such as age and gender. The research of interplay between genetic effects and environments is of very importance. Van den Oord and Sneider [2002] propose a general model for locus effects to study the interplay of the multiple etiological factors and other genetic effects such as age dependency based on structural equation modeling (SEM). Due to the length of this paper, we do not pursue this important issue in depth.

## Acknowledgment

We greatly appreciate the help of two reviewers for very detailed and thoughtful input, which renders this paper much more clearer. We thank Dr. D. Gordon for his patience in handling this paper, and Dr. M. Knapp for helpful discussion. R. Fan was supported partially by a research fellowship from the Alexander von Humboldt Foundation, Germany, and an International Research Travel Assistant grant, Texas A&M University.

## Appendix A

Without loss of generality, assume that  $k = 2$  and  $n = 3$  in Figure 1. Let  $TM_1$  be the abbreviation of the “transmitted marker allele for child 1”, and  $NM_1$  be the abbreviation of the “non-transmitted marker allele for child 1”, from the heterozygous mother  $M_iM_j$  in Figure 1. Similarly, we define the notations  $TM_i, NM_i, i = 2, 3$ . Denote  $A = (TM_1 = M_i, NM_1 = M_j, TM_2 = M_i, NM_2 = M_j)$ . Let  $S_{7kl}$  be the state where two offspring share two identical trait alleles  $Q_k$  and  $Q_l$  by descent, and  $Q_l$  is from the heterozygous father and  $Q_k$  is from the mother;  $S_{8klr}$  be the state where two offspring share one identical trait allele  $Q_k$  by descent, and the other two alleles  $Q_l$  and  $Q_r$  are not identical by descent; and  $S_{9krts}$  be the state where two offspring share no identical trait alleles by descent, and two alleles  $Q_l, Q_s$  are from the heterozygous father, and the other two alleles  $Q_k, Q_r$  are from the mother. Then

$$\begin{aligned} \Sigma_{ij,ij} = & \left[ \sum_k \sum_l \mu_{kl}^2 P(A \cap S_{7kl}) + \sum_k \sum_l \sum_r \mu_{kl} \mu_{kr} P(A \cap S_{8klr}) \right. \\ & \left. + \sum_k \sum_l \sum_r \sum_s \mu_{kl} \mu_{rs} P(A \cap S_{9krts}) \right] / (p_i p_j / 2) - (\nu - \alpha_{i,j})^2 + \sigma_G^2 / 2, \end{aligned}$$

where

$$\begin{aligned} P(A \cap S_{7kl}) &= \frac{q_k}{2} \left( 2h_{li} p_j \frac{1-\theta}{2} \frac{1-\theta}{2} + 2h_{lj} p_i \frac{\theta}{2} \frac{\theta}{2} \right) = q_k \left( h_{li} p_j (1-\theta)^2 + h_{lj} p_i \theta^2 \right) / 4 \\ P(A \cap S_{8klr}) &= \frac{q_l q_r}{2} \left( 2h_{ki} p_j \frac{1-\theta}{2} \frac{1-\theta}{2} + 2h_{kj} p_i \frac{\theta}{2} \frac{\theta}{2} \right) + \frac{q_k}{2} (h_{li} h_{rj} + h_{ri} h_{lj}) 2\theta(1-\theta) / 4 \\ &= q_l q_r \left( h_{ki} p_j (1-\theta)^2 + h_{kj} p_i \theta^2 \right) / 4 + q_k (h_{li} h_{rj} + h_{ri} h_{lj}) \theta(1-\theta) / 4 \\ P(A \cap S_{9krts}) &= \frac{q_k q_r}{2} (h_{li} h_{sj} + h_{si} h_{lj}) 2\theta(1-\theta) / 4 = q_k q_r (h_{li} h_{sj} + h_{si} h_{lj}) \theta(1-\theta) / 4. \end{aligned}$$

Similarly, denote  $B = (TM_1 = M_i, NM_1 = M_j, TM_3 = M_j, NM_3 = M_i)$ . We can calculate the conditional covariance of offspring 1 and 3 in Figure 1

$$\begin{aligned} \Sigma_{ij,ji} &= \Sigma_{ji,ij} = \text{Cov}(y_1, y_3) \\ &= \left[ \sum_k \sum_l \mu_{kl}^2 P(B \cap S_{7kl}) + \sum_k \sum_l \sum_r \mu_{kl} \mu_{kr} P(B \cap S_{8klr}) \right. \\ & \quad \left. + \sum_k \sum_l \sum_r \sum_s \mu_{kl} \mu_{rs} P(B \cap S_{9krts}) \right] / (p_i p_j / 2) - (\nu - \alpha_{i,j})(\nu - \alpha_{j,i}) + \sigma_G^2 / 2, \end{aligned}$$

where

$$\begin{aligned} P(B \cap S_{7kl}) &= \frac{q_k}{2} (2h_{li} p_j + 2h_{lj} p_i) \theta(1-\theta) / 4 = q_k (h_{li} p_j + h_{lj} p_i) \theta(1-\theta) / 4 \\ P(B \cap S_{8klr}) &= \frac{q_l q_r}{2} (2h_{ki} p_j + 2h_{kj} p_i) \theta(1-\theta) / 4 + \frac{q_k}{2} (h_{li} h_{rj} + h_{ri} h_{lj}) \frac{\theta^2 + (1-\theta)^2}{4} \\ &= q_l q_r (h_{ki} p_j + h_{kj} p_i) \theta(1-\theta) / 4 + q_k (h_{li} h_{rj} + h_{ri} h_{lj}) \frac{\theta^2 + (1-\theta)^2}{8} \\ P(B \cap S_{9krts}) &= \frac{q_k q_r}{2} (h_{li} h_{sj} + h_{si} h_{lj}) \frac{\theta^2 + (1-\theta)^2}{4} = q_k q_r (h_{li} h_{sj} + h_{si} h_{lj}) \frac{\theta^2 + (1-\theta)^2}{8}. \end{aligned}$$

## Appendix B

Assume that the marker locus and the trait locus are in linkage equilibrium, i.e.,  $h_{ri} = q_r p_i$  for all  $r, i$ . Then we have

$$\begin{aligned}\alpha_{i,j} &= \sum_{r=1}^2 (\nu + \mu_r) q_r = \nu + \mu = \alpha \\ \sigma_{i,j}^2 &= \sigma_e^2 + \sigma_G^2 + \sum_{r=1}^2 \sum_{s=1}^2 (\nu + \mu_{rs} - \alpha)^2 q_r q_s = \sigma^2 \\ \Sigma_{ij,ij} &= \sum_k \sum_l \mu_{kl}^2 q_k q_l [(1-\theta)^2 + \theta^2]/2 + \sum_k \sum_l \sum_r \mu_{kl} \mu_{kr} q_k q_l q_r / 2 \\ &\quad + \sum_k \sum_l \sum_r \sum_s \mu_{kl} \mu_{rs} q_k q_r q_l q_s \theta (1-\theta) - (\nu - \alpha)^2 + \sigma_G^2 / 2 = \Sigma_{ts} \\ \Sigma_{ij,ji} &= \sum_k \sum_l \mu_{kl}^2 q_k q_l (1-\theta)\theta + \sum_k \sum_l \sum_r \mu_{kl} \mu_{kr} q_k q_l q_r / 2 \\ &\quad + \sum_k \sum_l \sum_r \sum_s \mu_{kl} \mu_{rs} q_k q_r q_l q_s [\theta^2 + (1-\theta)^2] / 2 - (\nu - \alpha)^2 + \sigma_G^2 / 2 = \Sigma_{td}.\end{aligned}$$

Notice that  $\alpha, \sigma^2, \Sigma_{ts}$  and  $\Sigma_{td}$  do not depend on subscripts  $i$  and  $j$ .

## Appendix C

Assume that the recombination fraction  $\theta \approx 0$ , i.e. there is tight linkage between the trait locus and the marker. Then  $P(Q_r M_i, M_j) \approx h_{ri} p_j$ . Therefore, we have

$$\begin{aligned}\alpha_{i,j} &\approx \sum_{r=1}^2 (\nu + \mu_r) h_{ri} / p_i = \alpha_i \\ \sigma_{i,j}^2 &\approx \sigma_e^2 + \sigma_G^2 + \sum_{r=1}^2 \sum_{s=1}^2 (\nu + \mu_{rs} - \alpha_i)^2 q_s h_{ri} / p_i = \sigma_e^2 + \sigma_G^2 + \Sigma_i^2 = \sigma_i^2.\end{aligned}$$

Note that  $\alpha_i$  and  $\Sigma_i^2$  only depend on subscript  $i$ . Besides, the covariances  $\Sigma_{ij,ij}$  and  $\Sigma_{ij,ji}$  can be approximated by

$$\begin{aligned}\Sigma_{ij,ij} &\approx \left[ \sum_k \sum_l \mu_{kl}^2 q_k h_{li} + \sum_k \sum_l \sum_r \mu_{kl} \mu_{kr} q_l q_r h_{ki} \right] / (2p_i) - (\nu - \alpha_i)^2 + \sigma_G^2 / 2 = \Sigma_{i,i} \\ \Sigma_{ij,ji} &\approx \left[ \sum_k \sum_l \sum_r \mu_{kl} \mu_{kr} q_k (h_{li} h_{rj} + h_{ri} h_{lj}) + \sum_k \sum_l \sum_r \sum_s \mu_{kl} \mu_{rs} q_k q_r (h_{li} h_{sj} + h_{si} h_{lj}) \right] / (4p_i p_j) \\ &\quad - (\nu - \alpha_i)(\nu - \alpha_j) + \sigma_G^2 / 2 = \Sigma_{i,j} = \Sigma_{j,i}.\end{aligned}$$

Notice that  $\Sigma_{i,i}$  only depends on subscript  $i$ , but  $\Sigma_{i,j} = \Sigma_{j,i}$  depends on both  $i$  and  $j$ .

Appendix D

Let  $TH$  denote abbreviation of “transmitted haplotype”. Then  $P(TH = Q_r M_i) = (1 - \theta)h_{ri} + \theta q_r p_i$ . Notice that  $h_{2i} - q_2 p_i = -h_{1i} + q_1 p_i = -\delta_i$ . Such as in Appendix A, Fan, Floros and Xiong (2002), one may show that

$$\begin{aligned} \beta_i &= E[Y|TM = M_i] \\ &= \left[ E[Y|TH = Q_1 M_i]P(TH = Q_1 M_i) + E[Y|TH = Q_2 M_i]P(TH = Q_2 M_i) \right] / p_i \\ &= (1 - \theta) \left[ (\nu + \mu_1)h_{1i} + (\nu + \mu_2)h_{2i} \right] / p_i + \theta \alpha \\ \frac{\beta_i - \alpha}{1 - \theta} &= \left[ (\nu + \mu_1)h_{1i} + (\nu + \mu_2)h_{2i} \right] / p_i - [(\nu + \mu_1)q_1 + (\nu + \mu_2)q_2] = (\mu_1 - \mu_2)\delta_i / p_i. \end{aligned}$$

To calculate the conditional variance, we first notice the conditional variances

$$\sigma_{Q_k}^2 = \text{Var}(Y|TQ = Q_k) = \sigma_e^2 + \sigma_G^2 + (\mu_{k1} - \mu_k)^2 q_1 + (\mu_{k2} - \mu_k)^2 q_2, k = 1, 2.$$

The conditional variance

$$\sigma_{ir}^2 = \text{Var}(Y|TM = M_i) = \sum_{k=1}^2 [\sigma_{Q_k}^2 + (\nu + \mu_k - \beta_i)^2] P(TH = Q_k M_i) / p_i.$$

For two different alleles  $M_i$  and  $M_j$ ,  $i \neq j$ , the conditional covariance

$$\Sigma_{i,jr} = \text{Cov}(Y_1, Y_2|TM_1 = M_i, TM_2 = M_j) = \Sigma_{ij,ji}.$$

Let  $C_i = (TM_1 = M_i, TM_2 = M_i)$ . Then  $P(C_i) = \sum_{j \neq i} 2p_i p_j \frac{1}{2} \frac{1}{2} + p_i^2 \cdot 1 \cdot 1 = p_i(1 + p_i)/2$ . Let  $S_{7kl}, S_{8klr}$  and  $S_{9krls}$  be similar notations as those in Appendix A. Then

$$\begin{aligned} \Sigma_{i,ir} &= \text{Cov}(Y_1, Y_2|TM_1 = M_i, TM_2 = M_i) \\ &= \left[ \sum_k \sum_l \mu_{kl}^2 P(C_i \cap S_{7kl}) + \sum_k \sum_l \sum_r \mu_{kl} \mu_{kr} P(C_i \cap S_{8klr}) \right. \\ &\quad \left. + \sum_k \sum_l \sum_r \sum_s \mu_{kl} \mu_{rs} P(C_i \cap S_{9krls}) \right] / P(C_i) - (\nu - \beta_i)^2 + \sigma_G^2 / 2, \end{aligned}$$

where

$$\begin{aligned} P(C_i \cap S_{7kl}) &= \frac{q_k}{2} \left( 2h_{li} \frac{1 - \theta}{2} \frac{1 - \theta}{2} + 2q_l p_i \frac{\theta}{2} \frac{\theta}{2} + 2h_{li} p_i \frac{1 - \theta}{2} \frac{\theta}{2} \right) \\ &= q_k \left( h_{li} (1 - \theta)^2 + q_l p_i \theta^2 + 2h_{li} p_i \theta (1 - \theta) \right) / 4 \\ P(C_i \cap S_{8klr}) &= \frac{q_l q_r}{2} \left[ 2h_{ki} \frac{1 - \theta}{2} \frac{1 - \theta}{2} + 2q_k p_i \frac{\theta}{2} \frac{\theta}{2} + 2h_{ki} p_i \frac{\theta}{2} \frac{1 - \theta}{2} \right] \\ &\quad + \frac{q_k}{2} \left[ 2h_{li} h_{ri} \frac{\theta^2 + (1 - \theta)^2}{4} + 2h_{ri} q_l \theta (1 - \theta) / 4 + 2h_{li} q_r \theta (1 - \theta) / 4 \right] \\ &= q_l q_r \left[ h_{ki} (1 - \theta)^2 + q_k p_i \theta^2 + 2h_{ki} p_i \theta (1 - \theta) \right] / 4 \\ &\quad + q_k \left[ h_{li} h_{ri} [\theta^2 + (1 - \theta)^2] + (h_{ri} q_l + h_{li} q_r) \theta (1 - \theta) \right] / 4 \\ P(C_i \cap S_{9krls}) &= \frac{q_k q_r}{2} \left[ 2h_{li} h_{si} \frac{\theta^2 + (1 - \theta)^2}{4} + 2h_{li} q_s \theta (1 - \theta) / 4 + 2h_{si} q_l \theta (1 - \theta) / 4 \right] \\ &= q_k q_r \left[ h_{li} h_{si} [\theta^2 + (1 - \theta)^2] + (h_{li} q_s + h_{si} q_l) \theta (1 - \theta) \right] / 4. \end{aligned}$$

Assume that the marker  $M$  and the trait locus  $Q$  are in linkage equilibrium, i.e.,  $h_{ri} = q_r p_i$  for  $r = 1, 2, i = 1, \dots, m$ . Then  $\beta_i = \alpha$ ,  $\sigma_{ir}^2 = \sigma^2$ ,  $\Sigma_{i,jr} = \Sigma_{td}$  and

$$\begin{aligned} \Sigma_{i,ir} &= \sum_k \sum_l \mu_{kl}^2 q_k q_l \frac{\theta^2 + (1-\theta)^2 + 2p_i \theta(1-\theta)}{2(1+p_i)} + \sum_k \sum_l \sum_r \mu_{kl} \mu_{kr} q_k q_l q_r / 2 \\ &+ \sum_k \sum_l \sum_r \sum_s \mu_{kl} \mu_{rs} q_k q_l q_r q_s \frac{[\theta^2 + (1-\theta)^2] p_i + 2\theta(1-\theta)}{2(1+p_i)} - (\nu - \beta_i)^2 + \sigma_G^2 / 2. \end{aligned}$$

Assume that there is tight linkage between the trait locus and the marker, i.e.,  $\theta \approx 0$ . Then  $\beta_i \approx \alpha_i$ ,  $\sigma_{ir}^2 \approx \sigma_i^2$ ,  $\Sigma_{i,jr} \approx \Sigma_{i,j}$  and

$$\begin{aligned} \Sigma_{i,ir} &\approx \left[ \sum_k \sum_l \mu_{kl}^2 q_k h_{li} + \sum_k \sum_l \sum_r \mu_{kl} \mu_{kr} [q_l q_r h_{ki} + q_k h_{li} h_{ri}] \right. \\ &\left. + \sum_k \sum_l \sum_r \sum_s \mu_{kl} \mu_{rs} q_k q_r h_{li} h_{si} \right] / [4P(C_i)] - (\nu - \alpha_i)^2 + \sigma_G^2 / 2. \end{aligned}$$

## Appendix E

The loglikelihood function of model (5) is

$l = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^I \log |\Gamma_i| - \frac{1}{2} \sum_{i=1}^I (\bar{y}_i - X_i \gamma)^\tau \Gamma_i^{-1} (\bar{y}_i - X_i \gamma)$ . Assume that the data consist of both singleton families and sib-pair families. Suppose there are  $k_i$  singleton offspring who receive allele  $M_i$  from their heterozygous parents,  $k_{ii}$  ( $i = 1, 2, \dots, m$ ) sib pairs in each of them both sibs receive allele  $M_i$  from their heterozygous parents, and  $k_{ij} = k_{ji}$ ,  $i \neq j$  sib pairs in each of them one sib receives allele  $M_i$  from his/her heterozygous parent and the other receives allele  $M_j$  from the same heterozygous parent.

Let us denote  $\rho^\tau = (\rho_1 = \sigma_1^2, \rho_2 = \sigma_2^2, \dots, \rho_m = \Gamma_m^2, \rho_{m+1} = \Sigma_{1,1}, \dots, \rho_{2m} = \Sigma_{m,m}, \rho_{2m+1} = \Sigma_{1,2}, \dots, \rho_{3m-1} = \Sigma_{1,m}, \dots, \rho_{2m+m(m-1)/2} = \Sigma_{m-1,m})$ . We may get the following expected second partial derivatives for  $i, j, k = 1, \dots, m, i \neq j, j \neq k$

$$\begin{aligned} \frac{\partial^2 l}{\partial \gamma \partial \gamma^\tau} &= -X^\tau \Gamma^{-1} X, \quad E\left(\frac{\partial^2 l}{\partial \gamma \partial \rho^\tau}\right) = 0, \\ E\left(\frac{\partial^2 l}{\partial \rho_i^2}\right) &= E\left(\frac{\partial^2 l}{\partial (\sigma_i^2)^2}\right) = -\frac{k_i}{2(\sigma_i^2)^2} - \frac{k_{ii}[(\sigma_i^2)^2 + \Sigma_{i,i}^2]}{[(\sigma_i^2)^2 - \Sigma_{i,i}^2]^2} - \sum_{j \neq i} \frac{k_{ij}(\sigma_j^2)^2}{2[\sigma_i^2 \sigma_j^2 - \Sigma_{i,j}^2]^2}, \\ E\left(\frac{\partial^2 l}{\partial \rho_{m+i}^2}\right) &= E\left(\frac{\partial^2 l}{\partial \Sigma_{i,i}^2}\right) = -\frac{k_{ii}[(\sigma_i^2)^2 + \Sigma_{i,i}^2]}{[(\sigma_i^2)^2 - \Sigma_{i,i}^2]^2}, \quad E\left(\frac{\partial^2 l}{\partial \Sigma_{i,j}^2}\right) = -\frac{k_{ij}(\sigma_i^2 \sigma_j^2 + \Sigma_{i,j}^2)}{(\sigma_i^2 \sigma_j^2 - \Sigma_{i,j}^2)^2}, \\ E\left(\frac{\partial^2 l}{\partial \rho_i \partial \rho_j}\right) &= E\left(\frac{\partial^2 l}{\partial \sigma_i^2 \partial \sigma_j^2}\right) = -\frac{k_{ij} \Sigma_{i,j}^2}{2(\sigma_i^2 \sigma_j^2 - \Sigma_{i,j}^2)^2}, \\ E\left(\frac{\partial^2 l}{\partial \rho_i \partial \Sigma_{i,i}}\right) &= E\left(\frac{\partial^2 l}{\partial \sigma_i^2 \partial \Sigma_{i,i}}\right) = \frac{2k_{ii} \sigma_i^2 \Sigma_{i,i}}{[(\sigma_i^2)^2 - \Sigma_{i,i}^2]^2}, \quad E\left(\frac{\partial^2 l}{\partial \rho_i \partial \Sigma_{j,j}}\right) = E\left(\frac{\partial^2 l}{\partial \sigma_i^2 \partial \Sigma_{j,j}}\right) = 0, \\ E\left(\frac{\partial^2 l}{\partial \rho_i \partial \Sigma_{i,j}}\right) &= E\left(\frac{\partial^2 l}{\partial \sigma_i^2 \partial \Sigma_{i,j}}\right) = \frac{k_{ij} \sigma_j^2 \Sigma_{i,j}}{(\sigma_i^2 \sigma_j^2 - \Sigma_{i,j}^2)^2}, \quad E\left(\frac{\partial^2 l}{\partial \rho_i \partial \Sigma_{j,k}}\right) = E\left(\frac{\partial^2 l}{\partial \sigma_i^2 \partial \Sigma_{j,k}}\right) = 0, \\ E\left(\frac{\partial^2 l}{\partial \Sigma_{i,i} \partial \Sigma_{j,j}}\right) &= E\left(\frac{\partial^2 l}{\partial \Sigma_{i,i} \partial \Sigma_{j,k}}\right) = E\left(\frac{\partial^2 l}{\partial \Sigma_{i,j} \partial \Sigma_{k,l}}\right) = 0, \quad (i, j) \neq (k, l). \end{aligned}$$

Assume that  $k_i, k_{ii}, k_{ij} \rightarrow \infty, i, j = 1, \dots, m$ . To make it simple, assume  $k_{mm} = \min\{k_i, k_{ii}, k_{ij}\}$ . Then we can show that  $-\frac{1}{k_{mm}} \frac{\partial^2 l}{\partial \gamma \partial \gamma^\tau}$  and  $-\frac{1}{k_{mm}} E \left( \frac{\partial^2 l}{\partial \rho \partial \rho^\tau} \right)$  are positive definite. Now we are in a position to use the method in Miller (1977) and Pinheiro (1994) according to the theory of Weiss (1971, 1973). Actually, taking  $k_{mm}$  to replace  $v_j$  we can see that the key condition, i.e., Assumption 3.1.7 of Pinheiro (1994), p28, holds. Then by the same arguments in Pinheiro (1994), Chapter 3, we can show that  $\sqrt{k_{mm}} \hat{\gamma}$  converges to normal in distribution. This implies that the test statistic  $F_{het}$  is asymptotically  $F_{m-1, n-m}$  by considering the denominator of  $F_{het}$  as the estimate of mean squared error, which is independent of the numerator of  $F_{het}$  (Pinheiro 1994, pp28-29; Graybill 1976). In above discussion, we assume that there are sufficiently large data which include both trio families and sib-pair families. In addition, suppose we have nuclear families with any number children. We can show that  $-\frac{1}{k_{mm}} \frac{\partial^2 l}{\partial \gamma \partial \gamma^\tau}$  and  $-\frac{1}{k_{mm}} E \left( \frac{\partial^2 l}{\partial \rho \partial \rho^\tau} \right)$  are positive definite. Then, we can keep on using the method of Pinheiro (1994), chapters 2-3, to show that  $\sqrt{k_{mm}} \hat{\gamma}$  is asymptotically normal. Hence, the statistic  $F_{het}$  is asymptotically  $F(m-1, n-m)$ -distributed.

## Appendix F

If  $n_i = 1$  for each family, then there is only one child in each family. Let  $k_i, i = 1, 2, \dots, m$  be the number of offspring who receive allele  $M_i$  from their heterozygous parents. Let  $I_k$  be identity  $k \times k$  matrix.

$$\text{The design matrix and the variance-covariance matrix can be written as } X = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix},$$

$\Gamma = \text{diag}(\sigma_1^2 I_{k_1}, \dots, \sigma_m^2 I_{k_m})$ . Then we have  $X^\tau \Gamma^{-1} X = \text{diag}(k_1/\sigma_1^2, k_2/\sigma_2^2, \dots, k_m/\sigma_m^2)$ . Using a fact of inverse matrix  $(A + ab^\tau)^{-1} = A^{-1} - (A^{-1}a)(b^\tau A^{-1}) / (1 + b^\tau A^{-1}a)$ , we can calculate

$$\begin{aligned} (H[X^\tau \Gamma^{-1} X]^{-1} H^\tau)^{-1} &= \left[ \begin{pmatrix} \sigma_2^2/k_2 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & \sigma_m^2/k_m \end{pmatrix} + \frac{\sigma_1^2}{k_1} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} (1 \ \cdots \ 1) \right]^{-1} \\ &= \begin{pmatrix} k_2/\sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & k_m/\sigma_m^2 \end{pmatrix} - \begin{pmatrix} k_2/\sigma_2^2 \\ \vdots \\ k_m/\sigma_m^2 \end{pmatrix} \frac{(k_2/\sigma_2^2, \dots, k_m/\sigma_m^2)}{\frac{k_1}{\sigma_1^2} + \cdots + \frac{k_m}{\sigma_m^2}}. \end{aligned}$$

Therefore, the non-centrality parameter  $\lambda_{het, singleton} \approx (H\gamma)^\tau [H(X^\tau \Gamma^{-1} X)^{-1} H^\tau]^{-1} H$   
 $= \sum_{i=2}^m (\alpha_1 - \alpha_i)^2 k_i / \sigma_i^2 - \left[ \sum_{i=2}^m (\alpha_1 - \alpha_i) k_i / \sigma_i^2 \right]^2 / \left[ \sum_{i=1}^m k_i / \sigma_i^2 \right]$ .

## Appendix G

Such as in Appendix F, let us denote the variance-covariance matrix of the  $\sum_{i=1}^m k_i$  singleton offspring by  $\Gamma_1$ , and the related design matrix by  $X_1$ . Now let  $\Gamma_2$  denote the variance-covariance matrix of the  $\sum_{i=1}^m k_{ii}$  sib-pairs, in each of them both sibs receive the same allele from their heterozygous parents, and  $X_2$  the related design matrix. Then the form of  $X_2$  is similar to  $X_1$  given in Appendix F with different numbers of rows and  $\Gamma_2 = \text{diag}\left(\left(\begin{smallmatrix} \sigma_1^2 & \Sigma_{1,1} \\ \Sigma_{1,1} & \sigma_1^2 \end{smallmatrix}\right), \dots, \left(\begin{smallmatrix} \sigma_1^2 & \Sigma_{1,1} \\ \Sigma_{1,1} & \sigma_1^2 \end{smallmatrix}\right), \dots, \left(\begin{smallmatrix} \sigma_m^2 & \Sigma_{m,m} \\ \Sigma_{m,m} & \sigma_m^2 \end{smallmatrix}\right), \dots, \left(\begin{smallmatrix} \sigma_m^2 & \Sigma_{m,m} \\ \Sigma_{m,m} & \sigma_m^2 \end{smallmatrix}\right)\right)$ . Let  $\Gamma_3$  denote the variance-covariance matrix of the  $\sum_{i=1}^m \sum_{j>i} k_{ij}$  sib pairs, in each of them one sib receives one allele (i.e.,  $M_i, i = 1, 2, \dots, m$ , respectively) from his/her heterozygous parent and the other receives the other allele (i.e.,  $M_j, j \neq i, j = 1, 2, \dots, m$ , respectively) from the same heterozygous parent, and  $X_3$  be the related design matrix. The variance-covariance matrix is

$$\Gamma_3 = \text{diag}\left(\left(\begin{smallmatrix} \sigma_1^2 & \Sigma_{1,2} \\ \Sigma_{1,2} & \sigma_2^2 \end{smallmatrix}\right), \dots, \left(\begin{smallmatrix} \sigma_1^2 & \Sigma_{1,2} \\ \Sigma_{1,2} & \sigma_2^2 \end{smallmatrix}\right), \dots, \left(\begin{smallmatrix} \sigma_{m-1}^2 & \Sigma_{m-1,m} \\ \Sigma_{m-1,m} & \sigma_m^2 \end{smallmatrix}\right), \dots, \left(\begin{smallmatrix} \sigma_{m-1}^2 & \Sigma_{m-1,m} \\ \Sigma_{m-1,m} & \sigma_m^2 \end{smallmatrix}\right)\right).$$

The related design matrix is  $X_3 =$

$$\begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

In the same manner of Appendix F, we may obtain that

$$\begin{aligned} X_1^T \Gamma_1^{-1} X_1 &= \text{diag}\left(\frac{k_1}{\sigma_1^2}, \frac{k_2}{\sigma_2^2}, \dots, \frac{k_m}{\sigma_m^2}\right) \\ X_2^T \Gamma_2^{-1} X_2 &= \text{diag}\left(\frac{2k_{11}}{\sigma_1^2 + \Sigma_{1,1}}, \frac{2k_{22}}{\sigma_2^2 + \Sigma_{2,2}}, \dots, \frac{2k_{mm}}{\sigma_m^2 + \Sigma_{m,m}}\right). \end{aligned}$$

After some calculation, one may obtain that

$$X_3^T \Gamma_3^{-1} X_3 = \begin{pmatrix} \sum_{i \neq 1} \frac{k_{1i} \sigma_i^2}{\sigma_1^2 \sigma_i^2 - \Sigma_{1,i}^2} & -\frac{k_{12} \Sigma_{1,2}}{\sigma_1^2 \sigma_2^2 - \Sigma_{1,2}^2} & \cdots & -\frac{k_{1m} \Sigma_{1,m}}{\sigma_1^2 \sigma_m^2 - \Sigma_{1,m}^2} \\ -\frac{k_{12} \Sigma_{1,2}}{\sigma_1^2 \sigma_2^2 - \Sigma_{1,2}^2} & \sum_{i \neq 2} \frac{k_{2i} \sigma_i^2}{\sigma_2^2 \sigma_i^2 - \Sigma_{2,i}^2} & \cdots & -\frac{k_{2m} \Sigma_{2,m}}{\sigma_2^2 \sigma_m^2 - \Sigma_{2,m}^2} \\ \vdots & \vdots & \vdots & \vdots \\ -\frac{k_{1m} \Sigma_{1,m}}{\sigma_1^2 \sigma_m^2 - \Sigma_{1,m}^2} & -\frac{k_{2m} \Sigma_{2,m}}{\sigma_2^2 \sigma_m^2 - \Sigma_{2,m}^2} & \cdots & \sum_{i \neq m} \frac{k_{mi} \sigma_i^2}{\sigma_m^2 \sigma_i^2 - \Sigma_{m,i}^2} \end{pmatrix}.$$

## References

- Abecasis GR, Cardon LR, Cookson WOC (2000) A general test of association for quantitative traits in nuclear families. *Am J Hum Genet* 66:279–292.
- Allison DB (1997) Transmission-disequilibrium tests for quantitative traits. *Am J Hum Genet* 60:676–690.
- Cardon LR (2000) A sib-pair regression model of linkage disequilibrium for quantitative traits. *Hum Hered* 50:350–358.
- Cookson W, Abecasis G (2001) Oxford genome screen for asthma-associated traits. *Genet Epidemiol* 21 (Suppl 1):S1–S3.
- Daniel SE, Bhattacharya S, James A, et al (1996) A genome-wide search for quantitative trait loci underlying asthma. *Nature* 383:247–250.
- Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, ed 4, pp 15–19. London and New York: Longman.
- Fan R, Floros J, Xiong M (2002) Models and tests of linkage and association studies of QTL for multi-allele marker loci. *Hum Hered* 53:130–145.
- Fan RZ, Xiong MM (2003) Linkage and association studies of QTL for nuclear families by mixed models. *Biostatistics* (in press). [http://stat.tamu.edu/~rfan/paper.html/nuclear\\_family.pdf](http://stat.tamu.edu/~rfan/paper.html/nuclear_family.pdf).
- 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.
- Graybill FA (1976) Theory and Application of the Linear Model. Pacific Grove, California.
- Haseman JK, Elston RC (1972) The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* 2:3–19.
- Miller JJ (1977) Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance. *Ann Statistics* 5:746–762.
- Pinheiro JC (1994) Topics in Mixed-effects models. Ph.D thesis, University of Wisconsin-Madison.
- Pinheiro JC, Bates DM (2000) Mixed-effects models in S and S-plus. Springer, New York.
- Rabinowitz D (1997) A transmission disequilibrium test for quantitative trait loci. *Hum Hered* 47:342–350.
- 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.
- Sham PC, Curtis D (1995) An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. *Ann Hum Genet* 59:323–336.
- Spielman RS, Ewens WJ (1996) The TDT and other family-based tests for linkage disequilibrium and association. *Am J Hum Genet* 59:983–989.
- Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 52:506–516.
- Van den Oord EJCG, Sneider H (2002) Including measured genotypes in statistical models to study the interplay of multiple factors affecting complex traits. *Behavior Genetics* 32:1–22.
- Weiss L (1971) Asymptotic properties of maximum likelihood estimators in some nonstandard cases I. *J Am Stat Ass* 66:345–350.
- Weiss L (1973) Asymptotic properties of maximum likelihood estimators in some nonstandard cases II. *J Am Stat Ass* 68:428–430.
- Xiong MM, Krushkal J, Boerwinkle E (1998) TDT statistics for mapping quantitative loci. *Ann Hum Genet* 62:431–452.
- Zhu XF, Elston RC (2000) Power comparison of regression methods to test quantitative traits for association and linkage. *Genet Epidemiol* 18:322–330.
- Zhu XF, Elston RC (2001) Transmission/disequilibrium tests for quantitative traits. *Genet Epidemiol* 20:57–74.