

Models and Tests of Linkage and Association Studies of Quantitative Trait Locus for Multi-Allele Marker Loci

Ruzong Fan^a Joanna Floros^b Momiao Xiong^c

^aDepartment of Statistics, Texas A&M University, College Station, Tex., ^bDepartments of Cellular and Molecular Physiology and Pediatrics, Pennsylvania State University, Hershey, Pa., ^cHuman Genetics Center, University of Texas-Houston, Houston, Tex., USA

Key Words

Linkage · Linkage disequilibrium mapping · Quantitative trait locus · Transmission disequilibrium test statistic

Abstract

In this paper, we explore models and tests for association and linkage studies of a quantitative trait locus (QTL) linked to a multi-allele marker locus. Based on the difference between an offspring's conditional trait means of receiving and not receiving an allele from a parent at marker locus, we propose three statistics T_m , $T_{m,row}$ and $T_{m,col}$ to test association or linkage disequilibrium between the marker locus and the QTL. These tests are composite tests, and use the offspring marginal sample means including offspring data of both homozygous and heterozygous parents. For the linkage study, we calculate the offspring's conditional trait mean given the allele transmission status of a heterozygous parent at the marker locus. Based on the difference between the conditional means of a transmitted and a nontransmitted allele from a heterozygous parent, we propose statistics T_{parisi} , T_{satur} , T_{gen} and $T_{m,het}$ to perform composite tests of linkage between the marker locus and the quantitative trait locus in the presence of association. These tests only use the offspring data that are related to the heterozygous parents at the marker locus. T_{parisi} is a parsimonious or allele-wise statistic, T_{satur} and T_{gen} are saturated or

genotype-wise statistics, and $T_{m,het}$ compares the row and column sample means for offspring data of heterozygous parents. After comparing the powers and the sample sizes, we conclude that T_{parisi} has higher power than those of the bi-allele tests, T_{satur} , T_{gen} and $T_{m,het}$. If there is tight linkage between the marker and the trait locus, T_{parisi} is powerful in detecting linkage between the marker and the trait locus in the presence of association. By investigating the goodness-of-fit of T_{parisi} , we find that T_{satur} does not gain much power compared to that of T_{parisi} . Moreover, T_{parisi} takes into account the pattern of the data that is consistent with linkage and linkage disequilibrium. As the number of alleles at the marker locus increases, T_{parisi} is very conservative, and can be useful even for sparse data. To illustrate the usefulness and the power of the methods proposed in this paper, we analyze the chromosome 6 data of the Oxford asthma data, Genetic Analysis Workshop 12.

Copyright © 2002 S. Karger AG, Basel

Introduction

In the last decades, much research on qualitative traits has been done by using the methods of linkage mapping, haplotype mapping, and linkage disequilibrium mapping. Spielman et al. [1993] proposed a transmission disequilibrium test (TDT) to detect linkage between a bi-allele

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2002 S. Karger AG, Basel
0001-5652/02/0533-0130\$18.50/0

Accessible online at:
www.karger.com/journals/hhe

Dr. Ruzong Fan
Department of Statistics, Texas A&M University
3143 TAMUS College Station, TX 77843 (USA)
Tel. +1 979 845 3156 or 3141 (main office), Fax +1 979 845 3144
E-Mail rfan@stat.tamu.edu

marker locus and a disease locus in the presence of association. TDT has been a popular method for analysis of genetic data of complex diseases. Many authors have generalized the TDT to fit different situations [Boehnke and Langefeld, 1998; Kaplan et al., 1997; Martin et al., 1997]. Gordon et al. [2001] extended the TDT of qualitative traits to the TDT that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. For a multi-allele marker locus, one may use a direct generalization of TDT for data analysis [Bickeboller et al., 1995]. However, this method does not take into consideration the pattern of the data consistent with linkage and linkage disequilibrium [Sham and Curtis, 1995]. Moreover, the generalized TDT is χ^2 -distributed with big $t(t - 1)/2$ degrees of freedom for a marker locus with t alleles when t is relatively large. To overcome these drawbacks, Sham and Curtis [1995] derived transmission probabilities for a model with a multi-allele marker locus linked to a single disease locus. Then they justified the validity of this model by using logistic regression for data analysis of a random sample of affected individuals from a random mating population. A key point in Sham and Curtis [1995] is to assume that the recombination fraction between the disease locus and the marker locus is close to 0, i.e., tight linkage. Under this assumption, Sham and Curtis [1995] have been using logistic regression correctly to decrease the number of degrees of freedom to $t - 1$ for a marker locus with t alleles by proposing an allele-wise likelihood ratio test. Moreover, the logistic regression methods employed take into account the pattern of the data consistent with linkage and linkage disequilibrium.

Although much work on statistical analysis of linkage disequilibrium mapping of qualitative traits has been performed, the principle as well as the statistical methods of linkage disequilibrium required for mapping quantitative trait loci (QTL) are not well developed. Several investigators have studied TDT analysis for QTL in recent years. Allison [1997] proposed TDT-type tests. Rabinowitz [1997] did simulation analysis. Xiong et al. [1998] proposed a test statistic based on theoretical inference. Abecasis et al. [2000] proposed a general test of association for quantitative traits in nuclear families. Moreover, regression methods are used in linkage disequilibrium mapping of QTL. Allison et al. [1999] proposed a mixed-effect model for QTL data analysis to test linkage and association. George and Elston [1987] and George et al. [1999] in their pioneering work used a linear regression method for linkage and association analysis of transmission/disequilibrium test, between a bi-allele marker locus and a disease locus. Xiong et al. [2000], and Zhao et al. [2001] pro-

posed a method for mapping QTL in humans based on regression that deals with a bi-allele marker and can use both population data and family data.

So far, research in this promising field has been mainly focused on the bi-allele marker loci. To our knowledge, few TDT analyses for quantitative traits have seriously considered multiple allele markers. For multi-allele markers, one way to analyze data is to collapse them to bi-allele markers. Then one may use bi-allele marker test statistics to perform analysis. However, this may not be a good method because some information may be lost in the process of collapsing alleles. Also, the use of different ways to collapse a multi-allele marker may give different results, and hence may further complicate the interpretation of results. It would be better to perform overall tests that include all the alleles instead of several individual bi-allele tests.

In this paper, we investigate models and composite tests to detect association or linkage disequilibrium and linkage between a quantitative trait locus and a multi-allele marker locus. We first calculate the conditional expected trait mean given that a marker allele is transmitted or not transmitted from a parent. From the expression of the differences between the conditional and population means, we propose three composite tests T_m , $T_{m,row}$ and $T_{m,col}$ to test the association. These tests use all the data from the offspring of both homozygous and heterozygous parents. Moreover, $T_{m,row}$ and $T_{m,col}$ are composite tests from standard linear regression analysis.

In the presence of association, we explore methods to detect linkage between a quantitative trait locus and a marker locus. To achieve this, we first investigate a linear regression model for bi-allele marker locus. Then, we explore models to detect linkage between a multi-allele locus and a disease locus. We propose a direct generalization T_{gen} of TDT from a bi-allele marker locus to a multi-allele marker locus, a parsimonious test statistic T_{pars} , a saturated test statistic T_{satur} , and a marginal test statistic $T_{m,het}$. If there are only two alleles at the marker locus, the tests T_{gen} , T_{pars} , T_{satur} and $T_{m,het}$ are the same as the bi-allele tests of Xiong et al. [1998].

To show the advantage of the test statistics proposed in this paper over those available in the literature, we compare the powers of the tests with those in Allison [1997] and Xiong et al. [1998]. We find that the composite test statistic T_{pars} has the highest power, and it takes into account data patterns consistent with linkage and linkage disequilibrium such as the logistic regression described in Sham and Curtis [1995]. Moreover, we show that the composite test statistics are very conservative as the num-

Table 1. Summary of transmission data of a quantitative trait for a multi-allele marker

Transmitted allele	Nontransmitted allele					Row mean
	M_1	M_2	...	M_{t-1}	M_t	
M_1	$Y_{11}^{(k)}$	$Y_{12}^{(k)}$...	$Y_{1,t-1}^{(k)}$	$Y_{1t}^{(k)}$	$\bar{Y}_{1\cdot}^{(\cdot)}$
M_2	$Y_{21}^{(k)}$	$Y_{22}^{(k)}$...	$Y_{2,t-1}^{(k)}$	$Y_{2t}^{(k)}$	$\bar{Y}_{2\cdot}^{(\cdot)}$
.
.
.
M_{t-1}	$Y_{t-1,1}^{(k)}$	$Y_{t-1,2}^{(k)}$...	$Y_{t-1,t-1}^{(k)}$	$Y_{t-1,t}^{(k)}$	$\bar{Y}_{t-1,\cdot}^{(\cdot)}$
M_t	$Y_{t1}^{(k)}$	$Y_{t2}^{(k)}$...	$Y_{t,t-1}^{(k)}$	$Y_{tt}^{(k)}$	$\bar{Y}_{t\cdot}^{(\cdot)}$
Column mean	$\bar{Y}_{\cdot 1}^{(\cdot)}$	$\bar{Y}_{\cdot 2}^{(\cdot)}$...	$\bar{Y}_{\cdot,t-1}^{(\cdot)}$	$\bar{Y}_{\cdot t}^{(\cdot)}$	$\bar{Y}_{\cdot\cdot}^{(\cdot)}$

ber of alleles at the marker locus increases and can be useful even for sparse data, which is a merit observed by Terwilliger [1995] for qualitative traits based on likelihood methods.

To illustrate the usefulness and the power of the methods proposed in this paper, we analyze the chromosome 6 data of the Oxford asthma data, Genetic Analysis Workshop 12 [Cookson and Abecasis, 2001]. The data set is composed of 16 markers on chromosome 6, and each of the markers has 4 alleles. Using the SAS procedures and performing composite tests proposed in this paper, we confirm the findings of Daniel et al. [1996], i.e., markers of one region of chromosome 6 are potentially linked to an asthma phenotype. Moreover, markers in the region show association with the asthma phenotype. In addition, markers in other regions show significance of association or linkage, which do not show significance in the analysis of Daniel et al. [1996]. This provides evidence for increased detection sensitivity of significant association or linkage by the methods proposed in the present paper.

Models and Tests for Association Studies

Consider one quantitative trait locus L with alleles Q and q occurring with frequencies p_Q and p_q . Let Y be the phenotypic trait variable of an offspring, and the expected phenotypic trait value of a person with genotypes QQ , Qq , and qq be μ_{QQ} , $\mu_{Qq} = \mu_{qQ}$, and μ_{qq} , respectively. Suppose that a marker locus M is linked to the trait locus L . Denote the recombination fraction between the marker locus M and the trait locus L by θ . Assume that t alleles M_1, \dots, M_t

are typed at the marker locus M occurring with frequencies p_{M_1}, \dots, p_{M_t} . The haplotype frequencies are denoted by $h_{QM_1}, \dots, h_{QM_t}$, and $h_{qM_1}, \dots, h_{qM_t}$ for haplotypes QM_1, \dots, QM_t , and qM_1, \dots, qM_t , respectively. The measure of linkage disequilibrium between the trait allele Q and the marker allele M_i is defined by $\delta_i = h_{QM_i} - p_Q p_{M_i}$, $i = 1, \dots, t$. Then $\delta_i \neq 0$ refers to the presence of association or linkage disequilibrium between the trait allele Q and the marker allele M_i . Notice that the measure of linkage disequilibrium between the trait allele q and the marker allele M_i is $h_{qM_i} - p_q p_{M_i} = -h_{QM_i} + p_Q p_{M_i} = -\delta_i$, $i = 1, \dots, t$.

Consider a sample of pedigree data. For each family, there are two parents and one child. For each child, assume that an interesting quantitative trait is influenced by the gene locus L such as a blood pressure controlling gene locus. The data can be organized as in table 1. The row of table 1 represents the transmitted allele and the column represents the nontransmitted allele at the marker locus M . For a child in a trio family, one can determine which cell the child belongs to. Subscript notation ij indicates that allele M_i is transmitted and allele M_j is not transmitted. For a cell on row i and column j , one assumes that there are n_{ij} observations, and lets the trait value of a child of the k -th observation be $Y_{ij}^{(k)}$. The row sample size is denoted $n_{i\cdot} = \sum_{j=1}^t n_{ij}$, the column sample size $n_{\cdot j} = \sum_{i=1}^t n_{ij}$, and the total sample size $n = \sum_{i=1}^t \sum_{j=1}^t n_{ij}$. Then the cell sample mean is defined by $\bar{Y}_{ij}^{(\cdot)} = \sum_{k=1}^{n_{ij}} Y_{ij}^{(k)} / n_{ij}$, the row sample mean $\bar{Y}_{i\cdot}^{(\cdot)} = \sum_{j=1}^t \sum_{k=1}^{n_{ij}} Y_{ij}^{(k)} / n_{i\cdot}$, the column sample mean $\bar{Y}_{\cdot j}^{(\cdot)} = \sum_{i=1}^t \sum_{k=1}^{n_{ij}} Y_{ij}^{(k)} / n_{\cdot j}$, and the overall sample mean $\bar{Y}_{\cdot\cdot}^{(\cdot)} = \sum_{i=1}^t \sum_{j=1}^t \sum_{k=1}^{n_{ij}} Y_{ij}^{(k)} / n$.

Let TQ denote the abbreviation of ‘transmitted quantitative trait allele’. Then we have the conditional means $\mu_Q = E[Y|TQ = Q] = \mu_{QQ}p_Q + \mu_{Qq}p_q$ and $\mu_q = E[Y|TQ = q] = \mu_{qQ}p_Q + \mu_{qq}p_q$. Let TM denote the abbreviation of ‘transmitted marker allele’, and NM of ‘nontransmitted marker allele’. Then the expected row mean $\mu_i = E[Y|TM = M_i] = E[\bar{Y}_{i\cdot}^{(\cdot)}]$, and the expected column mean $v_i = E[Y|NM = M_i] = E[\bar{Y}_{\cdot i}^{(\cdot)}]$. The difference between the expected row and column means can be expressed as in Appendix A

$$\mu_i - v_i = (1 - \theta)(\mu_Q - \mu_q)\delta_i / [p_{M_i}(1 - p_{M_i})]. \quad (1)$$

The population trait mean can be expressed as $\mu = E[Y] = \mu_{QQ}p_Q^2 + 2\mu_{Qq}p_Qp_q + \mu_{qq}p_q^2 = \mu_Qp_Q + \mu_qp_q$, and the difference between the expected row mean and population mean can be expressed as in Appendix A

$$\mu_i - \mu = (1 - \theta)(\mu_Q - \mu_q)\delta_i / p_{M_i}. \quad (2)$$

The difference between the expected column mean and population mean can be expressed as in Appendix A

$$v_i - \mu = -(1 - \theta)(\mu_Q - \mu_q)\delta_i / (1 - p_{M_i}). \quad (3)$$

From equations (1), (2) and (3), we know that the presence of association between the trait locus L and the marker locus M for allele M_i (i.e., $\delta_i \neq 0$) is equivalent to $\mu_i = v_i$ or $\mu_i = \mu$ or $v_i = \mu$. Notice that the parameter estimates of μ , μ_i , and v_i are $\bar{Y}^{(\cdot)}$, $\bar{Y}_i^{(\cdot)}$ and $\bar{Y}_{\cdot i}^{(\cdot)}$, respectively. Therefore, we may construct the following statistic based on equation (1) to test association between the trait locus L and the marker locus M , i.e., to test the null hypothesis $H_0: \delta_1 = \dots = \delta_t = 0$

$$T_m = \frac{t-1}{t} \frac{\sum_{i=1}^t (\bar{Y}_i^{(\cdot)} - \bar{Y}^{(\cdot)})^2 / (1/n_i + 1/n_{\cdot i})}{\sum_{i=1}^t \sum_{j=1}^t [\sum_{k=1}^{n_{ij}} (Y_{ij}^{(k)} - \bar{Y}_i^{(\cdot)})^2 + \sum_{k=1}^{n_{ij}} (Y_{ij}^{(k)} - \bar{Y}_{\cdot i}^{(\cdot)})^2] / (2n - 2t)}$$

T_m has nearly a χ^2 distribution with $t - 1$ degrees of freedom under the null hypothesis H_0 (see Discussion, caution is noted). For qualitative traits, several authors have proposed statistics to test association and linkage between a disease locus and a multi-allele marker locus [Bickeböllner and Clerget-Darpourx, 1995; Jin et al., 1994; Spielman and Ewens, 1996; Kaplan et al., 1997]. One may view that T_m is the counterpart for a quantitative trait to test association. To construct a statistic for testing association between the quantitative trait locus L and the marker locus M based on equation (2), we consider the following linear regression models

$$Y_{ij}^{(k)} = \mu_i + e_{ijk}, \text{ full model;}$$

$$Y_{ij}^{(k)} = \mu + e_{ijk}, \text{ reduced model;}$$

where $i, j = 1, 2, \dots, t$, and $k = 1, 2, \dots, n_{ij}$, μ_i are row means, μ is overall mean, and e_{ijk} are normal variables with mean 0 and variance σ_e^2 . The parameter σ_e^2 can be estimated by error mean square

$$MSE_{m,row} = \sum_{i=1}^t \sum_{j=1}^t \sum_{k=1}^{n_{ij}} (Y_{ij}^{(k)} - \bar{Y}_i^{(\cdot)})^2 / (n - t).$$

The regression mean square is given by

$$MSR_{m,row} = \frac{\sum_{i=1}^t \sum_{j=1}^t \sum_{k=1}^{n_{ij}} (Y_{ij}^{(k)} - \bar{Y}^{(\cdot)})^2 - \sum_{i=1}^t \sum_{j=1}^t \sum_{k=1}^{n_{ij}} (Y_{ij}^{(k)} - \bar{Y}_i^{(\cdot)})^2}{(n-1) - (n-t)}$$

$$= \sum_{i=1}^t n_i (\bar{Y}_i^{(\cdot)} - \bar{Y}^{(\cdot)})^2 / (t-1).$$

The null hypothesis $H_0: \delta_1 = \dots = \delta_t = 0$ can be tested by statistics based on (2)

$$F_{m,row} = \frac{MSR_{m,row}}{MSE_{m,row}} = \frac{\sum_{i=1}^t n_i (\bar{Y}_i^{(\cdot)} - \bar{Y}^{(\cdot)})^2 / (t-1)}{\sum_{i=1}^t \sum_{j=1}^t \sum_{k=1}^{n_{ij}} (Y_{ij}^{(k)} - \bar{Y}_i^{(\cdot)})^2 / (n-t)}$$

and

$$T_{m,row} = (t-1)F_{m,row}.$$

Similarly, the following test statistics can be constructed based on (3)

$$F_{m,col} = \frac{\sum_{i=1}^t n_i (\bar{Y}_{\cdot i}^{(\cdot)} - \bar{Y}^{(\cdot)})^2 / (t-1)}{\sum_{i=1}^t \sum_{j=1}^t \sum_{k=1}^{n_{ij}} (Y_{ij}^{(k)} - \bar{Y}_{\cdot i}^{(\cdot)})^2 / (n-t)}$$

and

$$T_{m,col} = (t-1)F_{m,col}.$$

The statistics $F_{m,row}$ and $F_{m,col}$ are approximately F distributed with $(t-1, n-t)$ degrees of freedom under the null hypothesis H_0 . When the total number $n-t$ is large, $T_{m,row}$ and $T_{m,col}$ are approximately χ^2 distributed with $t-1$ degrees of freedom.

Models and Tests for Linkage Studies

Bi-Allele Marker Locus

Assume that there are two alleles $M = M_1$ and $m = M_2$ occurring with population frequencies p_M and p_m at the marker locus M . Let the haplotype frequencies of haplotypes QM , Qm , qM and qm be h_{QM} , h_{Qm} , h_{qM} and h_{qm} , respectively. The linkage disequilibrium coefficient is defined by $\delta = h_{QM} - p_Q p_M$. Then $\delta \neq 0$ refers to the presence of association between the marker locus M and the quantitative trait locus L . Let us list $Y_{12}^{(k)}$ and $Y_{21}^{(k)}$ as Y_i , $i = 1, 2, \dots, n_{12} + n_{21}$. In the case of no covariates, we consider a linear regression model

$$Y_i = \gamma + \alpha 1_{(TM=M, NM=M)} + e_i, \text{ full model,} \quad (4)$$

where γ is a mean parameter, α is a parameter to describe the expected trait difference of a child between the transmission of allele M and nontransmission of allele M , and e_i are normal variables with mean 0 and variance σ_e^2 . We show in Appendix B that

$$\alpha = (1 - 2\theta)\delta[(\mu_{QQ} - \mu_{Qq})p_Q + (\mu_{Qq} - \mu_{qq})p_q] / [p_M p_m]. \quad (5)$$

From equation (5) we know that in the presence of association between trait locus L and marker locus M (i.e., $\delta \neq 0$), the hypothesis of no linkage between trait locus L and marker locus M (i.e., $H_0: \theta = 1/2$) is equivalent to $\alpha = 0$. Hence, we use the standard linear regression method to test linkage simply by testing if the coefficient $\alpha = 0$. Some calculation gives the coefficient estimates $\hat{\gamma} = \bar{Y}_m$, $\hat{\alpha} = \bar{Y}_M - \bar{Y}_m$, where $\bar{Y}_m = \bar{Y}_{21}^{(\cdot)}$ is the sample mean of values Y_i which corresponds to allele m transmitted from a heterozygous parent, and $\bar{Y}_M = \bar{Y}_{12}^{(\cdot)}$ is the sample mean of values which corresponds to allele M transmitted from a heterozygous parent, respectively. To get a valid test statistic of no linkage, we need to calculate the error mean square

$$MSE = \frac{1}{n_{12} + n_{21} - 2} \sum_{i=1}^{n_{12} + n_{21}} (Y_i - \hat{\gamma} - \hat{\alpha} 1_{(TM=M, NM=m)})^2$$

$$= \frac{1}{n_{12} + n_{21} - 2} \left[\sum_{k=1}^{n_{12}} (Y_{12}^{(k)} - \bar{Y}_M)^2 + \sum_{k=1}^{n_{21}} (Y_{21}^{(k)} - \bar{Y}_m)^2 \right].$$

Consider a simplified linear model

$$Y_i = \beta + e_i, \text{ reduced model.} \quad (6)$$

The coefficient estimate $\hat{\beta} = \sum_{i=1}^{n_{12}+n_{21}} Y_i / (n_{12} + n_{21}) = \bar{Y}$.
The regression mean square of models (4) and (6) is

$$MSR = \frac{\sum_{i=1}^{n_{12}+n_{21}} (Y_i - \bar{Y})^2 - \sum_{i=1}^{n_{12}+n_{21}} (Y_i - \hat{\gamma} - \hat{\alpha}1_{(TM=M, NM=m)})^2}{(n_{12} + n_{21} - 1) - n_{12} + (n_{21} - 2)}$$

$$= \frac{(\bar{Y}_M - \bar{Y}_m)^2}{1/n_{12} + 1/n_{21}}$$

The test statistic for null hypothesis $H_0: \theta = 1/2$ or $\alpha = 0$ can be constructed by

$$T = \frac{MSR}{MSE} = \frac{(\bar{Y}_M - \bar{Y}_m)^2 / (1/n_{12} + 1/n_{21})}{[\sum_{k=1}^t (Y_{12}^{(k)} - \bar{Y}_M)^2 + \sum_{k=1}^t (Y_{21}^{(k)} - \bar{Y}_m)^2] / (n_{12} + n_{21} - 2)} \quad (7)$$

which is the same as TDT_1 in Xiong et al. [1998]. For a multiallele marker M , one may collapse the alleles to make a bi-allele marker, and use the bi-allele T in equation (7) to test linkage between the marker and the trait loci. Let us denote the heterozygous row sample size by $n_{i \cdot, het} = \sum_{j=1}^t n_{ij} - n_{ii}$, and the heterozygous column sample size by $n_{\cdot, j, het} = \sum_{i=1}^t n_{ij} - n_{ij}$. Moreover, we denote the heterozygous row sample mean $\bar{Y}_{i \cdot, het}^{(\cdot)} = \sum_{j=1, j \neq i}^t \sum_{k=1}^{n_{ij}^{(k)}} Y_{ij}^{(k)} / n_{i \cdot, het}$, and the heterozygous column sample mean $\bar{Y}_{\cdot, j, het}^{(\cdot)} = \sum_{i=1, i \neq j}^t \sum_{k=1}^{n_{ij}^{(k)}} Y_{ij}^{(k)} / n_{\cdot, j, het}$. Considering the bi-alleles M_i and $\cup_{j \neq i} M_j$ which collapses all but allele M_i , the individual test statistic can be constructed as

$$T_{i, clps} = \frac{(\bar{Y}_{i \cdot, het}^{(\cdot)} - \bar{Y}_{\cdot, i, het}^{(\cdot)})^2 / (1/n_{i \cdot, het} + 1/n_{\cdot, i, het})}{\sum_{j=1, j \neq i}^t [\sum_{k=1}^{n_{ij}^{(k)}} (Y_{ij}^{(k)} - \bar{Y}_{i \cdot, het}^{(\cdot)})^2 + \sum_{k=1}^{n_{ji}^{(k)}} (Y_{ji}^{(k)} - \bar{Y}_{\cdot, i, het}^{(\cdot)})^2] / [n_{i \cdot, het} + n_{\cdot, i, het} - 2]}$$

Similarly, the TDT_{Q1} in Allison [1997] of bi-alleles M_i and $\cup_{j \neq i} M_j$ can be written as

$$TDT_{Q1, i, clps} = \frac{(\bar{Y}_{i \cdot, het}^{(\cdot)} - \bar{Y}_{\cdot, i, het}^{(\cdot)})^2 / [2 / (n_{i \cdot, het} + n_{\cdot, i, het})]}{\sum_{j=1, j \neq i}^t \left[\frac{\sum_{k=1}^{n_{ij}^{(k)}} (Y_{ij}^{(k)} - \bar{Y}_{i \cdot, het}^{(\cdot)})^2}{n_{i \cdot, het} - 1} + \frac{\sum_{k=1}^{n_{ji}^{(k)}} (Y_{ji}^{(k)} - \bar{Y}_{\cdot, i, het}^{(\cdot)})^2}{n_{\cdot, i, het} - 1} \right]}$$

If there are covariates such as age and sex, we may use multiple regression models to test the null hypothesis $H_0: \theta = 1/2$ or $\alpha = 0$ by adding more terms to linear regressions (4) and (6).

Multi-Allele Marker Locus, Saturated Models and Generalized Tests

For a multi-allele marker locus M with t alleles M_1, \dots, M_t , we consider a saturated or genotype-wise linear regression model for $i, j = 1, 2, \dots, t, i \neq j$ of direct generalization of bi-allele marker model (4)

$$Y_{ij}^{(k)} = \gamma_{ij} + \alpha_{ij}1_{(TM=M_i, NM=M_j)} + e_{ijk}, \quad (8)$$

where $\gamma_{ij} = \gamma_{ji}$ are mean parameters, α_{ij} is the expected trait difference between cell ij and cell ji , and e_{ijk} are normal variables with 0 and variance σ_e^2 . We show in Appendix B that

$$\alpha_{ij} = (1 - 2\theta)\delta_{ij}[(\mu_{Q0} - \mu_{Qq})p_Q + (\mu_{Qq} - \mu_{qq})p_q] / [p_M p_{M_j}], \quad (9)$$

where $\delta_{ij} = h_{QM} p_{M_j} - h_{QM} p_{M_i}$. From equation (9) we know that in the presence of association between the trait locus L and the marker locus M for alleles M_i and M_j (i.e., $\delta_{ij} \neq 0$), the hypothesis of no linkage between trait locus L and marker locus M (i.e., $H_0: \theta = 1/2$) is equivalent to $\alpha_{ij} = 0$. Hence, We use the linear regression method to test linkage in the presence of association simply by testing if the coefficient $\alpha_{ij} = 0$. In the following, assume $\delta_{ij} \neq 0$ for all $i \neq j$. Then the null hypothesis $H_0: \theta = 1/2$ is equivalent to $\alpha_{ij} = 0$ for all $i < j$. Assume that there are no covariates in model (8), then the coefficient estimates are given by $\hat{\gamma}_{ij} = \bar{Y}_{ij}^{(\cdot)}$ and $\hat{\alpha}_{ij} = \bar{Y}_{ij}^{(\cdot)} - \bar{Y}_{ji}^{(\cdot)}$. We may then estimate the variance by the error mean square of model

$$MSE_{satur} = MSE_{gen} = \frac{\sum_{i=1}^t \sum_{j \neq i} \sum_{k=1}^{n_{ij}^{(k)}} (Y_{ij}^{(k)} - \hat{\gamma}_{ij} - \hat{\alpha}_{ij}1_{(i < j)})^2}{\sum_{i=1}^t \sum_{j \neq i} (n_{ij} - 1)}$$

$$= \frac{\sum_{i=1}^{t-1} \sum_{j=i+1}^t [\sum_{k=1}^{n_{ij}^{(k)}} (Y_{ij}^{(k)} - \bar{Y}_{ij}^{(\cdot)})^2 + \sum_{k=1}^{n_{ji}^{(k)}} (Y_{ji}^{(k)} - \bar{Y}_{ji}^{(\cdot)})^2]}{n - \sum_{i=1}^t n_{ii} - t(t-1)}$$

where $n_{\cdot, \cdot, het} = n - \sum_{i=1}^t n_{ii}$ is the total number of sample size except the data on the diagonal of table 1. If $\alpha_{ij} = 0$ for all $i < j$, model (8) becomes a simplified linear regression model for $i, j = 1, 2, \dots, t, i \neq j, k = 1, \dots, n_{ij}, \beta_{ij} = \beta_{ji}$,

$$Y_{ij}^{(k)} = \beta_{ij} + e_{ijk}.$$

The coefficient estimates are

$$\hat{\beta}_{ij} = \frac{n_{ij} \bar{Y}_{ij}^{(\cdot)} + n_{ji} \bar{Y}_{ji}^{(\cdot)}}{n_{ij} + n_{ji}}$$

The regression mean square is

$$MSR_{gen} = \frac{\sum_{i=1}^t \sum_{j \neq i} \sum_{k=1}^{n_{ij}^{(k)}} (Y_{ij}^{(k)} - \hat{\beta}_{ij})^2 - \sum_{i=1}^t \sum_{j \neq i} \sum_{k=1}^{n_{ij}^{(k)}} (Y_{ij}^{(k)} - \hat{\gamma}_{ij} - \hat{\alpha}_{ij}1_{(i < j)})^2}{[n - \sum_{i=1}^t n_{ii} - t(t-1)/2] - [n - \sum_{i=1}^t n_{ii} - t(t-1)]}$$

$$= \sum_{i=1}^{t-1} \sum_{j=i+1}^t [\bar{Y}_{ij}^{(\cdot)} - \bar{Y}_{ji}^{(\cdot)}]^2 / [(1/n_{ij} + 1/n_{ji})t(t-1)/2].$$

If all disequilibrium coefficients δ_{ij} are not equal to 0, we may construct a direct generalized test statistics

$$F_{gen} = \frac{MSR_{gen}}{MSE_{gen}}$$

$$= \sum_{i=1}^{t-1} \sum_{j=i+1}^t \frac{[\bar{Y}_{ij}^{(\cdot)} - \bar{Y}_{ji}^{(\cdot)}]^2 / [(1/n_{ij} + 1/n_{ji})t(t-1)/2]}{\sum_{i=1}^t \sum_{j \neq i} \sum_{k=1}^{n_{ij}^{(k)}} (Y_{ij}^{(k)} - \bar{Y}_{ij}^{(\cdot)})^2 / [n - \sum_{i=1}^t n_{ii} - t(t-1)]}$$

$$T_{gen} = [t(t-1)/2] F_{gen}$$

to test the hypothesis of no linkage $\theta = 1/2$. The statistic F_{gen} is approximately F distributed with $(t(t-1)/2, n -$

$\sum_{i=1}^t n_{ii} - t(t-1)$) as degrees of freedom. When the number $n - \sum_{i=1}^t n_{ii} - t(t-1)$ is large, T_{gen} is approximately χ^2 -distributed with $t(t-1)/2$ degrees of freedom. When the number of alleles (i.e., t) is relatively large, the degrees of freedom $t(t-1)/2$ can be too large. Although T_{gen} is a valid test statistic to detect linkage in the presence of association, the large number of degrees of freedom would be a drawback. Moreover, it treats each marker genotype separately although each marker allele can be present in $t-1$ heterozygous genotypes. If a marker locus is closely linked to the trait locus, some marker alleles may be more likely to be transmitted with the trait allele Q or q than other marker alleles. Hence certain patterns of data may be more consistent with linkage and linkage disequilibrium than other patterns. However, T_{gen} does not take into account this possibility sufficiently.

Multi-Allele Marker Locus, Parsimonious Models

If the trait locus and marker locus are not close, the recombination fraction θ is not small. In this case, tight linkage is absent, and the degree of linkage disequilibrium decreases very quickly after a small number of generations in a randomly mating population [Falconer and Mackay, 1996], due to the recombination between the trait locus and the marker locus. Hence the presence of linkage disequilibrium is usually the result of tight linkage between the trait and the marker loci. Assume that the trait locus L and the marker locus M are very close, and thus the recombination fraction $\theta \approx 0$. Then the equation (16) in Appendix C implies that for $i \neq j$, $E[Y_{ij}^{(k)}] = E[Y|M_i \text{ is transmitted but } M_j \text{ is not transmitted}] \approx [\mu_Q h_{QM_i} + \mu_q h_{qM_i}] / p_{M_i}$, which does not depend on allele/haplotype frequencies related to the nontransmitted marker allele M_j . This observation prompts one to propose the following parsimonious or allele-wise linear regression to analyze the data

$$Y_{ij}^{(k)} = \mu_{i,het} + e_{ijk}, \text{ full model,} \quad (10)$$

where $i, j = 1, 2, \dots, t, i \neq j, k = 1, 2, \dots, n_{ij}$, $\mu_{i,het}$ are mean parameters, and e_{ijk} are sampling errors which are distributed as normal with mean 0 and variance σ_e^2 . In model (10), we assume that there are no covariates. The parameter estimates for linear regression (10) are given by the row means

$$\hat{\mu}_{i,het} = \frac{\sum_{j \neq i} \sum_k Y_{ij}^{(k)}}{\sum_{j \neq i} n_{ij}} = \bar{Y}_{i, \cdot, het}^{(\cdot)}$$

Moreover, the variance σ_e^2 can be estimated

$$MSE_{parisi} = \frac{\sum_{i=1}^t \sum_{j \neq i} \sum_{k=1}^{n_{ij}} (Y_{ij}^{(k)} - \bar{Y}_{i, \cdot, het}^{(\cdot)})^2}{n - \sum_{i=1}^t n_{ii} - t}$$

If marker locus M is not linked to the trait locus L , then $E[Y_{ij}^{(k)}] = E[Y] = \mu_{QQ}p_Q^2 + 2\mu_{Qq}p_Qp_q + \mu_{qq}p_q^2$, which is the expected trait value in a population and does not depend on both marker alleles M_i and M_j . Hence coefficients $\mu_{i,het}$ should not depend on subscript i . Therefore, the null hypothesis of no linkage between the trait and the marker loci is equivalent to $\mu_{i,het} = \dots = \mu_{t,het}$. Under this assumption, model (10) is simplified to

$$Y_{ij}^{(k)} = \mu_{het} + e_{ijk}, i \neq j, \text{ reduced model,} \quad (11)$$

where μ_{het} is mean parameter. The estimate of μ_{het} is the overall sample mean based on heterozygous data

$$\hat{\mu}_{het} = \frac{\sum_{i=1}^t \sum_{j \neq i} \sum_k Y_{ij}^{(k)}}{n - \sum_{i=1}^t n_{ii}} = \bar{Y}_{\cdot, \cdot, het}^{(\cdot)}$$

The regression mean square

$$MSR_{parisi} = \frac{\sum_{i=1}^t \sum_{j \neq i} \sum_k (Y_{ij}^{(k)} - \bar{Y}_{i, \cdot, het}^{(\cdot)})^2 - \sum_{i=1}^t \sum_{j \neq i} \sum_k (Y_{ij}^{(k)} - \bar{Y}_{\cdot, \cdot, het}^{(\cdot)})^2}{(n - \sum_{i=1}^t n_{ii} - 1) - (n - \sum_{i=1}^t n_{ii} - t)}$$

$$= \frac{\sum_{i=1}^t n_{i, \cdot, het} (\bar{Y}_{i, \cdot, het}^{(\cdot)} - \bar{Y}_{\cdot, \cdot, het}^{(\cdot)})^2}{t - 1}$$

To test if there is linkage in the presence of association between the trait locus L and the marker locus M , one may use standard test statistics from analysis of variance (ANOVA)

$$F_{parisi} = \frac{MSR_{parisi}}{MSE_{parisi}} = \frac{\sum_{i=1}^t n_{i, \cdot, het} (\bar{Y}_{i, \cdot, het}^{(\cdot)} - \bar{Y}_{\cdot, \cdot, het}^{(\cdot)})^2 / (t - 1)}{\sum_{i=1}^t \sum_{j \neq i} \sum_{k=1}^{n_{ij}} (Y_{ij}^{(k)} - \bar{Y}_{i, \cdot, het}^{(\cdot)})^2 / (n - \sum_{i=1}^t n_{ii} - t)}$$

$$T_{parisi} = (t - 1)F_{parisi}$$

F_{parisi} is approximately F -distributed with $(t - 1, n - \sum_{i=1}^t n_{ii} - t)$ as degrees of freedom. When the number $n - \sum_{i=1}^t n_{ii} - t$ is large, T_{parisi} is approximately χ^2 -distributed with $t - 1$ degrees of freedom.

The regression mean square between saturated model (8) and reduced model (11) is

$$MSR_{satur} = \frac{\sum_{i=1}^t \sum_{j \neq i} \sum_k (Y_{ij}^{(k)} - \bar{Y}_{\cdot, \cdot, het}^{(\cdot)})^2 - \sum_{i=1}^t \sum_{j \neq i} \sum_k (Y_{ij}^{(k)} - \hat{y}_{ij} - \hat{\alpha}_{ij}1_{(i < j)})^2}{[n - \sum_{i=1}^t n_{ii} - t] - [n - \sum_{i=1}^t n_{ii} - t(t - 1)]}$$

$$= \frac{\sum_{i=1}^{t-1} \sum_{j=i+1}^t [n_{ij} (\bar{Y}_{ij}^{(\cdot)} - \bar{Y}_{\cdot, \cdot, het}^{(\cdot)})^2 + n_{ij} (\bar{Y}_{ji}^{(\cdot)} - \bar{Y}_{\cdot, \cdot, het}^{(\cdot)})^2]}{t(t - 2)}$$

To judge whether the intermediate test statistic T_{parisi} fits the data sufficiently, we may use the following test statistic

$$F_{satur} = \frac{MSR_{satur}}{MSE_{satur}}$$

$$= \frac{\sum_{i=1}^{t-1} \sum_{j=i+1}^t [n_{ij} (\bar{Y}_{ij}^{(\cdot)} - \bar{Y}_{\cdot, \cdot, het}^{(\cdot)})^2 + n_{ji} (\bar{Y}_{ji}^{(\cdot)} - \bar{Y}_{\cdot, \cdot, het}^{(\cdot)})^2] / t(t - 2)}{\sum_{i=1}^{t-1} \sum_{j=i+1}^t [\sum_{k=1}^{n_{ij}} (Y_{ij}^{(k)} - \bar{Y}_{ij}^{(\cdot)})^2 + \sum_{k=1}^{n_{ji}} (Y_{ji}^{(k)} - \bar{Y}_{ji}^{(\cdot)})^2]}{n - \sum_{i=1}^t n_{ii} - t(t - 1)}$$

$$T_{satur} = [t(t - 1) - 1]F_{satur}$$

When the number $n - \sum_{i=1}^t n_{ii} - t(t-1)$ is large, T_{satur} is approximately χ^2 -distributed with $t(t-1) - 1$ degrees of freedom. To test the goodness of fit of the parsimonious linear model (10), we first calculate the regression mean square

$$\begin{aligned} MSR_{fit} &= \frac{\sum_{i=1}^t \sum_{j \neq i} \sum_k (Y_{ij}^{(k)} - \bar{Y}_{i,\cdot,het}^{(\cdot)})^2 - \sum_{i=1}^t \sum_{j \neq i} \sum_k (Y_{ij}^{(k)} - \hat{Y}_{ij} - \hat{\alpha}_{ij}1_{(i < j)})^2}{[n - \sum_{i=1}^t n_{ii} - t] - [n - \sum_{i=1}^t n_{ii} - t(t-1)]} \\ &= \frac{\sum_{i=1}^{t-1} \sum_{j=i+1}^t [n_{ij} (\bar{Y}_{ij}^{(\cdot)} - \bar{Y}_{i,\cdot,het}^{(\cdot)})^2 + n_{ji} (\bar{Y}_{ji}^{(\cdot)} - \bar{Y}_{j,\cdot,het}^{(\cdot)})^2]}{t(t-2)}. \end{aligned}$$

The test statistic of the goodness of fit can be constructed as

$$\begin{aligned} F_{fit} &= \frac{MSR_{fit}}{MSE_{satur}} \\ &= \frac{\sum_{i=1}^{t-1} \sum_{j=i+1}^t [n_{ij} (\bar{Y}_{ij}^{(\cdot)} - \bar{Y}_{i,\cdot,het}^{(\cdot)})^2 + n_{ji} (\bar{Y}_{ji}^{(\cdot)} - \bar{Y}_{j,\cdot,het}^{(\cdot)})^2] / t(t-2)}{\sum_{i=1}^{t-1} \sum_{j=i+1}^t [\sum_{k=1}^{n_{ij}} (Y_{ij}^{(k)} - \bar{Y}_{ij}^{(\cdot)})^2 + \sum_{k=1}^{n_{ji}} (Y_{ji}^{(k)} - \bar{Y}_{ji}^{(\cdot)})^2]}{n - \sum_{i=1}^t n_{ii} - t(t-1)} \end{aligned}$$

$$T_{fit} = t(t-2)F_{fit}.$$

When the number $n - \sum_{i=1}^t n_{ii} - t(t-1)$ is large, T_{fit} is approximately χ^2 -distributed with $t(t-2)$ degrees of freedom.

Marginal Test Based on Heterozygous Data

In Appendix D, we calculate the conditional mean given that allele M_i is transmitted from a heterozygous parent $\mu_{i,het} = E[Y|TM = M_i, NM = \cup_{j \neq i} M_j]$, and the conditional mean given that allele M_i is not transmitted from a heterozygous parent $\nu_{i,het} = E[Y|NM = M_i, TM = \cup_{j \neq i} M_j]$. Then, we get the conditional mean difference

$$\mu_{i,het} - \nu_{i,het} = (1 - 2\theta) \frac{\delta_i (\mu_q - \mu_p)}{p_{M_i} (1 - p_{M_i})}. \quad (12)$$

From equation (12) we know that in the presence of association between trait locus L and marker locus M for alleles M_i (i.e., $\delta_i \neq 0$), the hypothesis of no linkage between trait locus L and marker locus M (i.e., $H_0: \theta = 1/2$) is equivalent to $\mu_{i,het} = \nu_{i,het}$. Hence we can construct statistics to test linkage in the presence of association simply by testing $\mu_{i,het} = \nu_{i,het}$. In the following, we assume $\delta_i \neq 0$ for all $i = 1, 2, \dots, t$. Then the null hypothesis $H_0: \theta = 1/2$ is equivalent to $\mu_{i,het} = \nu_{i,het}$ for all $i = 1, 2, \dots, t$. Notice that the parameter estimates of $\mu_{i,het}$ and $\nu_{i,het}$ are $\bar{Y}_{i,\cdot,het}^{(\cdot)}$ and $\bar{Y}_{\cdot,i,het}^{(\cdot)}$, respectively. Therefore, the following statistic can be constructed based on equation (12) to test the linkage in the presence of association between trait locus L and marker locus M , i.e., to test the null hypothesis $H_0: \theta = 1/2$

$$T_{m,het} = \frac{t-1 \sum_{i=1}^t (\bar{Y}_{i,\cdot,het}^{(\cdot)} - \bar{Y}_{\cdot,i,het}^{(\cdot)})^2 / (1/n_{i,\cdot,het} + 1/n_{\cdot,i,het})}{MSE_{m,het}},$$

where $MSE_{m,het}$ is the parameter estimate of variance σ_e^2 given by

$$MSE_{m,het} = \frac{\sum_{i=1}^t \sum_{j=1, j \neq i}^t [\sum_{k=1}^{n_{ij}} (Y_{ij}^{(k)} - \bar{Y}_{ij}^{(\cdot)})^2 + \sum_{k=1}^{n_{ji}} (Y_{ji}^{(k)} - \bar{Y}_{ji}^{(\cdot)})^2]}{2n - 2 \sum_{i=1}^t n_{ii} - 2t}.$$

Such as T_m in 'Models and Tests for Association Studies' above $T_{m,het}$ has nearly a χ^2 distribution with $t-1$ degrees of freedom under the null hypothesis H_0 (see 'Discussion', caution is noted).

Calculation of Noncentrality Parameters

In this section, we assume that there are no covariates. The statistic T_{gen} is distributed approximately as a central χ^2 with $t(t-1)/2$ degrees of freedom under the null hypothesis $H_0: \theta = 1/2$ in the presence of association between the trait locus and the marker locus for all marker alleles M_i and M_j , $i \neq j$. Under the alternative hypothesis, it is distributed approximately as a noncentral χ^2 with noncentrality parameter λ_{gen} given in Appendix E. Similarly, we calculate the noncentrality parameter λ_{fit} of T_{fit} in Appendix E. Under the null hypothesis H_0 , the test statistic T_{parisi} is distributed approximately as a χ^2 with $t-1$ degrees of freedom. Otherwise, it is distributed approximately as a noncentral χ^2 with noncentrality parameter (Appendix E)

$$\lambda_{parisi} = \frac{\sum_{i=1}^t n_{i,het} (\mu_{i,het} - \mu_{het})^2}{\sum_{i=1}^t (n_{i,het} - 1) \sigma_{ir,het}^2 / (n - \sum_{i=1}^t n_{ii} - t)},$$

where the conditional variance $\sigma_{ir,het}^2$ and mean μ_{het} are given in Appendix E. $T_{m,het}$ has nearly a χ^2 distribution with $t-1$ degrees of freedom under the null hypothesis H_0 . Under the alternative hypothesis, it is nearly distributed as a noncentral χ^2 with noncentrality parameter (Appendix E)

$$\lambda_{m,het} = \frac{t-1 \sum_{i=1}^t (\mu_{i,het} - \nu_{i,het})^2 / (1/n_{i,\cdot,het} + 1/n_{\cdot,i,het})}{t \sum_{i=1}^t [(n_{i,\cdot,het} - 1) \sigma_{ir,het}^2 + (n_{\cdot,i,het} - 1) \sigma_{ic,het}^2] / (2n - 2 \sum_{i=1}^t n_{ii} - 2t)},$$

where $\sigma_{ic,het}^2$ is given in Appendix E. Moreover, we calculate the noncentrality parameters λ_{satur} , $\lambda_{i,clps}$, λ_m , $\lambda_{m,row}$, $\lambda_{m,col}$, and $\lambda_{Q1,i,clps}$ of T_{satur} , $T_{i,clps}$, T_m , $T_{m,row}$, $T_{m,col}$, and $TDT_{Q1,i,clps}$ in Appendix E.

To calculate the noncentrality parameters λ_{parisi} , λ_{satur} , λ_{gen} , λ_{fit} , $\lambda_{m,het}$, $\lambda_{i,clps}$, λ_m , $\lambda_{m,row}$, $\lambda_{m,col}$, and $\lambda_{Q1,i,clps}$, we need parameters such as the marker allele frequencies p_{M_i} , trait

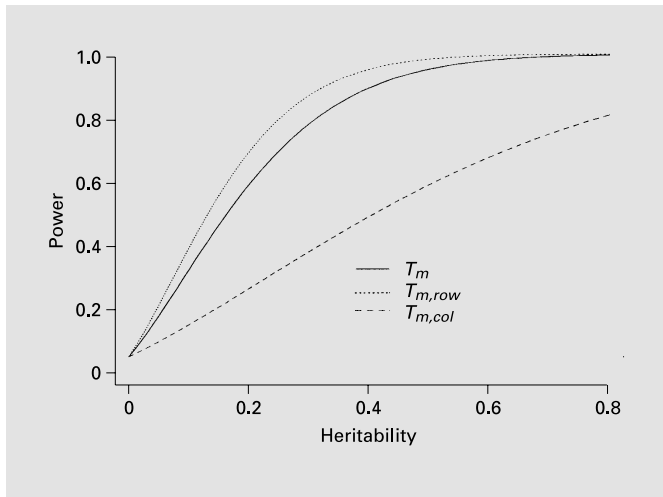


Fig. 1. Power curves of T_m , $T_{m,row}$ and $T_{m,col}$ when $p_Q = 0.15$, $\theta = 0.005$, $A = 20$, $t = 3$, $p_{M_1} = 0.40$, $p_{M_2} = p_{M_3} = 0.30$, and $n_{ij} = 40$, $i, j = 1, 2, 3$, for a dominant trait $a = d = 1.0$.

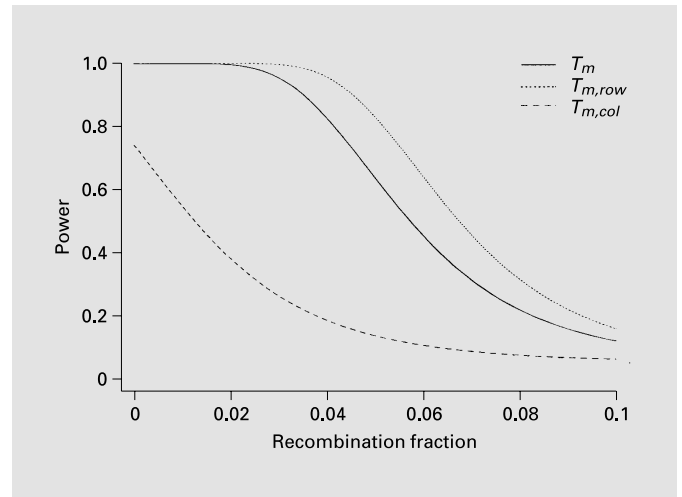


Fig. 2. Power curves of T_m , $T_{m,row}$ and $T_{m,col}$ when $p_Q = 0.15$, $h^2 = 0.80$, $A = 20$, $t = 4$, $p_{M_1} = \dots = p_{M_4} = 0.25$, and $n_{ij} = 25$, $i, j = 1, \dots, 4$, for a recessive trait $a = 1.0$, $d = -0.5$.

allele frequencies p_Q and p_q , haplotype frequencies h_{QM_i} and h_{qM_i} , recombination fraction θ , trait values μ_{QQ} , μ_{Qq} , μ_{qq} , and sampling error variance σ_e^2 . Due to the evolutionary process of the population, the haplotype frequencies h_{QM_i} , h_{qM_i} change from generation to generation. Under a Fisher-Wright model, the haplotype frequencies can be modeled by a diffusion process [Xiong and Guo, 1997]. The expected haplotype frequencies can be calculated by

$$E[h_{QM_i}] = h_{QM_i}(0)e^{-\theta A} + p_M p_Q (1 - e^{-\theta A}),$$

$$E[h_{qM_i}] = h_{qM_i}(0)e^{-\theta A} + p_M p_q (1 - e^{-\theta A}),$$

where A is the age of the most recent mutation at the trait locus, $h_{QM_i}(0)$ and $h_{qM_i}(0)$ are the initial haplotype frequencies of haplotypes QM_i and qM_i at the generation of occurrence of the mutation at the trait locus. If there is only a single mutation in the population, one may assume that $h_{QM_i}(0) = p_Q$, $h_{QM_i}(0) = 0$, $i = 2, \dots, t$, and $h_{qM_i}(0) = p_{M_i} - p_Q \geq 0$, $h_{qM_i}(0) = p_{M_i}$, $i = 2, \dots, t$. Replacing h_{QM_i} , h_{qM_i} in $P(QM_i, M_j)$, $P(qM_i, M_j)$ by $E h_{QM_i}$, $E h_{qM_i}$, we may calculate the approximations of the noncentrality parameters. If there are more than one mutations in a population, one needs more care about the fluctuation of haplotype frequencies. However, assuming a single copy mutation would be a sound assumption for a rare disease.

Power and Sample Size Comparison

To perform power and sample size calculation, we define the phenotypic values for genotypes QQ , Qq , qq by the traditional quantitative genetic analysis

$$\mu_{QQ} = a, \mu_{Qq} = \mu_{qQ} = d, \mu_{qq} = -a.$$

In addition, the additive variance

$$\sigma_a^2 = 2p_Q p_q (a + d(p_q - p_Q))^2$$

and the dominant variance $\sigma_d^2 = (2p_Q p_q d)^2$. Denote the heritability by h^2 , which is defined by $\sigma_a^2 / (\sigma_a^2 + \sigma_d^2 + \sigma_e^2)$. Using the non-centrality parameters given in 'Calculation of Noncentrality Parameters' above, we make power and sample size calculations by setting different values for the trait gene and the marker allele frequencies, heritability, recombination fraction, age of mutation for the trait allele, etc.

Figures 1 and 2 plot the power curves of T_m , $T_{m,row}$ and $T_{m,col}$ against the heritability for a 3-allele marker dominant trait, and the recombination fraction for a 4-allele marker recessive trait, respectively. From these 2 figures, we can see that T_m and $T_{m,row}$ have high power in detecting association when the recombination fraction is small and the heritability is high. The relatively low power of $T_{m,col}$ would be due to the choice of our parameters.

Figures 3 and 4 plot the power curves of T_{pars} , $T_{1,clps}$, $T_{m,het}$, T_{satur} , T_{gen} , $T_{2,clps}$ and T_{fit} against the recombination fraction for a 3-allele marker locus dominant and reces-

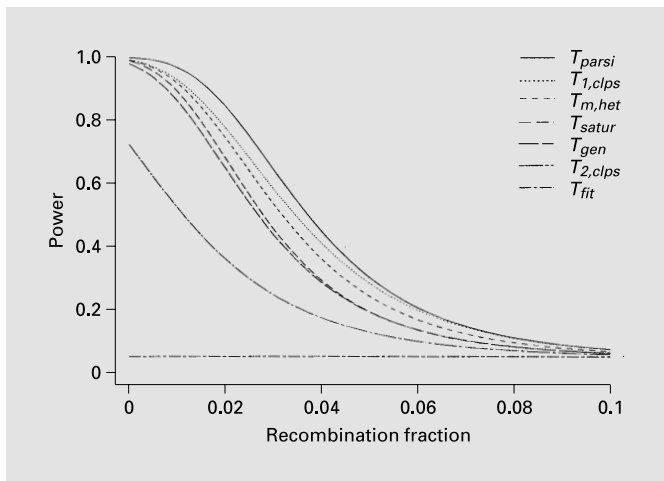


Fig. 3. Power curves of T_{parsl} , $T_{1,clps}$, $T_{m,het}$, T_{satur} , T_{gen} , $T_{2,clps}$ and T_{fit} when $p_Q = 0.15$, $h^2 = 0.80$, $A = 20$, $t = 3$, $p_{M_1} = 0.40$, $p_{M_2} = p_{M_3} = 0.30$, and $n_{12} = n_{13} = n_{21} = n_{23} = n_{31} = n_{32} = 40$ for a dominant trait $a = d = 1.0$.

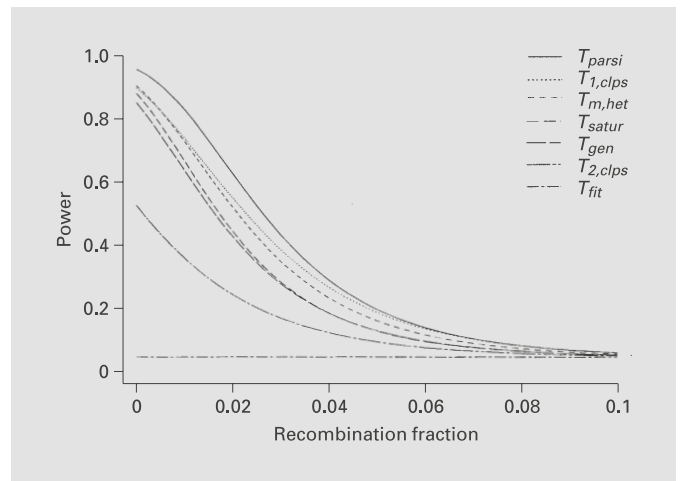


Fig. 4. Power curves of T_{parsl} , $T_{1,clps}$, $T_{m,het}$, T_{satur} , T_{gen} , $T_{2,clps}$ and T_{fit} when $p_Q = 0.15$, $h^2 = 0.50$, $A = 20$, $t = 3$, $p_{M_1} = 0.40$, $p_{M_2} = p_{M_3} = 0.30$, and $n_{12} = n_{13} = n_{21} = n_{23} = n_{31} = n_{32} = 40$ for a recessive trait $a = 1.0$, $d = -0.5$.

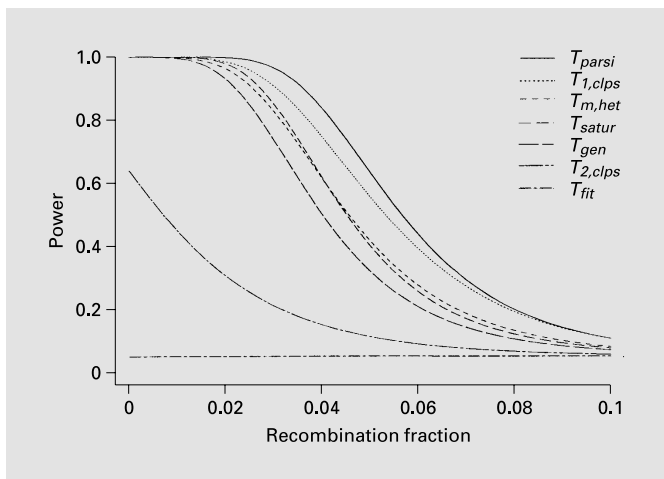


Fig. 5. Power curves of T_{parsl} , $T_{1,clps}$, $T_{m,het}$, T_{satur} , T_{gen} , $T_{2,clps}$ and T_{fit} when $p_Q = 0.15$, $h^2 = 0.80$, $A = 20$, $t = 4$, $p_{M_1} = \dots = p_{M_4} = 0.25$, and $n_{12} = n_{13} = n_{14} = n_{21} = n_{23} = n_{24} = n_{31} = n_{32} = n_{34} = n_{41} = n_{42} = n_{43} = 25$ for a dominant trait $a = d = 1.0$.

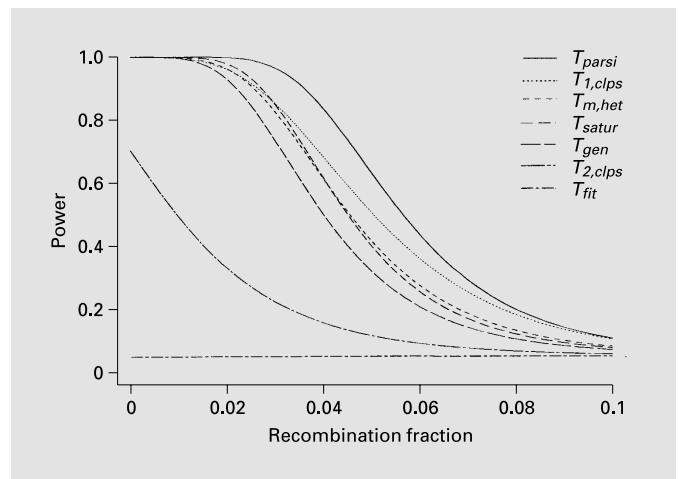


Fig. 6. Power curves of T_{parsl} , $T_{1,clps}$, $T_{m,het}$, T_{satur} , T_{gen} , $T_{2,clps}$ and T_{fit} when $p_Q = 0.15$, $h^2 = 0.80$, $A = 20$, $t = 4$, $p_{M_1} = \dots = p_{M_4} = 0.25$, and $n_{12} = n_{13} = n_{14} = n_{21} = n_{23} = n_{24} = n_{31} = n_{32} = n_{34} = n_{41} = n_{42} = n_{43} = 25$ for a recessive trait $a = 1.0$, $d = -0.5$.

sive models. Figures 5 and 6 plot the curves for a 4-allele marker locus dominant and recessive models. From these 4 figures, it is clear that T_{parsl} has higher power than those of the bi-allele tests $T_{1,clps}$ and $T_{2,clps}$, $T_{m,het}$, T_{satur} , and T_{gen} . Because of the low positions of the curve T_{fit} , using T_{satur} with higher degrees of freedom does not gain much power compared to T_{parsl} . The power curves are high if

there is tight linkage. Moreover, the power curves decrease quickly as the recombination increases. Therefore, the tests are powerful in detecting linkage between the marker and the trait loci in the presence of association if the quantitative trait locus is close to the marker locus. Figures 7 and 8 plot the power curves against the heritability h^2 . As the heritability increases, the power curves

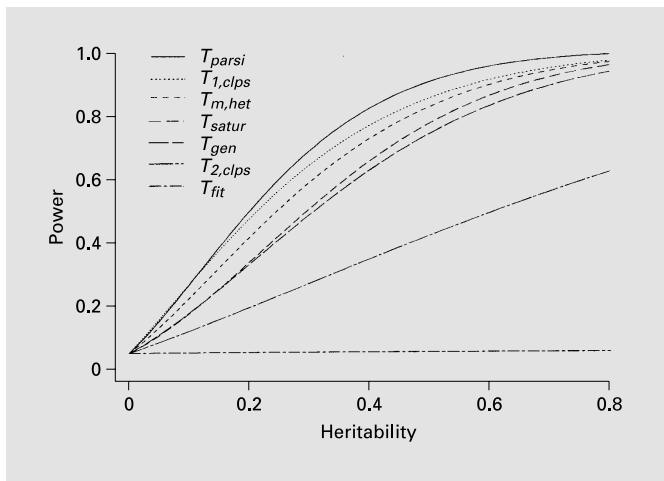


Fig. 7. Power curves of T_{parsis} , $T_{1,clps}$, $T_{m,het}$, T_{satur} , T_{gen} , $T_{2,clps}$ and T_{fit} when $p_Q = 0.15$, $\theta = 0.005$, $A = 20$, $t = 3$, $p_{M_1} = 0.40$, $p_{M_2} = p_{M_3} = 0.30$, and $n_{12} = n_{13} = n_{21} = n_{23} = n_{31} = n_{32} = 40$ for a dominant trait $a = d = 1.0$.

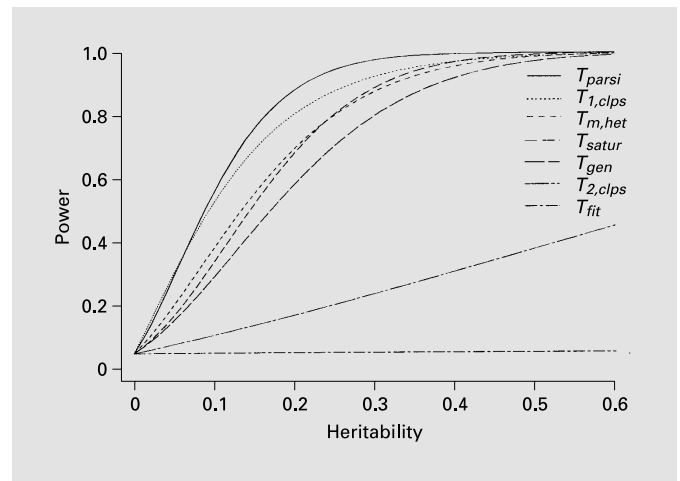


Fig. 8. Power curves of T_{parsis} , $T_{1,clps}$, $T_{m,het}$, T_{satur} , T_{gen} , $T_{2,clps}$ and T_{fit} when $p_Q = 0.15$, $\theta = 0.005$, $A = 20$, $t = 4$, $p_{M_1} = \dots = p_{M_4} = 0.25$, and $n_{12} = n_{13} = n_{14} = n_{21} = n_{23} = n_{24} = n_{31} = n_{32} = n_{34} = n_{41} = n_{42} = n_{43} = 25$ for a recessive trait $a = 1.0$, $d = -0.5$.

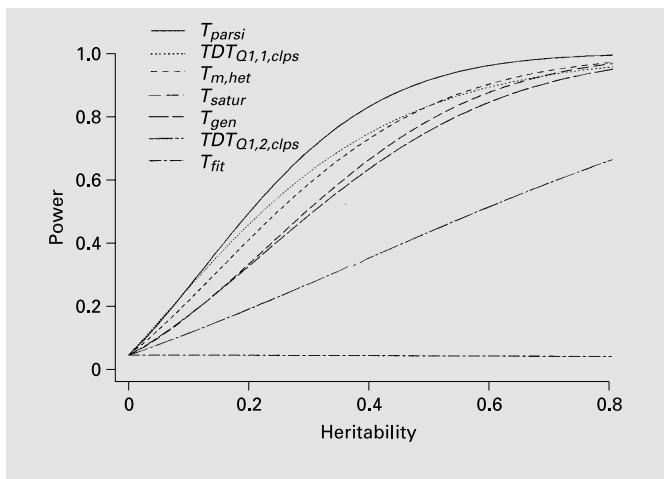


Fig. 9. Power curves of T_{parsis} , $TDT_{Q1,1,clps}$, $T_{m,het}$, T_{satur} , T_{gen} , $TDT_{Q1,2,clps}$ and T_{fit} when $p_Q = 0.15$, $\theta = 0.005$, $A = 20$, $t = 3$, $p_{M_1} = 0.40$, $p_{M_2} = p_{M_3} = 0.30$, and $n_{12} = n_{13} = n_{21} = n_{23} = n_{31} = n_{32} = 40$ for a recessive trait $a = 1.0$, $d = -0.5$.

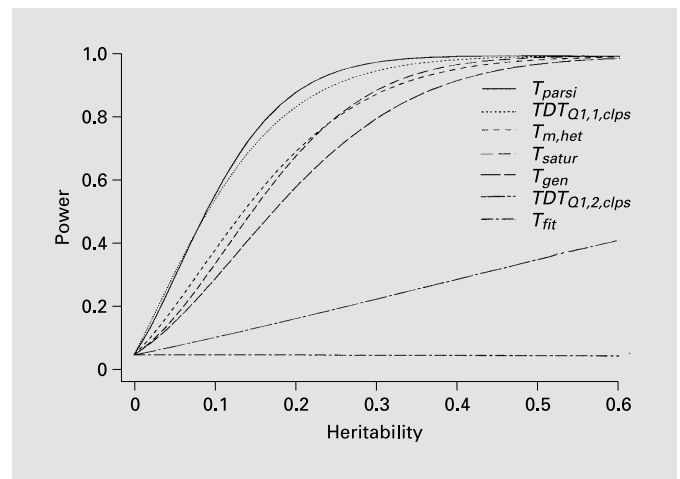


Fig. 10. Power curves of T_{parsis} , $TDT_{Q1,1,clps}$, $T_{m,het}$, T_{satur} , T_{gen} , $TDT_{Q1,2,clps}$ and T_{fit} when $p_Q = 0.15$, $\theta = 0.005$, $A = 20$, $t = 4$, $p_{M_1} = \dots = p_{M_4} = 0.25$, and $n_{12} = n_{13} = n_{14} = n_{21} = n_{23} = n_{24} = n_{31} = n_{32} = n_{34} = n_{41} = n_{42} = n_{43} = 25$ for a dominant trait $a = d = 1.0$.

increase as one expects. From the two figures, we can observe that T_{parsis} has higher power than those of bi-allele tests $T_{1,clps}$ and $T_{2,clps}$, $T_{m,het}$, T_{satur} , and T_{gen} . T_{satur} and T_{gen} do not gain much power although they have larger number of degrees of freedom. Figures 9 and 10 compare the powers of the composite tests introduced in this paper with $TDT_{Q1,1,clps}$ and $TDT_{Q1,2,clps}$ of Allison [1997]. The

parsimonious test statistic T_{parsis} has higher power than those of $TDT_{Q1,1,clps}$ and $TDT_{Q1,2,clps}$. Hence, it is better to perform overall test T_{parsis} instead of performing individual collapsed bi-allele tests.

Table 2 presents the cell sample size for dominant and recessive disease models at significance level 0.01 or 0.05 and 80% power with trait frequency $p_Q = 0.15$ and recom-

Table 2. Sample size for dominant and recessive disease models at significance level 0.01 or 0.05 and 80% power with trait frequency $p_Q = 0.15$ and recombination fractions $\theta = 0.005$, assuming $a = 1.0, d = 1.0$ for dominant disease and $a = 1.0, d = -0.5$ for recessive disease. When the number $t = 3$ of alleles at marker locus, marker allele frequencies $p_{M_1} = 0.4, p_{M_2} = p_{M_3} = 0.3$; when $t = 4, p_{M_1} = \dots = p_{M_4} = 0.25$; when $t = 5, p_{M_1} = \dots = p_{M_5} = 0.20$

Model	Para	AM	SL	T_{parsi}	$T_{m,het}$	T_{gen}	SL	T_{parsi}	$T_{m,het}$	T_{gen}
Dom	$t = 3$	10	0.05	$n_{ij} = 27$	34	42	0.01	$n_{ij} = 39$	49	59
	$h^2 = 0.50$	20		31	38	46		44	55	65
		30		34	42	51		49	61	72
Rec	$t = 3$	10	0.05	34	42	51	0.01	48	61	72
	$h^2 = 0.40$	20		37	47	57		54	68	80
		30		42	52	63		60	75	89
Dom	$t = 4$	10	0.05	$n_{ij} = 7$	11	13	0.01	$n_{ij} = 10$	16	18
	$h^2 = 0.50$	20		8	13	15		11	18	20
		30		9	14	17		13	20	23
Rec	$t = 4$	10	0.05	9	14	17	0.01	13	20	23
	$h^2 = 0.40$	20		10	16	19		14	22	26
		30		11	17	21		16	25	29
Dom	$t = 5$	10	0.05	$n_{ij} = 4$	6	7	0.01	$n_{ij} = 5$	8	10
	$h^2 = 0.50$	20		4	7	8		6	9	11
		30		5	7	9		6	10	13
Rec	$t = 5$	10	0.05	5	8	10	0.01	6	10	13
	$h^2 = 0.40$	20		5	8	11		7	12	14
		30		6	9	12		8	13	16

Para = Parameter; AM = age of mutation; SL = significance level.

bination fractions $\theta = 0.005$, assuming $a = 1.0, d = 1.0$ for dominant disease and $a = 1.0, d = -0.5$ for recessive disease. The sample sizes of T_{parsi} are smaller than those of $T_{m,het}$, and the sample sizes of $T_{m,het}$ are smaller than those of T_{gen} . Hence, T_{parsi} is more powerful than other tests. Moreover, the sample sizes decrease very rapidly as the number of alleles at the marker locus increases. For a marker with $t = 5$ alleles, a sample size of 5 for each cell in table 1 would be enough to achieve 80% power at a significant level 0.01 if one uses test statistic T_{parsi} . Hence, the composite tests are very conservative and these are similar to the likelihood method proposed by Terwilliger [1995] for qualitative traits.

An Example

To illustrate the usefulness and the power of the methods proposed in this paper, we analyze the Oxford asthma data [Cookson and Abecasis, 2001] from the Genetic Analysis Workshop 12. The data consist of 80 nuclear families with a total of 203 offspring. In these 80 families,

43 have 2 offspring, 31 have 3 offspring, and 6 have 4 offspring. On chromosome 6, 16 markers are typed, and each marker has 4 alleles. By using the tests of bi-allele markers available in the literature, one has to collapse the markers to be bi-allelic. Using the methods proposed in this paper, we are able to perform composite tests ($F_{m,row}, F_{m,col}, F_{parsi}, F_{gen}, F_{satur}$).

In Daniel et al. [1996], linkage to log blood eosinophil count ($\log_e eos$) and other quantitative traits was tested by the Haseman-Elston sib-pair technique [Haseman and Elston, 1972]. Three markers on chromosome 6 of potential linkage were detected with $\log_e eos$ [D6S260, $p < 0.05$; D6S276, $p < 0.0001$; D6S273, $p < 0.05$; Daniel et al., 1996]. We analyze the data using the methods described in the present paper and SAS procedures, and obtain the results given in table 3.

Four markers show association with $\log_e eos$ at significance level 0.05 (D6S276, D6S291, D6S262, and D6S264). Four markers show significance for TDT statistics F_{parsi} or F_{satur} at significance level 0.05 (D6S276, D6S273, TNFA, and D6S264). Because F_{parsi} is a valid test statistic of linkage in the presence of association, we

Table 3. Test results of asthma data

Marker locus	p values of association tests		p values of TDT statistics		
	$F_{m,row}$	$F_{m,col}$	F_{pars}	F_{satur}	F_{fit}
D6S276	0.004*		0.004*	0.22	1.0
D6S273			0.71	0.023	0.013
TNFA			0.49	0.0002*	0.0001*
D6S291		0.05*			
D6S262	0.04*				
D6S264	0.03*		0.05*	0.25	0.65

* p < 0.05.

are able to confirm that marker D6S276 is potentially linked to asthma phenotype $\log_e eos$ (p value 0.004), and show that marker D6S264 is also potentially linked to $\log_e eos$ (p value 0.05).

Discussion

In this paper, we explore models and tests to detect association and linkage between a multi-allele marker locus and a quantitative trait locus. The goal is to search for overall test statistics which are more powerful than those of bi-allele TDT statistics available in the literature by collapsing the multi-allele marker to a bi-allele marker. Using the homozygous and heterozygous data, we propose three composite tests T_m , $T_{m,row}$ and $T_{m,col}$ to carry out an association study. For linkage analysis in the presence of association, one may use the heterozygous data to perform composite tests based on statistics T_{pars} , T_{satur} , T_{gen} and $T_{m,het}$.

Based on a theoretical calculation of an offspring's conditional expectation of a trait, given the allele transmission status of a heterozygous parent at the marker locus, we propose a direct generalization T_{gen} of TDT from a bi-allele marker locus to a multi-allele marker locus. Although the generalized T_{gen} is valid in testing linkage in the presence of association between the quantitative trait locus and the multi-allele locus, it suffers from the same problems as the direct extension of TDT for a qualitative trait [Sham and Curtis, 1995]. One problem is the large number of degrees of freedom, and the other is lack of consideration of the data pattern consistent with linkage and linkage disequilibrium. As Sham and Curtis [1995], we observe that the conditional trait expectation only

depends on the transmitted marker allele when the marker locus is closely linked to the trait locus (Appendix C). Based on this observation, we propose a parsimonious or allele-wise linear regression model to analyze the data. The proposed model generalizes the case of a bi-allele marker locus. This parsimonious model leads to a test statistic T_{pars} , which is distributed approximately as a central χ^2 with $t - 1$ degrees of freedom for a marker locus with t alleles under the null hypothesis of no linkage between the marker and the trait locus. Moreover, this model takes into account the pattern of the data consistent with linkage and linkage disequilibrium. Based on the linear regression models, we propose test statistics T_{satur} and T_{fit} to investigate the goodness-of-fit of T_{pars} . T_{fit} tests how much power T_{satur} may gain by increasing the number of parameters from t of the parsimonious model to $t(t - 1)$ of the saturated model. To use the marginal heterozygous data for linkage study, we calculate an offspring's conditional trait means given an allele is transmitted or not transmitted from a heterozygous parent at the marker locus. Based on the difference between the conditional means of a transmitted and a nontransmitted allele from a heterozygous parent, we propose a statistic $T_{m,het}$ to perform composite tests of linkage between the marker locus and the quantitative trait locus in the presence of association. $T_{m,het}$ compares the row and column sample means for offspring data of heterozygous parents.

To compare the power and sample size of the tests, and investigate the goodness-of-fit of T_{pars} , we first calculate the noncentrality parameters for the tests under the alternative hypothesis, i.e., there is linkage between the marker and the trait locus in the presence of association. After comparing the power and sample size, we conclude that T_{pars} has higher power than those of the bi-allele tests from Allison [1997] and Xiong et al. [1998], $T_{m,het}$, T_{satur} , and T_{gen} . If there is tight linkage between the marker and the trait locus, T_{pars} is powerful in detecting linkage between the marker and the trait locus in the presence of association. By investigating the goodness-of-fit of T_{pars} , we find that T_{satur} does not gain much power compared to T_{pars} . Moreover, T_{pars} takes into account the data pattern consistent with linkage and linkage disequilibrium. Besides, the composite test statistics are very conservative when the number of alleles at the marker locus increases, a merit observed in the likelihood method proposed by Terwilliger [1995].

If there are no covariates, the tests except T_m and $T_{m,het}$ are equivalent to the tests in ANOVA models. If there are covariates, one may use linear regression to perform an analysis to adjust for the covariates. The tests except T_{satur}

and T_{gen} have few degrees of freedom. The tests T_{pars} , $T_{m,het}$, T_{satur} and T_{gen} proposed in this paper are the same as TDT_1 in Xiong et al. [1998] for a bi-allele marker. For a bi-allele marker, Xiong et al. [1998] showed that TDT_1 is more powerful than the test statistics TDT_{Q1} proposed in Allison [1997], Haseman and Elston [1972], and T1B1 proposed in Risch and Zhang [1996].

Although Allison et al. [1999] introduced a mixed effect model to perform a sibling-based test of QTL data, these investigators did not consider ways to reduce the parameters and as a result the number of parameters can be too large. In this paper, we studied both the parsimonious and saturated models based on our theoretical analysis, and conclude that the parsimonious model can be very useful in testing linkage between a multi-allele marker and a trait locus in the presence of association if there is tight linkage. In 'Bi-Allele Marker Locus' we used an indicator function $1_{(TM = M, NM = m)}$ for each heterozygous parent corresponding to each of his/her children. George et al. [1999] defined a similar variable for each child based on the transmission of allele M from a heterozygous parent. However, a heterozygous child of a double heterozygous mating was considered noninformative and hence discarded in George et al. [1999]. This may not be a valid treatment. In the reference of TDT by Spielman et al. [1993], the heterozygous child data of a double heterozygous mating were used in the analysis. This, although appropriate for TDT analysis, would be a problem in haplotype sharing analysis. In our analysis we use the same aspect of TDT that makes our analysis different from that of George et al. [1999]. In this paper, we simply assign that one parent transmits allele M and the other transmits allele m to the heterozygous child of a double heterozygous mating. Actually, the coefficient α in the equation (5) is different from that in George et al. [1999] due to the different treatments of data in this paper and in George et al. [1999].

From the power and sample size study, we observed that the test T_m (or $T_{m,het}$) has good power in detecting association (or linkage in the presence of association). However, one should be cautious in using T_m (or $T_{m,het}$). The reason is that the near χ^2 distribution assumption of T_m (or $T_{m,het}$) can be problematic due to its redundant utilization of data. For qualitative traits, several authors have noticed similar problems [Schaid, 1996; Sham, 1997; Lazzeroni and Lange, 1998]. A wise practice would be to carry out analysis by other test statistics, such as $T_{m,row}$ (or T_{pars}). Actually, $T_{m,row}$ (or T_{pars}) is more powerful than T_m (or $T_{m,het}$), and is a valid statistical test. We have included figures 1 and 2 only to show the effective-

ness of T_m , $T_{m,row}$, and $T_{m,col}$. One reason for the difference between T_m and $T_{m,het}$ ($T_{m,row}$ and T_{pars}) is: T_m and $T_{m,row}$ both use homozygous and heterozygous parent data for association studies, but $T_{m,het}$ and T_{pars} can only use the heterozygous data for linkage studies in the presence of association. In terms of sample size and power calculations, T_m is similar to $T_{m,het}$ and $T_{m,row}$ is similar to T_{pars} .

In this paper, we made the assumption that the parental genotypes are available. It would be interesting to extend this work to the situation when the parental genotypes are unavailable. One way to attack this problem is to do permutation tests [Spielman and Ewens, 1998], but this could be the focus of a further investigation. In the present study, we work on data consisting of trio families, each of the families has both parents and one child. If there are more than one child in a family, one may treat each child independently, and each child and his/her parent make up a trio. Hence, a family of k kids can be decomposed to be k independent trio families. Obviously, this treatment may have some limitations because the traits of children in each family are related to each other [Fan and Xiong, 2002]. However, our focus here is to obtain models and tests which lead to simple statistics with $t - 1$ degrees of freedom for a t allele marker locus. To accommodate the data with more complex family structures, one may need to use variance component models [Fan and Xiong, 2002] that use a bi-allele marker with better power than that in George et al. [1999], and have powers similar to those in Zhu and Elston [2000, 2001].

Acknowledgment

We are very grateful for the very helpful comments of two reviewers, which improve the paper a lot. We thank Miss Jeesun Jung for her generous assistance in the asthma data analysis by SAS procedures. Ruzong Fan was supported partially by a pilot project at Texas A&M University. Momiao Xiong was supported by NIH grant R01-GM56515, and MH59518.

Appendix A

In the following, we show equations (1), (2) and (3). Let us utilize the notations introduced in 'Models and Tests for Association Studies' above, such as TQ , TM and NM . Let TH denote abbreviation of 'transmitted haplotype'. Then $P(TH = QM_i) = (1 - \theta)h_{QM_i} + \theta p_{QP_{M_i}}$ and $P(TH = qM_i) = (1 - \theta)h_{qM_i} + \theta p_{qP_{M_i}}$. Given that a marker allele M_i is transmitted to an offspring, the conditional mean

$$\begin{aligned}\mu_i &= E[Y | TM = M_i] \\ &= [E[Y | TH = QM_i]P(TH = QM_i) \\ &\quad + E[Y | TH = qM_i]P(TH = qM_i)]/p_{M_i} \\ &= (1 - \theta)[\mu_Q h_{QM_i} + \mu_q h_{qM_i}]/p_{M_i} + \theta\mu.\end{aligned}$$

On the other hand, provided that allele M_i is not transmitted, the conditional mean

$$\begin{aligned}v_i &= E[Y | NM = M_i] = \sum_{j \neq i} E[Y | TM = M_j]P(TM = M_j)/P(NM = M_i) \\ &= \sum_{j \neq i} \mu_j p_{M_j} / (1 - p_{M_i}) \\ &= \sum_{j \neq i} [\mu_Q P(TH = QM_j) + \mu_q P(TH = qM_j)] / (1 - p_{M_i}) \\ &= (1 - \theta)[\mu_Q (p_Q - h_{QM_i}) + \mu_q (p_q - h_{qM_i})] / (1 - p_{M_i}) + \theta\mu \\ &= (1 - \theta)[\mu - \mu_i p_{M_i}] / (1 - p_{M_i}) + \theta\mu.\end{aligned}$$

Notice that $h_{qM_i} - p_{qP_{M_i}} = -h_{QM_i} = \delta_i$. Therefore, we can calculate the difference

$$\begin{aligned}\frac{\mu_i - v_i}{1 - \theta} &= \frac{\mu_Q [h_{QM_i}(1 - p_{M_i}) - (p_Q - h_{QM_i})p_{M_i}] + \mu_q [h_{qM_i}(1 - p_{M_i}) - (p_q - h_{qM_i})p_{M_i}]}{p_{M_i}(1 - p_{M_i})} \\ &= \frac{\mu_Q [h_{QM_i} - p_{QP_{M_i}}] + \mu_q [h_{qM_i} - p_{qP_{M_i}}]}{p_{M_i}(1 - p_{M_i})} \\ &= \frac{\mu_Q [h_{QM_i} - p_{QP_{M_i}}] + \mu_q [-h_{QM_i} - p_{QP_{M_i}}]}{p_{M_i}(1 - p_{M_i})} \\ &= \frac{(\mu_Q - \mu_q)(h_{QM_i} - p_{QP_{M_i}})}{p_{M_i}(1 - p_{M_i})} = (\mu_Q - \mu_q)\delta_i / [p_{M_i}(1 - p_{M_i})]\end{aligned}$$

$$\frac{\mu_i - \mu}{1 - \theta} = [\mu_Q h_{QM_i} + \mu_q h_{qM_i}] / p_{M_i} - (\mu_Q p_Q + \mu_q p_q) = (\mu_Q - \mu_q)\delta_i / p_{M_i}$$

$$\frac{v_i - \mu}{1 - \theta} = [\mu - \mu_i p_{M_i}] / (1 - p_{M_i}) - \mu = -(\mu_Q - \mu_q)\delta_i / (1 - p_{M_i}).$$

Appendix B

In the following, we show equations (5) and (9). Consider a heterozygous parent with genotype $M_i M_j$, $i < j$, at marker locus M . First of all, note that $P(TM = M_i, NM = M_j) = [1/2][2p_{M_i} p_{M_j}] = p_{M_i} p_{M_j}$ since the probability of a heterozygous parent possessing allele M_i at one copy of his/her chromosome and allele M_j at the other copy of chromosome is $2p_{M_i} p_{M_j}$, and the probability of giving one of his/her two

alleles to an offspring is 1/2. Similarly, we may show $P(TM = M_j, NM = M_i) = p_{M_i} p_{M_j}$. Let $P(QQ | TM = M_i, NM = M_j)$ be the conditional probability of a child who receives genotype QQ at the trait locus, given that he/she receives allele M_i from a heterozygous parent $M_i M_j$, and define other notations similarly such as $P(Qq | TM = M_i, NM = M_j)$. Let $\mu_{i,j} = E[Y | TM = M_i, NM = M_j] = \mu_{QQ} P(QQ | TM = M_i, NM = M_j) + \mu_{Qq} P(Qq | TM = M_i, NM = M_j) + \mu_{qq} P(qq | TM = M_i, NM = M_j)$. Then we have from Seber [1977]

$$\alpha_{ij} = \mu_{i,j} - \mu_{j,i} \quad (13)$$

To calculate the conditional probabilities in equation (13), we need to introduce some notations first. Let $P(QM_i, M_j)$ be the probability of a child who receives haplotype QM_i from his/her heterozygous parent but not alleles M_j . Similarly, we may define notations $P(qM_i, M_j)$. Then

$$P(QM_i, M_j) = (1 - \theta)h_{QM_i} p_{M_j} + \theta h_{QM_i} p_{M_i} \quad (14)$$

$$P(qM_i, M_j) = (1 - \theta)h_{qM_i} p_{M_j} + \theta h_{qM_i} p_{M_i}$$

Using the notations above, we can calculate the conditional probabilities

$$\begin{aligned}P(QQ | TM = M_i, NM = M_j) &= \frac{p_Q P(QM_i, M_j)}{p_{M_i} p_{M_j}} \\ P(Qq | TM = M_i, NM = M_j) &= \frac{p_Q P(qM_i, M_j) + p_q P(QM_i, M_j)}{p_{M_i} p_{M_j}} \\ P(qq | TM = M_i, NM = M_j) &= \frac{p_q P(qM_i, M_j)}{p_{M_i} p_{M_j}}\end{aligned} \quad (15)$$

Substituting the above conditional probabilities into equations (13) and using equation (14) and $\delta_{ij} = h_{QM_i} p_{M_j} - h_{qM_i} p_{M_j}$, we may get coefficients (9). For bi-allele case, $\delta_{12} = \delta$ if we use notations $M = M_1$ and $m = M_2$. Hence, equation (5) holds.

Appendix C

Assume that there is tight linkage between the trait locus L and the marker locus M . Then recombination fraction $\theta \approx 0$, and hence equations (14) imply $P(QM_i, M_j) \approx h_{QM_i} p_{M_j}$ and $P(qM_i, M_j) \approx h_{qM_i} p_{M_j}$. Provided that $\theta \approx 0$, the conditional probabilities given by equations (15) lead to the following conditional expectation

$$\begin{aligned}\mu_{i,j} &= [\mu_Q P(QM_i, M_j) + \mu_q P(qM_i, M_j)] / [p_{M_i} p_{M_j}] \\ &\approx [\mu_Q h_{QM_i} + \mu_q h_{qM_i}] / p_{M_i}\end{aligned} \quad (16)$$

which does not depend on the allele/haplotype frequencies related to allele M_j .

Appendix D

Let us denote $\cup_{j \neq i} M_j$ an allele that collapses all but allele M_i . To calculate the equation (12), we need the conditional means $\mu_{i,het} = E[Y | TM = M_i, NM = \cup_{j \neq i} M_j] = [\mu_Q P(QM_i, \cup_{j \neq i} M_j) + \mu_q P(qM_i, \cup_{j \neq i} M_j)] / [p_{M_i}(1 - p_{M_i})]$, $v_{i,het} = E[Y | NM = M_i, TM = \cup_{j \neq i} M_j] = [\mu_Q P(Q\cup_{j \neq i} M_j, M_i) + \mu_q P(q\cup_{j \neq i} M_j, M_i)] / [p_{M_i}(1 - p_{M_i})]$. We may calculate $P(QM_i, \cup_{j \neq i} M_j)$, $P(qM_i, \cup_{j \neq i} M_j)$ in a similar way as we have done for equation (14) in Appendix B, and then calculate the conditional mean difference (12).

Appendix E

First of all, we may calculate the conditional variances

$$\sigma_Q^2 = \text{Var}(Y|TQ = Q) = \sigma_e^2 + (\mu_{QQ} - \mu_Q)^2 p_Q + (\mu_{Qq} - \mu_Q)^2 p_q$$

$$\sigma_q^2 = \text{Var}(Y|TQ = q) = \sigma_e^2 + (\mu_{qQ} - \mu_q)^2 p_Q + (\mu_{qq} - \mu_q)^2 p_q.$$

$$\sigma_{i,j}^2 = \text{Var}[Y|TM = M_i, NM = M_j]$$

$$= E[(Y - \mu_{i,j})^2 | TM = M_i, NM = M_j]$$

$$= [\sigma_Q^2 + (\mu_Q - \mu_{i,j})^2] \frac{P(QM_i, M_j)}{P(M_i, M_j)} + [\sigma_q^2 + (\mu_q - \mu_{i,j})^2] \frac{P(qM_i, M_j)}{P(M_i, M_j)}.$$

Now the noncentrality parameters of T_{gen} and T_{fit} can be calculated by

$$\lambda_{gen} = \frac{\sum_{i=1}^{t-1} \sum_{j=i+1}^t (\mu_{i,j} - \mu_{j,i})^2 / (1/n_{ij} + 1/n_{ji})}{\sum_{i=1}^{t-1} \sum_{j=i+1}^t [(n_{ij} - 1) \sigma_{i,j}^2 + (n_{ji} - 1) \sigma_{j,i}^2] / (n - \sum_{i=1}^t n_{ii} - t(t-1))}$$

$$\lambda_{fit} = \frac{\sum_{i=1}^{t-1} \sum_{j=i+1}^t [n_{ij} (\mu_{i,j} - \mu_{i,het})^2 + n_{ji} (\mu_{j,i} - \mu_{j,het})^2]}{\sum_{i=1}^{t-1} \sum_{j=i+1}^t [(n_{ij} - 1) \sigma_{i,j}^2 + (n_{ji} - 1) \sigma_{j,i}^2] / (n - \sum_{i=1}^t n_{ii} - t(t-1))}$$

The conditional variances

$$\sigma_{ir,het}^2 = \text{Var}(Y|TM = M_i, NM = \cup_{j \neq i} M_j)$$

$$= [\sigma_Q^2 + (\mu_Q - \mu_{i,het})^2] P(QM_i, \cup_{j \neq i} M_j) / [p_{M_i}(1 - p_{M_i})] + [\sigma_q^2 + (\mu_q - \mu_{i,het})^2] P(qM_i, \cup_{j \neq i} M_j) / [p_{M_i}(1 - p_{M_i})].$$

$$\sigma_{ic,het}^2 = \text{Var}(Y|NM = M_i, TM = \cup_{j \neq i} M_j)$$

$$= [\sigma_Q^2 + (\mu_Q - v_{i,het})^2] P(Q\cup_{j \neq i} M_j, M_i) / [p_{M_i}(1 - p_{M_i})] + [\sigma_q^2 + (\mu_q - v_{i,het})^2] P(q\cup_{j \neq i} M_j, M_i) / [p_{M_i}(1 - p_{M_i})].$$

The noncentrality parameter of T_{paris} is

$$\lambda_{paris} = \frac{\sum_{i=1}^t n_{i \cdot, het} (\mu_{i,het} - \mu_{het})^2}{\sum_{i=1}^t (n_{i \cdot, het} - 1) \sigma_{ir,het}^2 / (n - \sum_{i=1}^t n_{ii} - t)}$$

where

$$\mu_{het} = E[\bar{Y}(\cdot)_{\cdot, het}] = \frac{\sum_{i=1}^t [\mu_Q P(QM_i, \cup_{j \neq i} M_j) + \mu_q P(qM_i, \cup_{j \neq i} M_j)]}{1 - \sum_i p_{M_i}^2}$$

is the conditional mean given that a marker allele is transmitted from a heterozygous parent. The non-centrality parameter of statistic T_{satur} is

$$\lambda_{satur} = \frac{\sum_{i=1}^{t-1} \sum_{j=i+1}^t [n_{ij} (\mu_{i,j} - \mu_{het})^2 n_{ij} (\mu_{j,i} - \mu_{het})^2]}{\sum_{i=1}^{t-1} \sum_{j=i+1}^t [(n_{ij} - 1) \sigma_{i,j}^2 + (n_{ji} - 1) \sigma_{j,i}^2] / (n - \sum_{i=1}^t n_{ii} - t(t-1))}$$

$T_{m,het}$ has nearly a χ^2 distribution with $t - 1$ degrees of freedom under the null hypothesis H_0 . Under the alternative hypothesis, it is distributed approximately as a noncentral χ^2 with noncentrality parameter

$$\lambda_{m,het} =$$

$$\frac{t-1}{t} \frac{\sum_{i=1}^t (\mu_{i,het} - v_{i,het})^2 / (1/n_{i \cdot, het} + 1/n_{\cdot, i, het})}{\sum_{i=1}^t [(n_{i \cdot, het} - 1) \sigma_{ir,het}^2 + (n_{\cdot, i, het} - 1) \sigma_{ic,het}^2] / (n - 2\sum_{i=1}^t n_{ii} - 2t)}$$

Similarly, the non-centrality parameters of $T_{i,clps}$, T_m , $T_{m,row}$, and $T_{m,col}$ are given by

$$\lambda_{i,clps} = \frac{(\mu_{i,het} - v_{i,het})^2 / (1/n_{i \cdot, het} + 1/n_{\cdot, i, het})}{[(n_{i \cdot, het} - 1) \sigma_{ir,het}^2 + (n_{\cdot, i, het} - 1) \sigma_{ic,het}^2] / (n_{i \cdot, het} + n_{\cdot, i, het} - 2)}$$

$$\lambda_m = \frac{t-1}{t} \frac{\sum_{i=1}^t (\mu_i - v_i)^2 / (1/n_{i \cdot} + 1/n_{\cdot, i})}{\sum_{i=1}^t [(n_{i \cdot} - 1) \sigma_{ir}^2 + (n_{\cdot, i} - 1) \sigma_{ic}^2] / (2n - 2t)}$$

$$\lambda_{m,row} = \frac{\sum_{i=1}^t n_{i \cdot} (\mu_i - \mu)^2}{\sum_{i=1}^t (n_{i \cdot} - 1) \sigma_{ir}^2 / (n - t)}$$

$$\lambda_{m,col} = \frac{\sum_{i=1}^t n_{\cdot, i} (v_i - \mu)^2}{\sum_{i=1}^t (n_{\cdot, i} - 1) \sigma_{ic}^2 / (n - t)}$$

where the conditional variances

$$\sigma_{ir}^2 = \text{Var}(Y|TM = M_i)$$

$$= [[\sigma_Q^2 + (\mu_Q - \mu_i)^2] P(TH = QM_i) + [\sigma_q^2 + (\mu_q - \mu_i)^2] P(TH = qM_i)] / p_{M_i}$$

$$\sigma_{ic}^2 = \text{Var}(Y|NM = M_i)$$

$$= (1 - \theta)[[\sigma_Q^2 + (\mu_Q - v_i)^2] (p_Q - h_{QM_i}) + [\sigma_q^2 + (\mu_q - v_i)^2] (p_q - h_{qM_i})] / (1 - p_{M_i}) + \theta[[\sigma_Q^2 + (\mu_Q - v_i)^2] p_Q + [\sigma_q^2 + (\mu_q - v_i)^2] p_q].$$

The non-centrality parameters of the statistic $TDT_{Q1,i,clps}$ from Allison [1997] can be calculated as

$$\lambda_{Q1,i,clps} = (\mu_{i,het} - v_{i,het})^2 / [2(\sigma_{ir,het}^2 + \sigma_{ic,het}^2) / (n_{i \cdot, het} + n_{\cdot, i, het})].$$

References

- Abecasis GR, Cardon LR, Cookson WOC: A general test of association for quantitative traits in nuclear families. *Am J Hum Genet* 2000;66: 279-292.
- Allison DB: Transmission-disequilibrium tests for quantitative traits. *Am J Hum Genet* 1997;60: 676-690.
- Allison DB, Heo M, Kaplan N, Martin ER: Sibling-based tests of linkage and association for quantitative traits. *Am J Hum Genet* 1999;64: 1754-1764.
- Bickeböllér H, Clerget-Darpoux F: Statistical properties of the allelic and genotypic transmission/disequilibrium test for multi-allelic markers. *Genet Epidemiol* 1995;12:865-870.
- Boehnke M, Langefeld CD: Genetic association mapping based on discordant sib pairs: The discordant-alleles test. *Am J Hum Genet* 1998; 62:950-961.
- Cookson W, Abecasis G: Oxford genome screen for asthma-associated traits. *Genet Epidemiol* 2001;21(suppl 1):S1-S3.
- Daniel SE, Bhattacharrya S, James A, et al: A genome-wide search for quantitative trait loci underlying asthma. *Nature* 1996;383:247-250.
- Falconer DS, Mackay TFC: *Introduction to Quantitative Genetics*, ed 4. London, Longman, pp 15-19.
- Fan RZ, Xiong MM: Linkage and association studies of QTL for nuclear families by mixed models. *Biostatistics*, 2002, in press.
- George V, Elston RC: Testing the association between polymorphic markers and quantitative traits in pedigrees. *Genet Epidemiol* 1987;4: 193-201.
- George V, Tiwari HK, Zhu XF, Elston RC: A test of transmission/disequilibrium for quantitative traits in pedigree data, by multiple regression. *Am J Hum Genet* 1999;65:236-245.
- Gordon D, Heath SC, Liu X, Ott J: A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. *Am J Hum Genet* 2001;69: 371-380.
- Haseman JK, Elston RC: The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* 1972;2:3-19.

- Jin K, Speed TP, Klitz W, Thomson G: Testing for segregation distortion in the HLA complex. *Biometrics* 1994;50:1189–1198.
- Kaplan NL, Martin ER, Weir BS: Power studies for the transmission/disequilibrium tests with multiple alleles. *Am J Hum Genet* 1997;60:691–702.
- Lazzeroni LC, Lange K: A conditional inference framework for extending the transmission/disequilibrium test. *Hum Hered* 1998;48:67–81.
- Martin ER, Kaplan NL, Weir BS: Tests for linkage and association in nuclear families. *Am J Hum Genet* 1997;61:439–448.
- Rabinowitz D: A transmission disequilibrium test for quantitative trait loci. *Hum Hered* 1997;47:342–350.
- Risch N, Zhang N: Mapping quantitative trait loci with extreme discordant sib pairs: Sample size considerations. *Am J Hum Genet* 1996;58:836–843.
- Seber GAF: *Linear Regression Analysis*. New York, Wiley, 1977, pp 1–8.
- Schaid DJ: General score tests for associations for genetic markers with disease using cases and their parents. *Genet Epidemiol* 1996;13:423–449.
- Sham P: The transmission/disequilibrium tests for multi-allele markers. *Am J Hum Genet* 1997;61:774–778.
- Sham PC, Curtis D: An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. *Ann Hum Genet* 1995;59:323–336.
- Spielman RS, Ewens WJ: The TDT and other family-based tests for linkage disequilibrium and association. *Am J Hum Genet* 1996;59:983–989.
- Spielman RS, Ewens WJ: A sibship test for linkage in the presence of association: The sib transmission/disequilibrium test. *Am J Hum Genet* 1998;62:450–458.
- Spielman RS, McGinnis RE, Ewens WJ: Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 1993;52:506–516.
- Terwilliger JD: A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *Am J Hum Genet* 1995;56:777–787.
- Xiong MM, Guo SW: Fine-Scale genetic mapping based on linkage disequilibrium: Theory and application. *Am J Hum Genet* 1997;60:1513–1531.
- Xiong MM, Jin L, Boerwinkle E: Linkage disequilibrium based regression: A method for mapping quantitative trait loci in humans, submitted.
- Xiong MM, Krushkal J, Boerwinkle E: TDT statistics for mapping quantitative loci. *Ann Hum Genet* 1998;62:431–452.
- Zhao J, Li W, Xiong M: Population based linkage disequilibrium mapping of QTL: An application to simulated data in an isolated population. *Genet Epidemiol* 2001;21(suppl 1):655–659.
- Zhu XF, Elston RC: Power comparison of regression methods to test quantitative traits for association and linkage. *Genet Epidemiol* 2000;18:322–330.
- Zhu XF, Elston RC: Transmission/disequilibrium tests for quantitative traits. *Genet Epidemiol* 2001;20:57–74.