DOI: 10.1534/genetics.105.046417

High-Resolution Association Mapping of Quantitative Trait Loci: A Population-Based Approach

Ruzong Fan,*,[1] Jeesun Jung[†] and Lei Jin*

**Department of Statistics, Texas A&M University, College Station, Texas 77843 and [†]Department of Human Genetics, University of Pittsburgh, Graduate School of Public Health, Pittsburgh, Pennsylvania 15261*

Manuscript received June 2, 2005
Accepted for publication September 19, 2005

ABSTRACT

In this article, population-based regression models are proposed for high-resolution linkage disequilibrium mapping of quantitative trait loci (QTL). Two regression models, the "genotype effect model" and the "additive effect model," are proposed to model the association between the markers and the trait locus. The marker can be either diallelic or multiallelic. If only one marker is used, the method is similar to a classical setting by Nielsen and Weir, and the additive effect model is equivalent to the haplotype trend regression (HTR) method by Zaykin *et al.* If two/multiple marker data with phase ambiguity are used in the analysis, the proposed models can be used to analyze the data directly. By analytical formulas, we show that the genotype effect model can be used to model the additive and dominance effects simultaneously; the additive effect model takes care of the additive effect only. On the basis of the two models, *F*-test statistics are proposed to test association between the QTL and markers. By a simulation study, we show that the two models have reasonable type I error rates for a data set of moderate sample size. The noncentrality parameter approximations of *F*-test statistics are derived to make power calculation and comparison. By a simulation study, it is found that the noncentrality parameter approximations of *F*-test statistics work very well. Using the noncentrality parameter approximations, we compare the power of the two models with that of the HTR. In addition, a simulation study is performed to make a comparison on the basis of the haplotype frequencies of 10 SNPs of angiotensin-1 converting enzyme (ACE) genes.

IN genetics research, one important goal is to locate and identify important genetic variants that are related to complex traits. With the development of dense maps such as single-nucleotide polymorphisms (SNPs) and high-resolution microsatellites in the human genome, enormous amounts of genetic data on human chromosomes are becoming available (International SNP Map Working Group 2001; Kong *et al.* 2002; International HapMap Consortium 2003; HapMap project, http://www.hapmap.org). The opportunities for a genomewide scan to map complex disease genes are tremendous. It is important to build appropriate models and useful algorithms in association mapping of complex diseases to identify important genetic variants of complex traits, for human, animal, or plant study.

In recent years, there has been great interest in linkage disequilibrium (LD) mapping (or association study) of quantitative traits of complex diseases. One way is to use diallelic markers such as SNPs in analysis. This approach has been receiving much attention and there are quite a lot of references to it in the literature (Fulker *et al.* 1999; George *et al.* 1999; Abecasis *et al.* 2000a,b, 2001; Sham *et al.* 2000; Fan *et al.* 2005). Another approach is to use haplotype data that may consist of a set of SNPs (Schaid *et al.* 2002; Zaykin *et al.* 2002; Schaid 2004). The haplotype data may provide more information on the relation between DNA variants and complex traits than that of any single SNP. Hence, it is important to investigate models and algorithms that are based on haplotype data. In Schaid *et al.* (2002) and Zaykin *et al.* (2002), score tests are proposed for association between complex traits and haplotypes, which can be ambiguous owing to the unknown linkage phase of different haplotypes. In Zaykin *et al.* (2002), the method is called haplotype trend regression (HTR), which is very close to the method of Schaid *et al.* (2002) (see Schaid 2004, p. 355, for further explanation). HTR does not assume that haplotype phases are known. Meuwissen and Goddard (2000) introduced a haplotype-based approach, which assumes that haplotype phases are known. In addition, mixed models are used to model the haplotype effect in Meuwissen and Goddard (2000). Morris *et al.* (2004) used a Markov chain Monte Carlo algorithm based on the shattered coalescent model for fine mapping.

On the other hand, the direct available information is genotypes by current genotyping technology, instead of haplotypes. Hence, it is interesting to build models by directly using genotype information; under these models, the main effects of each marker are modeled, which does not require phase information across the

[1]*Corresponding author:* Department of Statistics, Texas A&M University, 447 Blocker Bldg., College Station, TX 77843.
E-mail: rfan@stat.tamu.edu

markers. If phase is unknown, presumably the haplotypes would need to be estimated first, using a reconstruction algorithm such as PHASE or EM algorithms (Dempster *et al.* 1977; M. Stephens *et al.* 2001; Stephens and Donnelly 2003). This may introduce bias into the subsequent analysis, which would need to be investigated. It is of real interest in making comparison of the genotype-based models and the haplotype-based models. Interestingly, Morris *et al.* (2004) and Clayton *et al.* (2004) have observed that the haplotypes at SNPs may be only slightly more advantageous or even less powerful for fine mapping than the corresponding unphased genotypes.

Suppose that a quantitative trait locus (QTL) is located in a chromosome region. In the region, a marker (or two/multiple markers) is (or are) typed. In our previous research, the markers are assumed to be diallelic (Fan and Xiong 2002). In the current article, the markers can be either diallelic or multiallelic. Suppose that a population sample is available. For each individual in the sample, both trait value and genotypes at the markers are observed. We propose two regression models in association mapping of QTL based on population genetic data. One model is the "genotype effect model," and the other is the "additive effect model." These two models extend our previous research of high-resolution LD mapping of QTL using diallelic markers (Fan and Xiong 2002). The model can be very easily performed by using any statistical software in data analysis, or it can be easily implemented by widely used language such as C++. By analytical formulas, we show that the genotype effect model can be used to model the additive and dominance effects simultaneously; the additive effect model takes care of additive effect only. On the basis of the two models, *F*-test statistics are proposed to test association between the QTL and markers. To investigate the robustness of the proposed models and the related *F*-test statistics, simulation studies are performed to calculate the type I error rates. The noncentrality parameters of *F*-test statistics are derived to make power calculation and comparison. Moreover, the proposed models are compared with the haplotype trend regression method by simulation study and type I error rate analysis when two diallelic markers are used in the analysis (Zaykin *et al.* 2002). On the basis of the haplotype frequencies of 10 SNPs of angiotensin-1 converting enzyme (ACE) genes, a simulation study is performed to make power comparison of the proposed models with the haplotype trend regression method (Keavney *et al.* 1998).

A software, CLAM_QTL, is written in C++ to implement the proposed models and methods, which can be downloaded from http://www.stat.tamu.edu/~rfan/software.html/.

METHODS

As the first step, we present models and methods by using one marker. Here the marker can be either biallelic or multiallelic. This article extends our previous work (Fan and Xiong 2002). Similar results were worked out independently by colleagues at North Carolina State University, although their language and notations are slightly different (Weir and Cockerham 1977; Nielsen and Weir 1999, 2001). Then, the models and methods are extended to use two/multiple markers in analysis. On the basis of the models, *F*-test statistics are proposed, and the related noncentrality parameter approximations of the *F*-tests are derived.

Analysis by one marker: *Population models:* Consider a quantitative trait locus Q, which is located at an autosome. Suppose that there are two alleles Q_1 and Q_2 at the trait locus with frequencies q_1 and q_2, respectively. In a region of the QTL Q, suppose that one marker A is typed, which may be diallelic such as a single-nucleotide polymorphism or may be multiallelic such as a microsatellite marker. Let us denote the alleles of marker A by $A_1, \ldots, A_m$, where m is the number of alleles. Suppose that the marker A is in Hardy-Weinberg equilibrium (HWE). Let the frequency of A_i be $P_{A_i}, i = 1, 2, \ldots, m$. There are $J_A = m(m + 1)/2$ possible genotypes, which can be listed as $A_1A_1, \ldots, A_mA_m, A_1A_2, \ldots, A_1A_m, \ldots, A_{m-1}A_m$. Accordingly, let $\beta_{11}, \ldots, \beta_{mm}, \beta_{12}, \ldots, \beta_{1m}, \ldots, \beta_{m-1,m}$ be the corresponding effects of the listed genotypes on the quantitative trait. Let y be the trait value of an individual with genotype $G_A = A_iA_j$. Under an assumption of normality, the trait value can be modeled as

$$y = w\gamma + \beta_{ij} + e, \tag{1}$$

where w is a row vector of covariates such as sex and age, γ is a column vector of regression coefficients of w, and e is the error term. Assume that e is normal $N(0, \sigma_e^2)$. In addition to the covariate effects, there are $J_A = m(m + 1)/2$ parameters β_{ij} in model (1), where $\beta_{ij} = \beta_{ji}$. Model (1) treats each genotype effect as one parameter. Hence, we call it a genotype effect model. In practice, model (1) may lead to large number of parameters.

Now let us denote the effect of allele A_i as α_i, $i = 1, \ldots, m$. Suppose the genetic effect is additive in a sense of $\beta_{ij} = \alpha_i + \alpha_j$, $i, j = 1, \ldots, m$. If an individual has quantitative trait value y and genotype $G_A = A_iA_j$, model (1) can be modified as

$$y = w\gamma + \alpha_i + \alpha_j + e. \tag{2}$$

In addition to the covariate effects, there are m parameters α_i, $i = 1, \ldots, m$, in model (2). Compared with model (1), model (2) may significantly reduce the number of parameters. Since it models only the additive effect, we call it the additive effect model.

Property of model coefficients and association tests: As in the traditional quantitative genetics, let a be the effect of genotype Q_1Q_1, d be the effect of genotype Q_1Q_2, and $-a$ be the effect of genotype Q_2Q_2 (Falconer and Mackay 1996). Let $\alpha_Q = a + (q_2 - q_1)d$ be the average effect of gene substitution and $\delta_Q = 2d$ be the dominance

deviation. In addition, let $\mu = a(q_1 - q_2) + 2dq_1q_2$ be the aggregate effect of the QTL on the trait mean in the population. For $i = 1, 2, \ldots, m$, let us denote $D_{A_iQ} = P(Q_1A_i) - q_1P_{A_i}$, which are measures of LD between QTL Q and marker A. Here $P(Q_1A_i)$ is the frequency of haplotype Q_1A_i. In APPENDIX A, we show that the regression coefficients of model (1) are given by

$$\beta_{ij} = \mu + \alpha_Q[D_{A_iQ}/P_{A_i} + D_{A_jQ}/P_{A_j}] - \delta_Q D_{A_iQ}D_{A_jQ}/[P_{A_i}P_{A_j}]. \tag{3}$$

In APPENDIX B, we show that the regression coefficients of model (2) are given by

$$\alpha_i = \mu/2 + \alpha_Q D_{A_iQ}/P_{A_i}. \tag{4}$$

From Equations 3 and 4, it is clear that $\beta_{ij} = \alpha_i + \alpha_j$, when $\delta_Q = 0$, *i.e.*, no dominance effect. Suppose that the marker A and the QTL Q are in linkage equilibrium; *i.e.*, $D_{A_iQ} = 0, i = 1, 2, \ldots, m$. Then Equation 3 implies $\beta_{ij} = \mu$; Equation 4 implies that $\alpha_i = \mu/2$. Hence, models (1) and (2) are reduced to

$$y = w\gamma + \mu + e. \tag{5}$$

Assume that the additive genetic effect is significantly present, but the dominance genetic effect is not significantly present; *i.e.*, $\alpha_Q \neq 0$ but $\delta_Q = 0$. To test association between the marker A and the QTL Q, one may test hypotheses H_{a0}: $\alpha_1 = \cdots = \alpha_m$ *vs.* H_{a1}: at least two α_i's are not equal. To see this, note that the hypotheses H_{a0}: $\alpha_1 = \cdots = \alpha_m$ is equivalent to H_{a0}: $D_{A_1Q}/P_{A_1} = \cdots = D_{A_mQ}/P_{A_m}$, since α_Q is significantly different from 0. Thus, $0 = \sum_{i=1}^{m} D_{A_iQ} = D_{A_1Q}[1 + P_{A_2}/P_{A_1} + \cdots + P_{A_m}/P_{A_1}]$ implies $D_{A_1Q} = 0$ and so $D_{A_2Q} = \cdots = D_{A_mQ} = 0$ under H_{a0}. Hence, the hypotheses H_{a0}: $\alpha_1 = \cdots = \alpha_m$ *vs.* H_{a1}: at least two α_i's are not equal to each other are equivalent to H_{a0}: $D_{A_1Q} = \cdots = D_{A_mQ} = 0$ *vs.* H_{a1}: at least one D_{A_iQ} is not equal to 0. Model (2) can be used to map the QTL by an association analysis.

On the other hand, assume that both additive and dominance genetic effects are significantly present at the putative QTL Q; *i.e.*, $\alpha_Q \neq 0$ and $\delta_Q \neq 0$. To test association between the marker A and the QTL Q, one may test hypotheses H_{ad0}: $\beta_{11} = \cdots = \beta_{mm} = \beta_{12} = \cdots = \beta_{1m} = \cdots = \beta_{m-1,m}$ *vs.* H_{ad1}: at least two β_{ij}'s are not equal.

Relation to our previous work: If the marker A has only two alleles A_1 and A_2, FAN and XIONG (2002) proposed the following model in association mapping of the QTL Q,

$$y = w\gamma + \mu + x_A\alpha_A + z_A\delta_A + e, \tag{6}$$

where x_A and z_A are dummy random variables defined by

$$x_A = \begin{cases} 2P_{A_2} & \text{if } G_A = A_1A_1 \\ P_{A_2} - P_{A_1} & \text{if } G_A = A_1A_2, \\ -2P_{A_1} & \text{if } G_A = A_2A_2 \end{cases}$$
$$z_A = \begin{cases} -P_{A_2}^2 & \text{if } G_A = A_1A_1 \\ P_{A_2}P_{A_1} & \text{if } G_A = A_1A_2, \\ -P_{A_1}^2 & \text{if } G_A = A_2A_2 \end{cases} \tag{7}$$

and α_A and δ_A are regression coefficients of the dummy variables x_A and z_A. The regression coefficients are given by $\alpha_A = D_{A_1Q}\alpha_Q/(P_{A_1}P_{A_2})$ and $\delta_A = D_{A_1Q}^2\delta_Q/\left(P_{A_1}^2P_{A_2}^2\right)$ (FAN and XIONG 2002). It can be shown that model (6) is equivalent to model (1). Actually, the following relations of the regression coefficients of the two models can be shown: $\beta_{11} = \mu + 2P_{A_2}\alpha_A - P_{A_2}^2\delta_A$, $\beta_{12} = \mu + (P_{A_2} - P_{A_1})\alpha_A + P_{A_1}P_{A_2}\delta_A$, and $\beta_{22} = \mu - 2P_{A_1}\alpha_A - P_{A_1}^2\delta_A$. Similarly, model (2) is equivalent to $y = w\gamma + \mu + x_H\alpha_A + e$, and we have the following relations $2\alpha_1 = \mu + 2P_{A_2}\alpha_A$ and $2\alpha_2 = \mu - 2P_{A_1}\alpha_A$. The advantage of model (6) is that the association effect is decomposed into summations of additive and dominance effects if A is diallelic. If A has more than two alleles, model (1) extends model (6), and model (2) extends model $y = w\gamma + \mu + x_H\alpha_A + e$.

Regression models: Assume that N individuals from a population are available for study. Let us list their trait values as $y_1, \ldots, y_N$ and their genotypes as $G_{A1}, \ldots, G_{AN}$. For individual k, let $x_{ii}^{(k)}$ be the indicator function of genotype A_iA_i and $x_{ij}^{(k)}$ be the indicator function of genotype A_iA_j. That is, they are dummy variables defined by

$$x_{ii}^{(k)} = \begin{cases} 1 \text{ if } G_{Ak} = A_iA_i \\ 0 \text{ else,} \end{cases} \quad x_{ij}^{(k)} = \begin{cases} 1 \text{ if } G_{Ak} = A_iA_j \\ 0 \text{ else,} \end{cases}$$

where $i, j = 1, 2, \ldots, m$, $i \neq j$. Let $X_k = (x_{11}^{(k)}, \ldots, x_{mm}^{(k)}, x_{12}^{(k)}, \ldots, x_{1m}^{(k)}, \ldots, x_{m-1,m}^{(k)})^\tau$, $k = 1, 2, \ldots, N$; *i.e.*, X_k is a column vector of genotype indicator functions of individual k. Here the superscript τ denotes a vector/matrix transpose. Denote $\eta = (\beta_{11}, \ldots, \beta_{mm}, \beta_{12}, \ldots, \beta_{1m}, \ldots, \beta_{m-1,m})^\tau$. The corresponding regression of model (1) can be written as

$$y_k = w_k\gamma + X_k^\tau\eta + e_k, \tag{8}$$

where subscript k indicates the corresponding quantities of individual k.

Similarly, let $z_i^{(k)}$ be the number of alleles A_i of genotype G_{Ak}, $i = 1, 2, \ldots, m$, for individual k. That is, $z_i^{(k)}$ is a dummy variable defined by

$$z_i^{(k)} = \begin{cases} 2 \text{ if } G_{Ak} = A_iA_i \\ 1 \text{ if } G_{Ak} = A_iA_j, \quad j \neq i. \\ 0 \text{ else.} \end{cases}$$

Denote $Z_k = (z_1^{(k)}, \ldots, z_m^{(k)})^\tau$ and $\psi = (\alpha_1, \ldots, \alpha_m)^\tau$. To use model (2) for data analysis, the corresponding regression model is

$$y_k = w_k\gamma + Z_k^\tau\psi + e_k. \tag{9}$$

F-tests and noncentrality parameter approximations: It is well known that the additive variance $\sigma_{ga}^2 = 2q_1q_2\alpha_Q^2$ and the dominance variance $\sigma_{gd}^2 = (q_1q_2)^2\delta_Q^2$. Let $\sigma^2 = \sigma_{ga}^2 + \sigma_{gd}^2 + \sigma_e^2$ be the total variance. Assume that there are no covariates. Let us denote $X = (X_1, \ldots, X_N)^\tau$, $y = (y_1, \ldots, y_N)^\tau$, and $e = (e_1, \ldots, e_N)^\tau$. Then model (8) can be expressed as $y = X\eta + e$. By standard regression theory, the

coefficients can be estimated by $\hat{\eta} = (X^\tau X)^{-1} X^\tau Y$. Let H be a $(J_A - 1) \times J_A$ matrix defined by

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

Then, $(H\eta)^\tau = (\beta_{11} - \beta_{22}, \dots, \beta_{11} - \beta_{mm}, \beta_{11} - \beta_{12}, \dots, \beta_{11} - \beta_{1m}, \dots, \beta_{11} - \beta_{m-1,m})$. Hence, the hypothesis H_{ad0} is equivalent to $H\eta = (0, \dots, 0)^\tau$. From Graybill (1976), Chap. 6, the test statistic of a hypothesis H_{ad0} is noncentral $F(J_A - 1, N - J_A)$ defined by

$$F_{m,\mathrm{ad}} = \frac{(H\hat{\eta})^\tau [H(X^\tau X)^{-1} H^\tau]^{-1} (H\hat{\eta})}{Y^\tau [I_N - X(X^\tau X)^{-1} X^\tau] Y} \frac{N - J_A}{J_A - 1},$$

where I_N is the $N \times N$ identity matrix. The noncentrality parameter of the above F-statistic is $\lambda_{m,\mathrm{ad}} = (H\eta)^\tau [H(X^\tau X)^{-1} H^\tau]^{-1} (H\eta)/\sigma^2$. Under the assumption of large sample sizes N, we show in appendix c the approximation

$$\lambda_{m,\mathrm{ad}} \approx \frac{N}{\sigma^2} \left[\sigma^2_{\mathrm{ga}} R^2_{AQ} + \sigma^2_{\mathrm{gd}} R^4_{AQ} \right], \tag{10}$$

where R^2_{AQ} is a general measure of the degree of linkage disequilibrium between marker A and the QTL Q defined by $R^2_{AQ} = \sum_{j=1}^{m} \sum_{s=1}^{2} [P(Q_s A_j) - P_{A_j} q_s]^2 / [P_{A_j} q_s]$ (Crow and Kimura 1970; Hedrick 1987; Morton and Wu 1988; Sham *et al.* 2000). Note that R^2_{AQ} is the χ^2-statistic of the $m \times 2$ table of haplotype frequencies of the marker A and trait locus Q. Approximation (10) shows that the noncentrality parameter of test statistics of the null hypothesis of no genetic effects of model (1) is reduced by a factor of R^2_{AQ} for additive variance and by a factor of R^4_{AQ} for dominance variance.

Similarly, let us denote $Z = (Z_1, \dots, Z_N)^\tau$. Then model (9) can be expressed as $y = Z\psi + e$. The coefficients can be estimated by $\hat{\psi} = (Z^\tau Z)^{-1} Z^\tau Y$. Let K be a $(m - 1) \times m$ matrix defined by

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

Then, $(K\psi)^\tau = (\alpha_1 - \alpha_2, \dots, \alpha_1 - \alpha_m)$. Hence, the hypothesis H_{a0} is equivalent to $K\psi = (0, \dots, 0)^\tau$. From Graybill (1976), Chap. 6, the test statistic of the hypothesis H_{a0} is noncentral $F(m - 1, N - m)$ defined by

$$F_{m,\mathrm{a}} = \frac{(K\hat{\psi})^\tau [K(Z^\tau Z)^{-1} K^\tau]^{-1} (K\hat{\psi})}{Y^\tau [I_N - Z(Z^\tau Z)^{-1} Z^\tau] Y} \frac{N - m}{m - 1}.$$

The noncentrality parameter of the above F-statistic is $\lambda_{m,\mathrm{a}} = (K\psi)^\tau [K(Z^\tau Z)^{-1} K^\tau]^{-1} (K\psi)/\sigma^2$. Under an assumption of large sample sizes N, we show in appendix d the following approximation:

$$\lambda_{m,\mathrm{a}} = \frac{1}{\sigma^2} (K\psi)^\tau [K(Z^\tau Z)^{-1} K^\tau]^{-1} (K\psi) \approx \frac{N\sigma^2_{\mathrm{ga}}}{\sigma^2} R^2_{AQ}. \tag{11}$$

This approximation (11) shows that the noncentrality parameter $\lambda_{m,\mathrm{a}}$ is reduced by a factor of R^2_{AQ} for additive variance. The dominance variance is not present in $\lambda_{m,\mathrm{a}}$.

Analysis by two/multiple markers: *Population models and association tests:* If genetic data of two/multiple markers are available, models (1) and (2) can be extended for association study of QTL. Most importantly, the data of two/multiple markers may contain phase ambiguity, *i.e.*, phase unknown double heterozygotes. In the following, we generalize models (1) and (2) to directly analyze genetic data of two markers. The principle, actually, can be applied to multiple marker data.

In addition to marker A, assume that a second marker B is typed, which has n alleles denoted by $B_1, \dots, B_n$. Suppose that the marker B is also in Hardy-Weinberg equilibrium. Let the frequency of allele B_k be P_{B_k}, $k = 1, 2, \dots, n$. There are $J_B = n(n + 1)/2$ possible genotypes, which can be listed as $B_1B_1, \dots, B_nB_n, B_1B_2, \dots, B_1B_n, \dots, B_{n-1}B_n$. Let y be the trait value of an individual with genotype G_A at marker A and genotype G_B at marker B. Such as relations (7), define

$$\begin{aligned} x_{Ai} &= \begin{cases} 2 & \text{if } G_A = A_iA_i \\ 1 & \text{if } G_A = A_iA_j, \quad j \neq i \\ 0 & \text{else,} \end{cases} \\ z_{Aij} &= \begin{cases} -P^2_{A_j} & \text{if } G_A = A_iA_i \\ P_{A_i}P_{A_j} & \text{if } G_A = A_iA_j, \quad j \neq i \\ -P^2_{A_i} & \text{if } G_A = A_jA_j \\ 0 & \text{else,} \end{cases} \\ x_{Bk} &= \begin{cases} 2 & \text{if } G_B = B_kB_k \\ 1 & \text{if } G_B = B_kB_l, \quad l \neq k \\ 0 & \text{else,} \end{cases} \\ z_{Bkl} &= \begin{cases} -P^2_{B_l} & \text{if } G_B = B_kB_k \\ P_{B_k}P_{B_l} & \text{if } G_B = B_kB_l, \quad l \neq k \\ -P^2_{B_k} & \text{if } G_B = B_lB_l \\ 0 & \text{else.} \end{cases} \end{aligned} \tag{12}$$

If marker A has only two alleles A_1 and A_2, then x_{Ai} defined above is closely related to x_A, which is defined in (7). Actually, it is easy to see the following relation $x_A + 2P_{A_1} = x_{A1}$ since $P_{A_1} + P_{A_2} = 1$.

To extend model (2) by using two markers A and B in the analysis, consider the following model

$$y = w\gamma + \alpha + \sum_{i=1}^{m-1} x_{Ai}\alpha_{Ai} + \sum_{k=1}^{n-1} x_{Bk}\alpha_{Bk} + e. \tag{13}$$

In addition to the covariate effects, there are $m + n - 1$ parameters $\alpha, \alpha_{Ai}, \alpha_{Bk}$, $i = 1, \dots, m - 1, k = 1, \dots, n - 1$

in model (13). To see why model (13) extends model (2), it is worthwhile to note that model (2) is equivalent to $y = w\gamma + \alpha + \sum_{i=1}^{m-1} x_{Ai}\alpha_{Ai} + e$. Actually, the quantity $\sum_{i=1}^{m} x_{Ai} = 2$ implies that $y = w\gamma + \alpha + \sum_{i=1}^{m-1} x_{Ai}\alpha_{Ai} + e = w\gamma + \sum_{i=1}^{m-1} x_{Ai}[\alpha_{Ai} + \alpha/2] + x_{Am}\alpha/2 + e$ if only information of marker A is used in the analysis; thus, $\alpha_m = \alpha/2$, $\alpha_i = \alpha_{Ai} + \alpha/2$, $i = 1, \ldots, m-1$. Such as model (2), model (13) takes only the additive effect into account. Hence, we call it an additive effect model. Similarly, model (1) can be extended to

$$y = w\gamma + \alpha + \sum_{i=1}^{m-1} x_{Ai}\alpha_{Ai} + \sum_{k=1}^{n-1} x_{Bk}\alpha_{Bk} + \sum_{1\le i<j\le m} z_{Aij}\delta_{Aij} + \sum_{1\le k<l\le n} z_{Bkl}\delta_{Bkl} + e. \quad (14)$$

In addition to the covariate effects, there are $J_A + J_B - 1$ parameters α, α_{Ai}, α_{Bk}, δ_{Aij}, δ_{Bkl} in model (14). Model (14) takes both additive and dominance effects into account, and it is called the genotype effect model. Again, model (1) is equivalent to $y = w\gamma + \alpha + \sum_{i=1}^{m-1} x_{Ai}\alpha_{Ai} + \sum_{1\le i<j\le m} z_{Aij}\delta_{Aij} + e$.

Denote $X_A = (x_{A1}, \ldots, x_{A(m-1)})^\tau$, $X_B = (x_{B1}, \ldots, x_{B(n-1)})^\tau$, and $X_{A\cup B} = (X_A^\tau, X_B^\tau)^\tau$. Let us denote the additive variance–covariance matrix of the indicator variables x_{Ai}, x_{Bk} by $V_A = \mathrm{Cov}(X_{A\cup B}, X_{A\cup B}) = E(X_{A\cup B}X_{A\cup B}^\tau) - EX_{A\cup B}(EX_{A\cup B}^\tau)$. Similarly, let $Z_A = (z_{A12}, \ldots, z_{A1m}, z_{A23}, \ldots, z_{A2m}, \ldots, z_{A(m-1)m})^\tau$, $Z_B = (z_{B12}, \ldots, z_{B1n}, z_{B23}, \ldots, z_{B2n}, \ldots, z_{B(n-1)n})^\tau$, and $Z_{A\cup B} = (Z_A^\tau, Z_B^\tau)^\tau$. Let us denote the dominance variance–covariance matrix of the indicator variables z_{Aij}, z_{Bkl} by $V_D = \mathrm{Cov}(Z_{A\cup B}, Z_{A\cup B})$. For $k = 1, 2, \ldots, n$, let us denote $D_{B_kQ} = P(Q_1B_k) - q_1P_{B_k}$, which are measures of LD between QTL Q and marker B. In APPENDIX E, we show that the regression coefficients of models (13) and (14) are given by

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

$$\begin{pmatrix} \delta_{A12} \\ \vdots \\ \delta_{A(m-1)m} \\ \delta_{B12} \\ \vdots \\ \delta_{B(n-1)n} \end{pmatrix} = V_D^{-1} \begin{pmatrix} [P_{A_2}D_{A_1Q} - P_{A_1}D_{A_2Q}]^2 \\ \vdots \\ [P_{A_{m-1}}D_{A_mQ} - P_{A_m}D_{A_{m-1}Q}]^2 \\ [P_{B_2}D_{B_1Q} - P_{B_1}D_{B_2Q}]^2 \\ \vdots \\ [P_{B_{n-1}}D_{B_nQ} - P_{B_n}D_{B_{n-1}Q}]^2 \end{pmatrix} \delta_Q. \quad (15)$$

The elements of matrices V_A and V_D are provided in APPENDIX E. Equations 15 show that the parameters of LD (*i.e.*, D_{A_iQ} and D_{B_kQ}) and gene effect (*i.e.*, α_Q and δ_Q) are contained in the regression coefficients. Models (13) and (14) simultaneously take care of the LD and the effects of the putative trait locus Q. The gene substitution effect α_Q is contained only in α_{Ai}, α_{Bk}; and the dominance effect δ_Q is contained only in δ_{Aij}, δ_{Bkl}. Therefore, V_A is called the additive variance–covariance matrix; and V_D is called the dominance variance–covariance matrix. The model (14) orthogonally decomposes the genetic effect into a summation of additive and dominance effects.

In FAN and XIONG (2002), regression models are proposed for LD mapping of QTL by diallelic markers. Models (13) and (14) extend the models by using multiallelic markers in LD analysis. On the basis of Equations 15, we may use models (13) and (14) to test the association between the trait locus Q and the two markers A and B. Assume that the additive genetic effect is significantly present, but the dominance genetic effect is not significantly present; *i.e.*, $\alpha_Q \neq 0$ but $\delta_Q = 0$. To test association between the markers A and B and the QTL Q, one may test hypotheses $H_{AB\mathrm{a}0}$: $\alpha_{A1} = \cdots = \alpha_{A(m-1)} = \alpha_{B1} = \cdots = \alpha_{B(n-1)} = 0$ *vs.* $H_{AB\mathrm{a}1}$: at least one α_{Ai}, α_{Bk} is not equal to 0. To see this, note that the hypothesis $H_{AB\mathrm{a}0}$ is equivalent to $D_{A_1Q} = \cdots = D_{A_{m-1}Q} = D_{B_1Q} = \cdots = D_{B_{n-1}Q} = 0$, since α_Q is significantly different from 0. On the other hand, assume that both additive and dominance genetic effects are significantly present at the putative QTL Q; *i.e.*, $\alpha_Q \neq 0$ and $\delta_Q \neq 0$. To test association between the markers A and B and the QTL Q, one may test hypothesis $H_{AB\mathrm{ad}0}$: $\alpha_{A1} = \cdots = \alpha_{A(m-1)} = \alpha_{B1} = \cdots = \alpha_{B(n-1)} = \delta_{A12} = \cdots = \delta_{A1m} = \cdots = \delta_{A(m-1)m} = \delta_{B12} = \cdots = \delta_{B1n} = \cdots = \delta_{B(n-1)n} = 0$ *vs.* $H_{AB\mathrm{ad}1}$: at least one α_{Ai}, α_{Bk}, δ_{Aij}, δ_{Bkl} is not equal to 0, since both α_Q and δ_Q are significantly different from 0.

Regression models, F-tests, and noncentrality parameter approximations: Assume that N individuals from a population are available for study, whose trait values are listed as $y_1, \ldots, y_N$ and their genotypes as $G_{A1}, \ldots, G_{AN}$ at marker A and $G_{B1}, \ldots, G_{BN}$ at marker B. For individual s, let $x_{Ai}^{(s)}, z_{Aij}^{(s)}, x_{Bk}^{(s)}, z_{Bkl}^{(s)}$ be the corresponding coding functions of genotypes G_{As} and G_{Bs}. Let us denote $X_{A\cup B}^{(s)} = (1, x_{A1}^{(s)}, \ldots, x_{A(m-1)}^{(z)}, x_{B1}^{(s)}, \ldots, x_{B(n-1)}^{(s)})$ and $Z_{A\cup B}^{(s)} = (z_{A12}^{(s)}, \ldots, z_{A1m}^{(s)}, \ldots, z_{A(m-1)m}^{(s)}, z_{B12}^{(s)}, \ldots, z_{B1n}^{(s)}, \ldots, z_{B(n-1)n}^{(s)})$, $s = 1, 2, \ldots, N$. Denote $\alpha_{A\cup B} = (\alpha, \alpha_{A1}, \ldots, \alpha_{A(m-1)}, \alpha_{B1}, \ldots, \alpha_{B(n-1)})^\tau$, and $\delta_{A\cup B} = (\delta_{A12}, \ldots, \delta_{A(m-1)m}, \delta_{B12}, \ldots, \delta_{B(n-1)n})^\tau$. The corresponding regression of model (14) can be written as

$$y_s = w_s\gamma + X_{A\cup B}^{(s)}\alpha_{A\cup B} + Z_{A\cup B}^{(s)}\delta_{A\cup B} + e_s, \quad s = 1, 2 \ldots, N. \quad (16)$$

Let us denote $D_{AQ} = (D_{A_1Q}, \ldots, D_{A_{m-1}Q})^\tau$ and $D_{BQ} = (D_{B_1Q}, \ldots, D_{B_{n-1}Q})^\tau$; $\Delta_{AQ} = \left([P_{A_2}D_{A_1Q} - P_{A_1}D_{A_2Q}]^2, \ldots, [P_{A_{m-1}}D_{A_mQ} - P_{A_m}D_{A_{m-1}Q}]^2\right)^\tau$ and $\Delta_{BQ} = \left([P_{B_2}D_{B_1Q} - P_{B_1}D_{B_2Q}]^2, \ldots, [P_{B_{n-1}}D_{B_nQ} - P_{B_n}D_{B_{n-1}Q}]^2\right)^\tau$. On the basis of

regression (16), one may construct an F-test statistic $F_{AB,\mathrm{ad}}$ to test the null hypothesis $\mathrm{H}_{AB\mathrm{ad}0}$ in the same way as constructing $F_{m,\mathrm{ad}}$ or $F_{m,\mathrm{a}}$ (Graybill 1976, Chap. 6). Under the null hypothesis of $\mathrm{H}_{AB\mathrm{ad}0}$, $F_{AB,\mathrm{ad}}$ is central to $F(J_A + J_B - 2, N - J_A - J_B + 1)$. Assume the sample size N is large enough that the large sample theory applies. Under the alternative hypothesis of $\mathrm{H}_{AB\mathrm{ad}1}$, $F_{AB,\mathrm{ad}}$ is noncentral to $F(J_A + J_B - 2, N - J_A - J_B + 1)$, and it can be shown that the corresponding noncentrality parameter is approximated by

$$\lambda_{AB\mathrm{ad}} \approx \frac{N}{\sigma^2}\left[\begin{array}{l}(D_{AQ}^{\tau}, D_{BQ}^{\tau})(V_{\mathrm{A}}/2)^{-1}\begin{pmatrix}D_{AQ}\\ D_{BQ}\end{pmatrix}\sigma_{\mathrm{ga}}^2/(q_1 q_2)\\ \quad + (\Delta_{AQ}^{\tau}, \Delta_{BQ}^{\tau})V_{\mathrm{D}}^{-1}\begin{pmatrix}\Delta_{AQ}\\ \Delta_{BQ}\end{pmatrix}\sigma_{\mathrm{gd}}^2/(q_1^2 q_2^2)\end{array}\right].$$

Similarly, an F-test statistic $F_{AB,\mathrm{a}}$ used to test the null hypothesis $\mathrm{H}_{AB\mathrm{a}0}$ can be constructed. Under the null hypothesis of $\mathrm{H}_{AB\mathrm{a}0}$, $F_{AB,\mathrm{a}}$ is central to $F(m + n - 2, N - n - m + 1)$. Under the alternative hypothesis of $\mathrm{H}_{AB\mathrm{a}1}$, $F_{AB,\mathrm{a}}$ is noncentral to $F(m + n - 2, N - m - n + 1)$, and it can be shown that the corresponding noncentrality parameter is approximated by

$$\lambda_{AB\mathrm{a}} \approx \frac{N}{\sigma^2}(D_{AQ}^{\tau}, D_{BQ}^{\tau})(V_{\mathrm{A}}/2)^{-1}\begin{pmatrix}D_{AQ}\\ D_{BQ}\end{pmatrix}\sigma_{\mathrm{ga}}^2/(q_1 q_2).$$

The haplotype trend regression method: If only one marker A is used in the analysis, the proposed model (2) is equivalent to the HTR method of Zaykin *et al.* (2002). However, the proposed models are different from the haplotype trend regression method for two/multiple marker data. Assume that M markers are typed in a region of the trait locus Q. On the basis of the genotypes of the multiple markers, assume that J haplotypes can be determined as $h_1, \ldots, h_J$ with frequencies $P_{h_j}, j = 1, 2, \ldots, J$. For each individual, we may define an expected haplotype score vector as follows (Schaid *et al.* 2002; Zaykin *et al.* 2002). The expected haplotype score vector is a column vector of J elements $(c_1, \ldots, c_J)^{\tau}$ based on the genotype combination $(G_1, \ldots, G_M)$ at the markers of an individual. For instance, the score vector is $(1, 0, \ldots, 0)^{\tau}$ if haplotype pair h_1/h_1 is the only possible phase of the genotype combination $(G_1, \ldots, G_M)$. In general, c_j is the conditional probability of a haplotype h_j given genotype combination $(G_1, \ldots, G_M)$ at the markers; *i.e.*,

$$c_j = P(h_j|G_1, \ldots, G_M) = \frac{P_{h_j}\sum_{i=1}^{J} P(G_1, \ldots, G_M|h_j, h_i)P_{h_i}}{\sum_{i=1}^{J}\sum_{k=1}^{J} P(G_1, \ldots, G_M|h_i, h_k)P_{h_i}P_{h_k}}.$$

In the above equation, the conditional probability $P(G_1, \ldots, G_M|h_i, h_k)$ is 1 if haplotype pair h_i/h_k is a possible phase for the genotype combination $(G_1, \ldots, G_M)$, and $P(G_1, \ldots, G_M|h_k, h_j)$ is 0 otherwise. For each individual, the summation $\sum_{j=1}^{J} c_j$ of the expected haplotype scores is equal to 1.

TABLE 1

Expected scorings I_i, $i = 1, 2, 3, 4$ of haplotypes of model (17)

	Haplotype and related expected scoring			
Genotype (G_A, G_B)	A_1B_1, I_1	A_1B_2, I_2	A_2B_1, I_3	A_2B_2, I_4
(A_1A_1, B_1B_1)	1	0	0	0
(A_1A_1, B_1B_2)	$\frac{1}{2}$	$\frac{1}{2}$	0	0
(A_1A_1, B_2B_2)	0	1	0	0
(A_1A_2, B_1B_1)	$\frac{1}{2}$	0	$\frac{1}{2}$	0
(A_1A_2, B_1B_2)	c_1	c_2	c_2	c_1
(A_1A_2, B_2B_2)	0	$\frac{1}{2}$	0	$\frac{1}{2}$
(A_2A_2, B_1B_1)	0	0	1	0
(A_2A_2, B_1B_2)	0	0	$\frac{1}{2}$	$\frac{1}{2}$
(A_2A_2, B_2B_2)	0	0	0	1

The constants are given by $c_1 = P(A_1B_1|G_A = A_1A_2, G_B = B_1B_2) = P(A_1B_1)P(A_2B_2)/[2P(A_1B_1)P(A_2B_2) + 2P(A_1B_2)P(A_2B_1)]$ and $c_2 = \frac{1}{2} - c_1$.

For the purpose of explanation, consider two diallelic markers A and B. Let us denote the two alleles of marker A by A_1, A_2; and denote the two alleles of marker B by B_1, B_2. Table 1 gives the score vector for each genotype combination of markers A and B. To understand the entries of Table 1, it is worthwhile to take genotype combination $(G_A = A_1A_1, G_B = B_1B_1)$ as an example. Two copies of haplotype A_1B_1 can be formed from the genotype combination $(G_A = A_1A_1, G_B = B_1B_1)$. The score for haplotype A_1B_1 is 1 for this genotype combination; and scores for the other three haplotypes are all 0. Denote the genotype of an individual at marker A by G_A and the genotype at marker B by G_B. Let us denote $c_1 = P(A_1B_1|G_A = A_1A_2, G_B = B_1B_2) = P(A_1B_1)P(A_2B_2)/[2P(A_1B_1)P(A_2B_2) + 2P(A_1B_2)P(A_2B_1)] = c_4$; *i.e.*, c_1 is the conditional probability of a haplotype A_1B_1 given the double heterozygotes $(G_A = A_1A_2, G_B = B_1B_2)$; and $c_2 = c_3 = \frac{1}{2} - c_1$. For the double heterozygotes $(G_A = A_1A_2, G_B = B_1B_2)$, the expected scores are c_1, c_2, c_2, c_1 for haplotypes A_1B_1, A_1B_2, A_2B_1, A_2B_2. The scores of the other genotype combinations are provided in Table 1. Then the corresponding model of the haplotype trend regression method can be written as

$$y = w\gamma + \sum_{i=1}^{4} I_i\beta_i + e, \tag{17}$$

where β_i are regression coefficients, and I_i are expected scorings of haplotypes defined in Table 1. It can be seen that model (17) is not equivalent to either proposed model (13) or model (14).

In the general case of M markers, let I_j be the expected score of haplotype h_j, $j = 1, 2, \ldots, J$. In terms of conditional probabilities, I_j can be expressed as

$$I_j = \sum_{G_1}\cdots\sum_{G_M} P(h_j|G_1, \ldots, G_M)1_{(G_1,\ldots,G_M)}.$$

The corresponding model of the haplotype trend regression method can be written as

$$y = w\gamma + \sum_{j=1}^{J} I_j \beta_j + e. \tag{18}$$

For $j = 1, 2, \ldots, J$, let us denote $D_{h_j Q} = P(Q_1 h_j) - q_1 P_{h_j}$, which are measures of LD between QTL Q and the haplotypes. Here $P(Q_1 h_j)$ is the frequency of haplotype $Q_1 h_j$. In APPENDIX F, we show that the regression coefficients of model (18) satisfy the matrix equation

$$\begin{pmatrix} E(I_1^2) & E(I_1 I_2) & \ldots & E(I_1 I_J) \\ E(I_2 I_1) & E(I_2^2) & \ldots & E(I_2 I_J) \\ \vdots & \vdots & \ldots & \vdots \\ E(I_J I_1) & E(I_J I_2) & \ldots & E(I_J^2) \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_J \end{pmatrix} = \mu \begin{pmatrix} P_{h_1} \\ P_{h_2} \\ \vdots \\ P_{h_J} \end{pmatrix} + \alpha_Q \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_J \end{pmatrix} - \delta_Q \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_J \end{pmatrix}, \tag{19}$$

where $E(I_i I_k)$ are given in APPENDIX F, and

$$\begin{aligned} a_j &= \sum_{G_1} \cdots \sum_{G_M} P(h_j | G_1, \ldots, G_M) \\ &\quad \times \sum_{i=1}^{J} \sum_{k=1}^{J} P(G_1, \ldots, G_M | h_i, h_k) \left[P_{h_i} D_{h_k Q} + P_{h_k} D_{h_i Q} \right] \\ d_j &= \sum_{G_1} \cdots \sum_{G_M} P(h_j | G_1, \ldots, G_M) \\ &\quad \times \sum_{i=1}^{J} \sum_{k=1}^{J} P(G_1, \ldots, G_M | h_i, h_k) D_{h_i Q} D_{h_k Q}. \end{aligned}$$

From Equations 19, it is clear that model (18) models both the additive and dominance effects. Suppose that the haplotype and the QTL Q are in linkage equilibrium; *i.e.*, $D_{h_j Q} = 0, j = 1, 2, \ldots, J$. Then Equation 19 implies $\beta_1 = \cdots = \beta_J = \mu$, since $\sum_{j=1}^{J} I_j = 1$ and $EI_j = P_{h_j}$. Hence, model (18) is reduced to (5). To test association between the haplotypes and the trait locus, one may test a null hypothesis $\beta_1 = \cdots = \beta_J$, and the related F-test statistic can be constructed.

Again, assume that N individuals from a population are available for study with trait values and genotype information. On the basis of regression (18), one may construct an F-test statistic F_{HTR} to test the null hypothesis $\beta_1 = \cdots = \beta_J = \mu$ (GRAYBILL 1976). Under the null hypothesis, F_{HTR} is central to $F(J-1, N-J)$. Under the alternative hypothesis that at least two β_j's are not equal to each other, F_{HTR} is noncentral to $F(J-1, N-J)$. Assume the sample size N is large enough that the large sample theory applies. Then it can be shown that the corresponding noncentrality parameter is approximated by

$$\begin{aligned} \lambda_{\mathrm{HTR}} \approx \frac{N}{\sigma^2} &(\beta_1 - \beta_2, \ldots, \beta_1 - \beta_J)[HE^{-1}H^{\tau}]^{-1} \\ &\times (\beta_1 - \beta_2, \ldots, \beta_1 - \beta_J)^{\tau}, \end{aligned}$$

where

$$E = \begin{pmatrix} E(I_1^2) & E(I_1 I_2) & \ldots & E(I_1 I_J) \\ E(I_2 I_1) & E(I_2^2) & \ldots & E(I_2 I_J) \\ \vdots & \vdots & \ldots & \vdots \\ E(I_J I_1) & E(I_J I_2) & \ldots & E(I_J^2) \end{pmatrix}$$

$$H = \begin{pmatrix} 1 & -1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & \ldots & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ldots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \ldots & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 & \ldots & 0 & 0 & -1 \end{pmatrix}_{(J-1)\times J}.$$

The advantage of model (17) is that it may model the haplotype effect by parameters β_i. In practice, it is necessary to calculate the expected scorings or haplotype frequencies before building the haplotype trend regression model. Instead, the proposed models (13) and (14) may be used to analyze genetic data directly. Moreover, we have derived analytical formulas to calculate the regression coefficients of the HTR method and the related noncentrality parameter of the test statistic F_{HTR}. Note that the original article by ZAYKIN *et al.* (2002) did not work out this very useful information. Our analytical coefficient equations and related noncentrality parameter approximations can be readily utilized for power evaluation.

RESULTS

Type I error rates: To evaluate the robustness of the proposed models, we calculate type I error rates of test statistics $F_{m,\mathrm{ad}}$, $F_{m,\mathrm{a}}$, $F_{AB,\mathrm{ad}}$, $F_{AB,\mathrm{a}}$, and F_{HTR} at a 0.05 significance level. The results are presented in Tables 2 and 3. Four test cases are considered: null, no major gene effect $a = d = 0$; additive, additive mode of inheritance $a = 1$, but no dominant effect $d = 0$; dominant, dominant mode of inheritance $a = d = 1$; and recessive, recessive mode of inheritance $a = 1$ and $d = -0.5$. The total variance is fixed as $\sigma^2 = 1.0$ and the trait allele frequency is taken as $q_1 = q_2 = 0.5$ except for that in the null test case. In Table 2, only one marker A is used in analysis; the number m of alleles ranges from 2 to 6. The allele frequencies are given by: $P_{A_1} = P_{A_2} = 0.5$ when $m = 2$; $P_{A_1} = 0.4, P_{A_2} = P_{A_3} = 0.3$ when $m = 3$;

TABLE 2

Type I error rates (percentage) of test statistics $F_{m,ad}$ and $F_{m,a}$ at a 0.05 significance level when only one marker A is used in the analysis

No. of alleles	Sample size	Test case	Error rates $F_{m,ad}$	Error rates $F_{m,a}$
Diallele, $m = 2$	$N = 200$	Null	4.90	4.93
		Additive	5.10	4.89
		Dominant	4.75	4.98
		Recessive	5.03	5.09
Triallele, $m = 3$	$N = 200$	Null	4.94	5.18
		Additive	5.03	4.92
		Dominant	5.07	5.06
		Recessive	4.65	4.85
Quadriallele, $m = 4$	$N = 200$	Null	4.89	5.29
		Additive	4.72	4.69
		Dominant	5.03	4.92
		Recessive	4.86	4.85
Five alleles, $m = 5$	$N = 200$	Null	4.71	5.14
		Additive	4.96	4.49
		Dominant	5.02	4.94
		Recessive	5.04	4.76
Six alleles, $m = 6$	$N = 200$	Null	5.02	5.21
		Additive	5.23	4.92
		Dominant	9.11	5.16
		Recessive	7.04	4.97
Six alleles, $m = 6$	$N = 300$	Null	4.91	5.36
		Additive	5.08	4.98
		Dominant	5.39	4.91
		Recessive	5.32	5.11

The total variance is fixed as $\sigma^2 = 1.0$ and the trait allele frequency is taken as $q_1 = q_2 = 0.5$. The number m of alleles ranges from 2 to 6. The allele frequencies are given by: $P_{A_1} = P_{A_2} = 0.5$ when $m = 2$; $P_{A_1} = 0.4, P_{A_2} = P_{A_3} = 0.3$ when $m = 3$; $P_{A_1} = \cdots = P_{A_4} = 0.25$ when $m = 4$; $P_{A_1} = \cdots = P_{A_5} = 0.2$ when $m = 5$; and $P_{A_1} = P_{A_2} = 0.2, P_{A_3} = \cdots = P_{A_6} = 0.15$ when $m = 6$. Four test cases are considered: null, no major gene effect $a = d = 0$; additive, additive mode of inheritance $a = 1$, but no dominant effect $d = 0$; dominant, dominant mode of inheritance $a = d = 1$; recessive, recessive mode of inheritance $a = 1$ and $d = -0.5$. In each test case, linkage equilibrium is assumed between the QTL Q and the marker A; *i.e.*, $D_{A_iQ} = P(Q_1A_i) - q_1P_{A_i} = 0$.

$P_{A_1} = \cdots = P_{A_4} = 0.25$ when $m = 4$; $P_{A_1} = \cdots = P_{A_5} = 0.2$ when $m = 5$; and $P_{A_1} = P_{A_2} = 0.2, P_{A_3} = \cdots = P_{A_6} = 0.15$ when $m = 6$.

To calculate the type I error rates, 10,000 data sets are simulated for each test case. Each data set contains either 200 or 300 individuals. In each test case in Table 2, the data sets are generated under an assumption of linkage equilibrium between the QTL Q and the marker A; *i.e.*, $D_{A_iQ} = P(Q_1A_i) - q_1P_{A_i} = 0$. That is, there is no association between the QTL Q and marker A. Utilizing the data sets, we fit either model (8) or model (9), and then calculate the F-test $F_{m,ad}$ or $F_{m,a}$. Because the data sets are generated under the assumption of linkage equilibrium, an empirical test statistic that is larger than the cutting point of the related F-statistic at a 0.05 significance level is treated as a false positive. On the basis of the F-test of either $F_{m,ad}$ or $F_{m,a}$, type I error rates are calculated as the proportions of the 10,000 simulation data sets that give significant results at the 0.05 significance level.

For the test statistic $F_{m,a}$, the Table 2 results show that the type I error rates are around the 0.05 nominal significance level in all cases. Hence, the proposed model (9) is robust for data sets of a sample size $N = 200$. For test statistic $F_{m,ad}$, the type I error rates are around the 0.05 nominal significance level when $m \leq 5$ for data sets of sample size $N = 200$. For $m = 6$ and a sample size $N = 200$, the type I error rates of test $F_{m,ad}$ are too big for the dominant and recessive test cases (9.11 and 7.04%, respectively). This is partially due to the large degrees of freedom, $J_A - 1 = m(m + 1)/2 - 1 = 20$ of test $F_{m,ad}$ when $m = 6$; in addition, the high rate of type I error may be also caused by the mode of inheritance, *i.e.*, for the cases of dominant and recessive models. When the sample size increases to $N = 300$, the type I error rates of test $F_{m,ad}$ are around the 0.05 nominal significance level for $m = 6$. Model (8) is less robust than model (9).

In Table 3, two markers A and B are used in the analysis. The numbers m and n of alleles are equal to 2. The allele frequencies are given by $P_{A_1} = P_{A_2} = 0.5$ and $P_{B_1} = P_{B_2} = 0.5$. In each test case, linkage equilibrium is assumed between the QTL Q and the markers A and B; *i.e.*, $D_{A_iQ} = D_{B_iQ} = 0$. Denote $D_{A_1B_1} = P(A_1B_1) - P_{A_1}P_{B_1}$, which is the measure of LD between A and B. Here $P(A_1B_1)$ is the frequency of haplotype A_1B_1. Let

$$D_{AQB} = P(A_1Q_1B_1) - P_{A_1}D_{B_1Q} - q_1D_{A_1B_1} - P_{B_1}D_{A_1Q} - P_{A_1}q_1P_{B_1} \tag{20}$$

be the measure of the third-order LD (Thomson and Baur 1984). Here $P(A_1Q_1B_1)$ is the frequency of haplotype $A_1Q_1B_1$. Between marker A and marker B, two situations are considered: (1) linkage equilibrium, *i.e.*, $D_{A_1B_1} = 0$, and (2) linkage disequilibrium, *i.e.*, $D_{A_1B_1} = 0.08$. No linkage disequilibrium of third order is assumed among markers A and B and the QTL Q; that is, $D_{AQB} = 0$. Again, 10,000 data sets are simulated for each test case, and each data set contains 200 individuals. The simulation is done as follows. First, the haplotype frequencies are calculated on the basis of allele frequencies and LD coefficients by relation (20) (Thomson and Baur 1984). Then data sets are simulated using the haplotype frequencies. On the basis of the F-test of either $F_{AB,ad}$ or $F_{AB,a}$ or the HTR method, type I error rates are calculated as the proportions of the 10,000 simulation data sets that give significant results at the 0.05 significance level. The Table 3 results show that the type I error rates are around the 0.05 nominal

TABLE 3

Type I error rates (percentage) of test statistics $F_{AB,\mathrm{ad}}$, $F_{AB,\mathrm{a}}$, and F_{HTR} of the haplotype trend regression (HTR) method at a 0.05 significance level when two markers *A* and *B* are used in the analysis

LD measure $D_{A_1B_1} = P(A_1B_1) - P_{A_1}P_{B_1}$	Sample size	Test case	Error Rates		
			$F_{AB,\mathrm{ad}}$	$F_{AB,a}$	F_{HTR}
0	$N = 200$	Null	4.90	5.22	5.39
		Additive	5.09	4.75	4.77
		Dominant	4.62	4.87	4.79
		Recessive	5.36	5.12	4.81
0.08	$N = 200$	Null	5.09	5.23	5.55
		Additive	4.92	4.74	4.71
		Dominant	4.63	4.84	4.71
		Recessive	5.04	5.02	4.94

The total variance is fixed as $\sigma^2 = 1.0$ and the trait allele frequency is taken as $q_1 = q_2 = 0.5$. The numbers m and n of alleles = 2. The allele frequencies are given by $P_{A_1} = P_{A_2} = 0.5$ and $P_{B_1} = P_{B_2} = 0.5$. Four test cases are considered: null, no major gene effect $a = d = 0$; additive, additive mode of inheritance $a = 1$, but no dominant effect $d = 0$; dominant, dominant mode of inheritance $a = d = 1$; recessive, recessive mode of inheritance $a = 1$ and $d = -0.5$. In each test case, linkage equilibrium is assumed between the QTL *Q* and the markers *A* and *B*; *i.e.*, $D_{A_iQ} = D_{B_iQ} = 0$. No linkage disequilibrium of third order is assumed among markers *A* and *B* and the QTL *Q*; that is, $D_{AQB} = 0$.

significance level in all cases. Hence, the proposed models (13) and (14) and the HTR method are robust for data sets of a sample size $N = 200$.

Table 4 shows type I error rates (percentages) of test statistics $F_{ABC,\mathrm{ad}}$, $F_{ABC,\mathrm{a}}$, and F_{HTR} at a 0.05 significance level when three diallelic markers *A*, *B*, and *C* are used in the analysis. The measures D_{ABC}, D_{AQC}, and D_{BQC} of the third-order LD are defined as that of D_{AQB}; the measure of the fourth order is defined accordingly (Bennett 1954). Such as relation (20), the haplotype frequencies at the three markers *A*, *B*, and *C* and at QTL *Q* are calculated on the basis of allele frequencies and LD coefficients by Weir's (1996, p. 119) relation (3.14). Then data sets are simulated using the haplotype frequencies. Since this article is about population data, one individual may have two copies of haplotypes. Each haplotype is sampled according to the haplotype frequencies. From the Table 4 results, we can see that the proposed models and the HTR method give correct type I errors for data sets of a sample size $N = 200$.

TABLE 4

Type I error rates (percentage) of test statistics $F_{ABC,\mathrm{ad}}$, $F_{ABC,\mathrm{a}}$, and F_{HTR} of the haplotype trend regression (HTR) method at a 0.05 significance level when three diallelic markers *A*, *B*, and *C* are used in the analysis

LD measure $D_{A_1B_1} = D_{A_1C_1} = P_{B_1C_1}$	Sample size	Test case	Error rates		
			$F_{ABC,\mathrm{ad}}$	$F_{ABC,\mathrm{a}}$	F_{HTR}
0.08	$N = 200$	Null	5.2	5.35	5.43
		Additive	4.98	4.85	4.74
		Dominant	4.31	4.68	4.62
		Recessive	5.29	5.3	5.27
0.06	$N = 200$	Null	5.24	5.41	5.39
		Additive	5.15	4.89	4.71
		Dominant	4.61	5.0	5.03
		Recessive	5.09	4.94	5.08

The total variance is fixed as $\sigma^2 = 1.0$ and the trait allele frequency is taken as $q_1 = q_2 = 0.5$. The allele frequencies are given by $P_{A_1} = P_{A_2} = 0.5$, $P_{B_1} = P_{B_2} = 0.5$, and $P_{C_1} = P_{C_2} = 0.5$. Four test cases are considered: null, no major gene effect $a = d = 0$; additive, additive mode of inheritance $a = 1$, but no dominant effect $d = 0$; dominant, dominant mode of inheritance $a = d = 1$; recessive, recessive mode of inheritance $a = 1$ and $d = -0.5$. In each test case, linkage equilibrium is assumed between the QTL *Q* and the markers *A*, *B*, and *C*; *i.e.*, $D_{A_iQ} = D_{B_iQ} = D_{C_iQ} = 0$. Moreover, neither third- nor fourth-order linkage disequilibrium is assumed; *i.e.*, $D_{ABC} = D_{AQB} = D_{AQC} = D_{BQC} = D_{ABCQ} = 0$.

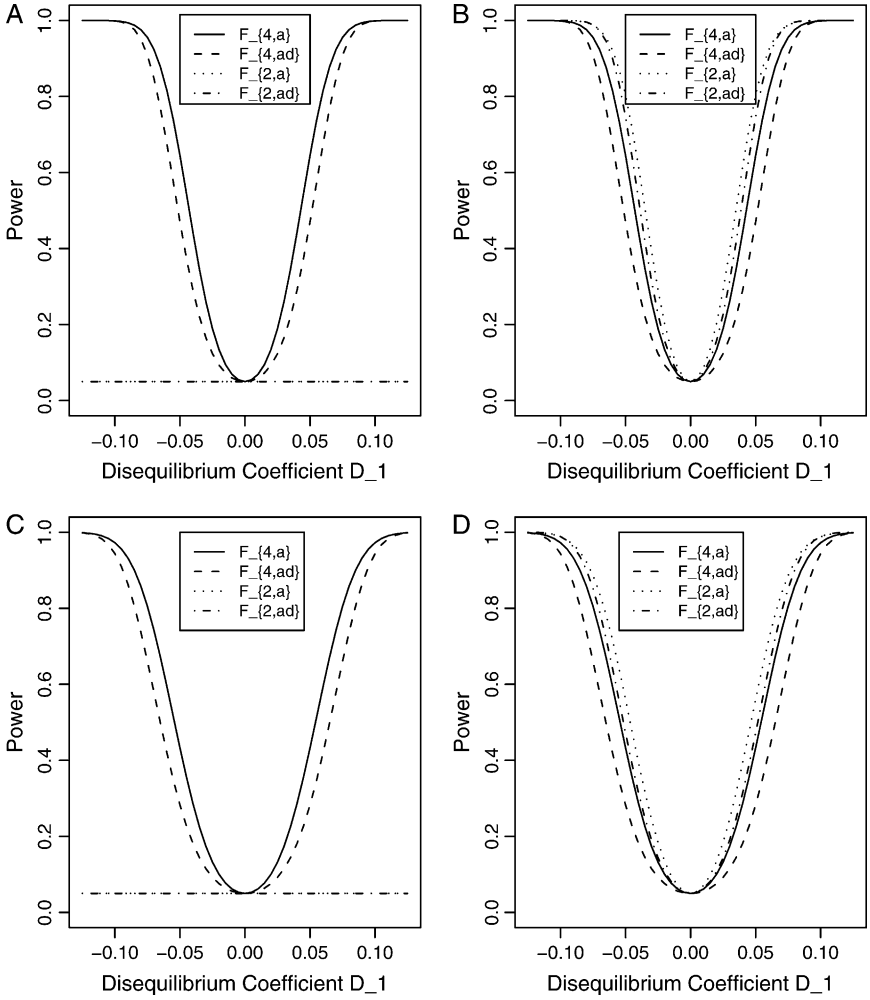

FIGURE 1.—Power curves of the test statistics $F_{4,ad}$, $F_{4,a}$, $F_{2,ad}$, and $F_{2,a}$ against the disequilibrium coefficient $D_1 = D_{A_1Q}$ for a dominant mode of inheritance $a = d = 1.0$ at a 0.05 significance level. $F_{4,ad}$ and $F_{4,a}$ are calculated when marker A has four equal frequency alleles; *i.e.*, $P_{A_1} = \cdots = P_{A_4} = 0.25$. The measures of LD are (A and B) $D_{A_2Q} = D_{A_4Q} = -D_{A_1Q}, D_{A_3Q} = D_{A_1Q}$ and (C and D) $D_{A_2Q} = -D_{A_1Q}, D_{A_3Q} = -D_{A_4Q} = D_{A_1Q}/2$. $F_{2,ad}$ and $F_{2,a}$ are calculated by collapsing two of the four alleles: (A and C) alleles A_1 and A_2 are collapsed as one allele, and alleles A_3 and A_4 are collapsed to be the other; (B and D) alleles A_1 and A_3 are collapsed to be one allele, and alleles A_2 and A_4 are collapsed to be the other. The other parameters are $q_1 = 0.50$, $h^2 = 0.25$, $N = 200$.

Power calculation and comparison: Let $h^2 = \sigma^2_{ga}/\sigma^2$ be the heritability. Figure 1 shows power curves of the test statistics $F_{4,a}$, $F_{4,ad}$, $F_{2,a}$, and $F_{2,ad}$ against the disequilibrium coefficient D_{A_1Q} for a dominant mode of inheritance $a = d = 1.0$ at a 0.05 significance level based on the approximations of noncentrality parameters $\lambda_{m,a}$ and $\lambda_{m,ad}$. $F_{4,a}$ and $F_{4,ad}$ are calculated when A has four equal frequency alleles; *i.e.*, $P_{A_1} = \cdots = P_{A_4} = 0.25$. In addition, the measures of LD are given as follows: Figure 1, A and B, $D_{A_2Q} = D_{A_4Q} = -D_{A_1Q}, D_{A_3Q} = D_{A_1Q}$, and Figure 1, C and D, $D_{A_2Q} = -D_{A_1Q}, D_{A_3Q} = -D_{A_4Q} = D_{A_1Q}/2$. $F_{2,a}$ and $F_{2,ad}$ are calculated by collapsing the four alleles to be two alleles: in Figure 1, A and C, alleles A_1 and A_2 are collapsed as one allele, and alleles A_3 and A_4 are collapsed to be the other; in Figure 1, B and D, alleles A_1 and A_3 are collapsed to be one allele, and alleles A_2 and A_4 are collapsed to be the other. For $F_{2,a}$ and $F_{2,ad}$, a simple calculation can show that the measures of LD in Figure 1A are 0, 0; the measures of LD in Figure 1B are $2D_{A_1Q}, -2D_{A_1Q}$; the measures of LD in Figure 1C are 0, 0; and the measures of LD in Figure 1D are $3D_{A_1Q}/2, -3D_{A_1Q}/2$. Hence, the QTL Q is in linkage equilibrium with the marker after collapsing the alleles in Figure 1, A and C. The other parameters are $q_1 = 0.50$, $h^2 = 0.25$, $N = 200$.

From Figure 1, we may see the following:

1. $F_{4,ad}$ is slightly less powerful than $F_{4,a}$, and $F_{2,ad}$ is slightly less powerful than $F_{2,a}$. This is because that test statistic $F_{m,ad}$ has larger degrees of freedom than those of $F_{m,a}$. Note that the noncentrality parameter approximation $\lambda_{m,ad}$ of $F_{m,ad}$ is given by Equation 10. The contribution of the dominance effect is $N\sigma^2_{gd}R^4_{AQ}/\sigma^2$, which depends on both dominance effect d and the magnitude of factor R^4_{AQ}; and it can be significant when both of them are large enough. Hence, including a dominance component in the model can improve the power of QTL detection only when the magnitude of $\sigma^2_{gd}R^4_{AQ}$ is large enough to compensate for the extra degrees of freedom. Note that the quantity $\sigma^2_{gd}R^4_{AQ}$ is the product of the dominance variance σ^2_{gd} and of the measure R^4_{AQ} of LD. The magnitude of $\sigma^2_{gd}R^4_{AQ}$ is the result of the dominance variance σ^2_{gd} reduced by a factor R^4_{AQ}. Even when σ^2_{gd} is large, $\sigma^2_{gd}R^4_{AQ}$ can be small when LD coefficients are not big; *i.e.*, R^4_{AQ} is small.

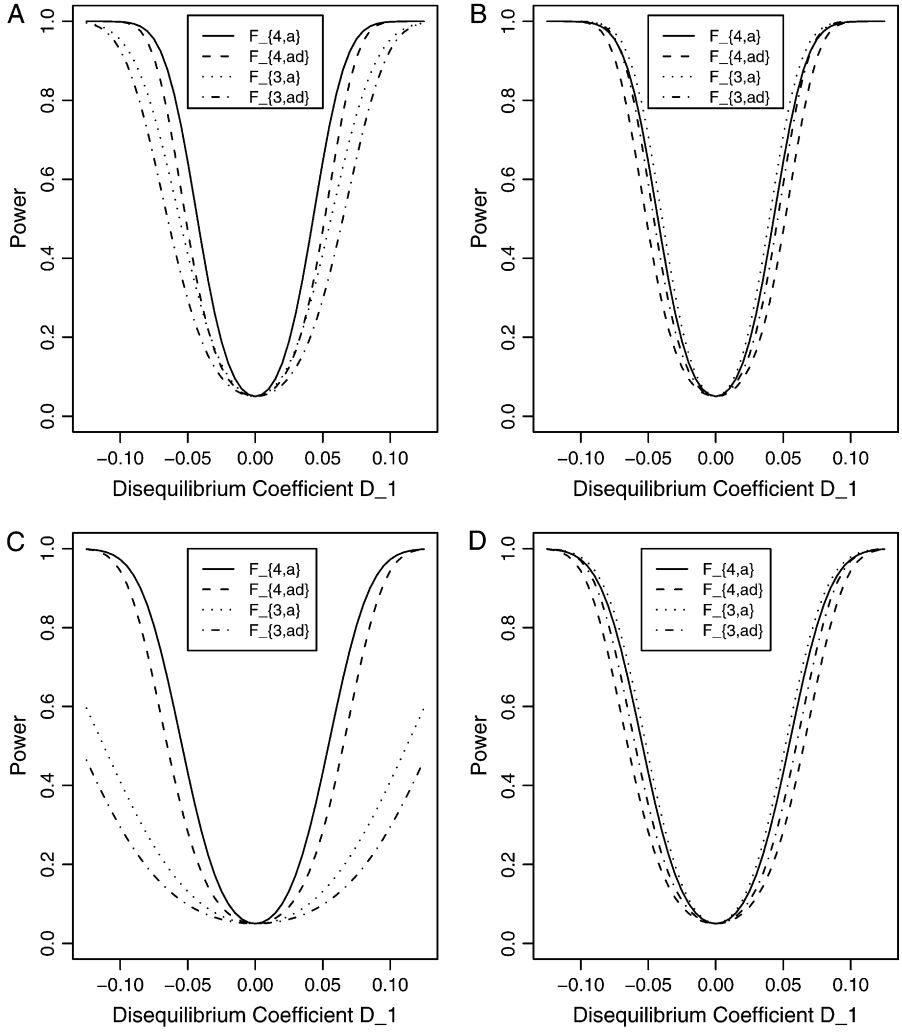

FIGURE 2.—Power curves of the test statistics $F_{4,a}$, $F_{4,ad}$, $F_{3,a}$, and $F_{3,ad}$, against the disequilibrium coefficient $D_1 = D_{A_1Q}$ for a dominant mode of inheritance $a = d = 1.0$ at a 0.05 significance level. $F_{4,ad}$ and $F_{4,a}$ are calculated when marker A has four equal frequency alleles; *i.e.*, $P_{A_1} = \cdots = P_{A_4} = 0.25$. The measures of LD are the same as those in Figure 1. $F_{3,ad}$ and $F_{3,a}$ are calculated by collapsing two of the four alleles: (A and C) alleles A_1 and A_2 are collapsed as a new one; (B and D) alleles A_1 and A_3 are collapsed to be a new one. The other parameters are $q_1 = 0.50$, $h^2 = 0.25$, $N = 200$.

2. When the measures of LD are high, the power of the test statistics is high. On the other hand, the power is minimal if all measures of LD are close to 0.
3. The dependence of power on measures of LD can also be observed by comparing Figure 1A with Figure 1C, 1B with 1D. The power of $F_{4,ad}$ and $F_{4,a}$ in Figure 1A is higher than that of $F_{4,ad}$ and $F_{4,a}$ in Figure 1C, respectively; the power of each test statistic in Figure 1B is higher than that of the same test statistic in Figure 1D. This is because the measures of LD in Figure 1A are equal to or higher than those in Figure 1C, and the measures of LD in Figure 1B are equal to or higher than those in Figure 1D.
4. In Figure 1B and Figure 1D, the power of $F_{4,ad}$ is slightly lower than that of $F_{2,ad}$; the power of $F_{4,a}$ is slightly lower than that of $F_{2,a}$.
5. In Figure 1A and Figure 1C, the power of $F_{2,ad}$ and $F_{2,a}$ is minimal. This is because measures of LD are 0 after collapsing the alleles in these two graphs.

Figure 2 shows power curves of the test statistics $F_{4,a}$, $F_{4,ad}$, $F_{3,a}$, and $F_{3,ad}$ against the disequilibrium coefficient D_{A_1Q} for a dominant mode of inheritance $a = d = 1.0$ at a 0.05 significance level. $F_{4,a}$ and $F_{4,ad}$ are calculated as those in Figure 1. $F_{3,a}$ and $F_{3,ad}$ are calculated by collapsing two of the four alleles to be a new alelle: in Figure 2, A and C, alleles A_1 and A_2 are collapsed as a new one; in Figure 2, B and D, alleles A_1 and A_3 are collapsed to be a new one. For $F_{3,a}$ and $F_{3,ad}$, a simple calculation can show that the measures of LD in Figure 2A are $0, D_{A_1Q}, -D_{A_1Q}$; the measures of LD in Figure 2B are $2D_{A_1Q}, -D_{A_1Q}, -D_{A_1Q}$; the measures of LD in Figure 2C are $0, D_{A_1Q}/2, -D_{A_1Q}/2$; and the measures of LD in Figure 2D are $3D_{A_1Q}/2, -D_{A_1Q}, -D_{A_1Q}/2$. Among the features shown in Figure 1, it can be seen that in Figure 2, A and C, the power of $F_{4,ad}$ is higher than that of $F_{3,ad}$, and the power of $F_{4,a}$ is higher than that of $F_{3,a}$. In Figure 2, B and D, the power of $F_{4,ad}$ is slightly lower than that of $F_{3,ad}$, and the power of $F_{4,a}$ is slightly lower than that of $F_{3,a}$. Hence, the way to collapse the alleles has impact on power.

From Figures 1 and 2, we may see that the power of $F_{4,a}$ and $F_{4,ad}$ is relatively stable although it may be slightly lower than that of $F_{3,a}$, $F_{3,ad}$, $F_{2,a}$, and $F_{2,ad}$ in

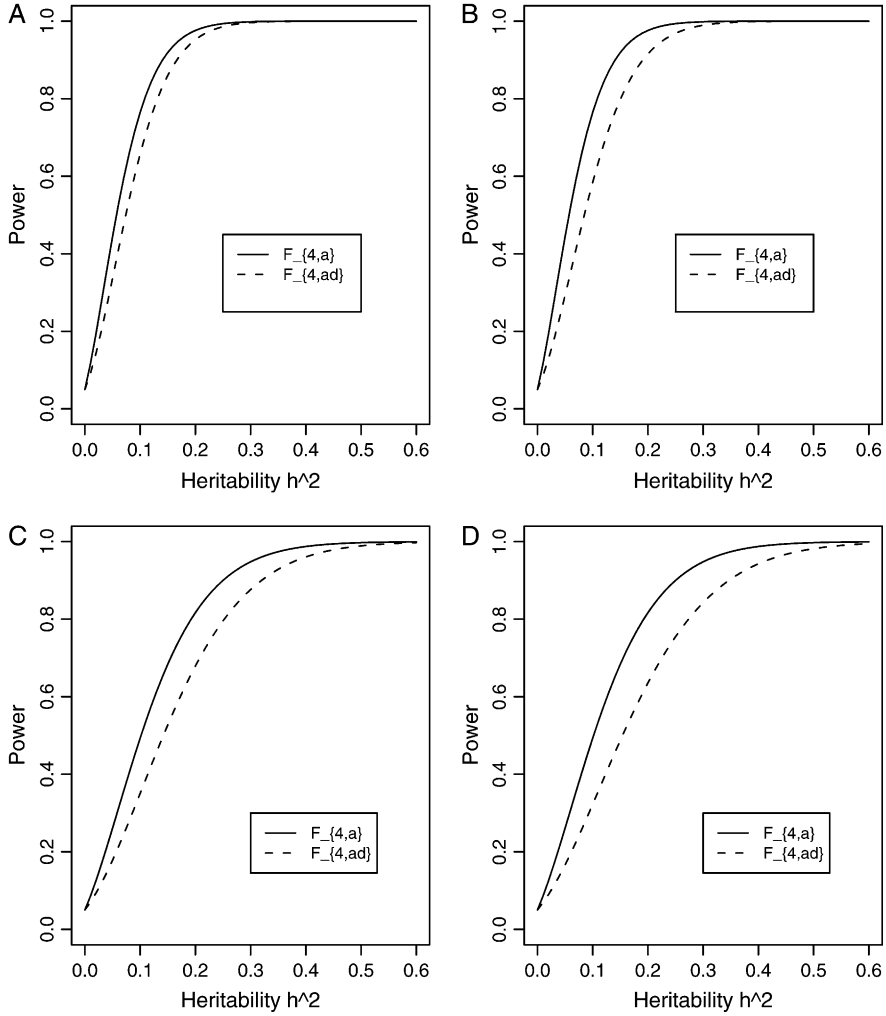

FIGURE 3.—Power curves of the test statistics $F_{4,a}$ and $F_{4,ad}$ against the heritability h^2 at a 0.05 significance level. (A and C) The curves are plotted for a dominant mode of inheritance $a = d = 1.0$; (B and D) the curves are plotted for a recessive mode of inheritance $a = 1.0$, $d = -0.5$. $F_{4,a}$ and $F_{4,ad}$ are calculated when marker A has four equal frequency alleles; *i.e.*, $P_{A_1} = \cdots = P_{A_4} = 0.25$. The measures of LD are given as follows: (A and B) $D_{A_1Q} = -D_{A_2Q} = D_{A_3Q} = -D_{A_4Q} = 0.08$; (C and D) $D_{A_1Q} = -D_{A_2Q} = D_{A_3Q} = -D_{A_4Q} = 0.06$. The other parameters are $q_1 = 0.50$ and $N = 250$.

certain circumstances. However, the power of $F_{3,a}$, $F_{3,ad}$, $F_{2,a}$, and $F_{2,ad}$ depends heavily on the way to collapse the alleles. This shows the advantage of using multiallelic markers in an association study of QTL detection. For multiallelic marker data, the proposed test statistics $F_{m,a}$ and $F_{m,ad}$ can be directly used to test if there is association between the marker and the QTL. As shown in Figures 1 and 2, the test statistic $F_{m,a}$ is usually more powerful than $F_{m,ad}$ due to the increase of degrees of freedom of test statistic $F_{m,ad}$.

Figure 3 shows power curves of the test statistics $F_{4,a}$ and $F_{4,ad}$ against the heritability h^2 at a 0.05 significance level for a dominant mode of inheritance $a = d = 1.0$ and for a recessive mode of inheritance $a = 1.0$, $d = -0.5$, respectively. As with Figures 1 and 2, Figure 3 is based on noncentrality parameter approximations (10) and (11). In Figure 3, A and B, the power can be high as the heritability $h^2 > 0.1$; in these two graphs, the measures of LD are given by $D_{A_1Q} = -D_{A_2Q} = D_{A_3Q} = -D_{A_4Q} = 0.08$. In Figure 3, C and D, the power can be high as the heritability $h^2 > 0.15$; in these two graphs, the measures of LD are given by $D_{A_1Q} = -D_{A_2Q} = D_{A_3Q} = -D_{A_4Q} = 0.06$. Figure 4 shows power curves of the test statistics $F_{4,a}$ and $F_{4,ad}$ against the trait allele frequency q_1 or marker allele frequency P_{A_1} at a 0.05 significance level. It can be seen that the power depends on both the measures of linkage disequilibrium and the trait allele frequency q_1 or marker allele frequency P_{A_1}.

Comparison with the haplotype trend regression method: Assume that the two diallelic markers A and B are used in the analysis. Figures 5 and 6 show power curves of the test statistics $F_{AB,a}$, F_{HTR}, and $F_{AB,ad}$ against the heritability h^2 at a 0.05 significance level. The related parameters are given in the figure legends. The power curves of the test statistics $F_{AB,a}$, F_{HTR}, and $F_{AB,ad}$ are calculated on the basis of approximations of noncentrality parameters λ_{ABa}, λ_{HTR}, and λ_{ABad}.

In Figure 5, no third-order linkage disequilibrium is assumed; *i.e.*, $D_{AQB} = 0$. In Figure 6, A and B, weak third-order linkage disequilibrium is assumed; *i.e.*, $D_{AQB} = 0.025$. It can be seen that the genotype effect model can be less powerful than the HTR method, and the HTR method can be less powerful than the additive effect model in the case of no or weak third-order linkage disequilibrium among the two markers and the QTL (Figure 5 and Figure 6, A and B). In Figure 6, C and D,

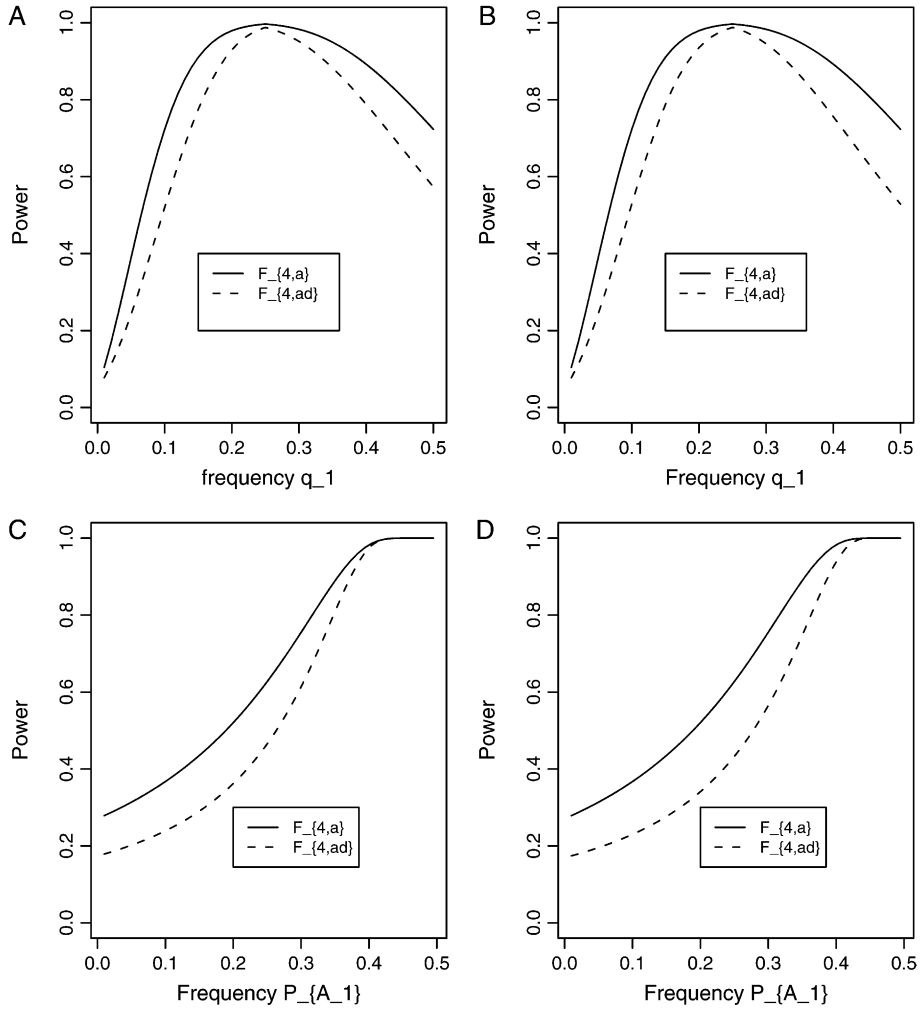

FIGURE 4.—Power curves of the test statistics $F_{4,a}$ and $F_{4,ad}$ against the trait allele frequency q_1 or allele frequency P_{A_1} at a 0.05 significance level. (A and C) The curves are plotted for a dominant mode of inheritance $a = d = 1.0$; (B and D) the curves are plotted for a recessive mode of inheritance $a = 1.0$, $d = -0.5$. (A and B) The parameters are given by $P_{A_1} = \cdots = P_{A_4} = 0.25$, $q_2 = 1 - q_1, D_{A_1Q} = (\min(q_1, P_{A_1}) - q_1 P_{A_1})/2 = -D_{A_2Q} = D_{A_3Q} = -D_{A_4Q}$; (C and D) the parameters are given by $P_{A_2} = 0.5 - P_{A_1}$, $P_{A_3} = P_{A_4} = 0.25$, $q_1 = 0.5, D_{A_1Q} = (\min(q_1, P_{A_1}) - q_1 P_{A_1})/2 = -D_{A_2Q}, D_{A_3Q} = -D_{A_4Q} = 0.05$. The other parameters are $h^2 = 0.15$ and $N = 250$.

strong third-order linkage disequilibrium is assumed; *i.e.*, $D_{AQB} = 0.065$. In the case that strong third-order linkage disequilibrium exists, the HTR method can be more powerful (Figure 6, C and D).

Note the following fact: in Figure 6, A and B, the maximum of D_{AQB} is 0.025; in Figure 6, C and D, the maximum of D_{AQB} is 0.065 (otherwise, some of the haplotype would have negative frequencies). Thus, the simulated power curves of the haplotype trend regression method in Figures 5 and 6 represent the two extreme situations: (1) no third-order linkage disequilibrium (Figure 5) and (2) strongest third-order linkage disequilibrium (Figure 6). In practice, the third-order linkage disequilibrium would exist in a more moderate way that is between the two extremes; and the power of the haplotype trend regression method should be between those of the two extremes. Note that the proposed genotype effect model and additive effect model utilize only the second-order linkage disequilibrium or pairwise linkage disequilibrium. Hence, the powers of $F_{AB,a}$ and $F_{AB,ad}$ are the same for Figures 5 and 6.

Figure 7 shows power curves of the test statistics $F_{ABC,a}$ and $F_{ABC,ad}$ and F_{HTR} against the heritability h^2 at a 0.05 significance level, when three diallelic markers *A*, *B*, and *C* are used in the analysis. The related parameters are given in the figure legend. From Figure 7, it can be seen that the power of F_{HTR} is the lowest. This is due to the large number of degrees of freedom of F_{HTR}, which is $F(7, N - 8)$, $N = 200$. In contrast, $F_{ABC,a}$ is $F(3, N - 4)$, $N = 200$; and $F_{ABC,a}$ is $F(6, N - 7)$, $N = 200$. The low power of F_{HTR} is most likely due to the biallelic QTL situation that we consider. In the situation of multiple QTL haplotypes and strong LD between QTL and marker haplotypes, the haplotype-based methods are expected to have good power.

Comparison based on ACE haplotype frequencies: To work on more realistic scenarios, we take the haplotype information of ACE genes as an example. Ten diallelic polymorphisms in the ACE gene spanning 26 kb were genotyped (KEAVNEY *et al.* 1998). The order of the 10 polymorphisms is T-5991C, A-5466C, T-3892C, A-240T, T-93C, T1237C, G2215A, I/D, G2350A, and 4656(CT)3/2. Table 5 lists 10 haplotypes, where the first 7 are the most frequent haplotypes (http://www.well.ox.ac.uk/~mfarrall/oxhap_freq.html). For the 10 haplotypes, allele I at marker I/D is always present with allele A at marker G2350A, and allele D at marker I/D is always present with allele G at marker G2350A. Hence,

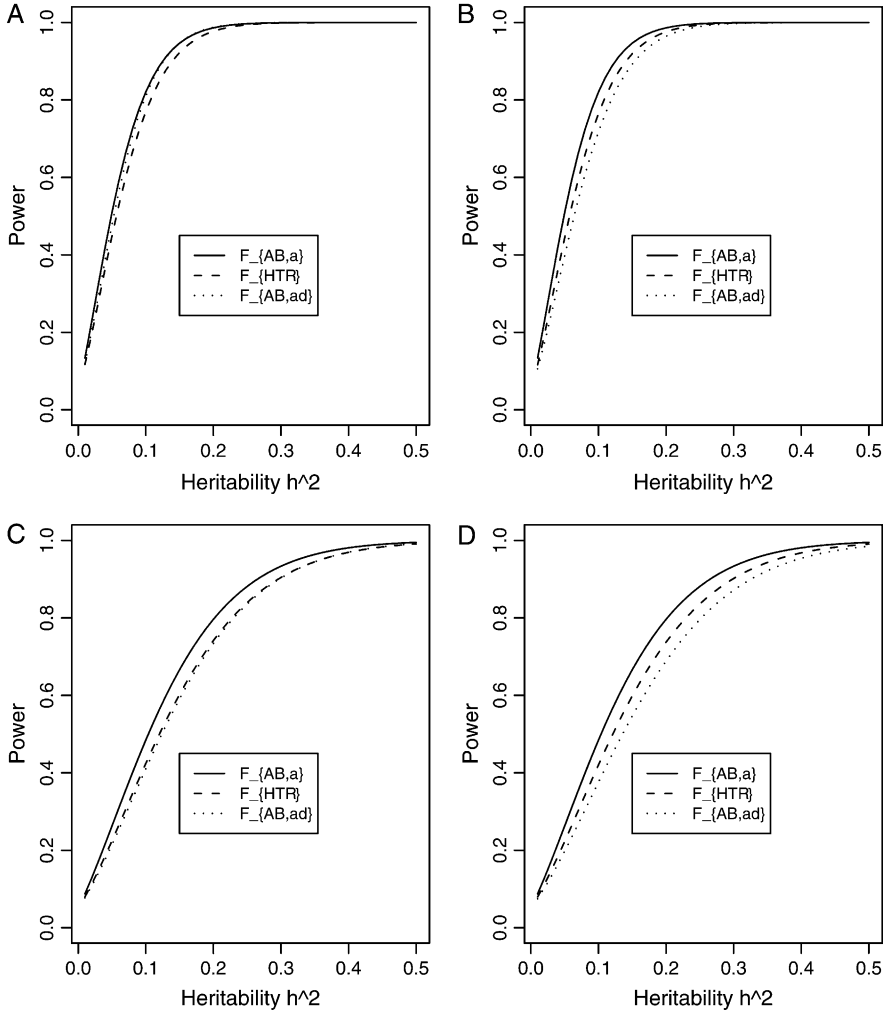

FIGURE 5.—Power curves of the test statistics $F_{AB,\mathrm{a}}$ and $F_{AB,\mathrm{ad}}$ and F_{HTR} of the haplotype trend regression method against the heritability h^2 at a 0.05 significance level, when two diallelic markers A and B are used in the analysis. (A and C) The curves are plotted for a dominant mode of inheritance $a = d = 1.0$; (B and D) the curves are plotted for an additive mode of inheritance $a = 1.0$, $d = 0$. (A and B) The parameters are given by $D_{A_1Q} = D_{B_1Q} = 0.15, D_{A_1B_1} = 0.10, D_{A_1QB_1} = 0$; (C and D) the parameters are given by $D_{A_1Q} = D_{B_1Q} = 0.10, D_{A_1B_1} = 0.08, D_{A_1QB_1} = 0$. The other parameters are $P_{A_1} = P_{A_2} = P_{B_1} = P_{B_2} = q_1 = q_2 = 0.5$ and $N = 200$.

the two markers can be treated as one. Similarly, markers T-5991C and A-5466C can be treated as one; and markers A-240T and T-93C can be treated as one. Therefore, the 10 haplotypes can be considered as containing seven markers.

In ABECASIS *et al.* (2000a,b) and FAN *et al.* (2005), it is found that that markers I/D and G2350A show strongest association with the circulating ACE level. Thus, markers I/D and G2350A are treated as a putative trait locus Q. A quantitative trait of the putative locus Q is simulated for each graph in Figure 8, A–D. The empirical power curves of the test statistics F_{HTR}, F_{a}, and F_{ad} are plotted against the heritability h^2 at a 0.05 significance level in Figure 8. Here F_{a} is the test statistic based on the additive effect model, and F_{ad} is the test statistic based on the genotype effect model. The empirical power curves SF_{HTR}, SF_{a}, and SF_{ad} in Figure 8 are calculated as follows. First, the interval (0.01, 0.25) of the heritability h^2 is divided into 24 subintervals. Correspondingly, the 24 subintervals lead to 25 end points. For each end point, there is a set of parameters for the power curve. Using the set of parameters, 2500 data sets are simulated for each end point. For each data set, empirical statistics of F_{HTR}, F_{a}, and F_{ad} are calculated. The simulated power is the proportion of the 2500 simulated data sets for which the empirical statistic is larger than the cutting point of the corresponding F-distribution at a 0.05 significance level.

In Figure 8, A and C, the curves are plotted for a dominant mode of inheritance $a = d = 1.0$; in Figure 8, B and D, the curves are plotted for an additive mode of inheritance $a = 1.0$, $d = 0$. In Figure 8, A and B, all 10 haplotypes are used in the simulations; in Figure 8, C and D, only the first 7 most frequent haplotypes are used. From Figure 8, A–D, it can be seen that the proposed additive effect model has similar power to that of the HTR method. In Figure 8, A and C, when the dominance effects are present, the genotype effect model has similar power to those of the additive effect model and the HTR method. In Figure 8, B and D, the genotype effect model is less powerful because of the absence of the dominance effect. Hence, the genotype effect model can be useful only if the dominance effect can compensate for the extra degrees of freedom.

Simulation study: To evaluate the accuracy of the noncentrality parameter approximations, we performed

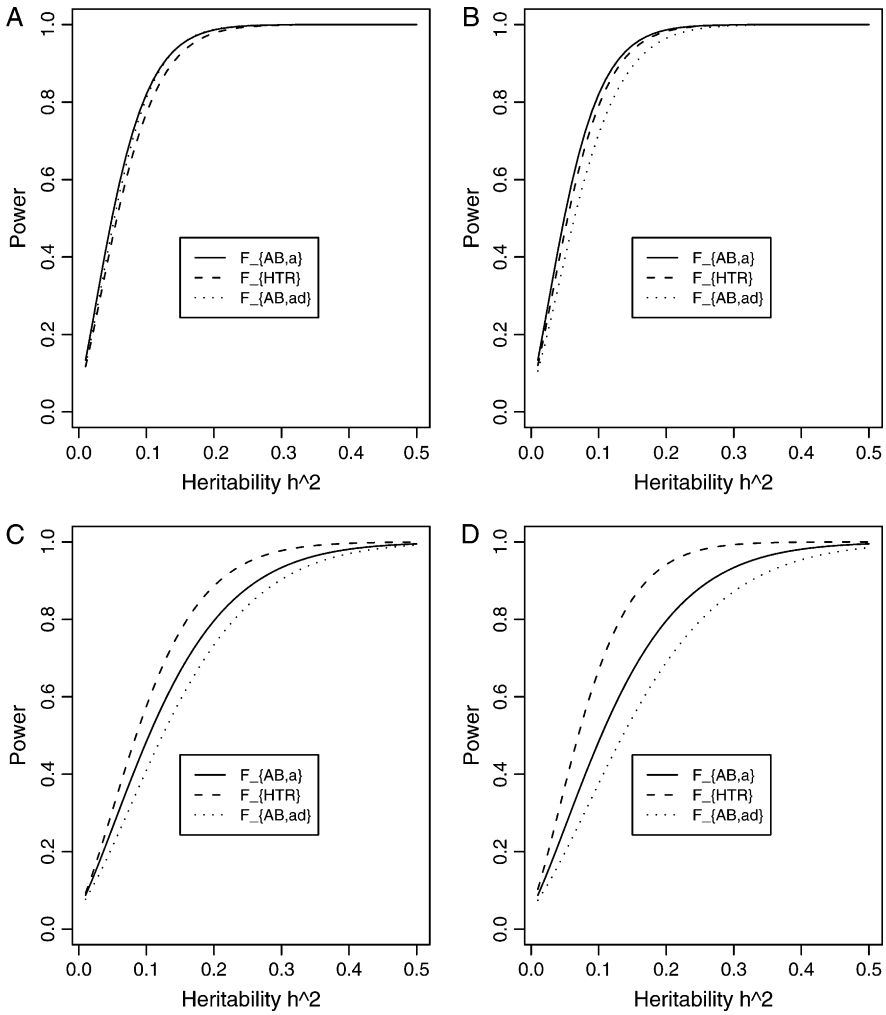

FIGURE 6.—Power curves of the test statistics $F_{AB,\mathrm{a}}$ and $F_{AB,\mathrm{ad}}$ and F_{HTR} of the haplotype trend regression method against the heritability h^2 at a 0.05 significance level, when two diallelic markers A and B are used in analysis. All parameters are the same as those in Figure 5 except that (A and B) $D_{A_1QB_1} = 0.025$ and (C and D) $D_{A_1QB_1} = 0.065$.

simulations for the power curves in Figures 1, 2, 5, 6, and 7. The results are presented as supplemental information (http://www.genetics.org/supplemental/). It can be seen that the approximations are excellent.

DISCUSSION

In this article, two models, the genotype effect model and the additive effect model, are proposed for high-resolution association mapping of QTL on the basis of population data. The two models extend our previous research, which is based on multiple diallelic markers (FAN and XIONG 2002, 2003; JUNG *et al.* 2005). The genotype effect model is closely linked to the measured genotype approach (BOERWINKLE *et al.* 1986). The very popular genetics software such as Mendel 5.0 is already capable of performing association mapping of QTL by the additive effect model (CANTOR *et al.* 2005; LANGE *et al.* 2005). Surprisingly, there is no research to theoretically show why these two models are valid methods in association mapping of QTL under normal distribution. There are no existing analytical formulas to evaluate the power of the related test statistics. This article shows that the model coefficients are functions of measures of LD; and thus related F-test statistics can be constructed for association study of QTL. In the presence of both additive and dominance effects of the QTL, either the $F_{m,\mathrm{ad}}$ (or $F_{AB,\mathrm{ad}}$) statistic or the $F_{m,\mathrm{a}}$ (or $F_{AB,\mathrm{a}}$) statistic can be used. Since the $F_{m,\mathrm{ad}}$ (or $F_{AB,\mathrm{ad}}$) test statistic has bigger degrees of freedom than those of $F_{m,\mathrm{a}}$ (or $F_{AB,\mathrm{a}}$), $F_{m,\mathrm{a}}$ (or $F_{AB,\mathrm{a}}$) can be more powerful. If the extra degrees of freedom of the $F_{m,ad}$ test can be compensated by magnitude $\sigma^2_{\mathrm{gd}} R^4_{AQ}$, it can be more powerful than $F_{m,\mathrm{a}}$.

The formulas of noncentrality parameter approximations (10) and (11) clearly indicate the dependence of the power on the quantity R^2_{AQ} for genetic data. That is, the noncentrality parameter of test statistics of the null hypothesis of no genetic effects is reduced by a factor of R^2_{AQ} for additive variance and by a factor of R^4_{AQ} for dominance variance. If only one diallelic marker A is used in the analysis, both our previous research and the work of colleagues have derived similar formulas to support this argument (SHAM *et al.* 2000; FAN and XIONG 2002, 2003; FAN and JUNG 2003; FAN *et al.* 2005; JUNG *et al.* 2005). This is a good example in the debate on appropriate measures of LD for markers or multiallelic

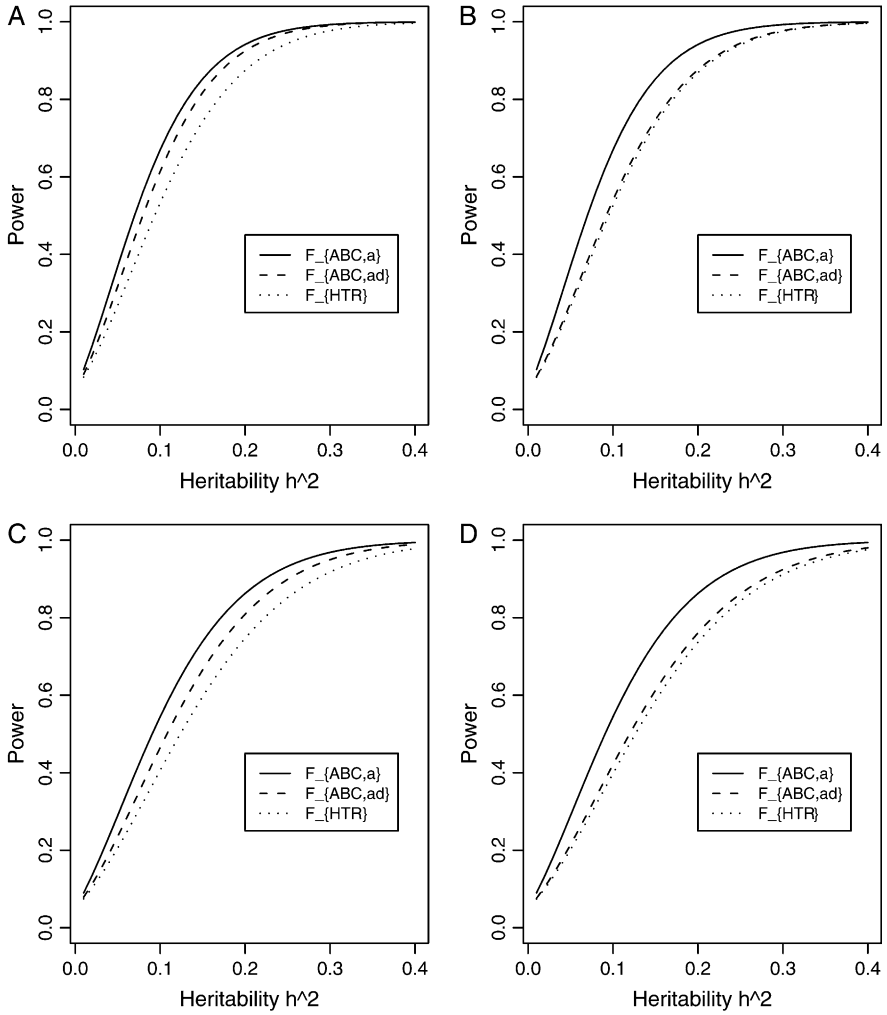

FIGURE 7.—Power curves of the test statistics $F_{ABC,\mathrm{a}}$ and $F_{ABC,\mathrm{ad}}$ and F_{HTR} of the haplotype trend regression method against the heritability h^2 at a 0.05 significance level, when three diallelic markers A, B, and C are used in the analysis. (A and C) The curves are plotted for a dominant mode of inheritance $a = d = 1.0$; (B and D) the curves are plotted for an additive mode of inheritance $a = 1.0$, $d = 0$. (A and B) The parameters are given by $D_{AQ} = D_{BQ} = D_{CQ} = D_{DQ} = 0.12$, $D_{AB} = D_{AC} = D_{BC} = 0.08$; (C and D) the parameters are given by $D_{AQ} = D_{BQ} = D_{CQ} = D_{DQ} = 0.10$, $D_{AB} = D_{AC} = D_{BC} = 0.06$. Neither third- nor fourth-order linkage disequilibrium is assumed among markers and the QTL. The other parameters are $P_{A_1} = P_{B_1} = P_{C_1} = q_1 = 0.5$ and $N = 200$.

markers (HEDRICK 1987; DEVLIN and RISCH 1995; PRITCHARD and PRZEWORSKI 2001; WEISS and CLARK 2002). For multiallelic markers or haplotypes, a satisfactory measure of LD has not been derived, as mentioned regarding p306 in ARDLIE *et al.* (2002). For two diallelic loci A and Q, ARDLIE *et al.* (2002) favor using $R^2_{AQ} = D^2_{A_1Q}/(P_{A_1}P_{A_2}q_1q_2)$, which is the correlation of alleles at the two loci. For multiallelic marker data, this article extends previous research by providing the definition of R^2_{AQ} and deriving Equations 10 and 11. HAYES *et al.* (2003) introduced a multilocus approach for estimating LD and past effective size and used chromosome segment homozygosity (CSH), which was introduced in SVED (1971). The dependence of the noncentrality parameter on the quantity R^2_{AQ} has been indicated by our study and also by SHAM *et al.* (2000).

In FULKER *et al.* (1999), ABECASIS *et al.* (2000a,b, 2001), and SHAM *et al.* (2000), an association between-family and association within-family ("AbAw") approach is proposed to decompose the genetic association into effects of between pairs and within pairs on the basis of variance component models. The AbAw approach is based on any single diallelic marker. Instead of using a single diallelic marker, we have proposed variance component models using multiple diallelic markers. In our models, the association is decomposed into additive and dominance components (FAN and XIONG 2002, 2003; FAN and JUNG 2003; FAN *et al.* 2005; JUNG *et al.* 2005). In FAN and JUNG (2003), FAN *et al.* (2005), and JUNG *et al.* (2005), we compare our method with the AbAw approach and find that our method is advantageous over the AbAw approach. In model (1) or (2), only one marker is used in model building. If multiple markers or multiallelic markers are available, it is very easy to generalize the models to analyze the data. For instance, model (14) generalizes model (1) if two markers are available in the analysis. Accordingly, model (13) generalizes model (2). If only one marker is used in analysis, the proposed model (2) is equivalent to the haplotype trend regression method by ZAYKIN *et al.* (2002), which is very close to the method of SCHAID *et al.* (2002). However, the proposed models are different from the haplotype trend regression method for two/multiple marker data. If both markers are diallelic markers, the genotype effect model can be less powerful than the HTR method, and the HTR method can be less powerful than the additive effect model in the case of no or weak third-order linkage disequilibrium among the two markers and the

TABLE 5

Ranked ACE haplotype frequencies

Haplotype rank	Haplotype identity	Haplotype code	Frequency
1	TATATTGIA3	1111112111	0.352113
2	CCCTCCADG2	2222221222	0.284507
3	TATATCADG2	1111121222	0.087324
4	TACATCADG2	1121121222	0.073239
5	TATATCGIA3	1111122111	0.050704
6	CCCTCCGDG2	2222222222	0.025352
7	TATATTAIA3	1111111111	0.025352
8	CCCTCCGIA3	2222222111	0.008451
9	CCCTCCADG3	2222221221	0.008451
10	TATATCGDG2	1111122222	0.008451

QTL. If strong third-order linkage disequilibrium exists, the HTR method can be more powerful.

Basically, the proposed models are genotype based. The models can be used to analyze directly any number of markers, and the markers can be either diallelic or multiallelic. By a simulation study based on ACE haplotype frequencies, we show that the proposed additive effect models have similar power to that of the haplotype-based HTR method. In the meantime, the proposed models enjoy the simplicity of not needing to estimate the expected haplotype scorings; in contrast, the HTR method needs to calculate the expected haplotype scorings before building the models. The proposed models decompose the main marker effects into a summation of additive and dominance effects. In the presence of haplotype effects, it is important to estimate the haplotype effects and haplotype-based methods are more relevant (Stram *et al.* 2003; Tregouet *et al.* 2004).

One potential problem of this generalization is that the number of parameters can be very big. Then, one needs to select important alleles in the analysis and search for important genetic variants that are truly associated with the genetic traits. At first glance, model (1), (2), (13), or (14) seems too complicated and contains too many terms. However, the models are not intimidating at all if one takes into account the recent discovery of haplotype structure in the human genome. Although a haplotype block may contain many SNPs, it takes only a few SNPs to uniquely identify each of the haplotypes in the block. Within a block, there are only two to four common haplotypes (Arnheim *et al.* 2003; Daly *et al.* 2001; Goldstein 2001; Patil *et al.* 2001;

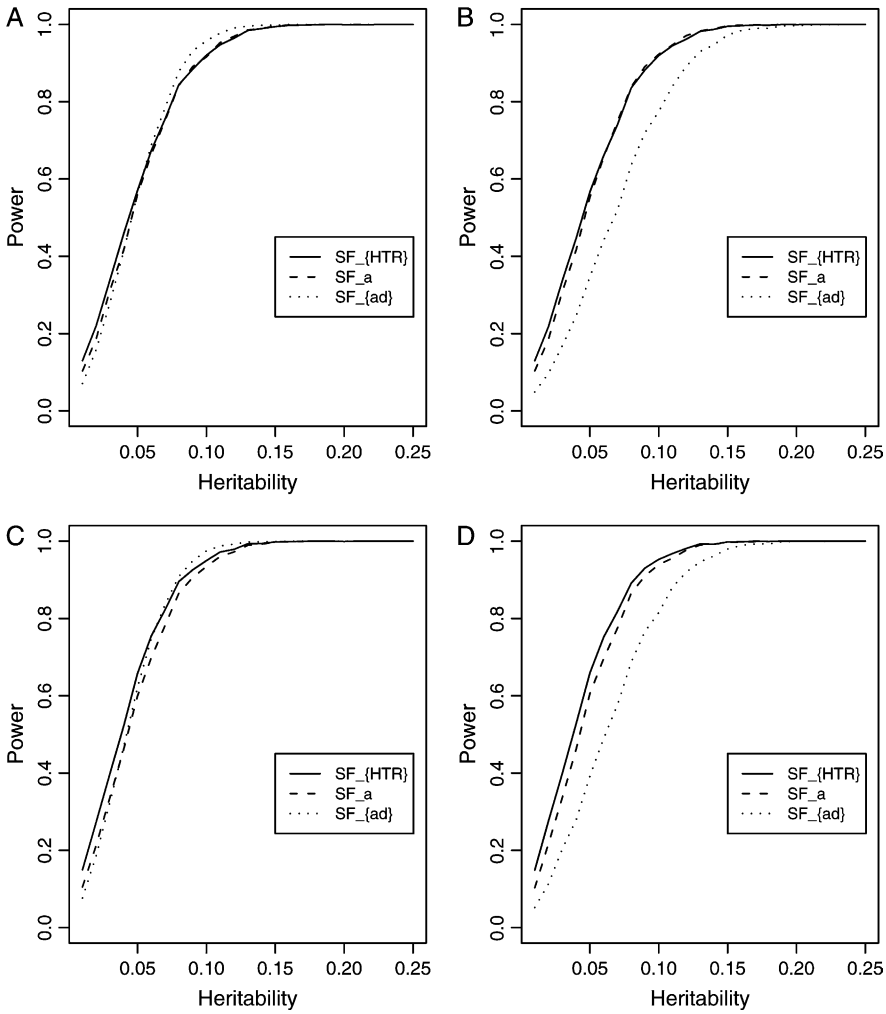

Figure 8.—Empirical power curves of the test statistics F_{HTR}, F_a, and F_{ad} against the heritability h^2 at a 0.05 significance level. The notation SF_a is the empirical power of the *F*-test statistic based on the additive effect model, SF_{ad} is the empirical power of the *F*-test statistic based on the genotype effect model, and SF_{HTR} is the empirical power of the *F*-test statistic based on HTR. (A and C) The curves are plotted for a dominant mode of inheritance $a = d = 1.0$; (B and D) the curves are plotted for an additive mode of inheritance $a = 1.0$, $d = 0$. (A and B) Ten haplotypes are used in the simulations; (C and D) 7 haplotypes are used.

Reich *et al.* 2001; Rioux *et al.* 2001; J. C. Stephens *et al.* 2001; Gabriel *et al.* 2002; Nordborg and Tavaré 2002; Phillips *et al.* 2003). This implies that model (1), (2), (13), or (14) contains a few terms and hence is manageable. Moreover, model (1) or (2) already takes the haplotype structure into account and is potentially more powerful. In practice, one may want to collapse some alleles to reduce the number of parameters. However, the collapsing process may decrease linkage disequilibrium and therefore result in loss of power. The proposed regression models can be fitted to alleviate the problem.

In the mathematical derivations, we make the assumption of HWE. It is unclear how to construct tests reflecting deviations from HWE and this requires further research. In addition, we illustrate that the false-positive rate of the genotype effect test is too high for more than five alleles in a sample of 200 individuals. This is obviously due to the large numbers of possible genotypes and hence to sparseness in the contingency table. This problem could be overcome by using exact tests or permutation procedures.

The models of this article are based on population data. Suppose that both population and pedigree data including sibships are available. Then, model (1) or (2) can be generalized to perform high-resolution combined LD mapping and a linkage study of QTL by variance component models in the spirit of our previous work. In fact, we may generalize regression (1) or (2) by adding the polygenic effect to fit the data. Moreover, log-likelihoods can be constructed on the basis of variance component models. This will generalize our research by using either diallelic/multiallelic markers or haplotypes in a combined analysis of population and pedigree data. It is well known that association study-based population data are prone to false positives, due to the population stratification and population history. A valid approach would be to find linkage information by using pedigree data to locate the QTL on a broad chromosome region. Then, a combined linkage and association mapping can be performed for fine mapping of the genetic traits on the basis of both population and pedigree data (Fan and Xiong 2003). This would be more likely to overcome the drawbacks of separate analysis of either a linkage study or association mapping: low resolution of linkage analysis and high false-positive rates in the association study. In the meantime, it is more likely to take advantage of the two methods: the low false-positive rates of linkage analysis and the high resolution of the association-mapping method.

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

LITERATURE CITED

Abecasis, G. R., L. R. Cardon and W. O. C. Cookson, 2000a A general test of association for quantitative traits in nuclear families. Am. J. Hum. Genet. **66:** 279–292.

Abecasis, G. R., W. O. C. Cookson and L. R. Cardon, 2000b Pedigree tests of linkage disequilibrium. Eur. J. Hum. Genet. **8:** 545–551.

Abecasis, G. R., W. O. C. Cookson and L. R. Cardon, 2001 The power to detect linkage disequilibrium with quantitative traits in selected samples. Am. J. Hum. Genet. **68:** 1463–1474.

Ardlie, K. G., L. Kruglyak and M. Seielsstad, 2002 Patterns of linkage disequilibrium in the human genome. Nat. Rev. Genet. **3:** 299–309.

Arnheim, N., P. Calabrese and M. Nordborg, 2003 Review article: hot and cold spots of recombination in the human genome: the reason we should find them and how this can be achieved. Am. J. Hum. Genet. **73:** 5–16.

Bennett, J. H., 1954 On the theory of random mating. Ann. Eugen. **18:** 311–317.

Boerwinkle, E., E. Chakraborty and C. F. Sing, 1986 The use of measured genotype information in the analysis of quantitative phenotype in man. I. Models and analytical methods. Ann. Hum. Genet. **50:** 181–194.

Cantor, R. M., G. K. Chen, P. Pajukanta and K. Lange, 2005 Association testing in a linked region using large pedigrees. Am. J. Hum. Genet. **76:** 538–542.

Clayton, D., J. Chapman and J. Cooper, 2004 The use of unphased multilocus genotype data in indirect association studies. Genet. Epidemiol. **27:** 415–428.

Crow, J. F., and M. Kimura, 1970 *An Introduction to Population Genetics Theory.* Harper & Row, New York.

Daly, M. J., J. D. Rioux, S. F. Schaffner, T. J. Hudson and E. S. Lander, 2001 High-resolution haplotype structure in the human genome. Nat. Genet. **29:** 229–232.

Dempster, A. P., N. M. Laird and D. B. Rubin, 1977 Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B **39:** 1–38.

Devlin, B., and N. Risch, 1995 A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics **29:** 311–322.

Falconer, D. S., and T. F. C. Mackay, 1996 *Introduction to Quantitative Genetics*, Ed. 4. Longman, London.

Fan, R. Z., and J. S. Jung, 2003 High resolution joint linkage disequilibrium and linkage mapping of quantitative trait loci based on sibship data. Hum. Hered. **56:** 166–187.

Fan, R. Z., and M. M. Xiong, 2002 High resolution mapping of quantitative trait loci by linkage disequilibrium analysis. Eur. J. Hum. Genet. **10:** 607–615.

Fan, R. Z., and M. M. Xiong, 2003 Combined high resolution linkage and association mapping of quantitative trait loci. Eur. J. Hum. Genet. **11:** 125–137.

Fan, R. Z., C. Spinka, L. Jin and J. S. Jung, 2005 Pedigree linkage disequilibrium mapping of quantitative trait loci. Eur. J. Hum. Genet. **13:** 216–231.

Fulker, D. W., S. S. Cherny, P. C. Sham and J. K. Hewitt, 1999 Combined linkage and association sib-pair analysis for quantitative traits. Am. J. Hum. Genet. **64:** 259–267.

Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy *et al.*, 2002 The structure of markers in the human genome. Science **296:** 2225–2229.

George, V., H. K. Tiwari, X. F. Zhu and R. C. Elston, 1999 A test of transmission/disequilibrium for quantitative traits in pedigree data, by multiple regression. Am. J. Hum. Genet. **65:** 236–245.

Goldstein, G. B., 2001 Islands of linkage disequilibrium. Nat. Genet. **29:** 109–111.

Graybill, F. A., 1976 *Theory and Application of the Linear Model.* Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA.

Hayes, B. J., P. M. Visscher, H. C. McPartlan and M. E. Goddard, 2003 Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Res. **13:** 635–643.

Hedrick, P. W., 1987 Gametic disequilibrium measures: proceed with caution. Genetics **117:** 331–341.

International HapMap Consortium, 2003 The International HapMap Project. Nature **426:** 789–796.

International SNP Map Working Group, 2001 A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature **409:** 928–933.

Jung, J. S., R. Z. Fan and L. Jin, 2005 Combined linkage and association mapping of quantitative trait loci by multiple markers. Genetics **170:** 881–898.

Keavney, B., C. A. McKenzie, J. M. Connell, C. Julier, P. J. Ratcliffe *et al.*, 1998 Measured haplotype analysis of the angiotension-1 converting enzyme gene. Hum. Mol. Genet. **7:** 1745–1751.

Kong, A., D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, S. A. Gudjonsson *et al.*, 2002 A high resolution recombination map of the human genome. Nat. Genet. **31:** 241–247.

Lange, K., J. S. Sinsheimer and E. Sobel, 2005 Association testing with Mendel. Genet. Epidemiol. **29:** 36–50.

Meuwissen, T. H. E., and M. E. Goddard, 2000 Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. Genetics **155:** 421–430.

Morris, A. P., J. C. Whittaker and D. J. Balding, 2004 Little loss of information due to unknown phase for fine-scale linkage-disequilibrium mapping with single-nucleotide-polymorphism genotype data. Am. J. Hum. Genet. **74:** 945–953.

Morton, N. E., and D. Wu, 1988 Alternative bioassays of kinship between loci. Am. J. Hum. Genet. **42:** 173–177.

Nielsen, D. M., and B. S. Weir, 1999 A classical setting for associations between markers and loci affecting quantitative traits. Genet. Res. **74:** 271–277.

Nielsen, D. M., and B. S. Weir, 2001 Association studies under general disease models. Theor. Popul. Biol. **60:** 253–263.

Nordborg, M., and S. Tavaré, 2002 Linkage disequilibrium: what history has to tell us. Trends Genet. **18:** 83–90.

Patil, N. P., A. J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi *et al.*, 2001 Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science **294:** 1719–1723.

Phillips, M. S., R. Lawrence, R. Sachidanandam, A. P. Morris, D. J. Balding *et al.*, 2003 Chromosome-wide distribution of markers and the role of recombination hot spots. Nat. Genet. **33:** 382–387.

Pritchard, J. K., and M. Przeworski, 2001 Linkage disequilibrium in humans: model and data. Am. J. Hum. Genet. **69:** 1–14.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, R. C. Sabett *et al.*, 2001 Linkage disequilibrium in the human genome. Nature **411:** 199–204.

Rioux, J. D., M. J. Daly, M. S. Silverberg, K. Lindblad, H. Steinhart *et al.*, 2001 Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. Nat. Genet. **29:** 223–228.

Schaid, D. J., 2004 Evaluating associations of haplotypes with traits. Genet. Epidemiol. **27:** 348–364.

Schaid, D. J., C. M. Rowland, D. E. Tines, R. M. Jacobson and G. A. Poland, 2002 Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am. J. Hum. Genet. **70:** 425–434.

Sham, P. C., S. S. Cherny, S. Purcell and J. K. Hewitt, 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.

Stephens, J. C., J. A. Schneider, D. A. Tanguay, J. Choi, T. Acharya *et al.*, 2001 Haplotype variation and linkage disequilibrium in 313 human genes. Science **293:** 489–493.

Stephens, M., and P. Donnelly, 2003 A comparison of Bayesian methods for haplotype reconstruction from population genotype data. Am. J. Hum. Genet. **73:** 1162–1169.

Stephens, M., N. Smith and P. Donnelly, 2001 A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. **68:** 978–989.

Stram, D. O., C. A. Haiman, J. N. Hirschhorn, D. Altshuler, L. N. Kolonel *et al.*, 2003 Choosing haplotype-tagging SNPs based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the multiethnic cohort study. Hum. Hered. **55:** 179–190.

Sved, J. A., 1971 Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theor. Popul. Biol. **2:** 125–141.

Thomson, G., and M. P. Baur, 1984 Third order linkage disequilibrium. Tissue Antigens **24:** 250–255.

Tregouet, D. A., S. Escolano, L. Tiret, A. Mallet and J. L. Golmard, 2004 A new algorithm for haplotype-based association analysis: the stochastic-EM algorithm. Ann. Hum. Genet. **68:** 165–177.

Weir, B. S., 1996 *Genetic Data Analysis II*, Ed. 2. Sinauer Associates, Sunderland, MA.

Weir, B. S., and C. C. Cockerham, 1977 Two-locus theory in quantitative genetics, pp. 247–269 in *Proceedings of the International Conference on Quantitative Genetics*, edited by E. Pollak, O. Kempthorne and T. B. Bailey. Iowa State University Press, Ames, IA.

Weiss, K. M., and A. G. Clark, 2002 Linkage disequilibrium and the mapping of complex traits. Trends Genet. **18:** 19–24.

Zaykin, D. V., P. H. Westfall, S. S Young, M. A. Karnoub, M. J. Wagner *et al.*, 2002 Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum. Hered. **53:** 79–91.

Communicating editor: G. Gibson

APPENDIX A

For an individual of a population with trait values y and genotype G_A at marker A, let x_{ii} be an indicator function of genotype A_iA_i and x_{ij} be an indicator function of genotype A_iA_j. That is, they are dummy variables defined by

$$x_{ii} = 1_{(A_iA_i)} = \begin{cases} 1 \text{ if } G_A = A_iA_i \\ 0 \text{ else,} \end{cases} \qquad x_{ij} = 1_{(A_iA_j)} = \begin{cases} 1 \text{ if } G_A = A_iA_j \\ 0 \text{ else,} \end{cases}$$

where $i, j = 1, 2, \ldots, m$, $i \neq j$. Then model (1) can be rewritten as

$$y = w\gamma + \sum_{i=1}^{m} x_{ii}\beta_{ii} + \sum_{1 \le i < j \le m} x_{ij}\beta_{ij} + e. \tag{A1}$$

Note that $E(x_{ii}) = P_{A_i}^2$. Given Equation A1, taking expectation of yx_{ii} leads to $E(yx_{ii}) = E(x_{ii})[w\gamma + \beta_{ii}] = P_{A_i}^2[w\gamma + \beta_{ii}]$. On the other hand, a true random-effect model describing the trait value is $y = w\gamma + g + e$, where

$$g = \begin{cases} a & \text{for genotype } Q_1Q_1 \\ d & \text{for genotype } Q_1Q_2 \\ -a & \text{for genotype } Q_2Q_2. \end{cases}$$

Utilizing $P(Q_1A_i) = D_{A_iQ} + P_{A_i}q_1$ and $P(Q_2A_i) = -D_{A_iQ} + P_{A_i}q_2$ gives

$$\begin{aligned} E(yx_{ii}) &= w\gamma E(x_{ii}) + E(gx_{ii}) \\ &= w\gamma P_{A_i}^2 + a[P(Q_1A_i)]^2 + d \cdot 2P(Q_1A_i)P(Q_2A_i) - a[P(Q_2A_i)]^2 \\ &= w\gamma P_{A_i}^2 + a[2D_{A_iQ} + P_{A_i}q_1 - P_{A_i}q_2]P_{A_i} + 2d(D_{A_iQ} + P_{A_i}q_1)(P_{A_i}q_2 - D_{A_iQ}) \\ &= w\gamma P_{A_i}^2 + \mu P_{A_i}^2 + 2D_{A_iQ}\alpha_Q P_{A_i} - \delta_Q D_{A_iQ}^2. \end{aligned} \tag{A2}$$

Equating the above quantity to $E(yx_{ii}) = P_{A_i}^2[w\gamma + \beta_{ii}]$ shows Equation 3 when $i = j$.

If $i \neq j$, $Ex_{ij} = 2P_{A_i}P_{A_j}$. Multiplying at both sides of Equation A1 by x_{ij} and taking the expectation lead to $E(yx_{ij}) = E(x_{ij})[w\gamma + \beta_{ij}]$. Again, utilizing $P(Q_1A_i) = D_{A_iQ} + P_{A_i}q_1$, $P(Q_2A_i) = -D_{A_iQ} + P_{A_i}q_2$, $P(Q_1A_j) = D_{A_jQ} + P_{A_j}q_1$, and $P(Q_2A_j) = -D_{A_jQ} + P_{A_j}q_2$ gives

$$\begin{aligned}
E(yx_{ij}) &= w\gamma E(x_{ij}) + E(gx_{ij}) \\
&= w\gamma \cdot 2P_{A_i}P_{A_j} + 2a[P(Q_1A_i)P(Q_1A_j) - P(Q_2A_i)P(Q_2A_j)] \\
&\quad + d[2P(Q_1A_i)P(Q_2A_j) + 2P(Q_2A_i)P(Q_1A_j)] \\
&= 2P_{A_i}P_{A_j}w\gamma + 2a[(D_{A_iQ} + P_{A_i}q_1)(D_{A_jQ} + P_{A_j}q_1) - (-D_{A_iQ} + P_{A_i}q_2)(-D_{A_jQ} + P_{A_j}q_2)] \\
&\quad + 2d[(D_{A_iQ} + P_{A_i}q_1)(P_{A_j}q_2 - {}_{A_jQ}) + (-D_{A_iQ} + P_{A_i}q_2)(D_{A_jQ} + P_{A_j}q_1)] \\
&= 2P_{A_i}P_{A_j}w\gamma + 2P_{A_i}P_{A_j}\mu + 2\alpha_Q[D_{A_iQ}P_{A_j} + D_{A_jQ}P_{A_i}] - 2\delta_Q D_{A_iQ}D_{A_jQ}. \qquad \text{(A3)}
\end{aligned}$$

Equating the above quantity to $E(yx_{ij}) = 2P_{A_i}P_{A_j}\,[w\gamma + \beta_{ij}]$ shows Equation 3 when $i \neq j$.

APPENDIX B

For an individual with trait values y and genotypes G_A at marker A, let z_i be the number of alleles A_i of genotype G_A, $i = 1, 2, \ldots, m$. That is, z_i is a dummy variable defined by

$$z_i = \begin{cases} 2 \text{ if } G_A = A_iA_i \\ 1 \text{ if } G_A = A_iA_j, \quad j \neq i \\ 0 \text{ else.} \end{cases}$$

Then model (2) can be rewritten as

$$y = w\gamma + \sum_{i=1}^{m} z_i\alpha_i + e. \qquad \text{(B1)}$$

Multiplying both sides of expression (B1) by z_i and taking the expectation lead to

$$w\gamma E\begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_m \end{pmatrix} + E\begin{pmatrix} z_1^2 & z_1z_2 & \ldots & z_1z_m \\ z_2z_1 & z_2^2 & \ldots & z_2z_m \\ \vdots & \vdots & \ldots & \vdots \\ z_mz_1 & z_mz_2 & \ldots & z_m^2 \end{pmatrix}\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{pmatrix} = E\begin{pmatrix} yz_1 \\ yz_2 \\ \vdots \\ yz_m \end{pmatrix}. \qquad \text{(B2)}$$

The elements of the matrix on the left-hand side of the above equation can be calculated as follows: $E(z_i) = 2P_{A_i}^2 + 2P_{A_i}\sum_{j\neq i}P_{A_j} = 2P_{A_i}$, $E(z_i^2) = 4P_{A_i}^2 + 2P_{A_i}\sum_{j\neq i}P_{A_j} = 2P_{A_i}^2 + 2P_{A_i}$. For $i \neq j$, the expectation $E(z_iz_j) = 2P_{A_i}P_{A_j}$. For the elements on the right-hand side, Equations A2 and A3 lead to $E(yz_i) = 2E(yx_{ii}) + \sum_{j\neq i}E(yx_{ij}) = 2P_{A_i}w\gamma + 2P_{A_i}\mu + 2\alpha_QD_{A_iQ}$, since $\sum_i D_{A_iQ} = 0$. Plugging the above quantities into matrix Equation B2 gives Equation 4 as

$$\begin{aligned}
\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{pmatrix} &= \left[2\,\mathrm{diag}(P_{A_1}, P_{A_2}, \ldots, P_{A_m}) + 2\begin{pmatrix} P_{A_1} \\ P_{A_2} \\ \vdots \\ P_{A_m} \end{pmatrix}(P_{A_1}, P_{A_2}, \ldots, P_{A_m})\right]^{-1}\begin{pmatrix} 2P_{A_1}\mu + 2\alpha_QD_{A_1Q} \\ 2P_{A_2}\mu + 2\alpha_QD_{A_2Q} \\ \vdots \\ 2P_{A_m}\mu + 2\alpha_QD_{A_mQ} \end{pmatrix} \\
&= \left[\mathrm{diag}(P_{A_1}^{-1}, P_{A_2}^{-1}, \ldots, P_{A_m}^{-1}) - \frac{1}{2}\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}(1, 1, \ldots, 1)\right]\begin{pmatrix} P_{A_1}\mu + \alpha_QD_{A_1Q} \\ P_{A_2}\mu + \alpha_QD_{A_2Q} \\ \vdots \\ P_{A_m}\mu + \alpha_QD_{A_mQ} \end{pmatrix} \\
&= \begin{pmatrix} \mu/2 \\ \mu/2 \\ \vdots \\ \mu/2 \end{pmatrix} + \alpha_Q\begin{pmatrix} D_{A_1Q}/P_{A_1} \\ D_{A_2Q}/P_{A_2} \\ \vdots \\ D_{A_mQ}/P_{A_m} \end{pmatrix},
\end{aligned}$$

where diag(...) denotes a diagonal matrix; *e.g.*, diag($P_{A_1}, \ldots, P_{A_m}$) is

$$\begin{pmatrix} P_{A_1} & 0 & \ldots & 0 \\ 0 & P_{A_2} & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & P_{A_m} \end{pmatrix}.$$

In the above calculation, we use a fact of the inverse matrix $(M + ab^\tau)^{-1} = M^{-1} - (M^{-1}a)(b^\tau M^{-1})/(1 + b^\tau M^{-1}a)$.

APPENDIX C

Denote a vector $v^\tau = (P^2_{A_2}, \ldots, P^2_{A_m}, 2P_{A_1}P_{A_2}, \ldots, 2P_{A_1}P_{A_m}, \ldots, 2P_{A_{m-1}}P_{A_m})$. If the sample size N is large enough, the large number law implies the approximation

$$X^\tau X/N = \frac{1}{N}\sum_{i=1}^{N} X_i X_i^\tau \approx E(X_1 X_1^\tau) = \mathrm{diag}(P^2_{A_1}, v), \tag{C1}$$

where diag($P^2_{A_1}, v$) is a diagonal matrix, whose elements on the diagonal are given by the elements of ($P^2_{A_1}, v$). That is, if $M = \mathrm{diag}(P^2_{A_1}, v)$, then $M[1,1] = P^2_{A_1}$, $M[J_A, J_A] = 2P_{A_{m-1}}P_{A_m}$. Let H be a $(J_A - 1) \times J_A$ matrix defined by

$$H = \begin{pmatrix} 1 & -1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & \ldots & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ldots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \ldots & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 & \ldots & 0 & 0 & -1 \end{pmatrix}_{(J_A-1)\times J_A}.$$

Then, $(H\eta)^\tau = (\beta_{11} - \beta_{22}, \ldots, \beta_{11} - \beta_{mm}, \beta_{11} - \beta_{12}, \ldots, \beta_{11} - \beta_{1m}, \ldots, \beta_{11} - \beta_{m-1,m})$. From approximation (C1), we have the approximation

$$\begin{aligned} H(X^\tau X)^{-1}H^\tau &\approx \frac{1}{N}H[\mathrm{diag}(P^2_{A_1}, v)]^{-1}H^\tau \\ &= \frac{1}{N}P_{A_1}^{-2}\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}(1, 1, \ldots, 1) + \frac{1}{N}\mathrm{diag}(u), \end{aligned}$$

where $u = (P^{-2}_{A_2}, \ldots, P^{-2}_{A_m}, [2P_{A_1}P_{A_2}]^{-2}, \ldots, [2P_{A_1}P_{A_m}]^{-2}, \ldots, [2P_{A_{m-1}}P_{A_m}]^{-2})$. Applying a fact of inverse matrix $(M + ab^\tau)^{-1} = M^{-1} - (M^{-1}a)(b^\tau M^{-1})/(1 + b^\tau M^{-1}a)$ again, we have

$$[H(X^\tau X)^{-1}H^\tau]^{-1} \approx N[\mathrm{diag}(v) - vv^\tau].$$

The noncentrality parameter is given by

$$\begin{aligned} \lambda_{m,\mathrm{ad}} &= \frac{1}{\sigma^2}(H\eta)^\tau[H(X^\tau X)^{-1}H^\tau]^{-1}(H\eta) \\ &\approx \frac{N}{\sigma^2}\sum_{i=2}^{m}(\beta_{11} - \beta_{ii})^2 P^2_{A_i} + \frac{N}{\sigma^2}\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} 2P_{A_i}P_{A_j}(\beta_{11} - \beta_{ij})^2 \\ &\quad - \frac{N}{\sigma^2}\left[\sum_{i=2}^{m}(\beta_{11} - \beta_{ii})P^2_{A_i} + \sum_{i=1}^{m-1}\sum_{j=i+1}^{m} 2P_{A_i}P_{A_j}(\beta_{11} - \beta_{ij})\right]^2 \\ &= \frac{N}{\sigma^2}\sum_{i=1}^{m}\sum_{j=1}^{m} P_{A_i}P_{A_j}(\beta_{11} - \beta_{ij})^2 - \frac{N}{\sigma^2}\left[\sum_{i=1}^{m}\sum_{j=1}^{m} P_{A_i}P_{A_j}(\beta_{11} - \beta_{ij})\right]^2. \end{aligned} \tag{C2}$$

From Equation 3, we have

$$\beta_{11} - \beta_{ij} = \alpha_Q[2D_{A_1Q}/P_{A_1} - D_{A_iQ}/P_{A_i} - D_{A_jQ}/P_{A_j}] - \delta_Q[D^2_{A_1Q}/P^2_{A_1} - D_{A_iQ}D_{A_jQ}/(P_{A_i}P_{A_j})].$$

Utilizing relation $\sum_{i=1}^m D_{A_iQ} = 0$, we have

$$\begin{aligned}\sum_{i=1}^m\sum_{j=1}^m P_{A_i}P_{A_j}(\beta_{11}-\beta_{ij})^2 &= 2\alpha_Q^2\left[2D^2_{A_1Q}/P^2_{A_1} + \sum_{i=1}^m D^2_{A_iQ}/P_{A_i}\right] - 4[D^3_{A_1Q}/P^3_{A_1}]\alpha_Q\delta_Q \\ &\quad + \delta_Q^2\left[(D^2_{A_1Q}/P^2_{A_1})^2 + \left(\sum_{i=1}^m D^2_{A_iQ}/P_{A_i}\right)^2\right], \\ \sum_{i=1}^m\sum_{j=1}^m P_{A_i}P_{A_j}(\beta_{11}-\beta_{ij}) &= [2D_{A_1Q}/P_{A_1}]\alpha_Q - [D^2_{A_1Q}/P^2_{A_1}]\delta_Q.\end{aligned}$$

Plugging the above equation into (C2), we have

$$\lambda_{m,\mathrm{ad}} \approx \frac{N}{\sigma^2}\left[2\alpha_Q^2 q_1 q_2 \sum_{i=1}^m D^2_{A_iQ}/[q_1q_2P_{A_i}] + \delta_Q^2 q_1^2 q_2^2\left(\sum_{i=1}^m D^2_{A_iQ}/[q_1q_2P_{A_i}]\right)^2\right].$$

Note that $P(Q_2A_i) - P_{A_i}q_2 = -D_{A_iQ}$, and so $R^2_{AQ} = \sum_{i=1}^m \left[D^2_{A_iQ}/[P_{A_i}q_1] + (-D_{A_iQ})^2/[P_{A_i}q_2]\right] = \sum_{i=1}^m D^2_{A_iQ}/[P_{A_i}\ q_1q_2]$. Hence, the noncentrality parameter approximation (10) is valid.

APPENDIX D

The large number law implies the following approximation:

$$Z^\tau Z/N = \frac{1}{N}\sum_{i=1}^N Z_iZ_i^\tau \approx E(Z_1Z_1^\tau) = 2\,\mathrm{diag}(P_{A_1}, P_{A_2}, \ldots, P_{A_m}) + 2\begin{pmatrix}P_{A_1}\\P_{A_2}\\\vdots\\P_{A_m}\end{pmatrix}(P_{A_1}, P_{A_2}, \ldots, P_{A_m}).$$

In the above approximation, the quantities $E(z_iz_j)$ in APPENDIX B are used. Applying a fact of inverse matrix $(M + ab^\tau)^{-1} = M^{-1} - (M^{-1}a)(b^\tau M^{-1})/(1 + b^\tau M^{-1}a)$, the inverse is

$$[Z^\tau Z/N]^{-1} \approx \frac{1}{2}\mathrm{diag}(P_{A_1}^{-1}, P_{A_2}^{-1}, \ldots, P_{A_m}^{-1}) - \frac{1}{4}\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix}(1,1,\ldots,1).$$

Let K be a $(m-1)\times m$ matrix defined by

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

Then, $(K\psi)^\tau = (\alpha_1 - \alpha_2, \ldots, \alpha_1 - \alpha_m)$. On the other hand, we have the approximation

$$K(Z^\tau Z)^{-1}K^\tau \approx \frac{1}{2N}K\,\mathrm{diag}(P_{A_1}^{-1}, P_{A_2}^{-1}, \ldots, P_{A_m}^{-1})K^\tau - \frac{1}{4N}K\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix}(1,1,\ldots,1)K^\tau$$

$$= \frac{1}{2N}P_{A_1}^{-1}\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix}(1,1,\ldots,1) + \frac{1}{2N}\mathrm{diag}(P_{A_2}^{-1}, P_{A_3}^{-1}, \ldots, P_{A_m}^{-1}),$$

whose inverse is given by

$$[K(Z^\tau Z)^{-1}K^\tau]^{-1} \approx 2N\,\mathrm{diag}(P_{A_2}, P_{A_3}, \ldots, P_{A_m}) - 2N\begin{pmatrix}P_{A_2}\\P_{A_3}\\\vdots\\P_{A_m}\end{pmatrix}(P_{A_2}, P_{A_3}, \ldots, P_{A_m}).$$

Therefore, an approximation of the noncentrality parameter is given by

$$\begin{aligned}\lambda_{m,\mathrm{a}} &= \frac{1}{\sigma^2}(K\psi)^\tau[K(Z^\tau Z)^{-1}K^\tau]^{-1}(K\psi)\\ &\approx \frac{2N}{\sigma^2}\sum_{i=1}^{m}P_{A_i}(\alpha_1-\alpha_i)^2 - \frac{2N}{\sigma^2}\left[\sum_{i=1}^{m}P_{A_i}(\alpha_1-\alpha_i)\right]^2.\end{aligned}$$

Equation 4 implies that $\alpha_1 - \alpha_i = \alpha_Q\left[D_{A_1Q}/P_{A_1} - D_{A_iQ}/P_{A_i}\right]$. Thus, the noncentrality parameter

$$\begin{aligned}\lambda_{m,\mathrm{a}} &\approx \frac{2N\alpha_Q^2}{\sigma^2}\left[\sum_{i=1}^{m}P_{A_i}(D_{A_1Q}/P_{A_1} - D_{A_iQ}/P_{A_i})^2 - \left(\sum_{i=1}^{m}P_{A_i}(D_{A_1Q}/P_{A_1} - D_{A_iQ}/P_{A_i})\right)^2\right]\\ &= \frac{2N\alpha_Q^2}{\sigma^2}\left[D_{A_1Q}^2/P_{A_1}^2 + \sum_{i=1}^{m}D_{A_iQ}^2/P_{A_i} - D_{A_1Q}^2/P_{A_1}^2\right] = \frac{2N\alpha_Q^2 q_1 q_2}{\sigma^2}R_{AQ}^2 = \frac{N\sigma_{\mathrm{ga}}^2}{\sigma^2}R_{AQ}^2.\end{aligned}$$

APPENDIX E

For $i = 1, 2, \ldots, m$, $k = 1, \ldots, n$, let us denote $D_{A_iB_k} = P(A_iB_k) - P_{A_i}P_{B_k}$, which are measures of LD between markers A and B. Here $P(A_iB_k)$ is frequency of haplotype A_iB_k. It can be shown that for $i \neq j$, $k \neq l$, $j \neq j'$, $l \neq l'$, $(i, j) \neq (i', j')$, $(k, l) \neq (k', l')$,

$$\begin{aligned}
Ex_{Ai} &= 2P_{A_i}, \quad Ex_{Ai}^2 = 2P_{A_i}^2 + 2P_{A_i}, \quad E(x_{Ai}x_{Aj}) = 2P_{A_i}P_{A_j},\\
Ex_{Bk} &= 2P_{B_k}, \quad Ex_{Bk}^2 = 2P_{B_k}^2 + 2P_{B_k}, \quad E(x_{Bk}x_{Bl}) = 2P_{B_k}P_{B_l},\\
Ez_{Aij} &= 0, \quad Ez_{Aij}^2 = P_{A_i}^2P_{A_j}^2[P_{A_i} + P_{A_j}]^2, \quad Ez_{Bkl} = 0, \quad Ez_{Bkl}^2 = P_{B_k}^2P_{B_l}^2[P_{B_k} + P_{B_l}]^2,\\
E[x_{Ai}x_{Bk}] &= 2D_{A_iB_k} + 4P_{A_i}P_{B_k}, \quad E[x_{Ai}z_{Aij}] = E[x_{Ai}z_{Ajj'}] = E[x_{Ai}z_{Bkl}] = 0,\\
E[x_{Bk}z_{Aij}] &= E[x_{Bk}z_{Bkl}] = E[x_{Bk}z_{Bll'}] = 0, \quad E[z_{Aij}z_{Aij'}] = (P_{A_i}P_{A_j}P_{A_j'})^2,\\
E[z_{Aij}z_{Ai'j'}] &= 0, \quad E[z_{Bkl}z_{Bkl'}] = (P_{B_k}P_{B_l}P_{B_l'})^2, \quad E[z_{Bkl}z_{Bk'l'}] = 0,\\
E[z_{Aij}z_{Bkl}] &= [P_{A_j}(P_{B_l}D_{A_iB_k} - P_{B_k}D_{A_iB_l}) - P_{A_i}(P_{B_l}D_{A_jB_k} - P_{B_k}D_{A_jB_l})]^2,\\
E[yx_{Ai}] &= 2P_{A_i}(w\gamma + \mu) + 2\alpha_QD_{A_iQ}, \quad E[yx_{Bk}] = 2P_{B_k}(w\gamma + \mu) + 2\alpha_QD_{B_kQ},\\
E[yz_{Aij}] &= \delta_Q[P_{A_i}D_{A_jQ} - P_{A_j}D_{A_iQ}]^2, \quad E[yz_{Bkl}] = \delta_Q[P_{B_k}D_{B_lQ} - P_{B_l}D_{B_kQ}]^2.
\end{aligned} \tag{E1}$$

The quantities in (E1) imply that

$$V_{\mathrm{A}} = 2\begin{pmatrix} P_{A_1}(1-P_{A_1}) & -P_{A_1}P_{A_2} & \dots & -P_{A_1}P_{A_{m-1}} & D_{A_1B_1} & \dots & D_{A_1B_{n-1}} \\ -P_{A_1}P_{A_2} & P_{A_2}(1-P_{A_2}) & \dots & -P_{A_2}P_{A_{m-1}} & D_{A_2B_1} & \dots & D_{A_2B_{n-1}} \\ \vdots & \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ -P_{A_1}P_{A_{m-1}} & -P_{A_2}P_{A_{m-1}}) & \dots & P_{A_{m-1}}(1-P_{A_{m-1}}) & D_{A_{m-1}B_1} & \dots & D_{A_{m-1}B_{n-1}} \\ D_{A_1B_1} & D_{A_2B_1} & \dots & D_{A_{m-1}B_1} & P_{B_1}(1-P_{B_1}) & \dots & -P_{B_1}P_{B_{n-1}} \\ \vdots & \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ D_{A_1B_{n-1}} & D_{A_2B_{n-1}} & \dots & D_{A_{m-1}B_{n-1}} & -P_{B_1}P_{B_{n-1}} & \dots & P_{B_{n-1}}(1-P_{B_{n-1}}) \end{pmatrix}.$$

Since $EZ_{A\cup B}$ is a vector of 0's by the quantities in (E1), it can be shown that $V_{\mathrm{D}} = \mathrm{Cov}(Z_{A\cup B}, Z_{A\cup B}) = E(Z_{A\cup B}Z_{A\cup B}^{\top})$. Moreover, the quantities in (E1) imply that the covariance matrix $\mathrm{Cov}(X_{A\cup B}, Z_{A\cup B})$ is a 0 matrix.

Taking variance–covariance between y and x_{Ai}, x_{Bk}, z_{Aij}, z_{Bkl} on the basis of relation (14), we may get the regression coefficients (15) of models (13) and (14).

APPENDIX F

Multiplying both sides of expression (18) by I_j and taking the expectation lead to

$$w\gamma E\begin{pmatrix} I_1 \\ I_2 \\ \vdots \\ I_J \end{pmatrix} + E\begin{pmatrix} I_1^2 & I_1I_2 & \dots & I_1I_J \\ I_2I_1 & I_2^2 & \dots & I_2I_J \\ \vdots & \vdots & \dots & \vdots \\ I_JI_1 & I_JI_2 & \dots & I_J^2 \end{pmatrix}\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_J \end{pmatrix} = E\begin{pmatrix} yI_1 \\ yI_2 \\ \vdots \\ yI_J \end{pmatrix}. \tag{F1}$$

The elements of the matrix on the left-hand side of the above equation can be calculated as follows:

$$E(I_j) = E(1_{h_j}) = P_{h_j}, E(I_kI_j) = \sum_{G_1}\dots\sum_{G_M} P(h_k|G_1,\dots,G_M)P(h_j|G_1,\dots,G_M)P(G_1,\dots,G_M).$$

The elements on the right-hand side are given by

$$\begin{aligned} E(yI_j) &= w\gamma E(I_j) + E(gI_j) \\ &= P_{h_j}w\gamma + \sum_{G_1}\dots\sum_{G_M} P(h_j|G_1,\dots,G_M)E[g1_{(G_1,\dots,G_M)}], \end{aligned}$$

where

$$\begin{aligned} E[g1_{(G_1,\dots,G_M)}] &= a\sum_{i=1}^{J}\sum_{k=1}^{J} P(G_1,\dots,G_M|h_i,h_k)[P(Q_1h_k)P(Q_1h_i) - P(Q_2h_k)P(Q_2h_i)] \\ &\quad + d\sum_{i=1}^{J}\sum_{k=1}^{J} P(G_1,\dots,G_M|h_i,h_k)[P(Q_1h_i)P(Q_2h_k) + P(Q_2h_i)P(Q_1h_k)] \\ &= \mu P(G_1,\dots,G_M) + \alpha_Q\sum_{i=1}^{J}\sum_{k=1}^{J} P(G_1,\dots,G_M|h_i,h_k)[P_{h_i}D_{h_kQ} + P_{h_k}D_{h_iQ}] \\ &\quad - \delta_Q\sum_{i=1}^{J}\sum_{k=1}^{J} P(G_1,\dots,G_M|h_i,h_k)D_{h_iQ}D_{h_kQ}. \end{aligned}$$

Plugging the above quantities into matrix Equation F1 gives Equation 19.