Modeling Linkage Disequilibrium Between a Polymorphic Marker Locus and a Locus Affecting Complex Dichotomous Traits in Natural Populations

Z. W. Luo*,† and Chung-I Wu‡

School of Biosciences, The University of Birmingham, Edgbaston, Birmingham B15 2TT, England, †Laboratory of Population and Quantitative Genetics, Institute of Genetics, Fudan University, Shanghai 200433, People's Republic of China and ‡Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637

Manuscript received August 8, 2000
Accepted for publication April 30, 2001

ABSTRACT

Linkage disequilibrium is an important topic in evolutionary and population genetics. An issue yet to be settled is the theory required to extend the linkage disequilibrium analysis to complex traits. In this study, we present theoretical analysis and methods for detecting or estimating linkage disequilibrium (LD) between a polymorphic marker locus and any one of the loci affecting a complex dichotomous trait on the basis of samples randomly or selectively collected from natural populations. Statistical properties of these methods were investigated and their powers were compared analytically or by use of Monte Carlo simulations. The results show that the disequilibrium may be detected with a power of 80% by using phenotypic records and marker genotype when both the trait and marker variants are common (30%) and the LD is relatively high (40–100% of the theoretical maximum). The maximum-likelihood approach provides accurate estimates of the model parameters as well as detection of linkage disequilibrium. The likelihood method is preferred for its higher power and reliability in parameter estimation. The approaches developed in this article are also compared to those for analyzing a continuously distributed quantitative trait. It is shown that a larger sample size is required for the dichotomous trait model to obtain the same level of power in detecting linkage disequilibrium as the continuous trait analysis. Potential use of these estimates in mapping the trait locus is also discussed.

LINKAGE disequilibrium (LD) between genes at different loci is a topic of historical importance in evolutionary and population genetics theory. Analysis of LD was shown to be a very powerful approach in distinguishing between alternative evolutionary models (Lewontin 1974). There was a recent resurgence in interest in linkage-disequilibrium analysis because of the abundance of genetic polymorphisms at the DNA level. These data made it possible to use the disequilibrium measure to map disease genes (Hastbacka *et al.* 1992; Kruglyak 1999) or to optimize breeding schemes for marker-assisted selection (Lande and Thompson 1990; Luo *et al.* 1997). From the statistical point of view, inferences about linkage disequilibrium involve two aspects: detecting its presence and estimating its magnitude once the disequilibrium has been confirmed (Weir 1979). Many studies focused on inferring linkage disequilibrium between alleles at two or more loci under the circumstance where gametic or genotypic data are available at the involved loci (Hill 1974, 1975; Brown 1975; Weir and Cockerham 1978; Spielman *et al.* 1993; Kaplan *et al.* 1995).

Many characters of economic or evolutionary importance are complex traits for which a one-to-one relationship between genotype and phenotype does not exist (Darvasi 1998). In many instances, a complex phenotype can be assessed continuously or discontinuously. The key difficulty encountered in modeling linkage disequilibria involved with complex traits is mainly caused by the incomplete information on the genotype of these traits. A theoretical analysis was carried out in Luo (1998) to explore the statistical power for detecting the disequilibrium between a polymorphic marker locus and any one of the loci contributing to quantitative genetic variation in natural populations. Applying the model to association studies between a marker and a trait, Nielsen and Weir (1999) showed that there is a simple relationship between the marker being examined and the trait loci. More recently, Luo and Suhai (1999) developed a likelihood approach to estimate the level of linkage disequilibrium between a polymorphic marker locus and any one of the loci affecting a continuous quantitative trait in natural populations. These methods are appropriate for analyzing the marker genotype and trait phenotype data that are directly observable from experiments. Luo *et al.* (2000) discussed how these analyses are relevant to mapping the major genetic effect of continuous quantitative variation.

Most intensively studied characters in evolutionary biology are complex traits that show discontinuous phenotypic variation. These include, for example, male fer-

Corresponding author: Z. W. Luo, School of Biosciences, The University of Birmingham, Edgbaston, Birmingham B15 2TT, England. E-mail: z.luo@bham.ac.uk

TABLE 1

Joint distribution of marker and disease genotype in a random mating population

Marker genotype	Disease genotypes		
	AA	*Aa*	*aa*
MM	$(D + pq)^2$	$2(D + pq)[p(1 - q) - D]$	$[D - p(1 - q)]^2$
Mm	$2(D + pq)[(1 - p)q - D]$	$2[2D^2 + (1 - 2p)(1 - 2q)D + 2pq(1 - p)(1 - q)]$	$2[D + (1 - p)(1 - q)][p(1 - q) - D]$
mm	$[D - (1 - p)q]^2$	$2[(1 - p)q - D][D + (1 - p)(1 - q)]$	$[D + (1 - p)(1 - q)]^2$
Genotypic value	$\mu + a - d/2$	$\mu + d/2$	$\mu - a - d/2$

D, *p*, and *q* are the coefficient of linkage disequilibrium between the marker and disease loci and the frequency of marker allele *M* and disease allele *A*, respectively; *a* and *d* are the additive and dominance effects of the disease gene.

tility (usually scored as fertile or sterile) in interspecific hybrids (Wu and Palopoli 1994), pheromonal types between different species of Drosophila (Coyne *et al.* 1994), and the presence or absence of wing spots in butterflies (Brakefield 1996). To extend our previous study to analyze complex traits of this sort, this article develops the theory and method to detect and estimate linkage disequilibrium between a polymorphic marker locus and any one of the loci affecting a complex binary trait in natural populations. Analyses with dichotomous traits may be theoretically more challenging than with continuous traits because the former requires modeling the link between the observable phenotype and the corresponding latent variable. We study various statistical properties of the theoretical model and compare it with the analysis involved with continuous traits for their statistical powers. We also discuss the implications of the methods in locating genes underlying complex binary traits by use of samples from natural populations.

MODEL AND THEORETICAL ANALYSES

Model and notations: We consider cosegregation of genes at two autosomal loci: One affects a dichotomous trait whereas the other is a codominant marker locus that is devoid of effect on the trait. The two alleles are denoted by *M* and *m* at the marker locus and by *A* and *a* at the trait locus. The association between genes at the two loci is quantified by *D*, the coefficient of the disequilibrium defined as $D = f_{MA} - pq$, where f_{MA} is the frequency of the *MA* haplotype and *p* and *q* denote frequencies of alleles *M* and *A* accordingly. With the assumption of random mating, the joint distribution of genotypes at the marker locus and QTL can be expressed as a function *D*, *p*, and *q* and is summarized in Table 1. It should be noted that the joint distribution implies random mating of the population. The model has been discussed elsewhere (Luo *et al.* 2000).

The phenotype of the trait is assumed to be distributed as a binary variable. If a random sample is collected from the population as described above, individuals sampled may be grouped according to their marker genotypes. The phenotype of the *j*th individual within the *i*th marker group is modeled by

$$y_{ij} = \begin{cases} 1 & \text{if } z_{ij} \geq \theta \\ 0 & \text{if } z_{ij} < \theta, \end{cases}$$

where θ is the threshold for the underlying liability of the trait z_{ij}, which is formulated as

$$z_{ij} = \mu + g_{ij} + \varepsilon_{ij}, \qquad (1)$$

in which μ is the overall population mean and g_{ij} is the genotypic value of the individual at the trait locus. Three genotypes at this locus, say *AA*, *Aa*, and *aa*, are assumed to have genotypic values $a - d/2$, $d/2$, and $-a - d/2$, respectively, where *a* and *d* indicate additive and domi-

TABLE 2

Layout and notation for sample marker genotype and disease phenotype frequencies in a 2 × 3 table

Disease phenotype	Marker genotypes			
	MM	*Mm*	*mm*	
Affected ($y_{ij} = 1$)	n_{11}	n_{12}	n_{13}	n_{1+}
Normal ($y_{ij} = 0$)	n_{21}	n_{22}	n_{23}	n_{2+}
	n_{+1}	n_{+2}	n_{+3}	n

nance effects. ε_{ij} is a normally distributed residual variable with mean zero and standard deviation 1.0, which accounts for the variation of polygenes that are in linkage equilibrium with the marker alleles and for environmental variation. Thus, the genetic effects of the trait locus are measured in units of the residual standard deviation of the liability.

The conditional probability of $y_{ij} = 1$ given the individual's genotype at the trait locus, say $G_{ij} = k$, is obtained by

$$\begin{aligned} f_k &= \Pr\{y_{ij} = 1 | G_{ij} = k, \theta\} = \Pr\{z_{ij} \geq \theta | G_{ij} = k, \theta\} \\ &= 1 - \Pr\{z_{ij} < \theta | G_{ij} = k, \theta\} \\ &= 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\theta} \exp\left[-\frac{(z - \mu - (2 - k)a - (-1)^k d/2)^2}{2} \right] dz, \end{aligned} \tag{2}$$

where $i = 1, 2, 3$ referring, correspondingly, to the marker genotypes *MM*, *Mm*, and *mm*; and $k = 1, 2, 3$, referring to the genotypes *AA*, *Aa*, and *aa* at the trait locus. f_k is also referred to as the penetrance of the kth trait genotype.

Statistical analyses: In this section, three approaches are presented for testing the presence of linkage disequilibrium using the random sample described above.

Independence test: The data can be sorted into a 2 × 3 contingency table as illustrated in Table 2. In Table 2, π_{ij} (or n_{ij}) is the frequency (or number of counts) of the individuals with the ith phenotype and jth marker genotype. The marginal frequencies are denoted by $\pi_{i+} = \Sigma_j \pi_{ij}$ and $\pi_{+j} = \Sigma_i \pi_{ij}$. A statistical test for the null hypothesis H_0, $\pi_{ij} = \pi_{i+}\pi_{+j}$ with $i = 1, 2$ and $j = 1, 2, 3$, is performed by using Pearson's chi-square test statistic

$$\chi^2_{\text{d.f.}} = \sum_{i=1}^{2} \sum_{j=1}^{3} n \frac{(\pi_{ij} - \pi_{i+}\pi_{+j})^2}{\pi_{i+}\pi_{+j}} = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(n_{ij} - m_{ij})^2}{m_{ij}}, \tag{3}$$

where m_{ij} represents the expected value of n_{ij} and can be replaced by their sampling estimates $\hat{m}_{ij} = n\pi_{ij}$ without affecting the distribution of χ^2 (Agresti 1990). For a large sample size and under H_0, χ^2 has a central chi-square distribution with d.f. $= (2 - 1) \times (3 - 1)$. Thus, the above test statistic provides a simple significance statistical test for the presence of linkage disequilibrium between the marker gene and the gene at the trait locus. In fact, it is shown in appendix a that the noncentrality parameter of the test statistic is given by

$$\begin{aligned} \lambda &= n \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{[\pi_{ij} - \pi_{ij}(M)]^2}{\pi_{ij}(M)} \\ &= \frac{nD^2\{(f_1 - 2f_2 + f_3)^2 D^2 + 2p(1 - p)[qf_1 + (1 - 2q)f_2 - (1 - q)f_3]^2\}}{p^2(1 - p)^2[1 - q^2 f_1 - 2q(1 - q)f_2 - (1 - q)^2 f_3][q^2 f_1 + 2q(1 - q)f_2 + (1 - q)^2 f_3]} \end{aligned} \tag{4}$$

when alleles at the marker and trait loci are in linkage disequilibrium as described in Table 1, where $\pi_{ij}(M)$ are the corresponding cell probabilities under the model given in Table 1, and f_k, $k = 1, 2, 3$ are given by Equation 2. The expected value of the $\chi^2_{\text{d.f.}}$ given in Equation 3 is λ + d.f.

Statistical power of the independence test can thus be formulated as

$$\beta_\chi = \Pr\{\chi^2_{(2)}(\lambda) \geq \chi^2_{(2,\alpha)}\}, \tag{5}$$

in which $\chi^2_{(2)}(\lambda)$ stands for a noncentral chi-square variate with 2 d.f. and the noncentrality parameter λ as given by Equation 4, $\chi^2_{(2,\alpha)}$, represents the up-α percentile of a central chi-square distribution with the same degrees of freedom. It should be noted that the independent test suggested above is based on the marker genotype. The test may be also on the marker allele. However, it was pointed out in Nielson and Weir (1999) that the genotype test tests for both additive and dominance effects, whereas the allele test only tests for additive effect.

Regression analysis: If the trait phenotype and the marker genotype data are used to fit a simple regression model,

$$y_{ij} = \alpha + \beta T_{ij} + \delta, \tag{6}$$

where T_{ij} refers to the number of marker alleles, for example, *M*, carried by the individual *ij*, and δ is the normally distributed residual variable. appendix b shows that expectation of the regression parameters is obtained by

$$\alpha = q^2 f_1 + 2q(1 - q)f_2 + (1 - q)^2 f_3 - 2\beta p \tag{7}$$

and

$$\beta = \frac{D[qf_1 + (1 - 2q)f_2 - (1 - q)f_3]}{p(1 - p)}. \tag{8}$$

It can be seen from the above equation that significance of the regression coefficient can be used to infer the presence of linkage disequilibrium. A statistical test for significance of the regression coefficient requires its variance. When the two variables (*i.e.*, *Y* and *T* in the present context) in the regression analysis are normally distributed, the variance of the regression coefficient is simply calculated as

$$\sigma^2_\beta = \frac{(1 - r^2)\sigma^2_Y}{n\sigma^2_T}, \tag{9}$$

where r is the correlation coefficient between Y and T, and σ_Y^2 and σ_T^2 are variances of Y and T, respectively. They are given in APPENDIX B. However, the two regression variables are in fact not normally distributed because their median and arithmetic mean may not be consistent. A general formula for the sampling variance of the regression coefficient without the normality assumption has a form given in KENDALL *et al.* (1983, p. 325) as

$$\sigma_\beta^2 = b^2\left\{\frac{\mathrm{Var}(\sigma_{TT}^2)}{E^2(\sigma_{TY}^2)} + \frac{\mathrm{Var}(\sigma_T^2)}{E^2(\sigma_T^2)} - \frac{2\ \mathrm{Cov}(\sigma_{TY}^2, \sigma_T^2)}{\sigma_{TY}^2\sigma_T^2}\right\}, \qquad (10)$$

where b is the estimate of the regression coefficient, σ_T^2 and σ_{YT}^2 are the sample variance of T and the sample covariance between T and Y, respectively, and $\mathrm{Var}(X)$ and $\mathrm{Cov}(X, Y)$ represent operators of sample variance and covariance. Expectations of the variance and covariance are also given in APPENDIX B. Appropriate use of Equation 10 requires that the sample variance and covariance are of order n^{-1}. This, together with the difference between the sample variance of the regression coefficient predicted by Equation 9 and calculated by the Equation 10, is discussed in the following numerical analysis.

The test statistic, *i.e.*, $t = |\beta|/\sigma_\beta$, is expected to follow a central t-distribution under the null hypothesis $D = 0$. Power of the test at a significant level of α can thus be calculated as the probability

$$\beta_t = \Pr\{t_\nu(\delta_t) > t_{\alpha/2;\nu}\}, \qquad (11)$$

where $t_{\alpha/2;\nu}$ is the upper $\alpha/2$ point of a central t variable with $\nu = n - 2$ d.f. and $t_\nu(\delta_t)$ represents a noncentral Student's t variable with the same degrees of freedom and the noncentral parameter is given by

$$\delta_t = \frac{\Gamma[\nu/2]b}{\sqrt{\nu/2}\Gamma[(\nu - 1)/2]\sigma_b} \qquad (12)$$

(JOHNSON and KOTZ 1970, p. 201), where $\Gamma()$ is a gamma function.

Maximum-likelihood analysis: The log-likelihood of the observed phenotype of the trait and the marker genotype data given the model parameters can be written as

$$L(Y, \Omega) = \sum_{i=1}^{3}\sum_{j=1}^{n_i}\log\left[\sum_{k=1}^{3} h_{ik}\Pr\{y_{ij} = 1|G_{ij} = k, \Omega\}^{y_{ij}}\Pr\{y_{ij} = 0|G_{ij} = k, \Omega\}^{(1-y_{ij})}\right]$$
$$= \sum_{i=1}^{3}\sum_{j=1}^{n_i}\log\left[\sum_{k=1}^{3} h_{ik} f_k^{y_{ij}}(1 - f_k)^{(1-y_{ij})}\right]. \qquad (13)$$

The likelihood may be analyzed under two models: (i) the penetrance model, which involves unknown parameters $\Omega = (p, q, D, f_1, f_2, f_3)$, where f_i ($i = 1, 2, 3$) refer to the penetrance of the three genotypes *AA*, *Aa*, and *aa* at the trait locus; or (ii) the liability model in which the unknown parameters $\Omega = (p, q, D, \theta, \mu, a, d)$.

Note that h_{ik} is the joint frequency of the ith marker genotype and the kth genotype at the trait locus, and $h_i = \Sigma_{k=1}^3 h_{ik}$ is the frequency of the ith marker genotype. The mixing proportion h_{ik} is the probability of an individual having the kth trait genotype and the ith marker genotype and is a function of the parameters p, q, and D (Table 1). Searching the observed likelihood function for the maximum-likelihood estimates of the unknown parameters is difficult because the observed data are incomplete with information about the genotype at the trait locus being missed. The maximum-likelihood estimates (MLEs) can be appropriately formulated following the principles of missing data analysis (LITTLE and RUBIN 1987). In the present context, the expectation of the complete data log-likelihood function has a form as

$$L_c(Y, \Omega) = \sum_{i=1}^{3}\sum_{j=1}^{n_i}\sum_{k=1}^{3}\{w_{ijk}\log(h_{ik}) + w_{ijk}y_{ij}\log(f_k) + w_{ijk}(1 - y_{ij})\log(1 - f_k)\}$$
$$= \sum_{i=1}^{3}\sum_{k=1}^{3}\tilde{w}_{ik}\log(h_{ik})$$
$$+ \sum_{i=1}^{3}\sum_{j=1}^{n_i}\sum_{k=1}^{3} w_{ijk}[y_{ij}\log(f_k) + (1 - y_{ij})\log(1 - f_k)], \qquad (14)$$

where w_{ijk} is the posterior probability of an individual having the kth genotype at the trait locus given his phenotype y_{ij} and the marker genotype i. $\tilde{w}_{ik} = \sum_{j=1}^{n_i} w_{ijk}$ and n_i is the number of individuals within the ith marker genotype group.

The MLEs of the unknown parameters in Equation 12 can be calculated by use of the expectation-maximization (EM) algorithm (DEMPSTER *et al.* 1977). Implementation of the EM algorithm in the present context involves iteration of the following two steps.

E-step: Calculate the posterior probability w_{ijk} using the parameter estimates $\Omega^{(t)}$ at the tth iteration for both the penetrance and liability models:

$$w_{ijk} = \frac{h_{ik}\Pr\{y_{ij} = 1|G_{ij} = k, \Omega\}^{y_{ij}}\Pr\{y_{ij} = 0|G_{ij} = k, \Omega\}^{(1-y_{ij})}}{\sum_{l=1}^{3} h_{il}\Pr\{y_{ij} = 1|G_{ij} = l, \Omega\}^{y_{ij}}\Pr\{y_{ij} = 0|G_{ij} = l, \Omega\}^{(1-y_{ij})}}. \qquad (15)$$

M-step: Substitute w_{ijk} and $\Omega^{(t)}$ into the complete data log-likelihood function (Equation 14) and search new estimates of the parameters that increase the likelihood. The marker allele frequency contains no missing information, and the MLE for this parameter is calculated directly from $p = (2n_1 + n_2)/2n$. The updated estimates, $D^{(t+1)}$ and $q^{(t+1)}$, can be obtained from numerically solving

$$\sum_{i=1}^{3}\sum_{j=1}^{n_i}\sum_{k=1}^{3} w_{ijk}\frac{\partial}{\partial D}[\log h_{ik}] = \sum_{i=1}^{3}\sum_{k=1}^{3}\tilde{w}_{ik}\frac{\partial}{\partial D}[\log h_{ik}]$$
$$= \frac{2\tilde{w}_{11}}{D + pq} + \frac{[2D - p(1 - 2q)]\tilde{w}_{12}}{(D + pq)[D - p(1 - q)]}$$
$$+ \frac{2\tilde{w}_{13}}{D - p(1 - q)} + \frac{[2D - (1 - 2p)q]\tilde{w}_{21}}{(D + pq)[D - (1 - p)q]}$$
$$+ \frac{[4D + (1 - 2p)(1 - 2q)]\tilde{w}_{22}}{2D^2 + (1 - 2p)(1 - 2q)D + 2pq(1 - p)(1 - q)}$$
$$+ \frac{[2D + (1 - 2p)(1 - 2q)]\tilde{w}_{23}}{[D + (1 - p)(1 - q)][D - p(1 - q)]}$$

$$+\frac{2\bar{w}_{31}}{D-(1-p)q}+\frac{[2D+(1-p)(1-2q)]\bar{w}_{32}}{[D+(1-p)(1-q)][D-(1-p)q]}$$

$$+\frac{2\bar{w}_{33}}{D+(1-p)(1-q)}=0 \tag{16}$$

for D with the constraint $\max\{-pq, -(1-p)(1-q)\} \le D \le \min\{p(1-q), (1-p)q\}$ (Weir 1990) and solving the equation

$$\sum_{i=1}^{3}\sum_{j=1}^{n_i}\sum_{k=1}^{3} w_{ijk}\frac{\partial}{\partial q}[\log h_{ik}] = \sum_{i=1}^{3}\sum_{k=1}^{3}\bar{w}_{ik}\frac{\partial}{\partial q}[\log h_{ik}]$$

$$=\frac{2p\bar{w}_{11}}{D+pq}+\frac{p[2D-p(1-2q)]\bar{w}_{12}}{(D+pq)[D-p(1-q)]}$$

$$+\frac{2p\bar{w}_{13}}{D-p(1-q)}-\frac{[(1-2p)D+2p(1-p)q]\bar{w}_{21}}{(D+pq)[D-(1-p)q]}$$

$$-\frac{2[(1-2p)D-p(1-p)(1-2q)]\bar{w}_{22}}{2D^2+(1-2p)(1-2q)D+2pq(1-p)(1-q)}$$

$$-\frac{[(1-2p)D-2p(1-p)(1-q)]\bar{w}_{23}}{[D+(1-p)(1-q)][D-p(1-q)]}+\frac{2(1-p)\bar{w}_{31}}{D-(1-p)q}$$

$$-\frac{(1-p)[2D+(1-p)(1-2q)]\bar{w}_{32}}{[D+(1-p)(1-q)][D-(1-p)q]}$$

$$+\frac{2(1-p)\bar{w}_{33}}{D+(1-p)(1-q)}=0 \tag{17}$$

for q with the constraint $0 < q < 1$.

The other parameters are calculated in different ways depending on which model is considered: For the penetrance model, the updated estimates of the three penetrance parameters can be directly obtained as

$$f_k=\frac{\sum_{i=1}^{3}\sum_{j=1}^{n_i} w_{ijk}y_{ij}}{\sum_{i=1}^{3}\sum_{j=1}^{n_i} w_{ijk}}. \tag{18}$$

For the reliability model, the new estimates of the unknown parameters θ, μ, a, and d can be calculated from the following procedure.

Let $\hat{\xi}_0$ be the proportion of the unaffected individuals in the sample. Thompson (1972) suggested use of $\hat{\xi}_0$ to calculate the maximum-likelihood estimate of the threshold θ. In the present context, the following equation

$$\hat{\xi}_0=\sum_{i=1}^{3}\left[\sum_{j=1}^{n_i}\frac{n_j}{n}h_{ij}\right]\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\theta}\exp\left[-\frac{(z-\mu-(2-k)a-(-1)^k d/2)^2}{2}\right]dz \tag{19}$$

is searched numerically for the MLE of θ on the basis of current estimates of the other model parameters. Derivation of the equation is described in appendix c.

The MLEs of μ, a, and d can be obtained from solving the following equations:

$$\frac{\partial}{\partial\mu}L_c(Y,\Omega)=\sum_{i=1}^{3}\sum_{j=1}^{n_i}\sum_{k=1}^{3}\frac{w_{ijk}(y_{ij}-f_k)}{\sqrt{2\pi}f_k(1-f_k)}\times\exp\left[-\frac{(\theta-\mu-(2-k)a-(-1)^k d/2)^2}{2}\right] \tag{20}$$

$$\frac{\partial}{\partial a}L_c(Y,\Omega)=\sum_{i=1}^{3}\sum_{j=1}^{n_i}\sum_{k=1}^{3}\frac{w_{ijk}(y_{ij}-f_k)(2-k)}{\sqrt{2\pi}f_k(1-f_k)}\times\exp\left[-\frac{(\theta-\mu-(2-k)a-(-1)^k d/2)^2}{2}\right] \tag{21}$$

$$\frac{\partial}{\partial d}L_c(Y,\Omega)=\sum_{i=1}^{3}\sum_{j=1}^{n_i}\sum_{k=1}^{3}\frac{w_{ijk}(y_{ij}-f_k)(-1)^k}{2\sqrt{2\pi}f_k(1-f_k)}\times\exp\left[-\frac{(\theta-\mu-(2-k)a-(-1)^k d/2)^2}{2}\right]=0. \tag{22}$$

Details for derivation of these equations and implementation of the numerical algorithm for solving the equations are given in appendix d.

The algorithm is iterated until the sequence of the estimate of Ω converges, and the converged values are the maximum-likelihood estimates of the parameters. Use of the MLEs allows calculation of the likelihood $L(Y, \hat{\Omega})$ under the full model and the likelihood $L(Y, \hat{\Omega})|_{D=0}$. The likelihood ratio

$$\mathrm{LR}=2[L(Y,\hat{\Omega})-L(Y,\hat{\Omega})|_{D=0}], \tag{23}$$

which, in general, follows asymptotically a chi-square distribution with 1 d.f. under the null hypothesis H_0: $D = 0$, where $L(Y, \hat{\Omega})$ and $L(Y, \hat{\Omega})|_{D=0}$ represent values of the log-likelihood function of (13) evaluated at the MLEs of the parameters and at the MLEs but with $D = 0$, respectively.

SIMULATION STUDY AND NUMERICAL ANALYSES

Simulation model: The strategy is described elsewhere (Luo 1998) for simulating the linkage disequilibrium between a polymorphic marker locus and a trait locus in random mating natural populations. For any given set of parameters, n, p, q, D, μ, a, and d, random samples of the individual genotypes at the marker and trait loci were generated. The liability of an individual was determined by its genotype value at the trait locus plus a normally distributed random variate, and the trait phenotype for the individual depended if its liability value was greater than the simulated threshold, θ. In the simulation study, we considered 15 sets of parameters and these were listed in Table 3 in which genetic effects of the disease gene were represented in term of h^2, heritability of the liability of the disease trait, and dr, dominance ratio at the disease locus. Each of the simulated parameters was repeated 100 or 200 times depending on the purposes of the analysis.

Behavior of the test statistics under the null hypothesis: To apply the methods developed above in practice, a question remains to determine appropriate critical threshold values for significance of the test statistics. These require knowledge about the distributions of the test statistics under the null hypothesis. Under the null

TABLE 3

Parameter values for the 15 populations considered in the numerical analyses

Population	n	p	q	D	θ	h^2	dr
1	300	0.5	0.5	0.10	0.0	0.1	0.0
2	300	0.5	0.5	0.10	0.5	0.1	0.0
3	300	0.5	0.5	0.20	0.0	0.1	0.0
4	300	0.5	0.5	0.10	0.0	0.2	1.0
5	500	0.5	0.5	0.10	0.0	0.1	0.0
6	500	0.5	0.5	0.10	0.5	0.1	0.0
7	500	0.5	0.5	0.00	0.0	0.1	0.0
8	300	0.5	0.5	0.10	0.0	0.1	0.5
9	300	0.5	0.5	0.10	0.0	0.1	1.0
10	300	0.3	0.3	0.09	0.0	0.1	0.0
11	300	0.7	0.7	0.09	0.0	0.1	0.0
12	300	0.3	0.5	0.10	0.0	0.1	0.0
13	300	0.3	0.5	0.10	0.0	0.1	1.0
14	300	0.5	0.3	0.10	0.0	0.1	0.0
15	300	0.5	0.3	0.10	0.0	0.1	1.0

n is the sample size; p and q, the population frequencies of the marker allele M and the disease allele A; D, the coefficient of linkage disequilibrium; θ, the threshold in the reliability model; h^2 and dr, heritability of the reliability of the disease trait and dominance ratio of allele effects at the disease locus.

hypothesis, the test statistic of the contingency table analysis is expected to be a $\chi^2_{\text{d.f.}=2}$ variable with the mean = d.f. and the variance = 2 d.f. (Johnson and Kotz 1970, p. 134). The test statistic of the regression analysis under the null hypothesis is expected to have the mean deviation given by $\sqrt{\nu}\Gamma[(\nu-1)/2]/[\sqrt{\pi}\Gamma(\nu/2)]$ and the variance given by $\nu/(\nu-2)$ (Johnson and Kotz 1970, p. 96). In many cases, likelihood-ratio test statistics under a null hypothesis would have asymptotically a χ^2 distribution. In addition, the 95th percentiles of these distributions can be calculated from the relevant probability density functions of these expected distributions. We check the agreement between the observed distributions of these test statistics under the null hypothesis and their corresponding expected ones by comparing these expected distribution parameters with those calculated from repeated simulations.

Table 4 illustrates the expected and simulated values of the means, variances, and 95th percentiles of the four test statistics that were calculated from 200 repeated simulation trials under the condition $D = 0$. The population numbers that specify the simulation parameter sets except for the value of D are in accordance with those given in Table 3. Table 4 shows that the observed values of the distribution parameters for the contingency and regression test statistics were in good agreement with their corresponding theoretical expectations. The observed frequencies of these tests being significant are in the range of the given significant level. These suggest validity in approximating the distributions of the two test

TABLE 4

Theoretical and simulated distribution characteristics of the four test statistics under the null hypothesis ($D = 0$)

	χ^2-test				t-test			LR_1-test				LR_2-test			
Population	$E(\chi^2_2)$	$\text{Var}(\chi^2_2)$	95 pt	α	$E[\lvert t\rvert]$	$\text{Var}(t)$	95 pt	α	$E(\chi^2_1)$	$\text{Var}(\chi^2_1)$	95 pt	α	$E(\chi^2_1)$	$\text{Var}(\chi^2_1)$	95 pt	α
Experiment	2.00	4.00	5.99	0.05	0.80	1.00	1.96	0.05	1.00	2.00	3.84	0.05	1.00	2.00	3.84	0.05
1	2.00	3.56	5.51	0.05	0.77	0.97	2.13	0.07	0.57	0.73	2.83	0.01	0.66	0.75	2.39	0.01
2	1.81	3.82	5.76	0.03	0.73	1.06	1.76	0.02	0.44	0.36	1.83	0.00	0.68	0.63	2.02	0.01
4	1.93	3.84	5.61	0.05	0.83	0.92	2.04	0.06	0.54	0.80	2.17	0.02	0.54	0.73	2.16	0.02
5 (7)	1.96	3.55	5.90	0.05	0.78	0.83	1.86	0.03	0.52	0.66	2.42	0.01	0.63	0.73	2.43	0.00
6	2.02	3.58	4.88	0.04	0.81	1.13	1.98	0.05	0.52	0.62	2.00	0.02	0.68	0.72	2.07	0.01
8	1.79	4.13	5.75	0.04	0.73	0.95	1.89	0.03	0.55	0.86	2.09	0.02	0.66	0.88	2.11	0.02
9	1.94	3.39	5.23	0.07	0.81	0.88	2.04	0.04	0.61	0.64	2.25	0.00	0.75	0.64	2.29	0.00
10	1.97	3.85	4.55	0.02	0.88	0.92	1.94	0.05	0.51	0.38	2.08	0.00	0.64	0.43	2.11	0.00
11	1.71	2.88	5.20	0.03	0.81	1.20	1.99	0.05	0.54	0.61	2.24	0.00	0.61	0.63	2.29	0.00
12	2.00	3.29	5.03	0.06	0.82	0.89	2.03	0.04	0.54	0.67	1.99	0.01	0.67	0.74	2.04	0.01
13	1.95	3.86	5.33	0.05	0.81	1.13	1.87	0.03	0.66	0.93	2.32	0.02	0.73	0.92	2.05	0.02
14	1.84	2.97	5.34	0.02	0.73	0.86	1.79	0.03	0.46	0.49	1.95	0.00	0.55	0.51	1.99	0.00
15	1.93	4.11	5.90	0.05	0.75	1.23	1.84	0.05	0.53	0.77	2.26	0.02	0.71	0.76	2.34	0.02

The contingency table test (χ^2-test), the regression test (t-test), the likelihood-ratio test under the penetrance model (LR_1-test), and the likelihood-ratio test under the liability model (LR_2-test) are shown. $E(\chi^2_{\text{d.f.}})$ and $\text{Var}(\chi^2_{\text{d.f.}})$ are the mean and variance of the χ^2-test statistic with d.f. degrees of freedom, $E[|t|]$ and $\text{Var}(t)$ are the mean deviation and variance of the t-test statistic, 95 pt represents the 95th percentile of the test statistics, and α the significance level of the tests.

statistics with the expected ones. However, the simulated means, variances, and 95th percentiles of the likelihood-ratio test statistics under the penetrance model and the liability model may differ markedly from those expected for the χ^2 distribution with 1 d.f. Moreover, the frequencies of the tests detected as significant are substantially lower than the given significance level. These indicate that approximating distributions of the likelihood-ratio test statistics given in Equation 23 with χ^2 is inappropriate. The deviation of distributions of the likelihood-ratio test statistics from χ^2_1 is most likely due to slow convergence of the statistics to the expected distribution. The discrepancy may be improved when a larger sample size is used. Here, we suggest using the permutation test, as is done regularly for mapping QTL in planned experiments (Churchill and Doerge 1994), to empirically estimate the threshold from the data for the likelihood-ratio tests.

Contingency and regression analyses: The expected value of the χ^2-test statistic in the contingency analysis under each of the 15 sets of the simulated parameters was theoretically predicted using the formula d.f. + λ and calculated from the simulation study. In the regression analysis, the expectation of the regression coefficient was calculated according to Equation 8 and the corresponding simulated observation was obtained as the mean of the regression coefficients over 100 simulations. The expected standard deviation of the regression coefficient was evaluated using Equations 10 ($\sigma_b^{(1)}$) and 11 ($\sigma_b^{(2)}$). The simulated value of the standard deviation ($\hat{\sigma}_b$) was calculated by standard deviation of the coefficients computed from the repeated simulations. Moreover, powers of the statistical tests under the two analyses were evaluated both theoretically and by simulation at the significant level of $\alpha = 5\%$. The simulated values of the powers were obtained as the frequency of the significant tests in 100 simulation trials.

Table 5 summarizes results of the analyses. The table shows a good agreement between the expected values of the statistics investigated here and their corresponding simulated values under both the contingency and regression analyses. The standard deviations of the regression coefficients predicted using Equation 10 are almost identical to that calculated using Equation 11 for all but population 7 considered here, suggesting that the normality assumption of the data is not important in calculating the standard deviations. When the marker gene and allele at the trait locus were in linkage equilibrium (population 7), the covariance between the marker genotype and the phenotype of the trait was expected to be zero, and Equation 11 then failed to provide prediction of the standard deviation. As long as the power is concerned, it can be seen from Table 4 that the level of linkage disequilibrium plays a major role in both the analyses. In addition, the power is expected to be higher when allelic frequencies at the marker and trait loci are low (population 11) or high (population 12) than that when the frequencies take intermediate values (population 1). The power may be reduced slightly with increase in the threshold. The powers of both the tests decrease as the dominance ratio at the trait locus increases (populations 1, 8, and 9). Comparison of the power between the two tests suggests that the regression model is generally a more efficient method to detect the linkage disequilibrium than the contingency analysis.

Maximum-likelihood analyses: The likelihood equation (13) was used to calculate the maximum-likelihood estimates of the parameters defining the penetrance model and the liability model. Table 6 illustrates the means and the standard errors of the MLEs of the parameters of the penetrance model. It shows that the parameters are well estimated by their corresponding maximum-likelihood estimates. The standard errors of the estimates of the marker allele frequencies were much smaller than those of the estimates of the gene frequencies and the coefficients of the linkage disequilibrium, revealing the fact that the marker genotype data provide full information for estimating the marker gene frequencies, while the estimates of the trait gene frequencies and the disequilibrium coefficients were based on incomplete information from the data of the marker genotype and the trait phenotype. The penetrances of the three genotypes at the trait locus were adequately predicted by the maximum-likelihood approach. The penetrances of the heterozygote genotype were estimated with smaller standard errors than those of the other two homozygote genotypes. Tabulated in Table 6 were also the empirical powers, on the basis of 200 simulations, for detecting the linkage disequilibrium model. The critical value used to determine significance of the likelihood-ratio test was based on 200 permutations at the null hypothesis and may change for each simulation. Table 6 shows that among all parameters under question, the level of linkage disequilibrium between the marker and trait loci is the most important factor that determines the efficiency of the likelihood-ratio test. There is a trend toward decrease in the power as the dominance effect at the trait locus increases (comparisons among populations 1, 8, and 9 and between populations 12 and 13 as well as between populations 14 and 15). Comparison among populations 1, 10, and 11 shows that the test tends to be more efficient when both the marker and trait genes are at lower or higher frequencies given the other parameters. This may reflect that the contrast of difference in value of the penetrance between the three genotypes is enhanced when the genes are at low or high frequencies.

Tabulated in Table 7 are the means and their standard errors of the MLEs of the parameters used in the liability model as well as the empirical powers of the likelihood-ratio test under the model. It can be seen from the table that the parameters are well predicted by their corresponding maximum-likelihood estimates.

TABLE 5

Theoretical predictions and simulation estimates of the parameters of the test statistics in the contingency and regression analyses

	Contingency analysis				Regression analysis						
Population	$E(\chi^2)$	$\hat{\chi}^2$	β_{χ^2}	$\hat{\beta}_{\chi^2}$	b	$\hat{b}$	$\sigma_b^{(1)}$	$\sigma_b^{(2)}$	$\hat{\sigma}_b$	β_t	$\hat{\beta}_t$
1	5.16	5.32	0.34	0.37	0.073	0.079	0.040	0.041	0.041	0.43	0.47
2	4.89	5.06	0.31	0.35	0.065	0.071	0.038	0.038	0.037	0.40	0.41
3	14.63	14.46	0.90	0.88	0.145	0.144	0.039	0.040	0.039	0.96	0.93
4	6.58	7.30	0.47	0.45	0.075	0.080	0.040	0.041	0.040	0.46	0.43
5	7.26	7.00	0.53	0.55	0.075	0.071	0.031	0.032	0.033	0.64	0.63
6	6.81	6.98	0.49	0.53	0.065	0.068	0.029	0.029	0.032	0.60	0.60
7	2.00	2.02	0.05	0.03	0.000	0.002	0.032	—	0.031	0.05	0.04
8	4.85	5.48	0.31	0.38	0.067	0.077	0.040	0.041	0.042	0.38	0.42
9	4.25	4.61	0.25	0.31	0.056	0.061	0.040	0.041	0.040	0.28	0.35
10	6.52	6.69	0.46	0.50	0.094	0.097	0.043	0.044	0.041	0.57	0.58
11	6.52	6.26	0.46	0.43	0.094	0.094	0.043	0.045	0.046	0.57	0.56
12	5.76	6.02	0.39	0.44	0.086	0.089	0.044	0.045	0.043	0.50	0.51
13	4.71	4.76	0.29	0.23	0.067	0.075	0.044	0.045	0.044	0.32	0.37
14	5.80	6.38	0.40	0.40	0.079	0.083	0.040	0.041	0.040	0.50	0.53
15	5.30	4.92	0.35	0.33	0.090	0.083	0.040	0.041	0.040	0.60	0.55

$E(\chi^2)$ and $\hat{\chi}^2$ are the chi-square test statistics predicted theoretically and by simulation, respectively; β_{χ^2} and $\hat{\beta}_{\chi^2}$, theoretically predicted and simulation observed powers of the test for linkage disequilibrium under the contingency analysis; b and $\hat{b}$, the theoretical predicted and the corresponding mean of simulation estimates of the regression coefficient; $\sigma_b^{(1)}$, $\sigma_b^{(2)}$, and $\hat{\sigma}_b$, the standard deviation of the regression coefficient predicted using Equations 1 and 2 and from the simulation, respectively; β_t and $\hat{\beta}_t$, the theoretically predicted and observed powers of the linkage disequilibrium test under the regression analysis.

The standard errors of the parameter estimates under this model are comparable with those under the penetrance model, suggesting consistency of the two models in the parameter estimates. The powers of the linkage disequilibrium test whose significance threshold was also based on 200 permutations at the null hypothesis are approximately the same as those observed in the penetrance model analysis. Effects of the allele frequencies at the marker and trait loci and the dominance level of the gene at the trait locus on the power show the same trend as that discussed in the above penetrance model. These reveal that both the models are almost equally efficient in the parameter estimation and the disequilibrium test.

Comparison of the powers observed in the likelihood analyses to those illustrated in Table 4 shows that the likelihood-ratio test is probably more powerful in detecting the disequilibrium than the test based on the contingency or regression analyses.

Comparison of the power between this model and the normal data model: We previously investigated the statistical test for the linkage disequilibrium between a polymorphic marker locus and a locus underlying a quantitative trait whose phenotype is normally distributed (Luo *et al.* 2000). It was shown that the unbalanced nested ANOVA and the regression analyses provided an efficient test for the disequilibrium using samples from a natural population. To compare these approaches analyzing the continuous quantitative traits to those developed in the present study, the sample sizes required for detecting the linkage disequilibrium with power of 80% at the significance level of 0.001 were evaluated for the methods and were summarized in Table 8. It shows that the methods modeling the continuous trait phenotype are usually more powerful than the methods analyzing the binary phenotype, with the exception that the regression analysis under the reliability model is only slightly better off when the marker allele is at the intermediate frequency but the frequency of the trait gene is low and the gene shows complete dominance (population 12). The table reveals that the regression models require smaller samples for the given power than the alternative methods regardless of whether the trait phenotype is distributed continuously or dichotomously.

DISCUSSION

Whole-genome association studies were recently proposed as a powerful approach for investigating many fundamental questions in evolutionary biology (Clark *et al.* 1998; Paabo 1999) and for detecting the many subtle genetic effects that underlie susceptibility to common diseases (Lander and Schork 1994; Lander 1996; Risch and Merikangas 1996).

Appropriate modeling of linkage disequilibria between polymorphic markers and genes affecting complex genetic variation in natural populations is an essen-

TABLE 6

Means of the MLEs of the genetic parameters and their corresponding standard errors and the empirical statistical power for detecting linkage disequilibrium between the marker locus and the disease locus under the penetrance model

Population	f_1	f_2	f_3	$\hat{p}$	$\hat{q}$	$\hat{D}$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	β_p
1	0.6813	0.5000	0.3187	0.5043(0.0020)	0.4798(0.0155)	0.1095(0.0067)	0.6655(0.0205)	0.5107(0.0091)	0.3064(0.0111)	0.52
2	0.4886	0.3085	0.1657	0.4998(0.0023)	0.4776(0.0124)	0.0974(0.0074)	0.4817(0.0174)	0.3106(0.0112)	0.1600(0.0145)	0.46
3	0.6813	0.5000	0.3187	0.5002(0.0021)	0.5030(0.0089)	0.2007(0.0043)	0.6736(0.0102)	0.5082(0.0059)	0.3177(0.0093)	0.92
4	0.6136	0.6136	0.1932	0.4978(0.0020)	0.4894(0.0105)	0.1017(0.0061)	0.6075(0.0119)	0.6120(0.0070)	0.1985(0.0104)	0.64
5	0.6813	0.5000	0.3187	0.4986(0.0017)	0.5058(0.0086)	0.1059(0.0055)	0.6749(0.0089)	0.5003(0.0056)	0.3369(0.0089)	0.70
6	0.4886	0.3085	0.1657	0.4975(0.0015)	0.4721(0.0110)	0.1058(0.0067)	0.4856(0.0087)	0.3102(0.0056)	0.1680(0.0061)	0.66
7	0.6813	0.5000	0.3187	0.4982(0.0016)	0.5010(0.0120)	0.0024(0.0098)	0.6966(0.0058)	0.5012(0.0049)	0.3114(0.0059)	0.06
8	0.6306	0.5442	0.2893	0.4990(0.0020)	0.5021(0.0126)	0.1109(0.0071)	0.6352(0.0114)	0.5400(0.0085)	0.2908(0.0093)	0.49
9	0.5763	0.5763	0.2819	0.4986(0.0022)	0.5178(0.0204)	0.1081(0.0101)	0.5923(0.0153)	0.5779(0.0101)	0.2669(0.0134)	0.38
10	0.6965	0.5000	0.3035	0.3007(0.0018)	0.2874(0.0101)	0.1061(0.0058)	0.6656(0.0210)	0.5099(0.0103)	0.3034(0.0161)	0.63
11	0.6965	0.5000	0.3035	0.6989(0.0021)	0.6967(0.0087)	0.1060(0.0070)	0.6869(0.0189)	0.5064(0.0110)	0.3088(0.0148)	0.55
12	0.6813	0.5000	0.3187	0.2995(0.0018)	0.5018(0.0075)	0.1081(0.0048)	0.6833(0.0167)	0.4849(0.0085)	0.3438(0.0179)	0.53
13	0.5763	0.5763	0.2819	0.2991(0.0017)	0.4896(0.0107)	0.1045(0.0089)	0.5886(0.0109)	0.5612(0.0087)	0.2772(0.0095)	0.34
14	0.6965	0.5000	0.3035	0.4989(0.0020)	0.3066(0.0241)	0.1068(0.0111)	0.6942(0.0117)	0.5057(0.0100)	0.3111(0.0121)	0.55
15	0.5662	0.5662	0.3085	0.4976(0.0021)	0.3043(0.0269)	0.1052(0.0131)	0.5647(0.0200)	0.5654(0.0131)	0.3052(0.0181)	0.43

$\hat{p}$ and $\hat{q}$ are the estimates of the frequencies of alleles M and A, $\hat{D}$ is the estimate of the coefficient of the linkage disequilibrium, f_i and $\hat{f}_i$ (i = 1, 2, 3) are the actual and estimated penetrances of the three disease genotypes, and β_p denotes the empirical power. The simulated values of p, q, and D and the sample sizes are listed in Table 3. Standard errors are in parentheses.

TABLE 7

Means of the MLEs of the genetic parameters and their corresponding standard errors and the empirical statistical power for detecting linkage disequilibrium between the marker locus and the disease locus under the reliability model

Population	a	d	$\hat{q}$	$\hat{D}$	$\hat{\theta}$	$\hat{\mu}$	$\hat{a}$	$\hat{d}$	$\hat{\beta}_r$
1	0.4714	0.0000	0.4765(0.0134)	0.1095(0.0087)	0.0028(0.0065)	−0.0031(0.0021)	0.4847(0.0164)	0.0661(0.0411)	0.52
2	0.4714	0.0000	0.4834(0.0143)	0.0974(0.0072)	0.5062(0.0070)	−0.0086(0.0047)	0.4947(0.0191)	0.0439(0.0384)	0.50
3	0.4714	0.0000	0.4890(0.0075)	0.2010(0.0046)	−0.0069(0.0067)	−0.0010(0.0010)	0.4741(0.0150)	0.0333(0.0222)	0.90
4	0.5774	0.5774	0.5024(0.0085)	0.1024(0.0073)	0.0075(0.0068)	0.0058(0.0029)	0.5835(0.0193)	0.5788(0.0398)	0.66
5	0.4714	0.0000	0.4880(0.0076)	0.1047(0.0065)	−0.0071(0.0053)	0.0014(0.0021)	0.4693(0.0145)	−0.0161(0.0355)	0.65
6	0.4714	0.0000	0.4809(0.0104)	0.0983(0.0076)	0.4975(0.0052)	−0.0103(0.0030)	0.4765(0.0158)	0.0120(0.0321)	0.65
7	0.4714	0.0000	0.4919(0.0112)	0.0024(0.0095)	−0.0075(0.0044)	−0.0036(0.0030)	0.5052(0.0152)	−0.0084(0.0114)	0.04
8	0.4444	0.2222	0.5012(0.0126)	0.1091(0.0073)	0.0050(0.0058)	0.0040(0.0023)	0.4672(0.0169)	0.2063(0.0451)	0.52
9	0.3849	0.3849	0.5044(0.0104)	0.1086(0.0105)	−0.0030(0.0062)	0.0013(0.0015)	0.4104(0.0168)	0.3964(0.0316)	0.40
10	0.5143	0.0000	0.2931(0.0091)	0.0971(0.0067)	−0.0028(0.0064)	−0.0020(0.0086)	0.5028(0.0268)	0.0420(0.0464)	0.66
11	0.5143	0.0000	0.6875(0.0086)	0.1066(0.0069)	0.0076(0.0069)	−0.0089(0.0097)	0.5425(0.0283)	0.0656(0.0482)	0.58
12	0.4714	0.0000	0.4879(0.0090)	0.0974(0.0053)	0.0042(0.0062)	0.0096(0.0058)	0.4639(0.0184)	−0.0775(0.0524)	0.55
13	0.3849	0.3849	0.4798(0.0110)	0.1063(0.0078)	0.0155(0.0060)	0.0030(0.0018)	0.4188(0.0167)	0.3407(0.0372)	0.38
14	0.5143	0.0000	0.2930(0.0194)	0.0988(0.0105)	−0.0272(0.0062)	0.0038(0.0095)	0.5485(0.0318)	−0.0335(0.0497)	0.50
15	0.3334	0.3334	0.3013(0.0219)	0.1065(0.0110)	−0.0159(0.0066)	−0.0016(0.0098)	0.3358(0.0303)	0.3052(0.0492)	0.43

$\hat{q}$ and $\hat{D}$ are the estimates of the frequencies of allele A and coefficient of the linkage disequilibrium; $\hat{\theta}$ is the estimate of the threshold; $\hat{\mu}$, $\hat{a}$, and $\hat{d}$ are the estimates of population mean and the additive and dominance effects of the disease locus, respectively. $\hat{\beta}_r$ denotes the empirical power under the reliability model. The simulated values of p, q, D, θ, μ and the sample sizes are listed in Table 3. Standard errors are in parentheses.

TABLE 8

The sample sizes required for detecting the linkage disequilibrium with a power of 80% at the significance level of 0.001

Population	p	q	D	θ	h^2	dr	N_a	N_b	N_c	N_r
1	0.5	0.5	0.10	0.0	0.1	0.0	1200	1006	1839	1569
2	0.5	0.5	0.10	0.5	0.1	0.0	1200	1006	2010	1719
3	0.5	0.5	0.20	0.0	0.1	0.0	293	245	460	373
4	0.5	0.5	0.10	0.0	0.2	1.0	830	753	1267	1457
5	0.5	0.5	0.10	0.0	0.1	0.5	1325	1133	2037	1838
6	0.5	0.5	0.10	0.0	0.1	1.0	1670	1514	2583	2663
7	0.3	0.3	0.09	0.0	0.1	0.0	1043	875	1584	1348
8	0.7	0.7	0.09	0.0	0.1	0.0	1043	875	1584	1348
9	0.3	0.5	0.10	0.0	0.1	0.0	1008	844	1545	1314
10	0.3	0.5	0.10	0.0	0.1	1.0	1384	1270	2140	2233
11	0.5	0.3	0.10	0.0	0.1	0.0	1006	844	1528	1299
12	0.5	0.3	0.10	0.0	0.1	1.0	1176	1027	1760	987

ANOVA (N_a) and regression analysis (N_b) are used when the phenotype of the disease trait is normally distributed; the contingency test (N_c) and the regression test (N_r) are used when the phenotype is a binary variable. p and q are the population frequencies of the marker allele M and the disease gene A; D is the coefficient of the linkage disequilibrium. θ is the threshold of the liability model; h^2 and dr are the heritability and dominance ratio of the disease gene.

tial step and has proved to be an effective approach in improving resolution of gene mapping. Linkage disequilibrium analysis of gene mapping may provide a mapping resolution <1 cM so that molecular screening at the DNA sequence level for the candidate gene can be performed. This is an improvement over the traditional pedigree-based or crossing population-based linkage analysis where the candidate gene can hardly be narrowed down to such a resolution (de la Chapelle and Wright 1998; Kearsey and Farquhar 1998; Guo and Lange 2000). The theoretical analysis of this article offers insight into the degree of association between the marker and trait and serves as an important first step in the direction of analyzing complex dichotomous traits using population-level LD at linked markers.

Efficiency and statistical properties of three methods proposed here for detecting and estimating linkage disequilibrium are investigated analytically or by simulation. It is shown that linkage disequilibrium under a defined spectrum of inheritance models may be detected with 80% power using a sample size of a few hundred to 2000. The contingency analysis and the regression analysis developed in this study provide the disequilibrium test, and the maximum-likelihood approach presented here can be used to estimate the model parameters, *i.e.*, gene frequencies at both the marker and trait loci, the coefficient of linkage disequilibrium between the two loci, and the genetic effects of the genes. These estimates may be useful in interpreting the demographic history of the natural populations under question (Thompson and Neel 1996; Fay and Wu 1999) and, in turn, to extend the principle of linkage disequilibrium-based mapping to complex traits.

The major difficulty in linkage disequilibrium-based mapping is to quantify the relationship between recombination fraction and linkage disequilibrium measure. Since recombinant events are not observed, the recombination fraction between the marker and trait locus must be estimated on the basis of a population genetics model. Several methods were suggested to address this problem. One of these attempted to search for the reparameterization by which the disequilibrium measure can be directly related to the recombination fraction. For instance, Devlin and Risch (1995) found that, in the present notations, the measure of the disequilibrium

$$\delta = \frac{D}{q[(1-p)(1-q)+D]}$$

has some interesting properties. In certain situations, the disequilibrium measure is related to the recombination fraction r as $\delta = (1-r)^T$, where T represents the generation number since the creation of the initial disequilibrium and may be estimated either from epidemiological survey (*i.e.*, in Hastbacka *et al.* 1992) or directly from the sampled data (*i.e.*, Kaplan *et al.* 1995; Thompson and Neel 1996). Kaplan and Weir (1997) proposed a simulation-based approach, which allows the maximum-likelihood estimate of the recombination fraction and its confidence interval to be estimated entirely on the basis of the observation of linkage disequilibrium between a polymorphic marker locus and a simple monogenic disease locus. These analyses were confined to the circumstances where the genotypes at the trait locus can be observed. However, the basic idea may be extended to the case where the genotypes at the trait locus are not observed such as the circumstances considered in this study. In fact, the information can

be, at least partially, uncovered for the joint distribution of genotypes at both the marker and trait loci using the maximum-likelihood estimates of the parameters, p, q, and D. The predicted recombination fraction will bear a larger sampling variation since q and D have to be estimated from data with incomplete information.

Development of a dense map of biallelic single-nucleotide polymorphisms (SNPs) was suggested to be an effective strategy for genome-wide search for linkage disequilibrium between the polymorphisms and candidate genes. Under that setting, a proportion of SNP markers are considered within the candidate genes; thus the recombination fraction between the marker and trait loci may be assumed to be zero (Martin *et al.* 2000). However, efficiency of the genome scan scheme with SNPs is still controversial (Kruglyak 1999; Ott 2000).

In this study, we considered inference of linkage disequilibrium using a random sample from natural populations. In practice, selected samples may be favored in relevant genetics studies. In principle, this study can be extended to analyze the selected samples. We note that the joint genotypic distribution at the marker and trait loci within the selected sample (say, for example, ξ_{ij}) is related to that in the natural population (h_{ij} given in Table 1) from which the sample is collected, as

$$\xi_{ij} = \frac{[\phi f_j + (1 - \phi - f_j)\overline{\eta}]}{\overline{\eta}(1 - \overline{\eta})} h_{ij},$$

where ϕ represents the proportion of individuals with the phenotype ($y = 1$) in the sample and $\overline{\eta} = q^2 f_1 + 2q(1 - q)f_2 + (1 - q)^2 f_3$. It may take different values for different data sets in practice. Replacement of h_{ij} by ξ_{ij} in the theory presented above allows nonrandomly sampled data sets to be analyzed straightforwardly using the methods. In addition, the analyses based on choosing appropriate ϕ open a window of flexibility for screening the optimized sampling schemes, which may yield the most efficient detection of linkage disequilibrium for a given set of parameters.

The methods presented here differ from others in several aspects. Allison (1997) extended the transmission disequilibrium test (TDT) theory to detect association between the marker locus and a locus contributing to quantitative genetic variation using family data. As a member of the TDT family, Allison's method can detect linkage between the marker locus and the trait locus only if linkage disequilibrium is present. An important assumption made in the TDT model is that the marker locus is the trait locus itself and is not just in linkage disequilibrium with the trait locus, indicating that the genotype information at the trait locus is assumed. This does not apply to the methods proposed in this study where information on the trait locus is missing. Rabinowitz (1997) generalized the TDT test for analyzing quantitative traits, but no effort was made to estimate genetic parameters of the traits. However, it must be noted that a distinct feature of the TDT methods is their robustness to population stratification. The question remains how this analysis may be affected by stratification. This may be tackled readily if the dynamics and properties of linkage disequilibrium in admixed populations are taken into account as we recently showed in Tao *et al.* (2000).

In a recent study, Nielsen and Weir (1999) proposed the model for association studies between a marker and a trait and showed that there is a simple relationship between the marker being examined and the trait loci. Their model is a very useful framework for specifying the structure of linkage disequilibrium between the loci and for investigating the genetic context of many marker-based statistic tests. Our main purpose is, however, to directly estimate the linkage disequilibrium and the genetic parameters of the dichotomous trait locus. Sharing the same theoretical model of gene segregation at both marker and trait loci, our previous studies (Luo 1998; Luo *et al.* 2000) focused on inferring linkage disequilibrium between a polymorphic marker locus and a locus affecting continuous complex genetic traits. In this study, however, the trait locus was assumed to affect a complex dichotomous phenotype. We found that the disequilibrium test with the dichotomous traits is generally less efficient than the test where the traits display a continuous phenotypic distribution. Furthermore, a statistical method was suggested in Xu and Atchley (1996) for mapping quantitative trait loci underlying complex binary diseases using planned crossing experiments. The genetic structure of the experiment population under the quantitative trait loci (QTL) mapping consideration allows more information about the QTL genotype to be extracted from the flanking markers because the parental linkage phases of genes at the marker and QTL loci are assumed known, and thus it provides a direct prediction of cosegregation of marker genes and genes at the QTL. With natural populations, none of the information is available.

Our model considers a biallelic model at both marker and trait loci. For two reasons, multiple alleles at the marker loci may need to be taken into account. First, multiple alleles may be common in natural populations for molecular DNA markers like microsatellites. Second, analysis of haplotypes over several marker loci may be effectively reduced to a multiallelic model as demonstrated in Service *et al.* (1999) and Clayton and Jones (1999). In principle, our model may be integrated into the multiallelic analysis by taking into account the probability that each of the marker alleles is in linkage disequilibrium with the trait allele A. If, for example, the ith marker allele is in linkage disequilibrium with the allele A, the rest of marker alleles are sorted into one class. With this strategy, a likelihood-based modeling of the multiallelic marker such as that proposed in Terwilliger (1995) may thus be tractable. It is much more

challenging to consider multiple alleles at the trait locus because the individual allelic effect cannot be identified with certainty from the trait phenotype. The problem was usually addressed by the way that individuals with a specific phenotype were recognized as carrying one class of alleles and the rest without the phenotype were assigned as carrying other class of alleles (Lazzeron 1998; Service *et al.* 1999). In this study, the putative allele *A* may be considered as a set of alleles that increase the presence of the trait phenotype ($y = 1$) while the allele *a* is a set of alleles that decrease the presence of the phenotype.

Moreover, we recognize there are several multiloci linkage disequilibrium mapping methods. Some of them are based on comparison of pairwise linkage disequilibria between the single trait locus and a set of marker polymorphisms and use the peak value of disequilibrium measure over several marker loci as evidence for location of the hypothesized QTL (*e.g.*, Terwilliger 1995; Rannala and Slatkin 1998; Slatkin 1999). In addition, some theoretical effort has been made to combine the pairwise disequilibrium between the putative trait locus and each of a set of marker loci by use of the composite likelihood principle (Devlin *et al.* 1996; Collins and Morton 1998). Our study provides statistical inference of linkage disequilibrium between a single marker locus and any one of the loci underlying complex genetic variation; hence the multiple-loci analysis could be built upon the two-locus model. However, more work will be needed to extend the method to appropriately taking multiple marker information into account and analyzing multiple trait loci. Such an extension would certainly need to take into account the complex structure of the disequilibria among multiple loci (Weir 1979) or at least a major part of it. We hope this study and analysis will provide a basis for a multilocus analysis and serve as an important building block for our understanding of the genetic architecture underlying complex genetic variation.

We are indebted to Mary Sara McPeek for her comments on the manuscript. We thank Richard Hudson for the discussion on the topic of linkage disequilibrium mapping. The comments and criticisms made by an anonymous reviewer and Dr. Marjorie A. Asmussen were very helpful in improving presentation and clarifying several ambiguities in an earlier version of this article. Z.W.L. is supported by China's Basic Research Program "973," the National Science Foundation of China, the QiuShi Foundation, and the Changjiang Scholarship. C.I.W. is supported by U.S. National Institutes of Health and National Science Foundation grants.

LITERATURE CITED

Agresti, A., 1990 *Categorical Data Analysis*. John Wiley & Sons, New York.

Allison, D. B., 1997 Transmission disequilibrium tests for quantitative traits. Am. J. Hum. Genet. **60:** 676–690.

Brakefield, P. M., 1996 Development, plasticity and evolution of butterfly eyespot patterns. Nature **384:** 236–242.

Brown, A. D. H., 1975 Sample sizes required to detect linkage disequilibrium between two or three loci. Theor. Popul. Biol. **8:** 184–201.

Churchill, G. A., and R. W. Doerge, 1994 Empirical threshold for quantitative trait mapping. Genetics **128:** 963–971.

Clark, A. G., K. M. Weiss, D. A. Nickerson, S. L. Taylor, A. Buchanan *et al.*, 1998 Haplotype structure and population-genetics inferences from nucleotide-sequence variation in human lipoprotein lipase. Am. J. Hum. Genet. **63:** 595–612.

Clayton, D., and H. Jones, 1999 Transmission disequilibrium tests for extended marker haplotypes. Am. J. Hum. Genet. **65:** 1161–1169.

Collins, A., and N. E. Morton, 1998 Mapping a disease locus by allelic association. Proc. Natl. Acad. Sci. USA **95:** 1741–1745.

Coyne, J. A., A. O. Crittenden and K. Mah, 1994 Genetics of a pheromonal difference contributing to reproductive isolation in Drosophila. Science **265:** 1461–1464.

Darvasi, A., 1998 Experimental strategies for the genetic dissection of complex traits in animal models. Nat. Genet. **18:** 19–24.

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

de la Chapelle, A., and F. A. Wright, 1998 Linkage disequilibrium mapping in isolated populations: the example of Finland revisited. Proc. Natl. Acad. Sci. USA **95:** 12416–12423.

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

Devlin, B., N. Risch and K. Roeder, 1996 Disequilibrium mapping: composite likelihood for pairwise disequilibrium. Genomics **36:** 1–16.

Fay, J. C., and C.-I Wu, 1999 A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. Mol. Biol. Evol. **16:** 1003–1005.

Guo, S. W., and K. Lange, 2000 Genetic mapping of complex traits: promises, problems, and prospects. Theor. Popul. Biol. **57:** 1–11.

Hastbacka, J., A. de la Chapelle, I. Kaitila, P. Sistonen, A. Weaver *et al.*, 1992 Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nat. Genet. **2:** 204–211.

Hill, W. G., 1974 Estimation of linkage disequilibrium in randomly mating populations. Heredity **33:** 229–239.

Hill, W. G., 1975 Tests for association of gene frequencies at several loci in random diploid populations. Biometrics **31:** 881–888.

Johnson, N. L., and N. Kotz, 1970 *Distributions in Statistics: Continuous Univeriate Distributions*. Houghton Mifflin, Boston.

Kaplan, N., W. G. Hill and B. S. Weir, 1995 Likelihood methods for locating disease genes in nonequilibrium populations. Am. J. Hum. Genet. **56:** 18–32.

Kaplan, N. L., and B. S. Weir, 1997 The use of linkage disequilibrium for estimating the recombination fraction between a marker and a disease gene, pp. 207–219 in *Progress in Population and Human Evolution*, edited by P. I. Donnelly and S. Tavare. Springer-Verlag, New York.

Kendall, M. G., A. Stuart and J. K. Ord, 1983 *The Advanced Theory of Statistics vol. 1, Distribution Theory*. Charles Griffin & Company, London.

Kearsey, M. J., and A. G. L. Farquhar, 1998 QTL analysis in plants: Where are we now? Heredity **80:** 137–142.

Kruglyak, L., 1999 Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat. Genet. **22:** 139–144.

Lande, R., and R. Thompson, 1990 Efficiency of marker-assisted selection in improvement of quantitative traits. Genetics **124:** 743–756.

Lander, E. S., 1996 The new genomics: global views of biology. Science **274:** 536–539.

Lander, E. S., and N. L. Schork, 1994 Genetic dissection of complex traits. Science **265:** 2037–2048.

Lazzeron, R. C., 1998 Linkage disequilibrium and gene mapping: an empirical least-square approach. Am. J. Hum. Genet. **62:** 159–170.

Lewontin, R. C., 1974 *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York.

Little, R. J. A., and D. B. Rubin, 1987 *Statistical Analysis With Missing Data*. John Wiley & Sons, New York.

Luo, Z. W., 1998 Detecting linkage disequilibrium between a polymorphic marker locus and a trait locus in natural populations. Heredity **80:** 198–208.

Luo, Z. W., and S. Suhai, 1999 Estimating linkage disequilibrium between a polymorphic marker locus and a trait locus in natural populations. Genetics **151:** 359–371.
Luo, Z. W., R. Thompson and J. A. Woolliams, 1997 A population genetics model of marker-assisted selection. Genetics **146:** 1173–1183.
Luo, Z. W., S. H. Tao and Z-B. Zeng, 2000 Inferring linkage disequilibrium between a polymorphic marker locus and a trait locus in natural populations. Genetics **151:** 457–467.
Martin, E. R., E. R. Lai, J. R. Gilbert, A. R. Rogala, A. J. Afshari *et al.*, 2000 SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in Alzheimer disease. Am. J. Hum. Genet. **67:** 383–394.
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.
Ott, J., 2000 Predicting the range of linkage disequilibrium. Proc. Natl. Acad. Sci. USA **97:** 2–3.
Paabo, S., 1999 Human evolution. Trends Genet. **15:** 13–16.
Press, W. H., B. P. Flanner, S. A. Teukolsky and W. T. Vellerting, 1992 *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, New York.
Rabinowitz, D., 1997 A transmission disequilibrium test for quantitative trait loci. Hum. Hered. **47:** 342–350.
Rannala, B., and M. Slatkin, 1998 Likelihood analysis of disequilibrium mapping and related problems. Am. J. Hum. Genet. **62:** 459–473.
Risch, N., and K. Merikangas, 1996 The future of genetic studies of complex human diseases. Science **273:** 1516–1517.
Service, S. K., D. W. Lang, N. B. Freimer and L. A. Sandkuijl, 1999 Linkage disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. Am. J. Hum. Genet. **64:** 1728–1738.
Slatkin, M., 1999 Disequilibrium mapping of a quantitative trait locus in an expanding population. Am. J. Hum. Genet. **64:** 1765–1773.
Spielman, R. S., R. E. McGinnis and W. J. Ewens, 1993 Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am. J. Hum. Genet. **52:** 506–516.
Tao, S. H., X. M. Liu and Z. W. Luo, 2000 A population genetics model of linkage disequilibrium in admixed populations. Chin. Sci. Bull. **45:** 2041–2047.
Terwilliger, J. D., 1995 A powerful likelihood method for the analysis of linkage disequilibrium between trait locus and one or more polymorphic marker loci. Am. J. Hum. Genet. **49:** 31–41.
Thompson, E. A., and J. V. Neel, 1996 Allelic disequilibrium and allele frequency distribution as a function of social and demographic history. Am. J. Hum. Genet. **60:** 197–204.
Thompson, R., 1972 The maximum likelihood approach to the estimate of liability. Ann. Hum. Genet. Lond. **36:** 221–231.
Weir, B. S., 1979 Inferences about linkage disequilibrium. Biometrics **35:** 235–254.
Weir, B. S., 1990 *Genetic Data Analysis*. Sinauer, Sunderland, MA.
Weir, B. S., and C. C. Cockerham, 1978 Testing hypotheses about linkage disequilibrium with multiple alleles. Genetics **88:** 633–642.
Wu, C. I., and M. F. Palopoli, 1994 Genetics of postmating reproductive isolation in animals. Annu. Rev. Genet. **28:** 283–308.
Xu, S., and W. R. Atchley, 1996 Mapping quantitative trait loci for complex binary diseases using line crosses. Genetics **143:** 1417–1424.

Communicating editor: M. A. Asmussen

APPENDIX A: CALCULATION OF THE NONCENTRALITY PARAMETER OF THE χ^2-TEST STATISTIC

In Table 2, let $T = 1, 2, 3$ correspond to the three marker genotypes *MM*, *Mm*, and *mm*, respectively, and $Y = 1, 2$ accord to the affected or normal phenotype, and let G be an indicator of genotype at the trait locus. The probability

$$\begin{aligned}\pi_{ij} &= \Pr\{Y = i,\ T = j\} = \Pr\{T = j\}\Pr\{Y = i|T = j\}\\ &= \Pr\{T = j\}\sum_{k=1}^{3}\Pr\{G = k|T = j\}\Pr\{Y = i|G = k,\ Y = j\}\\ &= \Pr\{T = j\}\sum_{k=1}^{3}\Pr\{G = k|T = j\}\Pr\{Y = i|G = k\}\end{aligned}$$

since the marker genotype gives no extra information when the genotype at the trait locus is known, where conditional probability $\Pr\{G = k|T = j\} = h_{jk}$ can be found in Table 1 and $\Pr\{Y = i|G = k\}$ represents penetrance of the three trait genotypes, which, by definition, is f_k. Given the joint probability π_{ij}, the corresponding marginal probability can be calculated and denoted by π_{+j} and π_{i+}. Under the null hypothesis, $\pi_{ij}(M) = \pi_{i+}\pi_{+j}$, the noncentrality parameter is calculated, according to Agresti (1990, p. 239–250), as $\lambda = n\Sigma_i\Sigma_j[\pi_{ij} - \pi_{ij}(M)]^2/\pi_{ij}(M)$. Equation 4 is derived after substituting the probabilities with their expressions in terms of the model parameters and after simplification algebra.

APPENDIX B: CALCULATION OF THE REGRESSION PARAMETERS AND THE SAMPLE VARIANCE OF THE REGRESSION COEFFICIENT

Let Y and T be trait phenotype and number of marker allele M carried by sampled individual. Y takes a value of 0 or 1, and T could be 0, 1, or 2. Under the assumption of a random mating population, the basic statistics of T can be readily derived as

$$\begin{aligned}\mu_1 &= E(T) = 2p\\ \mu_2 &= E(T^2) = 2p(1 + p)\\ \mu_3 &= E(T^3) = 2p(1 + 3p)\\ \mu_4 &= E(T^4) = 2p(1 + 7p)\\ \sigma_T^2 &= 2p(1 - p),\end{aligned}$$

the basic statistics of the distribution of Y are

$$\begin{aligned}\nu_r &= E(Y^r) = \Pr\{Y = 1\} = \sum_{k=1}^{3}\Pr\{G = k\}\Pr\{Y = 1|G = k\}\\ &= q^2f_1 + 2q(1 - q)f_2 + (1 - q)^2f_3 \quad (r = 1, 2, \ldots)\\ \sigma_Y^2 &= E(Y^2) - E(Y)^2 = \Pr\{Y = 1\} - \Pr\{Y = 1\}^2\\ &= \Pr\{Y = 1\}(1 - \Pr\{Y = 1\}) = \Pr\{Y = 1\}\Pr\{Y = 0\}\\ &= \sum_{k=1}^{3}\Pr\{G = k\}\Pr\{Y = 1|G = k\}\sum_{k=1}^{3}\Pr\{G = k\}\Pr\{Y = 0|G = k\}\\ &= (q^2f_1 = 2q(1 - q)f_2 + (1 - q)^2f_3)\\ &\quad \times [1 - (q^2f_1 + 2q(1 - q)f_2 + (1 - q)^2f_3)],\end{aligned}$$

and the basic statistics for the joint distribution between Y and T are calculated by noting that

$$\begin{aligned}\Pr\{T=2,\ Y=1\} &= \Pr\{T=2\}\Pr\{Y=1|T=2\}\\ &= \Pr\{T=2\}\sum_{k=1}^{3}\Pr\{G=k|T=2\}\\ &\quad\times\Pr\{Y=1|G=k,\ T=2\}\\ &= \sum_{k=1}^{3}\Pr\{G=k,\ T=2\}\Pr\{Y=1|G=k\} = \sum_{k=1}^{3}h_{1k}f_k\end{aligned}$$

and

$$\Pr\{T=1,\ Y=1\} = \sum_{k=1}^{3}h_{2k}f_k,$$

where the joint probabilities h_{ik} are listed in Table 1, and then

$$\begin{aligned}w_{1r} &= E(TY^r) = 2\Pr\{T=2,\ Y=1\} + \Pr\{T=1,\ Y=1\}\\ &= 2\{[f_2 - f_3 + (f_1 - 2f_2 + f_3)q]D\\ &\quad + p[f_3(1-q)^2 + (f_1q + 2(1-q)f_2)q]\}\\ w_{2r} &= E(T^2Y^r) = \{(f_1 = 2f_2 + f_3)D^2\\ &\quad + (1+2p)[f_2 - f_3 + (f_1 - 2f_2 + f_3)q]D\\ &\quad = +p(1-p)[f_3(1-q)^2 + (2f_2 + f_1q - 2f_2q)q]\}\\ w_{3r} &= E(T^3Y^r) = 2\{3(f_1 - 2f_2 + f_3)D^2\\ &\quad + (1-6p)[f_2 - f_3 + (f_1 - 2f_2 + f_3)q]D\\ &\quad + p(1+3p)[f_3(1-q)^2 + (2f_2 + f_1q - 2f_2q)q]\}\\ &\quad (r = 1, 2, \ldots)\end{aligned}$$

$$\sigma^2_{TY} = E(TY) - E(T)E(Y) = 2D[qf_1 + (1-2q)f_2 - (1-q)f_3].$$

Therefore the coefficient of regression of Y on T is derived as the form given by Equation 8 in this article. The correlation coefficient between Y and T is $r = D[qf_1 + (1-2q)f_2 - (1-q)f_3]/[p(1-p)(1-q^2f_1 - 2q(1-q)f_2 - (1-q)^2f_3)(q^2f_1 + 2q(1-q)f_2 + (1-q)^2f_3)]$. The variance and covariance in Equation 10 are calculated as

$$\begin{aligned}\mathrm{Var}(\sigma^2_T) &= \frac{1}{n}(\mu_4 - \mu_2^2) + \frac{1}{n^3}[\mu_4 - \mu_2^2 + 4(n-1)(n-2)\mu_1^2\sigma^2_T\\ &\quad + 4(n-1)\mu_1(\mu_3 - \mu_1\mu_2)]\\ &\quad - \frac{2}{n^2}[\mu_4 - \mu_2^2 + 2(n-1)\mu_1(\mu_3 - \mu_1\mu_2)]\\ \mathrm{Var}(\sigma^2_{TY}) &= \frac{(n-1)^2}{n_3}(w_{22} - w_{11}^2)\\ &\quad + \frac{(n-1)}{n^3}[\mu_2\nu_2 - \mu_1^2\nu_1^2 + 2(w_{11}^2 + \mu_1^2\nu_1^2)\\ &\quad + (n-2)(\mu_1^2\sigma^2_Y + \nu_1^2\sigma^2_T + 2\mu_1\nu_1\sigma^2_{TY})]\\ &\quad - \frac{2(n-1)^2}{n^3}(\mu_1w_{12} - 2\mu_1\nu_1w_{11} + \nu_1w_{21})\end{aligned}$$

and

$$\begin{aligned}\mathrm{Cov}(\sigma^2_T, \sigma^2_{TY}) &= \frac{1}{n}(w_{31} - \mu_2w_{11})\\ &\quad - \frac{1}{n^2}[w_{31} - \mu_2w_{11} + 2(n-1)(\mu_1w_{21} - \mu_1^2w_{11})]\\ &\quad\times\frac{1}{n^2}[w_{31} - \mu_2w_{11} + (n-1)(\mu_3\nu_1 + \mu_1w_{21} - 2\mu_1\nu_1\mu_2)]\\ &\quad\times\frac{1}{n^3}\{w_{31} - \mu_2w_{11} + (n-1)(\mu_1w_{21} - \mu_1\nu_1w_{11})\\ &\quad + (n-1)(\mu_3\nu_1 + \mu_1w_{21} - 2\mu_1\nu_1\mu_2)\\ &\quad + 2(n-1)[\mu_2w_{11} - \mu_1^3\nu_1 + (n-2)(\mu_1^2\sigma^2_{TY} + \mu_1\nu_1\sigma^2_T)]\}.\end{aligned}$$

APPENDIX C: CALCULATION OF MLE OF THE THRESHOLD θ

Let ξ_0 be the proportion of unaffected individuals sampled. It can be used to approximate its population probability (Thompson 1972), and, in the present context,

$$\begin{aligned}\hat{\xi}_0 &= \Pr\{Y=0\} = \sum_{k=1}^{3}\Pr\{G=k\}\Pr\{Y=0|G=k\}\\ &= \sum_{k=1}^{3}\left[\sum_{j=1}^{3}\Pr\{M=j\}\Pr\{G=k|M=j\}\right]\Pr\{Y=0|G=k\}\\ &= \sum_{k=1}^{3}\left(\sum_{j=1}^{3}\hat{m}_jh_{jk}\right)\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\theta}\exp\left[-\frac{(z-\mu-(2-k)a-(-1)^kd/2)^2}{2}\right]dz,\end{aligned}$$

where $\hat{m}_j$ represents the observed frequency of the jth marker genotype and the conditional probability h_{jk} is given in Table 1. Given the other model parameters, the above equation can be numerically solved for θ.

APPENDIX D: SOLVING FOR THE MLES OF THE MODEL PARAMETERS

Let ξ be either μ, a, or d, then the first derivative of the expected complete data log-likelihood, which is defined by Equation 14, with respect to ξ can be

$$\begin{aligned}\frac{\partial}{\partial\xi}L_c(Y,\phi) &= \sum_{i=1}^{3}\sum_{j=1}^{n_i}\sum_{k=1}^{3}w_{ijk}\\ &\quad\times\left[y_{ij}\frac{\partial}{\partial\xi}\log(f_k) + (1-y_{ij})\frac{\partial}{\partial\xi}\log(1-f_k)\right]\\ &= \sum_{i=1}^{3}\sum_{j=1}^{n_i}\sum_{k=1}^{3}w_{ijk}\frac{(y_{ij}-f_k)}{f_k(1-f_k)}\frac{\partial}{\partial\xi}f_k.\end{aligned}$$

If we let $g_k = \mu + (2-k)a + (-1)k/2$, then

$$\begin{aligned}\frac{\partial}{\partial\xi}f_k &= -\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\theta}\frac{\partial}{\partial\xi}\left(\exp\left[-\frac{(z_{ij}-g_k)^2}{2}\right]\right)dz_{ij}\\ &= \frac{1}{\sqrt{2\pi}}\left\{\int_{-\infty}^{\theta}\exp\left[-\frac{(z_{ij}-g_k)^2}{2}\right]d\left[-\frac{(z_{ij}-g_k)^2}{2}\right]\right\}\frac{\partial}{\partial\xi}(g_k)\\ &= \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{(z_{ij}-g_k)^2}{2}\right]\frac{\partial}{\partial\xi}(g_k)\end{aligned}$$

with $(\partial/\partial\xi)(g_k) = 1$, $(2 - k)$ or $(-1)^k/2$ when $\xi = \mu$, a, or d correspondingly.

The second derivatives are more messy but a general form was found as

$$\frac{\partial^2}{\partial\xi\partial\eta} Lf_c(Y, \phi) = \sum_{i=1}^{3}\sum_{j=1}^{n_i}\sum_{k=1}^{3} \times w_{ijk}\left\{\frac{(y_{ij} - f_k)}{f_k(1 - f_k)}\frac{\partial^2}{\partial\xi\partial\eta} f_k - \frac{1}{f_k^2(1 - f_k^2)} \times [f_k(1 - f_k) + (y_{ij} - f_k)(1 - 2f_k)] \times \frac{\partial}{\partial\xi}(g_k)\frac{\partial}{\partial\eta}(g_k)\right\},$$

where

$$\frac{\partial^2}{\partial\xi\partial\eta} f_k = \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{(z_{ij} - g_k)^2}{2}\right](\theta - g_k)\frac{\partial}{\partial\xi}(g_k)\frac{\partial}{\partial\eta}(g_k) + \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{(z_{ij} - g_k)^2}{2}\right]\frac{\partial^2}{\partial\xi\partial\eta} g_k$$

and

$$\frac{\partial^2}{\partial\xi\partial\eta} g_k = \begin{pmatrix} \frac{\partial^2}{\partial\mu^2} & \frac{\partial^2}{\partial\mu\partial a} & \frac{\partial^2}{\partial\mu\partial d} \\ & \frac{\partial^2}{\partial a^2} & \frac{\partial^2}{\partial a\partial d} \\ & & \frac{\partial^2}{\partial d^2} \end{pmatrix} g_k = \begin{pmatrix} 1 & (2 - k) & (-1)^k/2 \\ & (2 - k)^2 & (-1)^k(2 - k)/2 \\ & & 1/4 \end{pmatrix} \times \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{(\theta - g_k)^2}{2}\right](\theta - g_k).$$

With these derivatives, the MLEs of parameters μ, a, and d can be obtained using the numerical algorithm, for example, in Press *et al.* (1992), which solve the equation

$$\frac{\partial}{\partial\xi} L_c(Y, \phi) = 0$$

for $\xi = \mu$, a, d accordingly.