Heredity **70** (1993) 245–253

Received 13 April 1992

Estimation of genetic parameters using linkage between a marker gene and a locus underlying a quantitative character in F_2 populations

Z. W. LUO & J. A. WOOLLIAMS

Institute of Animal Physiology and Genetics Research, Roslin, Edinburgh EH25 9PS, Scotland, U.K.

A maximum likelihood algorithm is developed for estimating the recombination frequency in a segregating population (F_2), between a marker gene and a locus affecting a quantitative trait as well as estimating the means and variances of the three genotypes of the quantitative trait. The maximum likelihood estimates are compared with the moment estimates of these parameters obtained from the algorithm described by Luo & Kearsey in 1989. It is concluded from computer simulation results that the maximum likelihood algorithm provides more accurate estimates and is more robust to changes in the value of the recombination frequency than the moment solutions, particularly with heterogenous variances. The difference between the genetic model considered here and by Luo & Kearsey and that by Darwasi & Weller, in 1992, is also discussed. Both methods for estimating r and gene effects become biased for high values of r and low values of heritability, but the results are better for data with complete dominance than for additive data.

Keywords: EM algorithm, linkage, marker gene, moment solution, QTL.

Introduction

Mapping quantitative trait loci (QTL) using polymorphic marker genes has received attention in both theoretical quantitative genetic studies and plant/animal breeding practice. The main objective of the theoretical analysis involved in QTL mapping concentrates on estimating linkage between marker genes and QTL. Many researchers have addressed the problem of marker-QTL linkage in different genetic backgrounds in which the number of marker genes of QTLs vary (Jayakar, 1970; Hill, 1975; Weller, 1986; Jensen, 1989; Lander and Botstein, 1989). In general, the problem deals with obtaining estimates of the recombination frequency between marker gene(s) and individual QTL, the expected effect and residual variation of the QTL genotypes under question. Luo & Kearsey (1989, 1991) developed a method to estimate linkage between a marker gene and a QTL which was recently criticized by Darvasi & Weller (1992), who demonstrated divergence of the estimates obtained from their true maximum likelihood estimates when the three QTL genotypes had different variances, even though this was beyond the scope of Luo & Kearsey (1989, 1991). We will discuss the problem in more detail and develop a method to derive a maximum likelihood solution to the problem.

Theoretical approach

The structure of a breeding population

Consider an F_2 family derived from crossing two inbred lines, one of which is homozygous for alleles M_1 and Q_1 of the locus of the genetic marker and the QTL respectively, and the other is homozygous for the alleles M_2 and Q_2. The marker alleles are assumed to be co-dominant, therefore the three genotypes at the marker locus are distinguishable. The QTL is linked to the marker with a recombination frequency of r. The means and variances of the quantitative trait among the three marker genotypes of the F_2 population are as shown in Table 1.

The analytical methods

Suppose that phenotypic values of the three QTL genotypes Q_1Q_1, Q_1Q_2 and Q_2Q_2 are distributed as $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$ and $N(\mu_3, \sigma_3^2)$ respectively, where $N(\mu_j, \sigma_j^2)$ represents a normal distribution with mean μ_j and

Table 1 Basic statistics of the quantitative trait of the three marker genotypes in an F_2

Statistics of quantitative trait	Marker genotypes		
	M_1M_1	M_1M_2	M_2M_2
Sample mean	x_{11}	x_{12}	x_{22}
Sample variance	S_{11}	S_{12}	S_{22}
Sample size	n_1	n_2	n_3

variance σ_j^2. The variance can include both environmental variation and genetical variation at other loci affecting the quantitative trait but segregating independently with the marker gene. In previous papers (Luo & Kearsey, 1989, 1991) it was assumed that the three QTL genotypes have the same variance, i.e. 'homoscedastic model'. A more general assumption that the variances of the three QTL genotypes are unequal, i.e. 'heteroscedastic model', was made by Darvasi & Weller (1992) and will be discussed later. In general, it can be easily verified that the expected means and variances of marker genotypic groups (X_{ij} and S_{ij} respectively) defined in Table 1 have the following forms (Luo & Kearsey, 1989):

$$X_{11} = (1-r)^2\mu_1 + 2r(1-r)\mu_2 + r^2\mu_3 \quad (1.1)$$

$$X_{12} = r(1-r)\mu_1 + [1-2r(1-r)]\mu_2 + r(1-r)\mu_3 \quad (1.2)$$

$$X_{22} = r^2\mu_1 + 2r(1-r)\mu_2 + (1-r)^2\mu_3 \quad (1.3)$$

and

$$S_{11} = (1-r)^2[\sigma_1^2 + (\mu_1 - X_{11})^2] + 2r(1-r)[\sigma_2^2 + (\mu_2 - X_{11})^2] + r^2[\sigma_3^2 + (\mu_3 - X_{11})^2] \quad (2.1)$$

$$S_{12} = r(1-r)[\sigma_1^2 + (\mu_1 - X_{12})^2] + [1-2r(1-r)][\sigma_2^2 + (\mu_2 - X_{12})^2] + r(1-r)[\sigma_3^2 + (\mu_3 - X_{12})^2] \quad (2.2)$$

$$S_{22} = r^2[\sigma_1^2 + (\mu_1 - X_{22})^2] + 2r(1-r)[\sigma_2^2 + (\mu_2 - X_{22})^2] + (1-r)^2[\sigma_3^2 + (\mu_3 - X_{22})^2]. \quad (2.3)$$

Let

$$e_1 = S_{11} - \{(1-r)^2(\mu_1 - X_{11})^2 + 2r(1-r)(\mu_2 - X_{11})^2 + r^2(\mu_3 - X_{11})^2\}$$

$$e_2 = S_{12} - \{r(1-r)(\mu_1 - X_{12})^2 + [1-2r(1-r)](\mu_2 - X_{12})^2 + r(1-r)(\mu_3 - X_{12})^2\}$$

$$e_3 = S_{22} - \{r^2(\mu_1 - X_{22})^2 + 2r(1-r)(m_2 - X_{22})^2 + (1-r)^2(\mu_3 - X_{22})^2\}$$

Simultaneous equations (1) and (2) can be solved uniquely for μ_j and σ_j^2, respectively, into the following expressions for any r $(0 \le r < 0.5)$:

$$\hat{\mu}_1 = \frac{1}{(1-2r)^2}\{(1-r)^2X_{11} - 2r(1-r)X_{12} + r^2X_{22}\}, \quad (3.1)$$

$$\hat{\mu}_2 = \frac{1}{(1-2r)^2}\{r(r-1)X_{11} + [(1-r)^2 + r^2]X_{12} + r(r-1)X_{22}\}, \quad (3.2)$$

$$\hat{\mu}_3 = \frac{1}{(1-2r)^2}\{r^2X_{11} - 2r(1-r)X_{12} + (1-r)^2X_{22}\}. \quad (3.3)$$

and substituting $\hat{\mu}_j$ into (2)

$$\hat{\sigma}_1^2 = \frac{1}{(1-2r)^2}\{(1-r)^2e_1 - 2r(1-r)e_2 + r^2e_3\}, \quad (4.1)$$

$$\hat{\sigma}_2^2 = \frac{1}{(1-2r)^2}\{r(r-1)e_1 + [(1-r)^2 + r^2]e_2 + r(r-1)e_3\}, \quad (4.2)$$

$$\hat{\sigma}_3^2 = \frac{1}{(1-2r)^2}\{r^2e_1 - 2r(1-r)e_2 + (1-r)^2e_3\}. \quad (4.3)$$

Under the homoscedastic model, the estimate of σ^2 can be solved as

$$\hat{\sigma}^2 = \frac{n_1e_1 + n_2e_2 + n_3e_3}{n_1 + n_2 + n_3}. \quad (5)$$

For a given sample of the F_2 family, the means (X_{11}, X_{12} and X_{22}) and variances (S_{11}, S_{12} and S_{22}) of the marker groups are known. The statistics at the left side of equations (3)–(5) are thus functions whose values are completely determined by just one common unknown parameter r, i.e. the recombination frequency between the marker gene and QTL. It can be shown that equations (3) and (4) are identical to equations (12)–(14) and (18)–(20) of Luo & Kearsey (1989).

Luo & Kearsey (1989) attempted to use the invariant property of maximum likelihood estimates (Mood *et al.*, 1974) and searched the log-likelihood function (7) for just one parameter, r because the QTL genotypic means (μ_j) and variance (σ_j^2 or σ^2) had been determined by the value of r for a given marker group mean and variance as shown in equations (3)–(5). However, these equations are not themselves likelihood solutions but represent solutions from equating moments. These estimates will thus be termed the moment solution. With increasing sample size it is expected that the errors might decrease and the QTL means and variance might approximate to the maximum likelihood estimates.

The distributions of the quantitative effects for the three marker genotypes can be expressed as

$$g_1(y)=(1-r)^2 f_{11}(y)+2r(1-r)f_{12}(y)+r^2 f_{13}(y), \quad (6.1)$$

$$g_2(y)=r(1-r)f_{21}(y)+[1-2r(1-r)]f_{22}(y) + r(1-r)f_{23}(y), \quad (6.2)$$

$$g_3(y)=r^2 f_{31}(y)+2r(1-r)f_{32}(1-r)^2 f_{33}(y), \quad (6.3)$$

where $f_{ij}(y)$ is a normal distribution density with a general form of

$$f_{ij}(y)=\frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\left\{-\frac{(y-\mu_j)^2}{2\sigma_j^2}\right\}.$$

For an F_2 sample with size $N=n_1+n_2+n_3$, as described in Table 1, the log-likelihood function, based on distributions (6), are

$$\mathscr{L}_M(y;\Phi)=\sum_{i=1}^{n_1}\ln[g_1(y_{1i})]+\sum_{j=1}^{n_2}\ln[g_2(y_{2j})]+\sum_{k=1}^{n_3}\ln[g_3(y_{3k})]. \quad (7)$$

The right side of equation (7) consists of three terms, each of which is the log-likelihood of the density function of a mixture of three normal distributions. These have the following general form

$$g_i(y_{ik};\Phi)=p_{i1}f_{i1}(y_{ik})+p_{i2}f_{i2}(y_{ik})+p_{i3}f_{i3}(y_{ik}), \quad (8)$$

where y_{ik} is a phenotypic value of the kth individual with the ith marker genotype ($i=1, 2, 3$) and p_{ij} is the proportion of the jth subpopulation in the ith marker group, which is completely determined by r as shown in equations (4). Therefore, the maximum likelihood estimation of linkage between the QTL and marker gene could be statistically defined by solving the unknown parameter estimates $\Phi^T=(r, \mu_1, \mu_2, \mu_3, \sigma_1^2, \sigma_2^2, \sigma_3^2)$ [or $\Phi^T=(r, \mu_1, \mu_2, \mu_3, \sigma^2)$] which maximize the log-likelihood function (7) under the constrained conditions:

$$0\le r<0.5 \text{ and } \sigma_j^2\ge 0 \text{ (or } \sigma^2\ge 0). \quad (9)$$

It was strongly suggested by Kiefer & Wolfowitz (1956) and shown by Basford & McLachlan (1985) that under the heteroscedastic model, the log-likelihood functions of mixed distributions (8) might be unbounded and so the maximum likelihood estimates may not exist. This is because under the heteroscedastic model each sample point could generate a singularity in the likelihood function, and similarly, any pair of sample points which are sufficiently close together would generate a local maximum, as would triplets, quadruplets, etc. Maximum likelihood could therefore be inapplicable (Day, 1969; Everitt & Hand, 1981). In the homoscedastic model considered by Luo & Kearsey (1989), however maximum likelihood estimates always exist and are strongly consistent (Kiefer & Wolfowitz, 1956; Redner, 1981).

The EM algorithm proposed by Dempster *et al.* (1977) can be developed to solve the problem. One problem encountered by use of the EM algorithm is its slow convergence. However, as the moment solutions of the QTL parameters provide estimates which may be close to their maximum likelihood estimates if a reasonably large sample size is considered, these estimates can be used as initial points of the iterating algorithm described below.

Description of the algorithm

In general, the EM algorithm is an iterative approach for analysing incomplete data (Dempster *et al.*, 1977; Titterington *et al.*, 1985; Little & Rubin, 1987). Each of the iterations of the algorithm consists of two steps: and E (exception) step and an M (maximization) step. The use of the algorithm to find the maximum likelihood estimates of a mixture of distributions has been considered (Aitkin & Wilson, 1980; Titterington, *et al.*, 1985; DerSimonian, 1986; McLaren *et al.*, 1991) but not in this context. In this section we follow the main principle of the EM algorithm to develop a computational approach for obtaining the maximum likelihood estimates of equation (7) for both homoscedastic and heteroscedastic models.

Let $p_{ij}(r)$ represent the prior probability of the individual with the ith marker genotype having the jth QTL genotype (i and $j=1, 2, 3$). These are from equations (6).

The moment solutions of the QTL genotypic means and variance ($\hat{\mu}_j^{(0)}$ and $\hat{\sigma}_j^{(0)^2}$ or $\hat{\sigma}^{(0)^2}$), together with the prior probabilities $p_{ij}(r)$, are used to initialize the following iterating algorithm.

The expectation step (E). The probability of the kth individual with the ith marker genotype having the jth genotype is expected as

$$\omega_{ijk}(y_{ik})=\frac{p_{ij}(r)f_{ij}(y_{ik})}{\sum_{j=1}^{3}p_{ij}(r)f_{ij}(y_{ik})}, \quad (10)$$

where $i, j=1, 2, 3$.

The maximization step (M). The new estimates of the QTL means, variance(s) and the mixture proportions

are estimated for each marker mixed distribution by using value of $\omega_{ijk}(y_{ik})$ obtained in the previous E-step as

$$\hat{m}_j = \sum_{i=1}^{3} \sum_{k=1}^{n_i} \omega_{ijk}(y_{ik}) \tag{11.1}$$

$$\hat{\mu}_j^{(1)} = \frac{1}{\hat{m}_j} \sum_{i=1}^{3} \sum_{k=1}^{n_i} \omega_{ijk}(y_{ik}) y_{ik} \tag{11.2}$$

$$\hat{\sigma}_j^{(1)^2} = \frac{1}{\hat{m}_j} \sum_{i=1}^{3} \sum_{k=1}^{n} \omega_{ijk}(y_{ik}) (y_{ik} - \hat{\mu}_j^{(1)})^2 \tag{11.3}$$

or

$$\hat{\sigma}^{(1)^2} = \frac{1}{n_2 + n_2 + n_3} \sum_{j=1}^{3} \sum_{i=1}^{3} \sum_{k=1}^{n_i} \omega_{ijk}(y_{ik})(y_{ik} - \hat{\mu}_j^{(1)})^2 \tag{11.4}$$

depending on whether three or one variances are estimated.

In the next step, the newly estimated QTL genotypic means ($\hat{\mu}_i^{(1)}$) and variance ($\hat{\sigma}_j^{(1)^2}$ or $\hat{\sigma}^{(1)^2}$) are used to start the E-step of the next iteration as if they were the true estimates of these parameters. In the same way, the E and M steps are repeated iteratively following equations (10) and (11), so that a sequence of estimates of the unknown distribution parameter vector, $\{\Phi^{(t)}\}_{t=1, 2, \ldots}$ will be generated. It is expected that the value of the log-likelihood function (7) will increase monotonically as the iteration is continued, i.e.

$$\mathcal{L}_M[y; \Phi^{(t-1)}] \leq \mathcal{L}_M[y; \Phi^{(t)}].$$

It has been verified by Wu (1983) that under the homoscedastic model, the sequence $\{?_M[y; \Phi^{(t)}]\}_{t=1, 2 \ldots}$ is bounded and the sequence $\{\Phi^{(t)}\}_{t=1, 2, \ldots}$ will converge to its limit value, denoted by $\Phi(*)$, which are the maximum likelihood estimates.

In this process, a value of r has been used to initial the algorithm in equation (10) of the E step. For any given estimates of the QTL genotypic means and variance, the log-likelihood function (7) has been checked to be a unimodal function of r by plotting the log-likelihood function against r with hundreds of different simulations. Therefore, the maximum likelihood esimate of r can be readily searched over the interval $0 \leq r \leq 0.5$ by use of the 'golden search' method described by Press *et al.* (1986).

Simulations

In order to check the convergence of the algorithm developed in the present paper, Monte Carlo simulations were carried out to simulate genetic models of linkage between a marker gene and a QTL in the F_2 population. Computer simulation of the F_2 population has been described elsewhere (Luo & Kearsey, 1989). Both homoscedastic and heteroscedastic models were simulated. The simulated means of the three QTL genotypes were 35.0, 30.0 and 25.0, respectively when gene effects at the QTL were additive, and 35.0, 35.0 and 25.0 when the increasing allele at the QTL is completely dominant. In the homoscedastic model, the residual variance of these genotypes took values of 12.50, 29.17 and 112.50, equivalent to heritabilities of 0.5, 0.3 and 0.1, respectively, for the quantitative character in the F_2 population. In the heteroscedastic model, the simulated means for the cases of additive and dominant gene effect were the same as those described above, but the extra-genetic variances of the three QTL genotypes were 36.0, 30.25 and 25.0 for additive gene effects and 36.0, 36.0 and 25.0 for the dominant gene effect. The parameters were the same as those used in the simulation studies of Darvasi & Weller (1992). The recombination frequency between the marker and QTL was also varied taking the values of 0.10, 0.20 and 0.30. Five hundred F_2 progenies were used for every simulation, and each parameter combination was replicated 100 times.

The simulation data have been used to check the convergence of the algorithm described in the previous section and it was found that for a given r, the log-likelihood function (7) increased monotonically with the estimates obtained from the consecutive iterations. The convergence of the algorithm to the maximum of the log-likelihood function was confirmed by extensive grid searching over all elements of unknown parameter vector Φ on a subset of the data. The iterative searching was continued as long as the difference of the log-likelihood values between two consecutive cycles was greater than 10^{-3}.

The results of the simulations were analysed by calculating the following genetic effects: (i) mean of the QTL homozygotes, $\hat{\mu} = (\hat{\mu}_1 + \hat{\mu}_3)/2$; (ii) the additive effect, $\hat{a} = (\hat{\mu}_1 - \hat{\mu}_3)/2$; (iii) the dominance deviation, $\hat{d} = (\hat{\mu}_2 - \hat{\mu})$ and (iv) the residual variance within QTL genotypes (Mather & Jinks, 1982).

Results

The results of the simulations are shown in Tables 2, 3, 4 and 5 for $\hat{r}$, $\hat{a}$, $\hat{d}$, and $\hat{\sigma}^2$ (or $\hat{\sigma}_j^2$), respectively, for both moment (M) and likelihood (L) solutions. The estimates of r, derived from counting simulated recombinant gametes, were not significantly different from their corresponding expected values ($\chi_1^2 = 0.5$, $\chi_1^2 = 0.6$ and $\chi_1^2 = 0.1$ for $r = 0.1$, 0.2 and 0.3 respectively) confirming the reliability of the simulation data. When $\hat{r}$

Table 2 Means of the estimates of recombination frequency between the marker gene and QTL and their corresponding standard errors based on 100 replicates. Where r and h^2 represent the recombination frequency between the marker and QTL and the heritability of the quantitative character controlled by the QTL respectively; M and L are abbreviations for the two methods from which the estimates were obtained, i.e. moment solutions and the maximum likelihood estimates; Hom, Het, Add, and Dom are abbreviations for homoscedastic, heteroscedastic models, additive and dominant gene effects respectively

Models	h^2	Methods	$r=0.1$	$r=0.2$	$r=0.3$
(Hom, Add)	0.5	L	0.107 ± 0.0052	0.201 ± 0.0059	0.289 ± 0.0074
(Hom, Add)	0.5	M	0.111 ± 0.0059	0.206 ± 0.0062	0.285 ± 0.0075*
(Hom, Add)	0.3	L	0.100 ± 0.0064	0.209 ± 0.0106	0.283 ± 0.0134
(Hom, Add)	0.3	M	0.088 ± 0.0066	0.212 ± 0.0109	0.232 ± 0.0138**
(Hom, Add)	0.1	L	0.121 ± 0.0123	0.184 ± 0.0136	0.213 ± 0.0145
(Hom, Add)	0.1	M	0.107 ± 0.0113	0.136 ± 0.0145**	0.163 ± 0.0145**
(Hom, Dom)	0.5	L	0.101 ± 0.0017	0.201 ± 0.0028	0.299 ± 0.0031
(Hom, Dom)	0.5	M	0.102 ± 0.0018	0.203 ± 0.0029	0.280 ± 0.0069*
(Hom, Dom)	0.3	L	0.099 ± 0.0031	0.193 ± 0.0046	0.297 ± 0.0038
(Hom, Dom)	0.3	M	0.099 ± 0.0030	0.192 ± 0.0049	0.274 ± 0.0084**
(Hom, Dom)	0.1	L	0.095 ± 0.0071	0.185 ± 0.0107	0.279 ± 0.0113
(Hom, Dom)	0.1	M	0.095 ± 0.0070	0.173 ± 0.0109**	0.197 ± 0.0132**
(Het, Add)	0.3†	L	0.087 ± 0.0078	0.176 ± 0.0093**	0.217 ± 0.0109**
(Het, Add)	0.3†	M	0.059 ± 0.0056**	0.106 ± 0.0076**	0.118 ± 0.0101**
(Het, Dom)	0.3†	L	0.097 ± 0.0035	0.184 ± 0.0064*	0.281 ± 0.0098
(Het, Dom)	0.3†	M	0.088 ± 0.0035*	0.132 ± 0.0078**	0.129 ± 0.0119**

†Approximate heritability, $*P<0.05$, $**P<0.01$.

was derived from searching the likelihood surface conditional on the sample QTL means and variances, the results were consistently unbiased.

Homoscedastic, additive model

With data simulated from a homoscedastic, additive model, L solutions provided unbiased estimates of r with the exception of the $h^2=0.1$. In this case $\hat{r}$ was relatively unresponsible to a change in r, with trends towards overestimation when $r=0.1$, and underestimation when $r=0.2$ and severe underestimation (by 29 per cent) when $r=0.3$. Corresponding to this, when $h^2=0.1$, $\hat{a}$ was overestimated (by 13 per cent) for $r=0.1$ and underestimated (by 19 per cent) for $r=0.3$. M solutions underestimated r when $r=0.3$, with the bias increasing from 5 to 46 per cent as h^2 decreased from 0.5 to 0.1. Severe underestimation of r was also apparent for $h^2=0.1$, $r=0.2$. The underestimation of $r=0.3$ was associated with an underestimation of $\hat{a}$ by 20–40 per cent and this contributed to an overestimation of $\hat{\sigma}^2$. Ignoring bias, the standard errors of the estimates were similar for L and M and increased as r increased and h^2 decreased.

Homoscedastic, non-additive model

L solutions were almost entirely unbiased. The bias in $\hat{r}$ when $h^2=0.1$ and $r=0.3$ was much reduced compared to the additive case and not significantly different from zero. When $h^2=0.1$ and $r=0.3$, $\hat{a}$ was underestimated. The residual variance was underestimated when $h^2=0.1$ and $r=0.3$ but by less than 4 per cent. M solutions underestimated r for the same combinations of h^2 and r as with additive data, with only a slight improvement (34 per cent error for $h^2=0.1$ and $r=0.3$); but, bias in $\hat{a}$ and $\hat{\sigma}$ was much less than in the additive case. $\hat{d}$ was underestimated when $r=0.3$ for h^2 less than 0.5. Standard errors for estimates were less than in the additive case, particularly for L solutions.

Heteroscedastic, additive model

L solutions displayed a tendency to underestimate r, the bias increasing from 13 to 28 per cent as r increased from 0.1 to 0.3. An estimate of a for $r=0.3$ was underestimated by 18 per cent and the corresponding variances were overestimated. M solutions severely underestimated r throughout (40 per cent), to

Table 3 Means of the estimates of the additive effects at QTL and their corresponding standard errors based on 100 replicates. Where r and h^2 represent the recombination frequency between the marker and QTL and the heritability of the quantitative character controlled by the QTL respectively; M and L are abbreviations for the two methods from which the estimates were obtained, i.e. moment solutions and the maximum likelihood estimates; Hom, Het, Add, and Dom are abbreviations for homoscedastic, heteroscedastic models, additive and dominant gene effects respectively

Models	h^2	Methods	$r=0.1$	$r=0.2$	$r=0.3$
(Hom, Add)	0.5	L	4.93 ± 0.0541	4.92 ± 0.0715	4.84 ± 0.1165
(Hom, Add)	0.5	M	4.87 ± 0.0712	4.77 ± 0.0960	4.39 ± 0.1216**
(Hom, Add)	0.3	L	5.10 ± 0.0932	4.70 ± 0.1195	4.58 ± 0.1372
(Hom, Add)	0.3	M	4.93 ± 0.0855	4.54 ± 0.1162	2.90 ± 0.1297**
(Hom, Add)	0.1	L	5.63 ± 0.2331*	5.10 ± 0.2066	4.04 ± 0.2153**
(Hom, Add)	0.1	M	5.81 ± 0.2188**	4.81 ± 0.1884	3.45 ± 0.2038**
(Hom, Dom)	0.5	L	5.00 ± 0.0267	5.03 ± 0.0393	4.94 ± 0.0645
(Hom, Dom)	0.5	M	5.01 ± 0.0267	5.06 ± 0.0454	4.81 ± 0.1051
(Hom, Dom)	0.3	L	4.94 ± 0.5080	4.97 ± 0.0633	4.96 ± 0.1081
(Hom, Dom)	0.3	M	4.94 ± 0.4979	4.98 ± 0.0688	4.81 ± 0.1355
(Hom, Dom)	0.1	L	5.01 ± 0.1104	4.84 ± 0.1644	4.22 ± 0.1962**
(Hom, Dom)	0.1	M	5.01 ± 0.1087	4.81 ± 0.1688	4.00 ± 0.2078**
(Het, Add)	0.3†	L	5.00 ± 0.1142	4.87 ± 0.1327	4.08 ± 0.1506
(Het, Add)	0.3†	M	4.60 ± 0.0782**	3.89 ± 0.0928**	2.89 ± 0.1097**
(Het, Dom)	0.3†	L	5.03 ± 0.0553	4.93 ± 0.0937	4.58 ± 0.1544*
(Het, Dom)	0.3†	M	4.91 ± 0.0544	4.26 ± 0.0990**	2.94 ± 0.1277**

†Approximate heritability, $*P<0.05$, $**P<0.01$.

Table 4 Means of the estimates of the dominance deviations at QTL and their corresponding standard errors based on 100 replicates. Where r and h^2 represent the recombination frequency between the marker and QTL and the heritability of the quantitative character controlled by the QTL respectively; M and L are abbreviations for the two methods from which the estimates were obtained, i.e. moment solutions and the maximum likelihood estimates; Hom, Het, Add and Dom are abbreviations for homoscedastic, heteroscedastic models, additive and dominant gene effects respectively

Models	h^2	Methods	$r=0.1$	$r=0.2$	$r=0.3$
(Hom, Add)	0.5	L	−0.03 ± 0.0443	−0.03 ± 0.0598	−0.04 ± 0.0761
(Hom, Add)	0.5	M	−0.03 ± 0.0510	0.05 ± 0.0734	0.37 ± 0.0845**
(Hom, Add)	0.3	L	0.10 ± 0.0796	0.03 ± 0.0889	0.01 ± 0.1390
(Hom, Add)	0.3	M	0.12 ± 0.0805	0.45 ± 0.0927**	0.10 ± 0.1389
(Hom, Add)	0.1	L	−0.17 ± 0.2011	0.02 ± 0.2773	0.13 ± 0.3797
(Hom, Add)	0.1	M	0.33 ± 0.2011	−0.51 ± 0.2690	0.21 ± 0.3791
(Hom, Dom)	0.5	L	4.99 ± 0.0433	4.93 ± 0.0604	5.01 ± 0.1195
(Hom, Dom)	0.5	M	4.97 ± 0.0442	4.88 ± 0.0710	4.83 ± 0.1890
(Hom, Dom)	0.3	L	4.96 ± 0.0777	4.87 ± 0.0997	4.66 ± 0.1775
(Hom, Dom)	0.3	M	4.97 ± 0.0795	4.85 ± 0.1086	4.36 ± 0.2263**
(Hom, Dom)	0.1	L	5.34 ± 0.2061	4.94 ± 0.3167	4.60 ± 0.3775
(Hom, Dom)	0.1	M	5.34 ± 0.2075	4.85 ± 0.3282	4.72 ± 0.4103
(Het, Add)	0.3†	L	−0.10 ± 0.0785	−0.21 ± 0.140	−0.20 ± 0.1745
(Het, Add)	0.3†	M	−0.12 ± 0.0700	−0.10 ± 0.1046	−0.14 ± 0.1227
(Het, Dom)	0.3†	L	4.81 ± 0.0901	4.45 ± 0.1368**	3.86 ± 0.2615**
(Het, Dom)	0.3†	M	4.67 ± 0.1040**	3.64 ± 0.1711**	1.96 ± 0.2150

†Approximate heritability, $*P<0.05$, $**P<0.01$.

Table 5 Means of the estimates of the residual variance within QTL genotypes and their corresponding standard errors based on 100 replicates. Where r represents the recombination frequency between the marker and QTL; M and L are abbreviations for the two methods from which the estimates were obtained, i.e. moment solutions and the maximum likelihood estimates; Hom, Het, Add and Dom are abbreviations for homoscedastic, heteroscedastic models, additive and dominant gene effects respectively

Model	σ^2	Methods	$r=0.1$	$r=0.2$	$r=0.3$
(Hom, Add)	12.50	L	12.79 ± 0.242	12.70 ± 0.325	12.87 ± 0.416
(Hom, Add)	12.50	M	12.56 ± 0.232	15.46 ± 0.348**	18.53 ± 0.440**
(Hom, Add)	29.17	L	29.57 ± 0.464	30.10 ± 0.498	34.68 ± 0.554**
(Hom, Add)	29.17	M	29.27 ± 0.473	33.07 ± 0.525**	36.07 ± 0.556**
(Hom, Add)	112.50	L	107.59 ± 1.638**	105.28 ± 1.667**	110.49 ± 1.611
(Hom, Add)	112.50	M	105.26 ± 1.598**	108.99 ± 1.652*	112.16 ± 1.668
(Hom, Dom)	12.50	L	12.40 ± 0.091	12.36 ± 0.093	12.36 ± 0.265
(Hom, Dom)	12.50	M	12.47 ± 0.095	12.31 ± 0.103	12.66 ± 0.430
(Hom, Dom)	29.17	L	29.19 ± 0.278	29.16 ± 0.316	28.90 ± 0.430
(Hom, Dom)	29.17	M	29.40 ± 0.278	29.33 ± 0.321	29.51 ± 0.572
(Hom, Dom)	112.50	L	108.00 ± 1.044**	109.67 ± 1.560	111.68 ± 1.525
(Hom, Dom)	112.50	M	108.76 ± 1.054**	110.64 ± 1.585	112.22 ± 1.637
(Het, Add)	36.00	L	34.45 ± 0.849	35.66 ± 0.953	36.89 ± 1.111
(Het, Add)	36.00	M	36.67 ± 0.706	39.87 ± 0.763**	41.07 ± 0.922**
(Het, Add)	30.25	L	28.66 ± 0.720	28.54 ± 0.887*	33.09 ± 0.963**
(Het, Add)	30.25	M	31.40 ± 0.516**	34.11 ± 0.662**	38.42 ± 0.688**
(Het, Add)	25.00	L	24.96 ± 0.590	25.52 ± 0.851	27.17 ± 0.914**
(Het, Add)	25.00	M	27.33 ± 0.539**	30.92 ± 0.805**	33.49 ± 0.858**
(Het, Dom)	36.00	L	36.24 ± 0.645	34.93 ± 0.860	32.35 ± 1.219**
(Het, Dom)	36.00	M	36.96 ± 0.626	35.29 ± 0.829	38.69 ± 0.956**
(Het, Dom)	36.00	L	35.97 ± 0.513	37.07 ± 0.741	37.96 ± 0.956
(Het, Dom)	36.00	M	36.32 ± 0.603	41.31 ± 0.826**	46.17 ± 0.820**
(Het, Dom)	25.00	L	24.88 ± 0.528	27.17 ± 1.114	28.76 ± 1.332*
(Het, Dom)	25.00	M	27.62 ± 0.840**	35.80 ± 1.484**	46.58 ± 1.624**

*$P<0.05$, **$P<0.01$.

the extent that the values of $\hat{r}$ for $r=0.2$ and 0.3 were very similar. Correspondingly, a was underestimated (8–42 per cent) and the residual variances were overestimated, with bias increasing with r.

Heteroscedastic, non-additive model

L solutions had a tendency to underestimate r slightly and a was underestimated by 6 per cent when $r=0.3$. Dominance, d, was underestimated for both $r=0.2$ and 0.3. M solutions underestimated r grossly for $r=0.2$ and 0.3 and this resulted in underestimation of a and d and an overestimation of the residual variances.

Correlations

Correlations between L estimates were examined about their observed means for all models. Positive correlations (0.1–0.3) were observed between $\hat{r}$ and $\hat{a}$, large negative correlations (-0.6 to -0.9) between $\hat{a}$ and $\hat{\sigma}$. The correlation between the estimates of r derived by the L and M methods was greater than 0.6.

Discussion

Accurate estimation of the parameters involved in analysis of linkage between a marker gene and QTL has been recently emphasized by Dekkers & Dentine (1991). The efficiency of marker-assisted selection is determined both by the amount of additive genetic variance that can be traced to the marker(s) linked with QTLs under selection and the accuracy of the marker-QTL linkage estimation (Soller, 1978).

In general, the L method is reliable with both the homoscedastic and the heteroscedastic data generated here. The M method was unreliable with the heteroscedastic data and its accuracy is more sensitive to the

value of r than the L method. While a larger sample would be more informative and reduce errors, this study has shown the nature of the biases that may occur and has examined the relative merits of the L and the M methods. The most common bias for both methods lies in the underestimation of r, particularly when heritability is low: thus the methods tend to suggest that linkage is tighter than it is with a corresponding underestimate of gene effect and overestimate of residual variance. As an example, for homoscedastic, additive data when $h^2 = 0.1$ and $r = 0.3$, $\hat{r} \leq 0.01$ for 30 L estimates of the 100 simulations. Checks were carried out to ensure $\hat{a} > 0$ for all simulations because sampling errors might have been large enough to make linkage appear in the opposite phase. If so, reparameterization of the model may have reduced the bias in $\hat{r}$ and $\hat{a}$. However, this was not found to have occurred.

Statistically, the problem, as discussed above, is equivalent to parameter estimation of a mixture of normal distributions. It has been commonly accepted that the ease of deriving solutions depends mainly on a difference between the sub-distribution means to be dissected relative to the common distribution variance (Everitt & Hand, 1981; Titterington *et al.*, 1985). The power of both methods declined as h^2 decreased, i.e. as the distance between the distributions to be dissected decreased relative to the residual standard deviation, although the presence and tightness of linkage between the marker and QTL makes the situation distinct from the previous studies. The dominance model improved the power of the estimation as might have been expected from the greater distinction between the mixed distributions and the ability conferred by the markers to separate homozygotes with the increasing allele from heterozygotes.

It was shown by the simulation results that the biased estimates were usually associated with loose marker-QTL linkage. This is because the power of a linkage analysis declines rapidly when the two loci are linked with a recombination frequency larger than or equal to 0.3 (Ott, 1985; Risch, 1991; Collins & Morton, 1991).

Weller (1986) and Darvasi & Weller (1992) suggested a full-dimensional search method to derive maximum likelihood estimates of the function (7). Having taken their recently published results as an example, the value of the log-likelihood function (7) was evaluated 78,125 times in order to search the likelihood surface for the maxima, as they suggested, for just one single analysis. The program took more than 24 h to finish running on a Macintosh II computer. However, it spent only 2 min (on average, because of the number of iterations in the EM algorithm can vary for different situations) in obtaining all the necessary results represented in the present paper for a single sample. The algorithm discussed here is clearly easier to use in practice.

In general, for both its estimation accuracy and robustness, we conclude that the algorithm described here is an improvement on the previous moment solution method of the marker-QTL linkage estimation analysis.

The FORTRAN and PASCAL source programs of the simulation and analysis programs are available to interested readers.

Acknowledgements

The authors wish to thank Dr Robin Thompson for his guidance and many constructive suggestions to this study. We thank Dr M. J. Kearsey for introducing us to the problem discussed here and his comments on an early draft of this paper.

References

AITKIN, M. AND WILSON, G. T. 1980. Mixture models, outliers, and the EM algorithm. *Technometrics*, **22**, 325–331.

BASFORD, K. E. AND McLACHLAN, G. J. 1985. Likelihood estimation with normal mixture models. *Appl. Stat.*, **34**, 282–289.

COLLINS, A. AND MORTON, N. E. 1991. Significance of maximal lods. *Ann. Hum. Genet.*, **55**, 39–41.

DARVASI, A. AND WELLER, J. I. 1992. On the use of the moments method of estimation to obtain approximate likelihood estimates of linkage between a genetic marker and a quantitative locus. *Heredity*, **68**, 43–46.

DAY, N. E. 1969. Estimating the components of a mixture of normal distributions. *Biometrika*, **56**, 463–474.

DEKKERS, J. C. M. AND DENTINE, M. R. 1991. Quantitative genetic variance associated with chromosomal marker in segregating populations. *Theor. Appl. Genet.*, **81**, 212–220.

DEMPSTER, A. P., LAIRD, N. M. AND RUBIN, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B.*, **39**, 1–22.

DERSIMONIAN, R. 1986. Maximum likelihood estimation of a mixing distribution. *App. Stat.*, **35**, 302–309.

EVERITT, B. S. AND HAND, D. J. 1981. *Finite Mixture Distributions.* Chapman and Hall, London.

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

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.

JENSEN, J. 1989. Estimation of recombination parameters between a quantitative trait locus (QTL) and two marker gene loci. *Theor. Appl. Genet.*, **78**, 613–618.

KEIFER, J. AND WOLFOWITZ, J. 1956. Consistency of the maximum likelihood estimates in the presence of infinitely

many incidental parameters. *Ann. Math. Stat.*, **27**, 887–906.

LANDER, E. S. AND BOTSTEIN, D. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **121**, 185–199.

LITTLE, R. J. A. AND RUBIN, D. B. 1987. *Statistical Analysis with Missing Data*. John Wiley, New York.

LUO, Z. W. AND KEARSEY, M. J. 1989. Maximum likelihood estimation of linkage between a marker gene and a quantitative locus. *Heredity*, **63**, 401–408.

LUO, Z. W. AND KEARSEY, M. J. 1991. Maximum likelihood estimation of linkage between a marker gene and a quantitative locus. II. Application to backcross and double haploid populations. *Heredity*, **66**, 117–124.

MATHER, K. AND JINKS, J. L. 1982. *Biometrical Genetics*, 3rd edn, Chapman and Hall, London.

McLAREN, C. E., WAGSTAFF, M., BRITTENHAM, G. M. AND JACOBS, A. 1991. Detection of two-component mixtures of lognormal distributions in grouped, doubly truncated data: analysis of red blood cell volume distributions. *Biometrics*, **47**, 607–622.

MOOD, A. M., GRAYBILL, F. A. AND BOES, D. C. 1974. *Introduction to the Theory of Statistics*. McGraw-Hill, New York.

OTT, J. 1985. *Analysis of Human Genetic Linkage*. The Johns Hopkins University Press, Baltimore, MA.

PRESS, W. H., FLANNERY, B. P., TEUKOLSKY, S. A. AND VETTERLING, W. T. 1986. *Numerical Recipes, The Art of Scientific Computing*. Cambridge University Press, Cambridge.

REDNER, R. 1981. Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *Ann. Stat.*, **9**, 225–228.

RISCH, N. 1991. A note on multiple testing procedures in linkage analysis. *Am. J. Hum. Genet.*, **48**, 1058–1064.

SOLLER, M. 1987. The use of loci associated with quantitative effects in dairy cattle improvement. *Anim. Prod.*, **27**, 396–404.

TITTERINGTON, D. M., SMITH, A. F. AND MAKOV, U. E. 1985. *Statistical Analysis of Finite Mixture Distributions*. John Wiley, New York.

WELLER, J. I. 1986. Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. *Biometrics*, **42**, 627–640.

WU, C. J. 1983. On the convergence properties of the EM algorithm. *The Annals of Statistics*, Vol. 11. No. 1, 95–103.