Heredity **80** (1998) 198–208 *Received 25 February 1997*

Detecting linkage disequilibrium between a polymorphic marker locus and a trait locus in natural populations

Z. W. LUO*

Laboratory of Population and Quantitative Genetics, Institute of Genetics, Fudan University, 220 Handan Road, Shanghai 200433, China

A novel statistical model was developed to test for linkage disequilibrium between a polymorphic genetic marker locus and a locus underlying a quantitative trait (QTL) in natural populations using principles of analysis of variance of unbalanced data and analysis of regression involving data of non-normal distribution. Powers of these statistical tests are formulated as functions of census population size, allelic frequencies at the marker locus and the trait locus, additive and dominance effects at the QTL as well as the coefficient of linkage disequilibrium. Theoretical predictions of the power are validated by extensive Monte Carlo simulations. Among all these factors examined, the amount of the disequilibrium and the size of effect of the QTL are of most importance in determining the power, and the dominance and the allele frequencies at the two loci have substantial effects on the power. Numerical analyses based upon the theoretical calculations and simulation studies favour use of regression of the number of marker alleles on the trait phenotypes as a measure of detection of linkage disequilibrium. Theoretical analysis is also performed to investigate robustness of the formula for predicting the variance of the regression coefficient, which requires normality of the regression variables, whereas normality may not be strictly warranted.

Keywords: linkage disequilibrium, marker, QTL, statistical power.

Introduction

Linkage disequilibrium has been of great value in two areas of theoretical genetics studies: mapping quantitative trait loci (QTL) (Lander & Schork, 1994) and marker-assisted selection (MAS) (Lande & Thompson, 1990). In principle, the objectives of QTL location and MAS could be essentially achieved by detecting significant linkage disequilibrium between genetic loci affecting quantitative variation and polymorphic genetic marker loci as the first step.

Many researchers have focused on statistical methods for detecting the presence of or estimating the coefficient of linkage disequilibrium between two or more loci, at each of which there may be two or more alleles segregating. Hill (1974) developed likelihood-based procedures for estimating the coefficient of linkage disequilibrium between two loci in a finite random mating population. Brown (1975) established a theoretical framework for the sample sizes required to detect the disequilibrium by the use of data on gametic and zygotic frequencies. When there are multiple alleles segregating at the loci, a statistical procedure was suggested in Weir & Cockerham (1978) for calculating the power of testing the linkage disequilibrium. Furthermore, Weir (1979) presented a comprehensive discussion of the efficiency of using different sorts of data and statistical strategies from which linkage disequilibrium could be detected or estimated. These theories or techniques for inferences about linkage disequilibrium are, however, restricted to the circumstance where gametic or genotypic frequencies are observed directly.

The difficulties encountered in modelling linkage disequilibria involved with quantitative trait loci are mainly caused by the unavailability of genotypic data on the traits. Hill & Robertson (1966, 1968) demonstrated the predictability of the expected dynamics

*E-mail: zwluo@ms.fudan.sh.cn

of linkage disequilibrium between a pair of linked QTLs in finite populations with or without selection. Since the abundance of genetic polymorphisms at the DNA molecular level was discovered in nearly all organisms, many statistical methods have been suggested for detecting linkage disequilibrium between a quantitative trait locus and a genetic marker locus segregating in populations with various structures (Soller & Genizi, 1978; Luo, 1993; Knott, 1994; Le Roy & Elsen, 1995). In these studies, linkage disequilibria between a QTL and a marker locus were assumed to be generated by hybridization between inbred lines or strains, i.e. the putative QTL was linked to the marker locus.

In addition to hybridization, linkage disequilibrium between a pair of loci can be produced by random drift, mutation, selection, merging of populations and by nonrandom mating, although the magnitude of the disequilibrium is maintained by the recombination frequency between the loci (e.g. Hartl & Clark, 1989). The aim of this paper is to develop two theoretical procedures with generality for detecting linkage disequilibrium between a QTL and a genetic marker locus and for formulating the statistical powers of the relevant statistical tests. Numerical analyses based upon intensive simulation studies are used to illustrate the validation of theoretical analyses and to confirm the accuracy of theoretical predictions.

Model

The method involves analysing a population of census size n. Two autosomal loci are assumed: one affects a quantitative trait (QTL) whereas the other is a codominant marker which has no direct effect on the trait. The two alleles are denoted by M and m at the marker locus and by A and a at the QTL. The phenotype of the trait (Z) is assumed to have an enviromental variance σ_e^2 and to be normally distributed, although this is only necessary where stated. Three genotypes at the QTL, say AA, Aa and aa, are assumed to affect the quantitative trait by d, h and $-d$, respectively. The frequencies of M and A in the population are denoted by p and q, respectively. Genotypic value at the marker locus is denoted by T which is the number of alleles M.

The distribution of the QTL genotypes within each of three possible marker genotypes is illustrated in Table 1, where Q (or R) is the frequency of allele A at the QTL among chromosomes carrying M (or m), which is a function of allelic frequencies at the marker and QTL and the linkage disquilibrium between the two loci, say for example D. The rela-

Table 1 Distribution of genotypes at the marker locus and QTL. p is the frequency of marker allele M, and Q (or R) represents the frequency of allele A at the QTL among chromosomes carrying the marker allele M (or m); d and h are additive and dominance effects at the QTL

Marker genotypes	MM			Mm			mm		
Genotypes at QTL	AA	Aa	aa	AA	Aa	aa	AA	Aa	aa
Frequencies (f_{ij})	p^2Q^2	$2p^2Q(1-Q)$	$p^2(1-Q)^2$	$2p(1-p)QR$	$2p(1-p)(Q+R-2QR)$	$2p(1-p)(1-Q)(1-R)$	$(1-p)^2R^2$	$2(1-p)^2R(1-R)$	$(1-p)^2(1-R)^2$
Genotypic values	d	h	$-d$	d	h	$-d$	d	h	$-d$

tionships among Q, R and D can be derived by simple algebra: $Q = q + D/p$, $R = q - D/(1-p)$ and $D = p(1-p)(Q-R)$.

It must be noticed that the theoretical model described in Table 1 implies random union of gametes with respect to the genotypes at both the marker locus and the QTL.

Theoretical analyses

The statistical model for the kth individual with the jth QTL genotype and the ith marker genotype in the population can be written as follows

$$y_{ijk} = \mu + \beta_i + \omega_{ij} + \varepsilon_{ijk}, \tag{1}$$

where μ is the population mean, β_i is the effect of marker genotype i, ω_{ij} is the effect of the jth QTL genotype within the ith marker genotype, and ε_{ijk} is the residual effect whose distribution is normal with mean zero and variance σ^2_e. The residual variance accounts for variation of polygenes which are in linkage equilibrium with the marker and for the environmental variation. Under model (1), the population can be analysed by either of following procedures.

Analysis of variance

In model (1), the between- and within-marker genotype effects might be regarded as random effects (e.g. Jayakar, 1970; Hill, 1975), and then it can be worked out that the expected variance component between the marker genotypes is

$$\sigma^2_\beta = \frac{D^2}{p(1-p)}\left\{d^2 + \left[1 - 4q + \frac{2(D^2 + 2pq(q-pq))}{p(1-q)}\right]h^2 + 2(1-2q)dh\right\} \tag{2.1}$$

and the expected within-marker genotype variance component is

$$\sigma^2_\omega = \left[q - \frac{D^2 + pq(q-pq)}{p(1-q)}\right]\times \left\{d^2 + \left[1 - 2q + \frac{2(D^2 + pq(1-pq))}{p(1-q)}\right]h^2 + 2(1-2p)dh\right\}. \tag{2.2}$$

It has been shown that expected mean squares in the analysis of variance model with random effects could be biased downwards because of a highly unbalanced hierarchical structure of the data (Soller & Genizi, 1978; Luo, 1993; Knott, 1994). An assumption of fixed QTL effect has already been widely made in these studies and in Knapp & Bridges (1990).

Under the fixed model of the QTL effect, it can be readily shown that the between-marker genotype effects are

$$\beta_1 = \frac{2D[pd - (D - p + 2pq)h]}{p^2} \tag{3.1}$$

$$\beta_2 = \frac{D[(1-2p)d + (2D + (1-2p)(1-2q))h]}{p(1-p)} \tag{3.2}$$

$$\beta_3 = \frac{-2D[(1-p)d + (D + (1-p)(1-2q))h]}{(1-p)^2}. \tag{3.3}$$

The eqns (2) and (3) illustrate that under either the random or fixed model of the QTL effect, significant variation between the marker genotypes in the QTL effect is an indicator of the presence of linkage disequilibrium between the marker and QTL. The expected mean squares between and within marker genotypes under the fixed model are given in Appendix I.

Regression analysis

Lande & Thompson (1990) suggested the use of regression of phenotypic records of the quantitative trait on the number of alleles of marker loci as a marker score in a selection index of marker-assisted selection of a quantitative trait. In the present model, the regression coefficient is

$$b = \frac{D[d + (1-2q)h]}{p(1-p)}. \tag{4}$$

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

$$\sigma^2_b = \frac{(1-r^2)\sigma^2_Z}{n\sigma^2_T}, \tag{5.1}$$

where r is the correlation coefficient between Z and T, σ^2_Z is the phenotypic variance of the trait and σ^2_T is the genetic variance at the marker locus. However, because the number of marker alleles (M)

carried by each individual is a discrete random variable, and its median and arithmetic mean may not be consistent if the frequencies of the marker alleles are not equal, the normality approximation of its distribution may not be appropriate. In addition, the distribution of Z is not normal but a mixture of three normal subpopulations as described in model (1). A general formula for calculation of the sample variance of the regression coefficient can be derived following Kendall *et al*. (1983, p. 325):

$$\sigma_b^2 = b^2\left\{\frac{\mathrm{Var}[\mathrm{Cov}(T,Z)]}{\mathrm{Cov}(T,Z)}+\frac{\mathrm{Var}[\sigma_T^2]}{\sigma_T^2}\right.$$
$$\left.-\frac{2(\mathrm{Cov}[\mathrm{Cov}(T,Z),\sigma_T^2])}{\mathrm{Cov}(T,Z)\sigma_T^2}\right\}, \qquad (5.2)$$

where $\mathrm{Cov}(T, Z)$ and σ_T^2 are the sample covariance between T and Z and the sample variance of T, respectively, and $\mathrm{Cov}[X, Y]$ and $\mathrm{Var}[X]$ represent operators of sample covariance and variance. Appropriate use of eqn (5.2) requires that the sample variance and covariance of $\mathrm{Cov}(T, Z)$ and σ_T^2 are of order n^{-1}, and this will be investigated in the following numerical analysis. Calculations of the variances and covariances involved in eqn (5.2) are demonstrated in Appendix II.

Prediction of power

In the analysis of variance, it has been shown that the linkage disequilibrium between the marker and QTL can be detected through testing the significance of the expected mean square between marker genotypes (EMS_β) against that within marker genotypes (EMS_ω). The power of the F statistical test can be predicted from the probability

$$\beta_F = \Pr\{F_{v_1,v_2}(\delta_F) > F_{\alpha;v_1,v_2}\}, \qquad (6.1)$$

where $F_{v_1,v_2}(\delta_F)$ represents a noncentral F-variable with degrees of freedom v_1 and v_2 and noncentral parameter δ_F, and $F_{\alpha;v_1,v_2}$ stands for the upper α-point of a central F-variable with the same degrees of freedom. These distribution parameters can be determined following Johnson & Kotz (1970, p. 189ff.) as $v_1 = 2$, $v_2 = n-1$ and

$$\delta_F = \left(\frac{\mathrm{EMS}_\beta}{\mathrm{EMS}_\omega}\right)\frac{v_1(v_2-1)}{v_2} - v_1. \qquad (6.2)$$

The power function (6.1) can be evaluated using the cumulative distribution of the noncentral F-distribution which is expressed in terms of an infinite series of multiples of incomplete beta functions as given in Johnson & Kotz (1970, p. 192).

When the linkage disequilibrium is detected by testing the significance of the regression coefficient given by eqn (4) the corresponding power can be predicted from the probability

$$\beta_t = \Pr\{t_v(\delta_t) > t_{\alpha/2;v}\}, \qquad (7.1)$$

where t_v (δ_t) represents a random variable with noncentral Student's t-distribution of v degrees of freedom and noncentrality parameter δ_t, and $t_{\alpha/2;v}$ is the upper $\alpha/2$ point of a central t-variable with the same degrees of freedom. The value of v equals $n-2$ and the noncentral parameter is given by

$$\delta_t = \frac{\Gamma[v/2]b}{\sqrt{v/2}\,\Gamma[(v/2)]\sigma_b} \qquad (7.2)$$

(Johnson & Kotz, 1970, p. 201ff.). In the expression above, $\Gamma(.)$ is a gamma function, b and σ_b are, respectively, the regression coefficient and its standard deviation, which could be estimated using either eqns (5.1) or (5.2). The influence of using these different variance predictors will be discussed in the following numerical studies. The power function (7.1) can be evaluated by calculating the cumulative distribution of the noncentral t-distribution in terms of confluent hypergeometric functions discussed in Amos (1964) or Owen (1968).

Numerical analyses

Simultion study

In order to confirm the previous theoretical predictions of statistical powers for detection of the linkage disequilibrium, populations were simulated for 12 different sets of parameters as summarized in Table 2. For each set of parameters, the joint genotypes at both the marker locus and the QTL for an individual were sampled from a multinomial distribution with the probability parameters as shown in Table 1 and the given sample size n. Once the marker–QTL joint genotype was determined, the phenotypic record for an individual was generated by its genotypic value of the QTL plus a random number sampled from a normal distribution of mean zero and variance σ_e^2.

The simulation program used in the present study can be easily run with different values of the allelic frequencies at both the marker locus and QTL, the additive effect and the dominance level at the QTL and the census population size. For simplicity, the QTL genotypic effects were expressed in terms of the QTL heritability (i.e. the proportion of genetic

variance at the QTL to a given magnitude of phenotypic variance of the trait, which was assigned a constant value of 100).

Each parameter set was repeated 1000 times. The statistics involved in the power calculation were estimated as the mean of the repeated simulations, and the corresponding standard error of these means. Each set of the simulation data was used to perform analysis of variance and regression analysis. Calculating the frequency of the significant statistical tests of these two different analyses in the repeated simulation trials gives simulated observations of the power, as has been carried out in Carbonell *et al*. (1992).

Table 2 Parameters defining the 12 populations considered in numerical analyses, where n is the census population size, p and q are the frequencies of alleles M and A, D is the coefficient of linkage disequilibrium between the marker locus and QTL, h^2 is the heritability of the QTL and ϕ is the dominance ratio at the QTL

Population	n	p	q	D	h^2	ϕ
1	100	0.5	0.5	0.1	0.1	0.0
2	200	0.5	0.5	0.1	0.1	0.0
3	200	0.5	0.5	0.2	0.1	0.0
4	200	0.5	0.5	0.1	0.2	0.0
5	200	0.5	0.5	0.1	0.1	0.5
6	200	0.5	0.5	0.1	0.1	1.0
7	200	0.3	0.3	0.1	0.1	0.0
8	200	0.7	0.7	0.1	0.1	0.0
9	200	0.3	0.5	0.1	0.1	0.0
10	200	0.5	0.3	0.1	0.1	0.0
11	200	0.4	0.6	0.1	0.2	1.0
12	200	0.6	0.4	0.1	0.2	1.0

Results

Tabulated in Table 3 are the average of the mean squares and their corresponding standard errors over 1000 replicates of simulations and the mean squares predicted from calculations based on the theoretical analyses developed in the present study. The theoretical predictions are in good agreement with the simulated observations, validating the theoretical model presented here. In Table 3, simulated observations of the powers of statistically testing for linkage disequilibrium between the marker and the QTL are also shown together with the theoretical predictions for all 12 populations. The theoretical calculations of the power provided adequate predictions to the corresponding simulated values.

Table 4 illustrates the estimates of sample variances of the regression coefficients calculated as the average of repeated simulations and by the use of the theoretical predictions. Among the three estimates of the variance observed from the simulation studies, $\widehat{\sigma}_b^2$ was the variance of the 1000 regression coefficients calculated from repeated simulations, whereas $\hat{\sigma}_b^2$ and σ_b^2 were the averages of the observed variances of the regression coefficient, which were calculated by use of eqns (5.1) and (5.2), respectively. Theoretical predictions of these variances were derived in correspondingly similar ways.

Table 3 Numerical results of analysis of variance: expected mean squares between marker genotypes (EMS_β) and within marker genotype (EMS_ω) estimated from simulations, together with their corresponding standard errors, and predicted from theoretical calculation, as well as the observed powers and their corresponding theoretical predictions

	Simulated			Predicted		
Population	EMS_β	EMS_ω	Power	EMS_β	EMS_ω	Power
1	180.77 ± 5.20	98.48 ± 0.44	0.19	177.60	98.40	0.18
2	249.02 ± 6.04	98.23 ± 0.33	0.31	257.60	98.40	0.34
3	731.34 ± 11.58	93.87 ± 0.30	0.92	730.40	93.60	0.91
4	413.68 ± 8.48	96.80 ± 0.31	0.61	415.20	96.80	0.62
5	247.02 ± 6.22	99.07 ± 0.31	0.31	242.89	98.55	0.31
6	208.17 ± 5.73	99.23 ± 0.32	0.24	213.47	98.85	0.25
7	314.84 ± 7.33	97.56 ± 0.32	0.44	324.78	97.92	0.46
8	322.83 ± 7.55	97.74 ± 0.32	0.46	324.78	97.92	0.46
9	284.93 ± 6.56	97.79 ± 0.31	0.41	286.53	98.11	0.39
10	292.76 ± 7.32	97.69 ± 0.32	0.40	287.62	98.10	0.39
11	446.86 ± 9.09	96.52 ± 0.31	0.67	440.54	96.54	0.65
12	393.09 ± 8.42	96.78 ± 0.31	0.58	403.12	96.92	0.60

Table 4 Variances of the regression coefficient calculated: (i) from the sample variance of 1000 observed regression coefficients ($\hat{\sigma}_b^2$); (ii) from the average of each simulated value, where $\hat{\sigma}_b^2$ and σ_b^2 were derived using eqns (5.1) and (5.2) in the text, respectively, together with their corresponding standard errors. These variances were also predicted using the corresponding theoretical formula

	Simulated			Predicted	
Population	$\hat{\sigma}_b^2$	$\hat{\sigma}_b^2$	σ_b^2	$\hat{\sigma}_b^2$	σ_b^2
1	1.875	1.932 ± 0.009	1.901 ± 0.012	1.968	1.953
2	0.984	0.962 ± 0.003	0.974 ± 0.004	0.984	0.980
3	0.940	0.939 ± 0.003	0.937 ± 0.004	0.936	0.936
4	0.965	0.958 ± 0.003	0.964 ± 0.004	0.968	0.968
5	0.976	0.975 ± 0.003	0.981 ± 0.003	0.986	0.951
6	1.022	0.984 ± 0.003	0.987 ± 0.004	0.989	1.017
7	1.201	1.150 ± 0.004	1.151 ± 0.006	1.164	1.170
8	1.200	1.170 ± 0.004	1.177 ± 0.006	1.164	1.173
9	1.166	1.152 ± 0.004	1.145 ± 0.006	1.168	1.154
10	1.021	0.967 ± 0.003	0.960 ± 0.004	0.981	0.977
11	1.016	0.995 ± 0.003	0.987 ± 0.004	1.006	0.992
12	1.012	0.997 ± 0.003	0.989 ± 0.004	1.010	1.027

Numerical calculation indicates that the sample variance and covariance of the covariance between the phenotypic record (Z) and the number of the marker allele M (T) were in the range of 0.0021 to 0.0659, which were about of the order of n^{-1}, for the circumstances considered here, suggesting the appropriateness of using the variance prediction based upon formula (5.2). It can be seen from Table 4 that theoretical prediction of the variance of the regression coefficient using either eqns (5.1) or (5.2) provides an adequate approximation for the simulated value in all 12 simulated populations. This demonstrates that possible violation of normality of these regression variables did not cause significant bias of the variance estimation and thus confirms the reliability of using eqn (5.1) as a simple predictor of the variance of the regression coefficient.

The averages of the regression coefficients over the 1000 replicates of simulations and their corresponding standard errors are shown together with the theoretical predictions of these coefficients in Table 5. Comparisons of the coefficient estimates between the theoretical values and simulation averages show a good agreement. Theoretical calculation of the power provides an accurate prediction of the corresponding simulated values, and the theoretical power predictions using the different estimates of the regression coefficient variance based on eqns (5.1) and (5.2) yielded an almost identical value.

Comparison between the regression analysis and the analysis of variance shows that the regression test had consistently higher power than the F-statistical test in the analysis of variance.

Discussion

Statistical inference about linkage disequilibrium between polymorphic genetic marker loci and the loci controlling quantitative genetic variation is essential in the identification of genes affecting traits of great economic value in plant/animal breeding schemes or of disease-susceptibility in humans. It has been shown by Lande & Thompson (1990) and Gimelfarb & Lande (1994) that substantial linkage disequilibrium between marker loci and QTL is a prerequisite for marker-assisted selection (MAS) to achieve extra genetic progress. Moreover, the efficiency of MAS is highly dependent on correctly determining the markers which are incorporated in a MAS index. A false positive or negative inference about the disequilibrium, and in turn an erroneous use of the marker information, will result in reducing instead of improving the efficiency (Luo *et al*., 1997). It has also been hoped that linkage disequilibrium between a marker and a trait locus will lead to the identification of a disease gene in the vicinity of the marker (Weeks & Lathrop, 1995) even though very much care must be paid in interpreting the data of linkage disequilibrium as an alternative measure for obtaining a fine map for a disease predisposing gene in human populations (Hill & Weir, 1994). Moreover, it is widely agreed that the

Table 5 Regression coefficients estimated from the average of repeated simulations, together with the standard errors, and predicted from theoretical calculations, as well as the observed statistical powers from simulations and those from theoretical prediction, where power[1] was predicted using the sample variance eqn (5.1) and power[2] was predicted using the sample variance eqn (5.2)

	Simulated		Predicted		
Population	*b*	Power	*b*	Power[1]	Power[2]
1	1.790 ± 0.136	0.25	1.789	0.20	0.20
2	1.739 ± 0.070	0.41	1.789	0.45	0.46
3	3.571 ± 0.069	0.96	3.578	0.96	0.96
4	2.538 ± 0.069	0.75	2.540	0.75	0.75
5	1.715 ± 0.070	0.40	1.687	0.40	0.42
6	1.409 ± 0.071	0.30	1.461	0.30	0.29
7	2.255 ± 0.077	0.57	2.324	0.60	0.60
8	2.341 ± 0.077	0.57	2.324	0.60	0.60
9	2.136 ± 0.076	0.52	2.130	0.53	0.53
10	1.952 ± 0.071	0.51	1.952	0.53	0.53
11	2.733 ± 0.071	0.78	2.690	0.79	0.79
12	2.460 ± 0.071	0.70	2.510	0.73	0.72

objectives of screening genes underlying human complex disease have been seriously limited by the difficulties involved in collecting large and informative pedigrees. However, data are more easily obtained from natural populations than from structured pedigrees in human or animal populations or segregating populations in plant or animal species. The present analysis provides a fast screening of individual markers which are significantly associated with the genetic variation for the purposes of using marker information either for improving selection efficiency of quantitative traits or for mapping genes underlying quantitative genetic variation. The linkage disequilibrium mapping of a gene is based on the study of association between the gene and the marker(s) whose map position is well known. This requires further knowledge about the magnitude of the disequilibrium and an appropriate parameterization of the decay of the disequilibrium in terms of the genetic distance between the target gene and the marker locus (Baret & Hill, 1997 for a comprehensive review).

A novel quantitative genetics model has been developed in the present paper to detect the presence of linkage disequilibrium between a marker locus and a locus contributing to quantitative genetic variation in natural populations. The model is appropriate for analysing linkage disequilibrium generated from all potential causes. Theoretical analyses demonstrated that this can be achieved by the methods based upon an appropriate statistical method of analysis of variance or analysis of regression. The powers of these statistical analyses were adequately predicted and the factors affecting the powers were investigated. The model differs from others in various respects: the two-loci models of Soller & Genizi (1978), Luo (1993) and Knott (1994) assumed the disequilibrium to be produced from crossing two lines in which the marker and trait loci were linked or completely linked. The MAS model proposed by Lande & Thompson (1990) suggested the use of the regression coefficient of the number of favourable marker alleles in MAS on the trait phenotype as a measure of the magnitude of the disequilibrium, but no attempt was made in their study to investigate the efficiency of the method. The model presented in this paper is appropriate for directly analysing the data of marker genotypes and phenotypic records of a quantitative trait without requiring knowledge of the haplotype frequencies at the two loci which was assumed to be available in Hill & Weir (1994) and in Terwilliger (1995). An important assumption made in the present analysis is random union of gametes with respect to the marker and QTL loci. Any violation of this assumption would result in a reduction of the test statistic and thus a lowering of the power of the disequilibrium test.

The present study has shown the following. (i) There is an important difference in power between the two approaches; the regression analysis is more powerful than the analysis of variance, particularly when the QTL has a low heritability. An examination of the calculation of the test statistics in the two

approaches indicates that the test statistic in the analysis of variance essentially tests the significance of the correlation ratio of a continuous quantitative variate Z (i.e. the phenotypic records of the trait) on a discrete variate T (i.e. the number of the marker allele M), whereas the test statistic in the regression analysis virtually tests linearity of the regression of the variate Z on the variate T. It has been pointed out in Kendall & Stuart (1961, pp. 296–300) that the regression test will have higher power than the correlation ratio test (i.e. the test of the analysis of variance) if the alternative hypothesis is that the regression of Z on T is linear. Moreover, it can be seen, from comparing the powers of populations 2, 5 and 6 in Table 3 with those of the corresponding populations in Table 5, that the superiority of the regression analysis to the analysis of variance tends to become less important as the dominance ratio at the QTL increases from zero to one. This, however, is paralleled with the trend that the powers of both the approaches decrease to a very small value (<30 per cent) as the dominance ratio increases. The difference between the two statistical tests will become trivial when both of the tests have very low powers. (ii) Although the variables in the regression analysis do not strictly follow a normal distribution, the variance estimate of the regression coefficient predicted from using formula (5.1), which requires the variables to be normal, was not significantly different from that derived from using prediction eqn (5.2) without the need of invoking normality. (iii) Which factors affect the efficiency of the statistical tests of the linkage disequilibrium. Among the parameters considered, the amount of disequilibrium and size of the QTL are most important in determining the powers. A comparison of the powers among populations 2, 5 and 6 indicates that the power decreases with an increase in the dominance ratio at the QTL. The allelic frequencies at the marker and QTL display an important effect on the power in both models (the analysis of variance and the regression analysis). When the allelic frequencies at the two loci are low (e.g. population 7 in which $p = q = 0.3$) or high (e.g. population 8 in which $p = q = 0.7$), the power is increased compared to intermediate values (e.g. population 2 in which $p = q = 0.5$). Moreover, comparison of the power for populations 9 and 10 shows that the frequencies p and q were interchangeable in determining the power. These agree with the fact that the two loci were symmetric in the theoretical model as described in Table 1. It was found in our previous study (Luo *et al*. 1997) that the allelic frequencies at the marker locus and QTL play an important role in determining the efficiency of MAS, but the effects of allelic frequencies at the two loci on MAS efficiency were not interchangeable. For a given amount of linkage disequilibrium between the two loci, an increase in the efficiency can be expected when both the frequencies p and q are low. Combining the findings of the present study with those of Luo *et al*., (1997) suggests that the allelic frequencies display a more important influence on the efficiency of MAS at the stage of selection than at the stage of screening the markers.

The present study has been focusing on modelling linkage disequilibrium between a single marker locus and a single QTL. This may seem distant from being completely realistic for polygenic inheritance of quantitative traits and for availability of marker linkage maps. In distinct comparison with the model present here, Lande & Thompson (1990) proposed a multiple regression approach in which an infinite loci model of quantitative genetic variation was assumed and the use of multiple markers was allowed. However, it has been shown in a simulation example given by Gimelfarb & Lande (1994) that either a false positive or a false negative inference about the linkage disequilibria between the marker loci and QTLs may be frequently made using multiple regression analysis because the marker-associated quantitative effects could be counterbalanced or could inflate each other among the linked marker loci. These problems, thus, leave the multiple regression model far from being conclusive for the theory of linkage disequilibria among marker loci and QTLs. A full understanding of the multidimensional marker-associated quantitative genetic effects requires a further reparameterization of the multiple regression coefficients in terms of genetic parameters such as the disequilibrium coefficients of different orders. The model studied here is increasingly likely to be a subunit of the sophisticated framework of multiple-loci disequilibria, and the study of the two-locus system in isolation is an important building block for an understanding of the system as a whole.

Acknowledgements

I wish to thank Drs M. J. Kearsey, W. G. Hill and C. C. Tan for their constructive comments and criticisms. I am indebated to Dr Terry Crawford and two anonymous reviewers for their comments and suggestions which have been very helpful in improving presentation and clarifying several ambiguities in an earlier version of the present paper. I am grateful to Dr P. V. Baret for sending me his manuscript

before publication and acknowledge the National Natural Science Foundation and the National Education Commission of China for funding this study.

References

AMOS, D. E. 1964. Presentations of the central and non-central *t*-distributions. *Biometrika*, **51**, 451–458.

BARET, P. V. AND HILL, W. G. 1997. Gametic disequilibrium mapping: potential applications in livestock. *Animal Breeding Abstracts* (in press).

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

CARBONELL, E. A., CRIG, T. M., BALANSARD, E. AND ASINS, M. J. 1992. Interval mapping in the analysis of nonadditive quantitative trait loci. *Biometrics*, **48**, 305–315.

GIMELFARB, A. AND LANDE, R. 1994. Simulation of marker assisted selection in hybrid populations. *Genet. Res.*, **63**, 39–47.

HARTL, D. L. AND CLARK, A. G. 1989. *Principles of Population Genetics*, 2nd edn. Sinauer Associates, Sunderland, MA.

HILL, A. P. 1975. Quantitative linkage: a statistical procedure for its detection and estimation. *Ann. Hum. Genet.*, **38**, 439–449.

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

HILL, W. G. AND ROBERTSON, A. 1966. The effect of linkage on limits to artificial selection. *Genet. Res.*, **8**, 269–294.

HILL, W. G. AND ROBERTSON, A. 1968. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.*, **38**, 226–231.

HILL, W. G. AND WEIR, B. S. 1994. Maximum-likelihood estimation of gene location by linkage disequilibrium. *Am. J. Hum. Genet.*, **54**, 705–714.

JAYAKAR, S. D. 1970. On the detection and estimation of linkage between a locus influencing a quantitative character and a marker locus. *Biometrics*, **26**, 451–464.

JOHNSON, N. L. AND KOTZ, N. 1970. *Distributions in Statistics: Continuous Univariate Distributions*. Houghton Mifflin, Boston.

KENDALL, M. G. AND STUART, A. 1961. *The Advanced Theory of Statistics*, vol. 2, *Inference and Relationship*. Butler & Tanner, Frome.

KENDALL, M. G., STUART, A. AND ORD, J. K. 1983. *The Advanced Theory of Statistics*, vol. 1, *Distribution Theory*. Charles Griffin & Company, London.

KNAPP, S. J. AND BRIDGES, W. C. 1990. Using molecular markers to estimate quantitative trait locus parameters: power and genetic variances for unreplicated and replicated progeny. *Genetics*, **126**, 769–777.

KNOTT, S. A. 1994. Prediction of the power of detection of marker–quantitative trait locus linkage using analysis of variance. *Theor. Appl. Genet.*, **89**, 318–322.

LANDE, R. AND THOMPSON, R. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics*, **124**, 743–756.

LANDER, E. S. AND SCHORK, N. J. 1994. Genetic dissection of complex traits. *Science*, **265**, 2037–2048.

LE ROY, P. AND ELSEN, J. M. 1995. Numerical comparison between powers of maximum likelihood and analysis of variance methods for QTL detection in progeny test designs. *Theor. Appl. Genet.*, **90**, 65–72.

LUO, Z. W. 1993. The power of two experimental designs for detecting linkage between a marker locus and a locus affecting a quantitative character in a segregating population. *Génét. Sél. Évol.*, **25**, 249–261.

LUO, Z. W., THOMPSON, R. AND WOOLLIAMS, J. A. 1997. A population genetics model of marker assisted selection. *Genetics*, **146**, 1173–1183.

OWEN, D. B. 1968. A survey of properties and applications of the noncentral *t*-distribution. *Technometrics*, **10**, 445–478.

SEARLE, S. R. 1987. *Linear Models for Unbalanced Data*. John Wiley & Sons, New York.

SOLLER, M. AND GENIZI, A. 1978. The efficiency of experimental designs for the detection of linkage between a marker locus and a locus affecting a quantitative trait in segregating populations. *Biometrics*, **34**, 47–55.

TERWILLIGER, J. D. 1995. A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *Am. J. Hum. Genet.*, **49**, 31–41.

WEEKS, D. E. AND LATHROP, G. M. 1995. Polygenic disease: methods for mapping complex disease traits. *Trends Genet.*, **11**, 463–524.

WEIR, B. S. 1979. Inferences about linkage disequilibrium. *Biometrics*, **35**, 235–254.

WEIR, B. S. AND COCKERHAM, C. C. 1978. Testing hypotheses about linkage disequilibrium with multiple alleles. *Genetics*, **88**, 633–642.

Appendix I

Calculation of expected mean squares in the analysis of variance model

Let f_i be the frequency of the ith marker genotype and f_{ij} be the frequency of the jth QTL genotype and the ith marker genotype ($i,j = 1, 2, 3$). These frequencies are related to the genetic parameters given in Table 1. The expected mean squares must be calculated following an analysis of variance under an unbalanced linear model as described in Searle (1987, pp. 111ff.) because n_{ij}, the number of individuals with the jth QTL genotype and the ith marker genotype, is not constant for different i and j. When n_{ij} is considered as a random

variate with a multinomial distribution of parameters f_{ij} and n (population sample size) (Knott, 1994), the expected mean square between the marker genotypes is

$$\begin{aligned}\mathrm{EMS}_{\beta} = \frac{1}{G-1}\Bigg\{ & \sum_{i=1}^{G} f_i n\beta_i^2 - \Bigg[\sum_{i=1}^{G} f_i(1+(n-1)f_i)\beta_i^2 + 2(n-1)\sum_{i<j\le 3} f_i f_j \beta_i \beta_j\Bigg] \\ & + \sum_{i=1}^{G}\frac{1}{f_i}\Bigg[\sum_{j=1}^{3} f_{ij}(1+(n-1)f_{ij})\omega_{ij}^2 + 2(n-1)\sum_{j<k\le 3} f_{ij}f_{ik}\omega_{ij}\omega_{ik}\Bigg] \\ & - \Bigg[\sum_{i=1}^{G}\sum_{j=1}^{3} f_{ij}[1+(n-1)f_{ij}]\omega_{ij}^2 + 2(n-1)\sum_{i<j\le 3}\sum_{k<l\le 3} f_{ij}f_{kl}\omega_{ij}\omega_{kl}\Bigg]\Bigg\} + \sigma_{\mathrm{e}}^2\end{aligned}$$

and the expected mean square within the marker genotype is

$$\mathrm{EMS}_{\omega} = \frac{1}{n-G}\Bigg\{\sum_{i=1}^{G}\sum_{j=1}^{3} f_{ij} n\omega_{ij}^2 - \sum_{i=1}^{G}\frac{1}{f_i}\Bigg[\sum_{j=1}^{3} f_{ij}(1+(n-1)f_{ij})\omega_{ij}^2 + 2(n-1)\sum_{j<k\le 3} f_{ij}f_{ik}\omega_{ij}\omega_{ik}\Bigg]\Bigg\} + \sigma_{\mathrm{e}}^2$$

where G is the number of marker genotypes and ω_{ij} is the effect of the jth QTL genotype within the ith marker genotype as defined in the model eqn (1) and is calculated as follows:

$$\omega_{11} = \frac{2[D-p(1-q)][(D+pq)h-pd]}{p^2}$$

$$\omega_{12} = \frac{-p[2D-p(1-2q)]d + [2D^2-2p(1-2q)D+(1-2q+2q^2)p^2]h}{p^2}$$

$$\omega_{13} = \frac{2(D+pq)[(D-p+pq)h-pd]}{p^2}$$

$$\omega_{21} = -\frac{[(1-2p)D-2p(1-p)(1-q)]d + [2D^2+(1-2p)(1-2q)D+2pq(1-p)(1-q)]h}{p(1-p)}$$

$$\omega_{22} = -\frac{[(1-2p)D-p(1-p)(1-2q)]d + [2D^2+(1-2p)(1-2q)D-p(1-p)(1-2q+2q^2)]h}{p(1-p)}$$

$$\omega_{23} = \frac{[(1-2p)D+2pq(1-p)]d + [2D^2+(1-2p)(1-2q)D+2pq(1-p)(1-q)]h}{p(1-p)}$$

$$\omega_{31} = \frac{2(1+D-p-q+pq)[(1-p)d+(D-q+pq)h]}{(1-p)^2}$$

$$\omega_{32} = \frac{(1-p)[2D+(1-p)(1-2q)]d + [2D^2+2(1-p)(1-2q)D+(1-2q+2q^2)(1-p)^2]h}{(1-p)^2}$$

$$\omega_{33} = \frac{2(D-q+pq)[(1-p)d+(D+(1-p)(1-q))h]}{(1-p)^2}.$$

Appendix II

Variances and covariances involved in the calculation of the variance of the regression coefficient

$$\mathrm{Var}[\sigma_T^2] = \frac{1}{n}(\mu_4-\mu_2^2)+\frac{1}{n^3}[(\mu_4-\mu_2^2)+2(n-1)(\mu_2^2-\mu_1^4)+4(n-1)(n-2)\mu_1^2\sigma_T^2$$

$$+4(n-1)\mu_1(\mu_3-\mu_1\,\mu_2)]-\frac{2}{n^2}[(\mu_4-\mu_2^2)+2(n-1)\mu_1(\mu_3-\mu_1\,\mu_2)]$$

$$\mathrm{Var}[\mathrm{Cov}(T,Z)] = \frac{(n-1)^2}{n^3}(\omega_{22}-\omega_{11}^2)+\frac{(n-1)}{n^3}[(\mu_2\,\nu_2-\mu_1^2\nu_1^2)+(n-2)\nu_1^2\sigma_T^2+\mu_1^2\sigma_Z^2)$$

$$+(2n-3)\mu_1\nu_1(\omega_{11}-\mu_1\nu_1)]-\frac{2(n-1)^2}{n^3}(\mu_1\omega_{12}+\nu_1\omega_{21}-2\mu_1\nu_1\omega_{11})$$

$$\mathrm{Cov}[\mathrm{Cov}(T,Z),\sigma_T^2] = \frac{1}{n}(\omega_{31}-\mu_2\omega_{11})-\frac{1}{n^2}[2(\omega_{31}-\mu_2\omega_{11})+2(n-1)\mu_1(\omega_{21}-\mu_1\omega_{11})]-\frac{(n-1)}{n^2}[(\mu_3-\mu_1\mu_2)\nu_1$$

$$+\mu_1(\omega_{21}-\mu_2\nu_1)]+\frac{1}{n^3}\{\omega_{31}-\mu_2\omega_{11}+(n-1)[(\mu_3-\mu_1\mu_2)\nu_1+\mu_1(\omega_{21}-\mu_2\nu_1)]$$

$$+2(n-1)[(\omega_{21}-\omega_{11}\mu_1)\mu_1+(\mu_2\omega_{11}-\nu_1\mu_1^3)+(n-2)(\mu_1\nu_1\sigma_T^2+\sigma_{TZ}\mu_1^2)]\},$$

where $\mu_r = \mathrm{E}(T^r)$, $\nu_r = \mathrm{E}(Z^r)$ and $\omega_{rs} = \mathrm{E}(T^r Z^s)$. They can be calculated as

$\mu_1 = 2p$

$\mu_2 = 2p(1+p)$

$\sigma_T^2 = 2p(1-p)$

$\mu_3 = 2p(1+3p)$

$\mu_4 = 2p(1+7p)$

$\nu_1 = (2q-1)d+2q(1-q)h$

$\nu_2 = [q^2+(1-q)^2]d^2+2q(1-q)h^2+\sigma_e^2$

$\sigma_Z^2 = 100.0$ (see the text)

$\omega_{11} = 2\{[D-p(1-2q)]d+[D(1-2q)+2pq(1-q)]h\}$

$\sigma_{TZ} = 2D[d+(1-2q)h]$

$\omega_{12} = 2\{[p(1-2q(1-q))-(1-2q)D]\,d^2+[(1-2q)D+2pq(1-q)]h^2\}+\mu_1\sigma_e^2$

$\omega_{21} = 2\{[(1-2p)D-p(1+p)(1-2q)]d+[(1+2p-4q-4pq)D-2D^2+2pq(1+p)(1-q)]h\}$

$\omega_{22} = 2\{[2D^2-(1+2p-2q-4pq)D+p(1+p)(1-2q+2q^2)]d^2-[2D^2-(1-2p+2q+4pq)D$
$-2pq(1+p)(1-q)]h^2\}+\mu_2\sigma_e^2$

$\omega_{31} = 2\{[(1+6p)D-p(1+3p)(1-2q)]d+[(1+6p-2q-12pq)D-6D^2+2pq(1+3p)(1-q)]h\}.$