Methodology and Accuracy of Estimation of Quantitative Trait Loci Parameters in a Half-Sib Design Using Maximum Likelihood

M. J. Mackinnon* and J. I. Weller†

**Institute of Cell, Animal and Population Biology, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom and* †*Institute of Animal Sciences, Agricultural Research Organization, The Volcani Centre, 50250, Israel*

Manuscript received February 15, 1995
Accepted for publication June 27, 1995

ABSTRACT

Maximum likelihood methods were developed for estimation of the six parameters relating to a marker-linked quantitative trait locus (QTL) segregating in a half-sib design, namely the QTL additive effect, the QTL dominance effect, the population mean, recombination between the marker and the QTL, the population frequency of the QTL alleles, and the within-family residual variance. The method was tested on simulated stochastic data with various family structures under two genetic models. A method for predicting the expected value of the likelihood was also derived and used to predict the lower bound sampling errors of the parameter estimates and the correlations between them. It was found that standard errors and confidence intervals were smallest for the population mean and variance, intermediate for QTL effects and allele frequency, and highest for recombination rate. Correlations among standard errors of the parameter estimates were generally low except for a strong negative correlation ($r = -0.9$) between the QTL's dominance effect and the population mean, and medium positive and negative correlations between the QTL's additive effect and, respectively, recombination rate ($r = 0.5$) and residual variance ($r = -0.6$). The implications for experimental design and method of analysis on power and accuracy of marker-QTL linkage experiments were discussed.

DETECTION of major genes influencing quantitative traits is now feasible with the explosion of genetic marker technology in recent years. DNA-level markers can be used to establish genetic maps and systematically screen genomes for quantitative trait loci (QTL) in many agriculturally important species (EDWARDS *et al.* 1987; PATERSON *et al.* 1988; FRIES *et al.* 1989; HALEY *et al.* 1990; KEIM *et al.* 1990; SOLLER 1990; GEORGES *et al.* 1995). Detection of QTL can be achieved by finding a difference in quantitative trait value of two groups of offspring inheriting alternative marker alleles from a common heterozygous parent. This procedure can be used for both inbred species (*e.g.*, experimental species such as Drosophila and mice) using crosses between inbred lines (SAX 1923; SOLLER *et al.* 1976) and for outbred species (*e.g.*, farm animals and fruit trees) where the analysis is carried out within large family groups (GELDERMANN 1975; NEIMANN-SØRENSEN and ROBERTSON 1961; SOLLER and GENIZI 1978).

Having detected a QTL, it is of interest to find its location in the genome by estimating its recombination distance from the marker and to determine the effects of the QTL in the population by estimating the magnitude of its effect, the degree of dominance, and its frequency in the population. All of these parameters are important for both assessing the potential usefulness of the QTL in genetic improvement programs and the way in which it will be used. Estimation of these parameters can be carried out using maximum likelihood methods. While the power to detect QTL has been described well for both inbred line cross designs (SOLLER *et al.* 1976; WELLER 1986, 1987; JENSEN 1989; LANDER and BOTSTEIN 1989; LUO and KEARSEY 1989; SIMPSON 1989; KNOTT and HALEY 1992a; DARVASI *et al.* 1993) and within-family designs (SOLLER and GENIZI 1978; WELLER *et al.* 1990; KNOTT and HALEY 1992b; BOVENHUIS and WELLER 1994), only a few studies have explored the accuracy of the estimates (KNOTT and HALEY 1992a,b; VAN OOIJEN 1992; DARVASI *et al.* 1993). All of the latter have relied on replicated stochastic simulations to measure accuracy empirically. As this can be tedious, it would be useful to be able to predict the accuracy of QTL parameters using theoretical methods, although the complex nature of the likelihood models has so far prevented this being done. In this study, a quasitheoretical numerical method is used to predict both power and accuracy of QTL parameters from the shape of the multidimensional expected likelihood surface. It is applied to the half-sib design in which six parameters are estimated simultaneously and is used to explore the effects of data structure, statistical model and inaccurate marker allele frequencies on power, accuracy and sampling correlations between parameter estimates. The procedure described here can be used to determine the optimum experimental design and form of analysis

Corresponding author: Margaret Mackinnon, Institute of Cell, Animal and Population Biology, University of Edinburgh, West Mains Rd., Edinburgh, EH9 3JT, UK. E-mail: eang20@festival.ed.ac.uk

to extract the best information about the QTL from marker-QTL linkage data.

THEORY

Experimental design: There are s sires each of which produce N progeny when mated to an outbred population of dams. The phenotype for a quantitative trait is determined by many genes of small effect and by random environmental factors, which together cause the trait to be normally distributed with a mean of μ and variance of σ^2. The trait is also affected by a relatively major QTL with two alleles, Q and q, at population frequencies of p and $1 - p$, respectively. Additive and dominance effects at the QTL are denoted a and d, respectively. Thus there are three possible genotypes, QQ, Qq and qq, which have quantitative trait values of $\mu_{QQ} = \mu + a$, $\mu_{Qq} = \mu + d$ and $\mu_{qq} = \mu - a$. A marker locus is situated adjacent to the QTL at a recombination distance of r. Only sires heterozygous for the genetic marker are included in the analysis. The notation will assume only two alleles at the genetic marker, M and m, although a situation of multiple marker alleles can be handled with only minor modifications. Thus sires have one of four possible marker-QTL genotypes, *viz:* MQ/mQ, MQ/mq, Mq/mQ or Mq/mq. Dams have one of 10 possible genotypes corresponding to all possible combinations of gametes MQ, Mq, mQ and mq. The frequency of allele M in the dam population is denoted t and is assumed to be known or estimated with a high degree of accuracy. The population is assumed to be at Hardy-Weinberg equilibrium with respect to the QTL and marker locus. Progeny with records are assumed to be a random sample of the sire's progeny with respect to their value for the quantitative trait and marker loci.

Likelihood of the model: The likelihood for the half-sib design, first given by WELLER (1990) and extended here to account for dominance, QTL allele frequency and marker allele frequency, is

$$L = \prod_{b}^{s} \Bigg[p^2 \prod_{i=1}^{3} \prod_{k=1}^{n_i} \sum_{j}^{3} c_{ij} f(x_k - \mu_j) + p(1-p) \prod_{i=4}^{6} \prod_{k=1}^{n_{i-3}} \sum_{j}^{3} c_{ij} f(x_k - \mu_j) + p(1-p) \prod_{i=7}^{9} \prod_{k=1}^{n_{i-6}} \sum_{j}^{3} c_{ij} f(x_k - \mu_j) + (1-p)^2 \prod_{i=10}^{12} \prod_{k=1}^{n_{i-9}} \sum_{j}^{3} c_{ij} f(x_k - \mu_j) \Bigg] \quad (1)$$

where $b = 1$ to s and denotes the family number, $i = 1$–12 for progeny marker genotypes of MM, Mm and mm nested within sire marker-QTL genotypes MQ/mQ, MQ/mq, Mq/mQ and Mq/mq, respectively; n_i = the number of progeny in the ith marker genotypic group and $n_i = n_{i+3}$; $j = 1$, 2 and 3 for progeny QTL genotypes QQ, Qq and qq, respectively; c_{ij} = the element in row i and column j from the matrix, C (Table 1), of progeny QTL genotype probabilities conditional on sire QTL genotype and progeny marker genotype; and $f(x_k - \mu_j)$ = the normal density function for the kth observation, x_k, ($k = 1, \ldots, n_i$) of the bth sire of the ith marker genotype group, conditional on the jth QTL genotype, and is abbreviated by f_j and calculated as:

$$f_j = f(x_k - \mu_j) = \frac{e^{-1/2(x_k - \mu_j/\sigma)^2}}{\sqrt{2\pi}\sigma}.$$

The four terms that are summed in the likelihood equation for each sire represent the four possible QTL genotypes of the sire, denoted h. That is, for each possible genotype of the sire, the statistical density of the sire's progeny's genotypes, conditional on this sire's genotype, is computed over all the observed data. These terms, called sublikelihoods from here on, are multiplied by the probabilities of the sire's genotype P_h [p^2 for $h = 1$, $p(1 - p)$ for $h = 2$ or 3 and $(1 - p)^2$ for $h = 4$] and summed to obtain the overall likelihood.

C, the matrix of progeny QTL genotype probabilities, conditional on sire QTL genotype and progeny marker genotypes and assuming linkage equilibrium is computed as shown in Table 1. The rows of C represent the progeny marker genotypes, MM, Mm, and mm, nested within sire marker-QTL genotype, as shown to the right of the matrix. The columns of C represent the progeny QTL genotypes QQ, Qq, and qq. Thus the probabilities for each row sum to unity.

Each sire passes on average half of his breeding value for the quantitative trait to his progeny. Equation 1 does not account for variance among sires due to genes other than the QTL. The likelihood was therefore modified by subtracting half the sire's polygenic breeding value, g_b, from each observation. Estimates of the sire's breeding value from genetic evaluations based on information from ancestors and descendants will include the sire's QTL genotype effect and are therefore biased estimates of g_b. Thus an alternative estimate of g_b is required. For the situation here, where each sire has many offspring, an appropriate estimate is $g_b = \bar{X}_b - G_b$ where $\bar{X}_b$ is the observed mean of the sire's progeny and G_b is the expected QTL genotype mean of the sire's progeny, which is a function of a, d and p (BOICHARD *et al.* 1990). This is likely to be an effective way of partitioning out the effects of polygenes from the QTL effect in the sires because it uses two pieces of good and almost independent information, namely the sire mean, which estimates the total genetic effect, and the between-marker contrast within this sire mean, which estimates the QTL effect. Because g_b is conditional on the sire's QTL genotype, it is different for each of the four summed terms of (1). The likelihood accounting for polygenic effects thus becomes

TABLE 1

Progeny QTL genotype probabilities conditional on sire marker-QTL genotype and own marker genotype

Progeny QTL genotype			Progeny marker genotype	Sire genotype
QQ	*Qq*	*qq*		
p	$1-p$	0	*MM*	
p	$1-p$	0	*Mm*	*MQ/mQ*
p	$1-p$	0	*mm*	
$p(1-r)$	$1-p-r+2pr$	$(1-p)r$	*MM*	
$[tpr$	$[t(p+r-2pr)$	$[t(1-p-r+rp)$		
$+(1-t)p(1-r)]$	$+(1-t)(1-p-r+2pr)]$	$+(1-t)(1-p)r]$	*Mm*	*MQ/mq*
pr	$p+r-2pr$	$1-p-r+pr$	*mm*	
pr	$p+r-2pr$	$1-p-r+pr$	*MM*	
$[tp(1-r)$	$[t(1-p-r+2pr)$	$[t(1-p)r$		
$+(1-t)pr]$	$+(1-t)(p+r-2pr)]$	$+(1-t)(1-p-r+rp)]$	*Mm*	*Mq/mQ*
$p(1-r)$	$1-p-r+2pr$	$(1-p)r$	*mm*	
0	p	$1-p$	*MM*	
0	p	$1-p$	*Mm*	*Mq/mq*
0	p	$1-p$	*mm*	

$$L = \prod_{b}^{s}\left[p^2\prod_{i=1}^{3}\prod_{k=1}^{n_i}\sum_{j}^{3} c_{ij}f(x_k-\mu_j-\bar{X}_b+[pa+(1-p)d])\right.$$
$$+\,p(1-p)\prod_{i=4}^{6}\prod_{k=1}^{n_{i-3}}\sum_{j}^{3} c_{ij}f(x_k-\mu_j-\bar{X}_b+{}^1/_2[(2p-1)a+d])$$
$$+\,p(1-p)\prod_{i=7}^{9}\prod_{k=1}^{n_{i-6}}\sum_{j}^{3} c_{ij}f(x_k-\mu_j-\bar{X}_b+{}^1/_2[(2p-1)a+d])$$
$$\left.+\,(1-p)^2\prod_{i=10}^{12}\prod_{k=1}^{n_{i-9}}\sum_{j}^{3} c_{ij}f(x_k-\mu_j-\bar{X}_b+[(p-1)a+pd])\right] \quad (2)$$

The properties of this likelihood were explored both by computation of its expectation as a function of parameter estimates (deterministic simulations) and by stochastic simulations. The situations for which simulations were performed are summarized in Table 2 and described in more detail below.

Deterministic simulations: *Expectation of the log likelihood:* A close approximation of the expected value of the natural log likelihood, $E(\log L)$, for a given set of parameter estimates is obtained by modifying (1) to give the following equation:

$$E(\log L) = s\sum_{h}^{4} P_h \log\left[\sum_{H}^{4}\exp\left[\log[\hat{P}_H]\right.\right.$$
$$\left.\left.+\int_{-\infty}^{\infty}\sum_{i=3h-2}^{3h} n_i\sum_{j}^{3} c_{ij}f_j\log\sum_{J}^{3}\hat{c}_{ij}\hat{f}_J\,dx\right]\right] \quad (3)$$

The derivation of this equation and the degree of approximation is explained in the APPENDIX. Note that now the lowercase subscripts h, i and j indicate that the parameters are the true values and the uppercase superscripts and the $\hat{}$ indicate that these parameters are estimates or functions of estimates on which the value of the likelihood is conditional. Thus the symbols in (3) take on the following meanings: h denotes the sire's true QTL genotype, and H the sire's hypothesized QTL genotype; P_h denotes the expected frequency of sires with genotype h in the data (depending on the real value of p, as described above), and $\hat{P}_H$ denotes the probability that the sire has genotype H, which is a function of the current estimate of p, denoted $\hat{p}$; $\hat{c}_{ij}$ is the conditional probability of falling into the jth QTL class and the ith marker-sire class given the current estimates of the parameters p and r, denoted $\hat{p}$ and $\hat{r}$, and the hypothesized sire genotype, H, where $I = i - 3h + 3H$; f_j is an abbreviation of the density function defined previously used in (1) and (2) depending on whether the model takes account of polygenic effects (see below) and $\hat{f}_J$ is the function f_J given the estimates of their component parameters *e.g.*, $\hat{c}_{43} = (1-\hat{p})\hat{r}$; and n_i is the expected number of observations in the ith marker genotypic group and is equal to ${}^1/_2Nt$, for $i = 1$, 4, 7, and 10; ${}^1/_2N$ for $i = 2$, 5, 8, and 11; and ${}^1/_2N(1-t)$ for $i = 3$, 6, 9, and 12, respectively. Without polygenic effects (Equation 1), $f_j = f(x_k - \mu_j)$ and $\hat{f}_J = f(x_k - \hat{\mu}_j)$. With polygenic effects (Equation 2), $f_j = f(x_k - \mu_j - g_b)$ and $\hat{f}_J = f(x_k - \hat{\mu}_J - \hat{g}_b)$ where $\hat{g}_b = g_b - G_b - \hat{G}_b$.

When sample size is infinite, this expected log L,

TABLE 2

Models, family structures and parameter values used for simulations

Model	No. of families	Family size	a	d	r	p	μ	σ	t
Deterministic									
$h^2 = 0$	20	250	0.5[a]	0.1[a]	0.1[a]	0.5[a]	0.0[a]	1.0[a]	0.3
	50	100	0.5[a]	0.1[a]	0.1[a]	0.5[a]	0.0[a]	1.0[a]	0.3
	40	50	0.5[a]	0.1[a]	0.1[a]	0.5[a]	0.0[a]	1.0[a]	0.3
	20	250	0.5	0.1	0.1	0.5	0.0	1.0	0.3[a]
$h^2 = 0.25$	20	250	0.5[a]	0.1[a]	0.1[a]	0.5[a]	0.0[a]	1.0[a]	0.3
Stochastic									
$h^2 = 0$	20	250	0.5	0.1	0.1	0.5	0.0	1.0	0.3
	20	250	0.5	0.1	0.2	0.5	0.0	1.0	0.3
	20	100	0.5	0.1	0.1	0.5	0.0	1.0	0.3
	40	50	0.5	0.1	0.1	0.5	0.0	1.0	0.3
$h^2 = 0.25$	20	250	0.5	0.1	0.1	0.5	0.0	1.0	0.3

$h^2 = 0$ indicates that the data were simulated without polygenes; $h^2 = 0.25$ indicates that the data were simulated with polygenes, which accounted for 20% of the residual variation.

[a] The likelihood was evaluated for a range of parameter estimates around these true values.

because it is based on infinite sample size theory, is equivalent to the average of the log L evaluated in an infinite number of replicates of stochastic data. It is therefore useful for examining the shape of the log L surface in infinite samples. Because it is the shape of the surface around the maximum which determines the accuracy of the parameter estimates, this method is thus a useful tool for exploring the properties of the likelihood surface without having to resort to many time-consuming replicates using stochastic data. The populations used in this study are large but not infinite, in which case the expected log L is an approximation to the true average likelihood. In the size of data sets used in this study, this approximation appears to be close enough to have no appreciable influence on the power and accuracy (see APPENDIX).

Shape of the likelihood surface: Equation 3 was evaluated for values of estimates $\hat{\mu}$, $\hat{a}$, $\hat{d}$, $\hat{r}$, $\hat{p}$, and $\hat{\sigma}$, which ranged from ± 0.25 of the true values $\mu = 0$, $a = 0.5$, $d = 0.1$, $r = 0.1$, $p = 0.5$ and $\sigma = 1$ in increments of 0.025 for three family structures ($N = 250$, $s = 20$; $N = 50$, $s = 100$; $N = 40$, $s = 50$) without a polygenic effect in the model and one structure ($N = 250$, $s = 20$) with a polygenic effect in the model (Table 2). These data structures and parameter values are considered to be typical of large-scale studies likely to be used for mapping QTL in forest trees or animals. The marker allele frequency, t, was 0.3 in all cases. For each parameter varied, all other parameters were held at their true values. In these simulations g_b was set to 0 because for deterministic evaluations, this parameter is, in effect, known. log L was then plotted against each parameter to give "uniparameter profile likelihoods" assuming all other parameters are known. Three-dimensional plots of log L as a function of two parameters were also produced for a few pairs of parameters ($\hat{a}$ and $\hat{d}$, $\hat{d}$ and $\hat{r}$, $\hat{\mu}$ and $\hat{d}$) to examine the shape of the likelihood surface with respect to these parameters for the presence of ridges or local maxima.

Accuracy of the estimates: From the uniparameter profile likelihoods, the 95% confidence intervals were calculated as the parameter value at which the log L equalled the maximum log L (log L_0) plus half the 95% chi-squared value (χ^2) on one degree of freedom (*i.e.*, log $L_{95\%}$ = log L_0 + 1.92). This test statistic is based on the fact that for large samples, the difference between log L maximized for all but one parameter (log L_1), and log L maximized for all parameters (log L_0) is asymptotically distributed as $-{}^1/_2\chi^2$ with one degree of freedom (WILKS 1938). Assuming normality of the estimate's sampling distribution, the standard error of the estimate is expected to be half the 95% confidence interval (KENDALL and STUART 1973). Approximate standard errors and the correlations between them were also calculated from the approximate variance-covariance matrix of the estimates, which is obtained from the inverse of the matrix of all the second partial derivatives of the log likelihood function (*i.e.*, the observed information matrix) (KENDALL and STUART 1973). These are lower bound standard errors according to the Cramer-Rao inequality. For the simulation conditions used above, these second derivatives were obtained by evaluating (3) at points on the surface at small distances (± 0.005) away from the true values for all possible pairwise combinations of the parameters. This matrix was then inverted to give the approximate variances and covariances of the parameter estimates and from these the standard errors of the estimates and correlations between them were calculated. When the matrix of numerically derived second derivatives was not positive definite, the interval at which they were evaluated was increased to 0.01. The standard errors and correlations were the same to the third decimal place for the intervals of 0.005 and 0.01

in cases where the matrix was positive definite for both intervals, indicating that the choice of interval size did not have detectable effects on the results. The estimates of standard errors derived from the information matrix differ from those calculated from the confidence intervals by the fact that they are not based on the assumption that all other parameters are known. To test the effect on the accuracy of a single parameter of fixing all other parameters *vs.* simultaneously estimating all parameters, standard errors were calculated from the reciprocal of only the diagonals of the information matrix instead of the inverse of the whole matrix. This effectively ignores any covariances between parameters. Standard errors derived in this way and those derived from the confidence intervals are called "uniparameter standard errors" from here on, and those derived from the full information matrix are called "multiparameter standard errors".

Effect of marker allele frequency: The effect of using wrong values of the marker allele frequency, t, was investigated by replacing t with $\hat{t}$ ranging from 0 to 1 in intervals of 0.1 in (3). Maximum likelihood estimates of each of the parameters were found while holding all other parameters constant. The bias in each parameter over the range of $\hat{t}$ was calculated.

Stochastic simulations: Stochastic simulations were performed to illustrate that estimates of all parameters could be obtained by numerical maximization in stochastic data. Also the results from stochastic simulations were used to compare with those from the deterministic simulations. Data on progeny from multiple half-sib families of equal size were generated by first randomly assigning QTL genotypes to sires according to probabilities p^2, $p(1 - p)$, $p(1 - p)$ and $(1 - p)^2$, and then generating progeny from each sire by random sampling from independent normal distributions with means of μ_j (j = 1, 2 or 3 for QTL genotypes *QQ*, *Qq* or *qq*) and variance σ^2 in the expected proportions given the sire's genotype and frequency of the QTL alleles in the dam population. Marker genotypes were then assigned at random according to expected frequencies given the progeny's QTL genotype. Three different family sizes (N = 50, 100, and 250), two numbers of families (s = 20 and 40) and two recombination rates (r = 0.1 and 0.2) were simulated (Table 2). The other parameters were set at $\mu = 0$, $a = 0.5$, $d = 0.1$, $p = 0.5$, $\sigma = 1$ and $t = 0.3$. The above simulations assumed that, except for the linked QTL, sires all had equal breeding values. A further simulation with $N = 250$, $s = 20$, $r = 0.1$, $\mu = 0$, $a = 0.5$, $d = 0.1$, $p = 0.5$, $\sigma = 1$ and $t = 0.3$, was performed accounting for polygenes by adding a half of a sire value randomly sampled from a normal distribution with mean of 0 and variance $^1/_4\sigma_a^2 = 0.0625$ (*i.e.*, $h^2 = 0.25$). Because random mating between sires and dams was assumed, it was not necessary to account for genetic variation among dams for polygenes.

Fifteen replicate populations were generated for each combination of values. For each replicate, maximum likelihood estimates of $(\mu + a)$, $(\mu + d)$, $(\mu - a)$, $\hat{r}$, $\hat{p}$ and $\hat{\sigma}$ were obtained by numerically maximizing the logarithm of the likelihood described in (1), [or (2) for the simulations with a polygenic effect] using the iterative numerical maximization subroutine GEMINI (LALOUEL 1979). This routine guarantees to find the global maximum and extensive testing using different starting values on same sets of data indicated that stopping at local maxima was extremely unlikely. Thus for the replicates reported here, prior values were chosen at random from a range spanning the possible parameter space given the limits imposed by the overall mean and variance of the observed data. Convergence was considered to be reached when the normalized gradient of the likelihood was $<10^{-5}$. Parameters were constrained to wide boundaries, and in the case of r and p, were allowed to go out of the theoretical parameter space during iterations to facilitate convergence. If convergence was reached at values outside the parameter space, iteration was restarted with a different set of priors. Standard errors were estimated from the inverted matrix of the numerically derived second derivatives of the function close to the maximum (LALOUEL 1979). Predicted standard errors of combined parameters were calculated from the predicted variances and covariances of individual parameters. For example, the standard error of $(\mu + a)$ was calculated as the square root of the sums of the squared standard errors of $\hat{a}$ and $\hat{\mu}$ plus twice the covariance between them.

For all replicates the likelihood was also maximized with r fixed at $r = 0.5$ to obtain a χ_1^2 statistic with which to test the null hypothesis of no linkage between the marker and the QTL. As stated above, the four terms which are summed in the likelihood equations (1) and (2) represent the four possible QTL genotypes for each sire. The most likely QTL sire genotype was designated as that corresponding to the maximum of the four sub-likelihoods.

RESULTS

Deterministic simulations: All of the deterministic simulations performed in this study took <5 min to run on a personal computer with a 80486 processor. Thus the method was very rapid and not limited by computer requirements.

Standard errors from confidence intervals: Figure 1 shows the log L profiles as functions of each of the six parameters $\hat{\mu}$, $\hat{a}$, $\hat{d}$, $\hat{r}$, $\hat{p}$ and $\hat{\sigma}$ for two family structures (20×250 and 50×100) with and without a polygenic effect. The likelihood always maximized at the true value of the parameters indicating that the expected value of the log likelihood as given in (3) calculated using deterministic simulation yielded unbiased maximum likelihood estimates, despite it being an approximation. Also shown in Figure 1 are the confidence intervals (*CI20* and *CI50* for the two family structures, respectively)

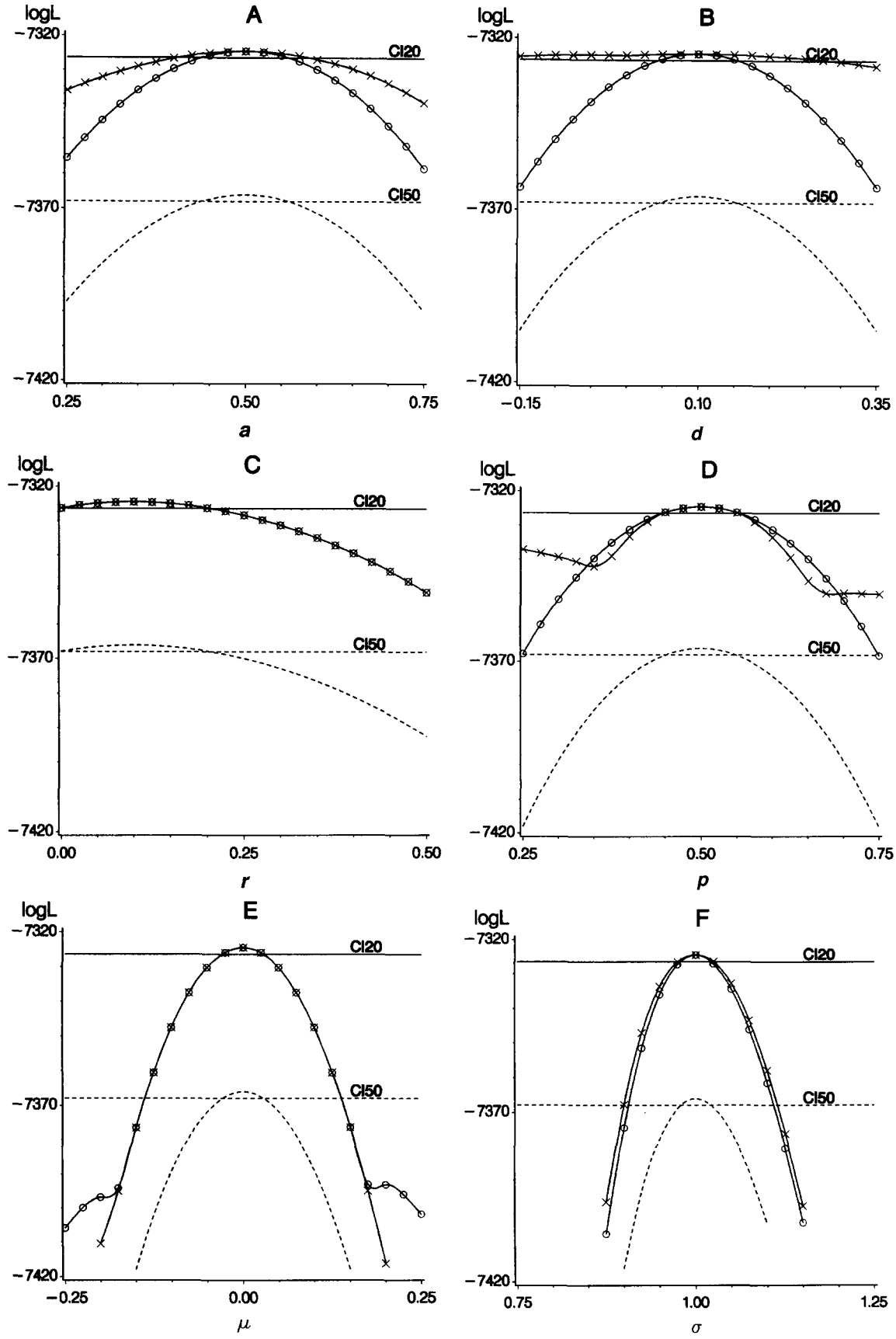

FIGURE 1.—Uniparameter profiles of the expected log likelihood (log L) with respect to parameter estimates assuming all other parameters known for data sets with 5000 records divided into either 20 families without (○——○) and with (×——×) a polygenic effect fitted in the model or 50 families (- - - -) without a polygenic effect. True values of parameters were $a = 0.5$, $d = 0.1$, $r = 0.1$, $p = 0.5$, $t = 0.3$, $\mu = 0$ and $\sigma = 1$. Confidence intervals for estimates from data with 20 families (CI20) and 50 families (CI50) are shown as horizontal lines.

TABLE 3

Standard errors and correlations from deterministic simulations

Method	No. of families	Family size	$\hat{a}$	$\hat{d}$	$\hat{r}$	$\hat{p}$	$\hat{\mu}$	$\hat{\sigma}$
				Standard errors				
Uniparameter								
From CIs	20	250	0.031	0.027	0.053	0.026	0.015	0.010
From matrix			0.031	0.028	0.053	0.027	0.015	0.011
From CIs	50	100	0.030	0.027	0.053	0.024	0.014	0.011
From matrix			0.031	0.028	0.053	0.025	0.015	0.011
From CIs	40	50	0.047	0.043	0.085	0.046	0.023	0.019
From matrix			0.048	0.045	0.086	0.046	0.023	0.019
$h^2 = 0.25$	20	250						
From CI's			0.045	0.098	0.103	0.022	0.014	0.012
From matrix			0.046	0.098	0.053	0.020	0.015	0.011
Multiparameter	20	250	0.042	0.100	0.061	0.077	0.063	0.013
	50	100	0.041	0.100	0.061	0.049	0.057	0.013
	40	50	0.064	0.159	0.097	0.056	0.086	0.021
$h^2 = 0.25$	20	250	0.147	0.104	0.122	0.027	0.020	0.029
				Correlations				
$\hat{a}$	20	250	1.00	−0.096	0.448	0.329	−0.128	−0.538
	50	100	1.00	−0.088	0.467	0.218	−0.022	−0.536
	40	50	1.00	−0.085	0.474	0.157	0.023	−0.535
$h^2 = 0.25$	20	250	1.00	−0.282	0.898	−0.064	0.039	−0.920
$\hat{d}$	20	250		1.00	−0.016	−0.057	−0.759	−0.117
	50	100		1.00	−0.016	−0.037	−0.863	−0.122
	40	50		1.00	0.017	−0.026	−0.908	−0.123
$h^2 = 0.25$	20	250		1.00	−0.243	0.012	−0.005	0.187
$\hat{r}$	20	250			1.00	−0.020	0.025	−0.354
	50	100			1.00	−0.013	0.020	−0.357
	40	50			1.00	−0.009	0.018	−0.358
$h^2 = 0.25$	20	250			1.00	−0.087	0.058	−0.850
$\hat{p}$	20	250				1.00	−0.564	−0.112
	50	100				1.00	−0.402	−0.072
	40	50				1.00	−0.298	−0.051
$h^2 = 0.25$	20	250				1.00	−0.677	0.069
$\hat{\mu}$	20	250					1.00	0.163
	50	100					1.00	0.140
	40	50					1.00	0.131
$h^2 = 0.25$	20	250					1.00	−0.043

Uniparameter and multiparameter lower bound standard errors of parameter estimates, and correlations between them, derived from deterministic simulations for various models and family structures. True parameter values were $\mu = 0.0$, $a = 0.5$, $d = 0.1$, $r = 0.1$, $p = 0.5$, $\sigma = 1$ and $h^2 = 0$, unless otherwise indicated.

around the maximum likelihood estimates. The standard errors derived from these confidence intervals (uniparameter standard errors), and those for the 40 sires × 50 progeny structure (not shown in Figure 1) are given in Table 3. Clearly, r is the most difficult parameter to estimate accurately because of the flatness of the likelihood surface across the range of values (Figure 1C) and largest standard error. The profile likelihoods for $\hat{a}$, $\hat{d}$ and $\hat{p}$ were all similar in curvature (Figure 1, A, B and D) and had similar confidence intervals, indicating that these parameters are estimated to approximately the same degree of accuracy. The profiles for $\hat{\mu}$ and $\hat{\sigma}$ were steepest reflecting the greater information contributing to these parameters (Figure 1, E and F). The effect of decreasing total experimental size from 5000 to 2000 was to increase the standard errors by ~50% for all parameters except for $\hat{p}$. The effect of increasing the number of families on accuracy of all parameters was negligible. Allowing for a polygenic effect in the model seemed to decrease the accuracy of $\hat{a}$ and $\hat{d}$, increase the accuracy of $\hat{p}$ and do nothing to the accuracy of $\hat{r}$, $\hat{\mu}$, and $\hat{\sigma}$, although these effects could not be tested for statistical significance. The reduction in accuracy of $\hat{a}$ and $\hat{d}$ can be explained by the fact that

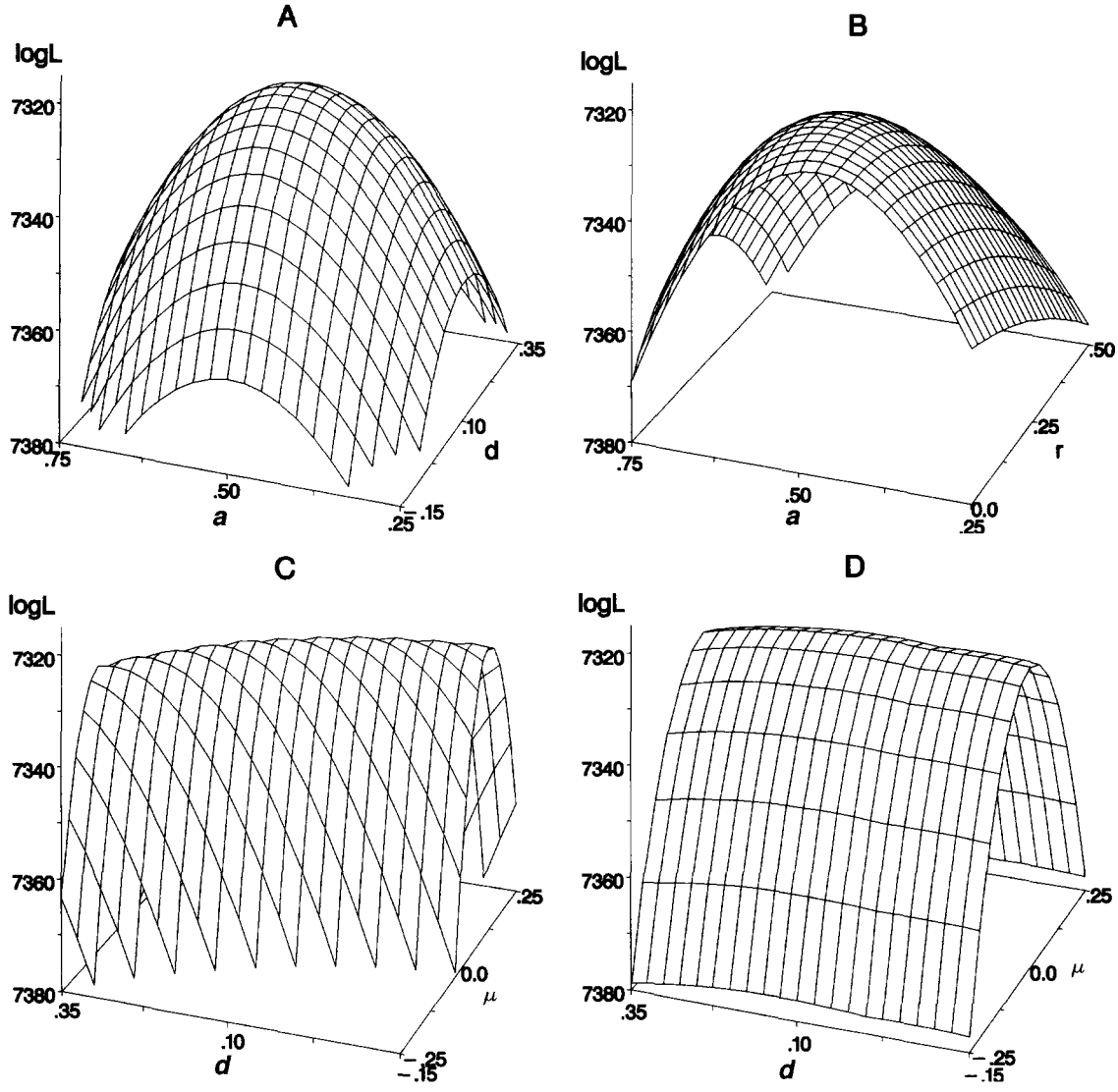

FIGURE 2.—Expected log likelihood surface with respect to estimates of two parameters at a time. (A–C): No polygenic effect was fitted whereas in D it was. Data sets comprised 5000 records divided into 20 families with true parameter values of $a = 0.5$, $d = 0.1$, $r = 0.1$, $p = 0.5$, $t = 0.3$, $\mu = 0$ and $\sigma = 1$.

by estimating the polygenic effect, $\hat{g}_b$, as a function of the sire's major genotype mean, $\hat{G}_b$, this conditions out some of the information on the component parameters of this mean, $\hat{a}$, $\hat{d}$, and $\hat{p}$. While this explains the decrease in accuracy of $\hat{a}$ and $\hat{d}$, it is not clear why the accuracy of $\hat{p}$ increases when this polygenic effect is included. Accuracies of $\hat{r}$, $\hat{\mu}$, and $\hat{\sigma}$ are not affected by the incorporation of a polygenic effect because these parameters are not components of the adjustment term, $\hat{g}_b$.

Standard errors and correlations from the information matrix: Approximate standard errors of the estimates under the four sets of deterministically simulated conditions (Table 2) and the matrix of correlations between them are given in Table 3. Uniparameter standard errors derived from the diagonals of the information matrix were similar to those derived from the confidence intervals from the profile likelihoods. The multiparameter standard errors were considerably higher than the uniparameter standard errors, indicating that confidence intervals derived from profile likelihoods generated by assuming all other parameters are known will overestimate the accuracy of the estimates from a multiparameter model when the parameter estimates are not independent, *i.e.*, correlated. While in general there were low correlations between parameter estimates across the range of conditions tested, there were some notable exceptions: a high correlation (−0.9) between $\hat{\mu}$ and $\hat{d}$, and medium correlations (±0.4 to 0.6) between $\hat{a}$, $\hat{r}$ and $\hat{\sigma}$. This first correlation is presumably because when $p = 0.5$, none of the between-marker contrasts are influenced by d so that the information on d relies on the mean of the *Mm* group of progeny from heterozygous sires. This mean has an expectation of $(\mu \dotplus {}^1/_2 d)$ so that $\hat{d}$ and $\hat{\mu}$ are confounded. This correlation is expected to be lower when p is not equal to 0.5 because the between-marker contrasts then become functions of d. When the polygenic effect is fitted,

this correlation is reduced to ~0 because the expectation of the mean becomes $\hat{\mu} + {}^1/_2\hat{d} + \bar{X}k - {}^1/_2\hat{d} = \hat{\mu} + \bar{X}_k$. Thus $\hat{d}$ and $\hat{\mu}$ are no longer confounded. The correlations between $\hat{a}$, $\hat{r}$ and $\hat{\sigma}$ arise from the fact that the contrast yielding most information on these parameters is *MM vs. mm*, which has an expected value of $(1 - 2r)a/\sigma$. Thus high values of $\hat{a}$ are consistent with high values of $\hat{r}$ and low values of $\hat{\sigma}$. These correlations become higher ($\approx \pm 0.9$) when much of the information on $\hat{a}$ is conditioned out by accounting for the polygenic effect. Even though these correlations exist and influence the accuracy of the estimates, the fact that they are not perfect indicates that there is some independent information contributing to each parameter and thus each of the parameters is, in theory, estimable. These results show that the accuracy of individual parameter estimates is partly dependent on the data structure and partly on the statistical model fitted to the data.

Figure 2 shows three-dimensional representations of the likelihood surface with respect to two parameters at a time ($\hat{a}$ and $\hat{d}$, $\hat{a}$ and $\hat{r}$, $\hat{\mu}$ and $\hat{d}$). The figures illustrate that the surface is relatively uniform in curvature and has a single global maximum. This indicates that, in theory at least, the approach to the maximum during the estimation procedure should be relatively unimpeded. Figure 2, C and D, shows the relationship between $\hat{\mu}$ and $\hat{d}$ with and without a polygenic effect fitted, respectively. It can be seen that while there is a ridge along the axis of $\hat{d}$ in both figures, that the orientation of the ridge when the polygenic effect is not fitted (Figure 1C) is on the diagonal of $\hat{\mu}$ and $\hat{d}$ whereas it is almost parallel with $\hat{d}$ when a polygenic effect is fitted (Figure 1D). This reflects the dependence between $\hat{\mu}$ and $\hat{d}$ in the first case and their independence in the second case. Similarly, the orientation of the profiles for $\hat{a}$ and $\hat{d}$ in the same plane as their corresponding axes in Figure 1A, and the slightly diagonal orientation of the profiles for $\hat{r}$ in Figure 1B, respectively, reflect the independence and moderate dependence between these pairs of parameters. Thus these three-dimensional figures are useful to illustrate the behavior of maximum likelihood estimators in multiparameter models.

Effect of wrong marker allele frequency: Figure 3 shows the bias in estimates of $\hat{a}$, $\hat{r}$ and $\hat{p}$ as functions of assumed marker allele frequency, $\hat{t}$, when the true frequency was $t = 0.3$. It was found that $\hat{r}$ was very sensitive to $\hat{t}$, always being inflated by wrong allele frequencies, $\hat{a}$ was moderately sensitive, and always deflated, and $\hat{p}$ was relatively unaffected. There were no effects on $\hat{d}$, $\hat{\mu}$ and $\hat{\sigma}$. Thus inaccuracies in estimated marker allele frequencies will cause the QTL to appear smaller and less tightly linked than it really is. The magnitude of the bias in recombination fraction was of the order of one standard error in the case presented here.

Stochastic simulations: Stochastic simulations were performed on a Sun SparcStation 2 computer. On this machine, which processes at 2.4×10^6 floating point operations per second, a typical run time per replicate for 30 iterations or ~350 evaluations of the likelihood on 5000 records was ~6 min of CPU. Generally only one run was required for a replicate to reach convergence, although ~10% of runs had to be restarted with different starting values because of parameters estimates exiting the parameter space.

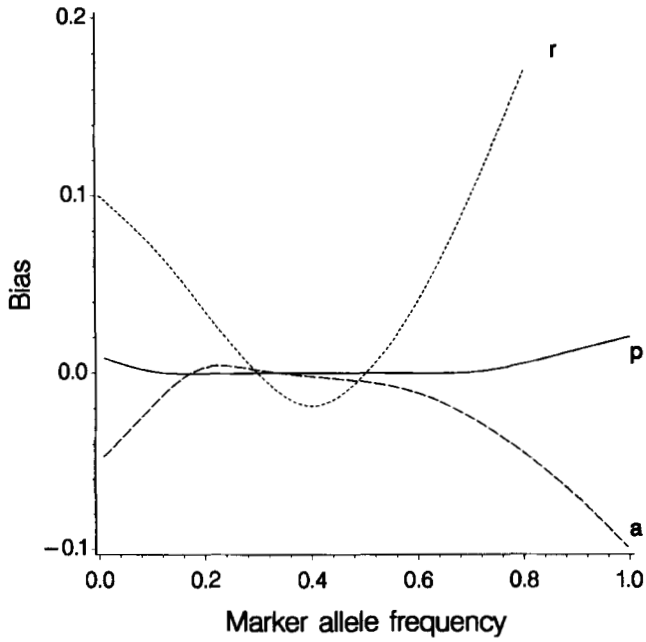

FIGURE 3.—Bias in estimates of $\hat{a}$, $\hat{r}$ and $\hat{p}$ as functions of assumed marker allele frequency, $\hat{t}$, for data sets of 5000 records divided into 20 families and true parameter values of $a = 0.5$, $d = 0.1$, $r = 0.1$, $p = 0.5$, $t = 0.3$, $\mu = 0$ and $\sigma = 1$.

Estimates and standard errors: The estimates from the simulations, averaged over replicates, are given in Table 4. The success in obtaining estimates similar to simulated values shows that the half-sib likelihood model presented in (1) and (2) can separate the effects of the individual parameters and that estimation by numerical maximization is computationally feasible. Close agreement between the means over replicates of simulated and estimated parameter values indicates that the estimates were unbiased or at least that any bias was negligible (Table 4). This applied equally in models with and without the polygenic effect, indicating that the method of approximating and adjusting for the polygenic effects was successful in separating the major (QTL) genotype effect from polygenic effects.

Standard errors of the estimates are also given in Table 4. These standard errors are derived from various sources. The standard error on the simulated data replicate mean (SSE) simply reflects the variation between replicates in the true parameters due to sampling in the simulated data and is based on only 15 observations. The standard error on the estimated parameter (OSE) is also based on 15 observations and is the observed between-replicate variation in estimates of the parameters. The estimated standard error (ESE) is that esti-

TABLE 4

Parameter estimates and standard errors from stochastic simulations

No. families × No. progeny/sire	Parameter estimated					
	$(\mu \hat{+} a)$	$(\mu \hat{+} d)$	$(\mu \hat{-} a)$	$\hat{r}$	$\hat{p}$	$\hat{\sigma}$
20 × 250						
Data ± SSE[a]	0.505 ± 0.036	0.100 ± 0.022	−0.503 ± 0.017	0.103 ± 0.007	0.495 ± 0.008	0.999 ± 0.013
Estimate ± OSE[b]	0.529 ± 0.074	0.075 ± 0.055	−0.488 ± 0.119	0.115 ± 0.057	0.500 ± 0.088	1.000 ± 0.013
ESE ± SE[c]	0.095 ± 0.036	0.074 ± 0.014	0.082 ± 0.021	0.073 ± 0.011	0.087 ± 0.023	0.014 ± 0.002
PSE[d]	0.071	0.067	0.080	0.061	0.077	0.013
20 × 250, $r = 0.2$						
Data ± SSE	0.503 ± 0.034	0.099 ± 0.024	−0.491 ± 0.023	0.199 ± 0.008	0.498 ± 0.010	0.998 ± 0.012
Estimate ± OSE	0.484 ± 0.094	0.073 ± 0.107	−0.461 ± 0.124	0.199 ± 0.078	0.529 ± 0.101	1.004 ± 0.014
ESE ± SE	0.106 ± 0.056	0.092 ± 0.027	0.097 ± 0.028	0.083 ± 0.019	0.088 ± 0.028	0.024 ± 0.034
PSE	0.073	0.069	0.081	0.059	0.077	0.013
20 × 250, $h^2 = 0.25$						
Data ± SSE	0.512 ± 0.032	0.105 ± 0.018	−0.496 ± 0.032	0.100 ± 0.004	0.502 ± 0.007	1.003 ± 0.011
Estimate ± OSE	0.491 ± 0.091	0.146 ± 0.146	−0.490 ± 0.096	0.151 ± 0.089	0.569 ± 0.083	1.004 ± 0.017
ESE ± SE	0.152 ± 0.057	0.165 ± 0.049	0.153 ± 0.046	0.120 ± 0.036	0.103 ± 0.026	0.027 ± 0.008
PSE	0.149	0.106	0.148	0.122	0.028	0.029
20 × 100						
Data ± SSE	0.483 ± 0.051	0.115 ± 0.039	−0.503 ± 0.065	0.100 ± 0.007	0.496 ± 0.009	0.994 ± 0.013
Estimate ± OSE	0.418 ± 0.208	0.119 ± 0.135	−0.468 ± 0.167	0.099 ± 0.124	0.508 ± 0.074	0.996 ± 0.020
ESE ± SE	0.160 ± 0.092	0.149 ± 0.068	0.162 ± 0.088	0.154 ± 0.040	0.096 ± 0.046	0.027 ± 0.011
PSE	0.113	0.093	0.110	0.097	0.079	0.021
40 × 50						
Data ± SSE	0.514 ± 0.035	0.096 ± 0.024	−0.506 ± 0.033	0.103 ± 0.011	0.494 ± 0.006	1.000 ± 0.014
Estimate ± OSE	0.411 ± 0.112	0.051 ± 0.098	−0.393 ± 0.199	0.138 ± 0.094	0.522 ± 0.060	1.017 ± 0.028
ESE ± SE	0.178 ± 0.044	0.155 ± 0.074	0.198 ± 0.111	0.205 ± 0.090	0.084 ± 0.038	0.027 ± 0.007
PSE	0.109	0.088	0.106	0.097	0.056	0.021

Replicate means (with replicate standard errors) of parameters in the data and their estimates from stochastic simulations, their maximum likelihood estimates, and estimated and predicted standard errors for various family sizes and number of families for true values of $\mu = 0.0$, $a = 0.5$, $d = 0.1$, $r = 0.1$, $p = 0.5$, $\sigma = 1$ and $h^2 = 0$, unless otherwise indicated.

[a] Mean and standard error (SSE) of the replicates of simulated data (see text).

[b] Estimates and observed standard error of parameters, OSE.

[c] Estimated standard errors (ESE) of parameters from information matrix using GEMINI.

[d] Multiparameter predicted standard errors (PSE) from deterministic simulations.

mated by GEMINI from the shape of the likelihood surface around the maximum within each replicate of stochastic simulations, averaged over 15 replicates. The predicted standard error (PSE) is the multiparameter standard error derived from the deterministic simulations, given previously in Table 3.

The standard errors for simulated data on 5000 observations were in general agreement with those predicted from the deterministic simulations, which are also shown in Table 4. Where discrepancies occurred between empirical (OSE and ESE) and predicted standard errors, the empirical standard errors (*i.e.*, from the stochastic simulations) were usually higher. However, for the data sets with 2000 observations, the empirical standard errors were always considerably higher than the predicted standard errors. This difference was even greater when smaller families were used. These results confirm that the predicted standard errors from the information matrix are lower bound standard errors. The underestimation is believed to be because, in the calculation of the expected log likelihood based on infinite sampling theory, sampling variation present in stochastic data due to sampling among sire genotypes and among progeny QTL and marker genotypes is not represented. Despite this underestimation, the predicted standard errors were not beyond the lower bound 95% confidence limit of the empirical standard errors and so are not likely to be misleading if they are correctly treated as lower bound estimates.

As r was increased from 0.1 to 0.2, the standard errors of all parameters changed very little. The effects of varying a, d and p were not investigated although this can readily be done by using the theoretical method based on the expected value of log L presented here. Since it is possible to fix σ at unity and measure a and d in units of σ, the effect of varying σ need not be considered.

The log likelihoods, under the full and reduced models, and probabilities for the likelihood ratio chi-squared tests, averaged over replicates, are given in Table 5 for all parameter combinations and data set sizes stochastically simulated. These figures give an average test statistic for the different set of simulated conditions and thereby an indication of the average power of the

TABLE 5

Log likelihood ratio tests and proportion of correctly predicted genotypes

Family structure	log L_0	log L_1	P	No. correct sire genotypes/no. sires
20 × 250	−7313.8	−7320.6	0.039	19.5/20
20 × 250, $r = 0.2$	−7320.2	−7323.5	0.010	18.9/20
20 × 250, $h^2 = 0.25$	−7311.3	−7324.8	0.004	12.7/20
20 × 100	−2917.1	−2921.0	0.028	16.3/20
40 × 50	−2923.4	−2921.1	0.194	26.5/40

Family structures are number of sires × number of progeny for sire. Replicate means of log likelihoods of the full model (log L_0) and the reduced model with $r = 0.5$ (log L_1), the probability, P, of the likelihood ratio test for linkage being significant, and the proportion of sire genotypes ascertained correctly, for true values of $\mu = 0.0$, $a = 0.5$, $d = 0.1$, $r = 0.1$, $p = 0.5$, $\sigma = 1$ and $h^2 = 0$, unless otherwise indicated.

experiment under these conditions. The likelihood ratio test was significant at the 5% level for all combinations, except for the simulations of 40 sires with 50 progeny per sire when it was significant at the 20% level, on average. The corresponding power of these data sets, calculated using the approximate method of WELLER *et al.* (1990) was >90% for all sets except the 40 × 50 data set.

The number of sire genotypes ascertained correctly, averaged over replicates, are also given in Table 5 for all parameter combinations and data set sizes simulated. Prediction of sire genotype was 95% correct for the data sets with 20 sires and 250 progeny per sire. As expected, the proportion of correct assignments decreased with a reduction in the number of records. For the same number of records, the proportion of correct predictions was higher with more progeny per sire. This corresponds to the results given above with respect to standard errors of parameter estimates and likelihood ratio test probabilities. When adjustment for a polygenic effect was included in the model, the number of correctly ascertained sire genotypes was much lower. Presumably, this is because a major source of information on sire genotype, the sire genotype mean, is removed by the adjustment. Thus, for the purposes of sire genotype determination, it may be better to fit a model without a polygenic effect in order to determine which sires have high probabilities of being heterozygous.

DISCUSSION

This study has demonstrated that it is feasible, both computationally and theoretically, to obtain unbiased estimates of the six parameters which describe a QTL in a segregating outbred population. Attention was focused on the accuracy of the estimates of these parameters because this is crucial to the interpretation of the QTL's characteristics and also how effectively the QTL is used in subsequent exploitative breeding programs. The development of a deterministic method based on infinite population sampling theory for predicting the lower bound standard errors of estimates was used to demonstrate the various factors influencing accuracy. These factors are summarized and discussed below. However, the value of the deterministic technique requires discussion first, because it forms part of the basis for the conclusions drawn about those factors affecting accuracy.

Deterministic simulation: The deterministic method was based on infinite sampling in that it predicted numbers of observations with a given value based on the density of the normal distribution and then used these "perfect" samples to perform a weighted analysis of the whole population. This implies that all the sampling variation in the data was due to random normal deviations within QTL genotypes within families. However, in finite populations, there is also expected to be a sampling effect on the numbers of animals that fall into each of the progeny marker-QTL genotype groups and also on the numbers of sires falling into each sire QTL genotype group. These multinomial sampling effects were not accounted for in the calculation of the expected log likelihood (3) because the proportions in each group were considered to be fixed as dictated by the frequency parameters, p and t. While in data sets of 5000 observations, this did not seem to be a problem, in data sets of 2000, it is the most likely cause of the clear underestimation of the predicted standard errors compared with the empirical standard errors, especially when the families were smaller. The intermediate values of p and t used in this study would have exacerbated this sampling variation problem. Modification of (3) to take these sampling effects into account could be performed: to do so would involve calculating the expected log likelihood for values of $n_i c_{ij}$ and sP_h taken over the whole range of the appropriate discrete multinomial distribution, weighting accordingly by the multinomial probabilities of observing each value, and then summing to obtain an overall weighted likelihood. This would be considerably more complex than the method used here but deserves further investigation to see whether it accounts for the underestimation of the standard errors found using deterministic simulation.

Even though the standard errors predicted from deterministic simulation were lower bound standard er-

rors, these estimates are considered to be useful for predicting the minimum size of experiment required to obtain a given level of accuracy of estimation of QTL parameters. Also, the fact that they are lower bound standard errors does not devalue the results concerning their behavior with respect to data structure, statistical model used, and their relationship with each other. This seems to be true because of the similar (though not always parallel) behavior of the empirical standard errors with changes in these conditions. Thus the conclusions drawn below on the factors influencing accuracy are considered valid.

Factors influencing accuracy of parameter estimates: The study identified five separate (though not independent) influences on standard errors of the parameters. The first was the nature and magnitude of the parameter itself. Recombination rate, r, was by far the most inaccurate parameter and, even with 5000 observations, could not be estimated to an accuracy that was able to exclude 0 recombination from the 95% confidence interval. In this study, only a single marker adjacent to the QTL was considered. However, most other studies where accuracy has been examined have also considered two markers which bracket the QTL (KNOTT and HALEY 1992a; VAN OOIJEN 1992; DARVASI *et al.* 1993) in which case standard errors are reduced, though are still large (0.05–0.15) relative to the recombination fraction itself. For example, DARVASI *et al.* (1993) estimated standard errors of 0.11 and 0.05 for a gene with an effect of 0.5 standard deviations located at a recombination distance of ~0.1 from a single marker and bracket markers space 20 cM apart, respectively, in a backcross population of 1000 informative individuals. The closest situation in this study is that for a gene effect of 0.5 standard deviations at a recombination distance of 0.1 in a population of 40 sires, only 20 of which were informative, with 50 progeny each, where the standard error was 0.10. Thus the results presented here concur with those of DARVASI *et al.* (1993) and VAN OOIJEN (1992) that for most feasible experimental sizes, even when two markers bracket the QTL, the resolution of mapping a QTL is near the limit of the marker map itself (10–20 cM) and therefore other techniques need to be employed for fine mapping of the gene.

In contrast, reasonably accurate estimates of the gene effects, a and d, can be obtained. Using the examples above, the standard errors on a were found to be 0.06 and 0.05 in this and the study of DARVASI *et al.* (1993), respectively. This is ~10% of the parameter itself, which is typical of the accuracy of many of the parameters used in routine genetic prediction analyses. Several authors (*e.g.*, SMITH and SIMPSON 1986) have pointed out the importance of the accuracy of QTL parameters in predicting breeding values based on marker-QTL linkages, but most of the loss in prediction accuracy is likely to arise from inaccurate estimates of recombination rate rather than gene effect, as the consequences of inferring a recombinant individual *vs.* a nonrecombinant individual are far greater than overestimating or underestimating by a small fraction the value of an individual. The influence of the magnitude of the gene effect on accuracy was not investigated directly. However, as the gene effect decreases, the power of the analysis also decreases because the between-marker contrast, which is the main source of information, is reduced also. This reduction in power is similar to that caused by a decrease in experimental size; this was investigated in this study and shown to cause a decrease in accuracy.

The population mean and residual standard deviation were estimated relatively accurately as expected from the fact that all individuals in the data contribute information to these parameters, not just those from heterozygous sires. Though not investigated, the accuracy of the QTL allele frequency, p, is expected to decrease as the value of p tends towards extremes. This is because the frequency of informative families will decrease. It is expected that the accuracy of a, d and r will also decrease for the same reason. It is worth noting that accurate estimation of p is very important in some situations because it determines how much scope for genetic improvement there is to bring the favorable QTL to fixation and therefore the value of a marker-assisted within-breed improvement program.

The second factor affecting accuracy of individual parameters is the correlation with other parameters. This study is the first to demonstrate that strong sampling correlations exist between some parameters. This result means that the accuracy of parameters cannot be considered in isolation, because correlations with other parameters cause a decrease in accuracy of the other parameters. The implications are that if one parameter is poorly estimated or biased for some reason, then correlated parameters will also be affected. The relationships between parameters must be considered in the prediction of response to breeding programs that attempt to incorporate information on several parameters simultaneously.

The third factor affecting accuracy is total experimental size, and this influence is large. When experimental size was increased by a factor of 2.5, the accuracy of all parameters improved. It was found that parameters slightly increased in accuracy with an increase in family size or decrease in number of families when the total number of progeny is held constant, reflecting the importance of within-family information to estimate most parameters. These results are concomitant with a decrease in power as family size decreases as also shown by SOLLER and GENIZI (1978) and WELLER *et al.* (1990). It is expected that the influence of family size on accuracy would be more noticeable when the total experimental size is smaller than that used in the present study.

The fourth influence found to affect accuracy was the choice of statistical model. When there are polygenes causing between-family differences, it is necessary to adjust for the average family effect because ignoring

this can lead to overestimation of the QTL effect (KNOTT *et al.* 1990; KNOTT and HALEY 1992b). Here the adjustment used was to subtract the sire's mean, which is similar to the "modal method" for adjusting for sire effects used by HOESCHELE (1988) and LE ROY *et al.* (1989) for segregation analyses to detect major genes without the aid of markers. In this study, it was shown to be an effective way of accounting for between-family variation without prior knowledge of polygenic heritability or simultaneous estimation of the between-family variation from the likelihood analysis. The latter approach was found to be necessary in the KNOTT and HALEY study (1992b), which examined a full-sib design in which family sizes were relatively small and therefore family means were inaccurate. The effect of fitting a polygenic model on parameter accuracy was to decrease the accuracy of the parameters (a and d); this contributed to the adjustment term because, in making the adjustment, much of the information on these parameters is lost. Thus for the purposes of obtaining very accurate estimates of the magnitude of gene effects, an alternative method for adjusting for between family variation, such as that used by KNOTT and HALEY (1992b) may be considered. Certainly, the effect of using the polygenic model on the correlations between parameters was to strengthen them, which is an undesirable consequence if several imperfectly estimated parameters are to be used simultaneosuly in breeding programs.

The final influence on parameter accuracy examined in this study was that of assumed marker allele frequency. When wrong marker allele frequencies were used in the likelihood equation, the QTL was estimated to be larger and more distant from the marker than it really was. This problem is likely to be reduced as the heterozygosity of the marker increases because more progeny will have marker alleles that can be assigned to the sire and dam with certainty. Increased heterozygosity will certainly occur with the widespread transition of use of diallelic markers such as RFLPs to polyallelic markers such as microsatellites. The model presented here can be adjusted to account for multiple alleles by grouping the progeny into three marker genotype groups: one for those with the same genotype as the sire and two for those having either one of the sire's alleles. In the current model, these groups would correspond to the *Mm*, *MM* and *mm* groups and the frequencies of the sire alleles in the dam population, t_1 and t_2 would then replace t and $(1 - t)$ in the matrix C. One problem that may arise is that with many alleles, marker frequencies estimated from the sample will be less accurate and wrongly assumed frequencies will influence the estimates, as shown in this study. One way to avoid this is to only use data from animals that have been unambiguously assigned a paternal and maternal allele, although this would result in loss of information and hence loss of accuracy of the estimates. Sensitivity to assumptions about marker allele frequency is a well known problem in linkage analysis of discrete traits.

So far, the information yielded by the likelihood on the QTL parameters has been discussed. However, the likelihood also yielded information on the genotypes of the sires and was found to very accurately predict the sire's genotype when the number of progeny was >100. This information is of great practical importance when the ultimate goal of the experiment is to select sires carrying favorable QTL alleles for commercial breeding via marker-assisted selection (WELLER and FERNANDO 1991). This prediction method may also be used to eliminate the uninformative QTL homozygous sires at the early stage of a QTL mapping experiment, although where there are multiple QTL to detect, there will be few sires that are homozygous for all QTL and therefore able to be discarded. The confidence with which sires can be classified into QTL genotypes needs further investigation, especially when there are many possible QTL genotypes, in which case the test of one genotype *vs.* the rest requires some knowledge of the properties of this multiple hypothesis test statistic.

The implications of violations of the assumptions of the model will be briefly considered. The model assumed underlying normality and equal variance among QTL genotypes and that the progeny with records were a random sample of the sire's progeny. Skewness and kurtosis may lead to false conclusions about major gene effects (GO *et al.* 1978). If there is skewness, the analysis can be performed on transformed data, as suggested by WELLER (1987), although overcorrection of the skewness can remove some of the information on the QTL from the mixture of distributions. DARVASI (1990) found that estimation of QTL parameters by ML under the assumption of equal genotype variance resulted in accurate estimates for genotype means even if this assumption was incorrect, provided that the actual variances were not radically different. If recorded progeny are a selected sample, estimates of QTL effects will be severely biased (MACKINNON and GEORGES 1992). Incorporation of a truncation parameter in the likelihood model presented here may be one way of accounting for the effects of selection. The degree of selection would, however, have to be known from an independent analysis.

Conclusions: The study showed both theoretically and empirically that unbiased estimates of the six parameters that determine a QTL effect in the half-sib design are able to be estimated by maximum likelihood methodology via linkage to a genetic markers. All parameters except for recombination frequency and the QTL's dominance effect are accurately estimated relative to the size of the parameter itself when there are several thousand progeny records. Further improvements in accuracy may be obtained by including other markers in the model that reduce the background variation (JANSEN 1993; ZENG 1994). Parameter estimates derived by this method could be used as input values to genetic evaluation schemes that include information

on QTL in addition to polygenic effects, pedigree information and fixed effects (FERNANDO and GROSSMAN 1989; BENTSEN and KLEMETSDAL 1991; KENNEDY *et al.* 1992). The high predictive power of individual sires' QTL can be directly applied to marker assisted selection. Computation of the log likelihood expectation is an efficient method for exploration of the accuracy and degree of independence of parameter estimates and power of QTL detection for different experimental designs and parameter values.

Extensive discussions with HENK BOVENHUIS and comments by journal referees are gratefully acknowledged. An anonymous referee is thanked for a formal derivation of Equation 3. Support was provided by the Australian Meat Research Council, CSIRO, and the Israeli Dairy Board.

LITERATURE CITED

BENTSEN, H. B., and G. KLEMETSDAL, 1991 The use of fixed effects models and mixed models to estimate single gene associated effects on polygenic traits. Genet. Sel. Evol. **23:** 407–420.

BOICHARD, D., J. M. ELSEN, P. LE ROY and B. BONAITI, 1990 Segregation analysis of fat content data in Holstein × European crossbred cattle. Poceedings 4th World Congress on Genetics Applied to Livestock Production. Edinburgh. **14:** 167–169.

BOVENHUIS, H., and J. I. WELLER, 1994 Mapping and analysis of dairy cattle quantitative trait loci by maximum likelihood methodology using milk protein genes as genetic markers. Genetics **136:** 267–280.

DARVASI, A., 1990 Analysis of genes affecting quantitative traits with the aid of bracketed pairs of genetic markers and maximum likelihood methodology. M.Sc. Thesis. The Hebrew University, Jerusalem.

DARVASI, A., A. WEINREB, V. MINKE, J. I. WELLER and M. SOLLER, 1993 Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. Genetics **134:** 943–951.

EDWARDS, M. D., C. W. STUBER and J. F. WENDEL, 1987 Molecular-marker-facilitated investigations of quantitative trait loci in maize. I. Numbers, genomic distribution and types of gene action. Genetics **116:** 113–125.

ELSEN J. M., J. VU TIEN KHANG and P. LE ROY, 1988 A statistical model for genotype determination at a major locus in a progeny test design. Genet. Sel. Evol. **20:** 211–226.

FERNANDO, R., and M. GROSSMAN, 1989 Marker assisted selection using best linear unbiased prediction. Genet. Sel. Evol. **21:** 467–477.

FRIES, R., J. S. BECKMANN, M. GEORGES, M. SOLLER and J. WOMACK, 1989 The bovine gene map. Anim. Genet. **20:** 3–29.

GELDERMANN, H., 1975 Investigations on inheritance of quantitative characters in animals by gene markers. Methods Theor. Appl. Genet. **46:** 319–330.

GEORGES, M., D. NIELSEN, M. J. MACKINNON, A. MISHRA, R. OKIMOTO *et al.*, 1995 Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. Genetics **139:** 907–920.

GO, R. C. P., R. C. ELSTON and E. B. KAPLAN, 1978 Efficiency and robustness of pedigree segregation analysis. Am. J. Hum. Genet. **30:** 28–37.

HALEY C. S., A. ARCHIBALD, L. ANDERSSON, A. A. BOSMA, W. DAVIES *et al.*, 1990 The pig gene mapping project—PiGMaP. Proceedings 4th World Congress on Genetics Applied to Livestock Production. Edinburgh. **13:** 67–70.

HOESCHELE, I., 1988 Statistical techniques for detection of major genes in animal breeding data. Theor. Appl. Genet. **76:** 311–319.

JANSEN, R. C., 1993 Interval mapping of multiple quantitative trait loci. Genetics **135:** 205–211.

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

KEIM, P., B. W. DIERS, T. C. OLSON and R. C. SHOEMAKER, 1990 RFLP mapping in soybean: association between marker loci and variation in quantitative traits. Genetics **126:** 735–742.

KENDALL, M. G., and A. STUART, 1973 *The Advanced Theory of Statistics. Vol. 2: Inference and Relationship.* Charles Griffin & Co. Ltd., London.

KENNEDY, B. W., M. QUINTON and J. A. M. VAN ARENDONK, 1992 Estimation of effects of single genes on quantitative traits. J. Anim. Sci. **70:** 2000–2012.

KNOTT, S. A., and C. S. HALEY, 1992a Aspects of maximum likelihood methods for the mapping of quantitative trait loci in line crosses. Genet. Res. **60:** 139–151.

KNOTT, S. A., and C. S. HALEY, 1992b Maximum likelihood mapping of quantitative trait loci using full-sib families. Genetics **132:** 1211–1222.

KNOTT, S. A., C. S. HALEY and R. THOMPSON, 1990 Approximations to segregation analysis for the detection of major genes. Proceedings 4th World Congress on Genetics Applied to Livestock Production. Edinburgh. **13:** 504–507.

LALOUEL, J. M., 1979 GEMINI—a computer program for optimization of general nonlinear functions. Technical Report No. 14. Dept. Human Genetics, Univ. Utah, Salt Lake City, UT.

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

LE ROY, P., J. M. ELSEN and S. A. KNOTT, 1989 Comparison of four statistical methods for detection of a major gene in a progeny test design. Genet. Sel. Evol. **21:** 341–357.

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

MACKINNON, M. J., and M. A. J. GEORGES, 1992 The effects of selection on linkage analysis for quantitative traits. Genetics **132:** 1177–1185.

NEIMANN-SØRENSEN, A., and A. ROBERTSON, 1961 The association between blood groups and several production characters in three Danish cattle breeds. Acta Agr. Scand. **11:** 163–196.

PATERSON, A. H., E. S. LANDER, J. D. HEWITT, S. PETERSON, S. E. LINCOLN *et al.*, 1988 Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. Nature **335:** 721–726.

SAX, K., 1923 Association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris.* Genetics **8:** 552–560.

SIMPSON, S. P., 1989 Detection of linkage between quantitative trait loci and restriction fragment length polymorphisms using inbred lines. Theor. Appl. Genet. **77:** 815–819.

SMITH, C., and S. P. SIMPSON, 1986 The use of genetic polymorphisms in livestock improvement. J. Anim. Breed. Genet. **103:** 205–217.

SOLLER, M., 1990 Genetic mapping of the bovine genome using DNA-level markers with particular attention to loci affecting quantitative traits of economic importance. J. Dairy Sci. **73:** 2628–2646.

SOLLER, M., and A. GENIZI, 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

SOLLER, M., A. GENIZI and T. BRODY, 1976 On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. Theor. Appl. Genet. **47:** 35–39.

VAN OOIJEN, J. W., 1992 Accuracy of mapping quantitative trait loci in autogamous species. Theor. Appl. Genet. **84:** 803–811.

WELLER, J. I., 1986 Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. Biometrics **42:** 627–639.

WELLER, J. I., 1987 Mapping and analysis of quantitative trait loci in *Lycopersicon.* Heredity **59:** 413–421.

WELLER, J. I., 1990 Experimental designs for mapping quantitative trait loci in segregating populations. Proceedings 4th World Congress on Genetics Applied to Livestock Production Edinburgh. **13:** 113–116.

WELLER, J. I., and R. L. FERNANDO, 1991 Strategies for the improvement of animal production using marker assisted selection, pp. 305–328 in *Gene Mapping: Strategies, Techniques and Applications,* edited by L. B. SCHOOK, H. A. LEWIN, and D. G. MCLAREN. Marcel Dekker, New York..

WELLER, J. I., Y. KASHI and M. SOLLER, 1990 Power of daughter and granddaughter designs for determining linkage between marker

loci an quantitative trait loci in dairy cattle. J. Dairy Sci. **73:** 2525–2537.

WILKS, S. S., 1938 The large sample distribution of the likelihood ratio for testing composite hypotheses. Annu. Math. Stat. **99:** 60–62.

ZENG, Z. B., 1994 Precision mapping of quantitative trait loci. Genetics **136:** 1457–1468.

Communicating editor: B. S. WEIR

APPENDIX

The theoretical value of a likelihood function, L, can be derived as follows. For a likelihood function defined by a vector of parameters, θ, and a density function $f(x|\theta)$, the log likelihood, log L, of a single observation, x, conditional on estimates of the parameters, $\hat{\theta}$, is log-$[f(x|\hat{\theta})]$. The theoretical number of observations with value x in a population of size N is $Nf(x|\theta)$ so that the log L over all observations with value x is $Nf(x|\theta)$log-$[f(x|\hat{\theta})]$. Summing over all values of x gives the expected total log likelihood of the distribution of observations as:

$$E(\log L) = N\int_{-\infty}^{\infty} f(x|\theta)\ \log f(x|\hat{\theta}) \cdot dx \qquad \text{(A1)}$$

This equation can be evaluated at different values of $\hat{\theta}$ by numerical integration to give log likelihood profiles for the parameters in $\hat{\theta}$.

For the model presented in this study, the likelihood involves a mixture of three normal distributions of observations (x_l, $l = 1, \ldots, n_i$) with means μ_j, density functions $f(x_l - \mu_j)$ and each with number of observations equal to:

$$n_i \sum_j^3 c_{ij} \qquad \text{(A2)}$$

where $i = 1,12$ depends on the sire's QTL genotype, as in (1). The likelihood of the model is comprised of four sublikelihoods, each conditional on the QTL genotype of the sire, H, and also dependent on the sire's actual genotype, h. For example, if the sire's actual genotype is MQ/mQ ($h = 1$), the expected value of the log of the sublikelihood of the model conditional on the sire having genotype QQ ($H = 1$) over all observations with value x_l is:

$$E_{QQ|\,QQ}(\log L_{x_l}) \approx \sum_{i=1}^{3} n_i \sum_j^3 c_{ij} f(x_l - \mu_j) \times \log \sum_j^3 \hat{c}_{ij} f(x_l - \hat{\mu}_j) \qquad \text{(A3)}$$

and if conditional on the sire being MQ/mq ($H = 2$) is:

$$E_{QQ|\,Qq}(\log L_{x_l}) \approx \sum_{i=1}^{3} n_i \sum_j^3 c_{ij} f(x_l - \mu_j) \times \log \sum_j^3 \hat{c}_{(i+3)j} f(x_l - \hat{\mu}_j) \qquad \text{(A4)}$$

Note that these expected log likelihoods are approximations, the explanation and justification for which are given at the end of this APPENDIX. Integrating over all values of x_l and adding together the antilogs (denoted by exp) of the conditional log sublikelihoods weighted by the prior probability of the sire's genotype, $\hat{P}_H$, gives:

$$E_{QQ}(\log L) \approx \sum_H^4 \exp\Bigg[\log[\hat{P}_H] + \int_{-\infty}^{\infty} \sum_{i=1}^{3} n_i \sum_j^3 c_{ij} f(x_l - \mu_j) \times \log \sum_j^3 \hat{c}_{rj} f_j(x_l - \hat{\mu}_j) \cdot dx\Bigg] \qquad \text{(A5)}$$

for $r = i + 3(H - 1) + 3(h - 1) = i + 3(H - 1)$ in this case.

If the sire's actual genotype is MQ/mq ($h = 2$) then the total number of observations with value x_l becomes:

$$\sum_{i=4}^{6} n_i \sum_j^3 c_{ij} f(x_l - \mu_j) \qquad \text{(A6)}$$

giving the expected value of the log likelihood for such a sire as:

$$E_{Qq}(\log L) \approx \sum_H^4 \exp\Bigg[\log[\hat{P}_H] + \int_{-\infty}^{\infty} \sum_{i=4}^{6} n_i \sum_j^3 c_{ij} f(x_l - \mu_j) \times \log \sum_j^3 \hat{c}_{rj} f_j(x_l - \hat{\mu}_j) \cdot dx\Bigg] \qquad \text{(A7)}$$

Note that now $r = i + 3(H - 1) + 3$ because $h = 1$.

If there are s sires, the number of sires expected to be QQ is $p^2 s$ and for Qq is $p(1 - p)s$, *etc.* Thus the expected value of the total log likelihood is:

$$E(\log L) = s \sum_h^4 P_h \log[E_h(\log L)] \qquad \text{(A8)}$$

which is equivalent to (3).

The approximations used to arrive at this equation are indicated in the derivation given below.

$$
\begin{aligned}
E(\log L) &= E\left[\sum_{b}^{s} \log\left(\sum_{H}^{4} P_H \prod_{i}^{3} \prod_{k}^{n_i} \sum_{J}^{3} c_{IJ} f_J\right)\right] = s \sum_{h}^{4} P_h E\left[\log\left(\sum_{H}^{4} P_H \prod_{i}^{3} \prod_{k}^{n_i} \sum_{J}^{3} c_{IJ} f_J\right)\right] \\
&\approx s \sum_{h}^{4} P_h \log\left(E\left[\sum_{H}^{4} P_H \prod_{i}^{3} \prod_{k}^{n_i} \sum_{J}^{3} c_{IJ} f_J\right]\right) = s \sum_{h}^{4} P_h \log\left(\sum_{H}^{4} E\left[P_H \prod_{i}^{3} \prod_{k}^{n_i} \sum_{J}^{3} c_{IJ} f_J\right]\right) \\
&= s \sum_{h}^{4} P_h \log\left(\sum_{H}^{4} E\left[\exp\left(\log(P_H) + \sum_{i}^{3} \sum_{k}^{n_i} \log\left(\sum_{J}^{3} c_{IJ} f_J\right)\right)\right]\right) \\
&\approx s \sum_{h}^{4} P_h \log\left(\sum_{H}^{4} \exp\left(E[\log(P_H)] + E\left[\sum_{i}^{3} \sum_{k}^{n_i} \log\left(\sum_{J}^{3} c_{IJ} f_J\right)\right]\right)\right) \\
&= s \sum_{h}^{4} P_h \log\left(\sum_{H}^{4} \exp\left(E[\log(P_H)] + \int_{-\infty}^{\infty} \sum_{i}^{3} n_i \sum_{j}^{3} c_{ij} f_j E\left[\log\left(\sum_{J}^{3} c_{IJ} f_J\right)\right]\right)\right) \\
&\approx s \sum_{h}^{4} P_h \log\left(\sum_{H}^{4} \exp\left(\log(P_H) + \int_{-\infty}^{\infty} \sum_{i}^{3} n_i \sum_{j}^{3} c_{ij} f_j \log\left(\sum_{J}^{3} c_{IJ} f_J\right) \cdot dx\right)\right)
\end{aligned}
\tag{A9}
$$

To determine the effect of the approximation on the results on standard errors and correlations presented in this paper, 50 replicates each of 100 progeny from each of the four possible sire genotypes (*i.e.*, 400 progeny per replicate population) were stochastically simulated using the parameters used in this study. Their likelihoods for the true values, and the second derivatives with respect to $\hat{a}$ and $\hat{r}$ were calculated for each replicate and then averaged to give average estimates of standard errors and the correlation between them. The average log L from simulations was -586.324 compared with -589.096 for the $E(\log L)$. Estimated lower bound standard errors of $\hat{a}$, $\hat{r}$ and their correlation from simulations averaged 0.128, 0.232 and -0.438, respectively, and were predicted to be 0.126, 0.226 and -0.380, respectively. Thus there was good agreement between the $E(\log L)$ from (3) and the realized average value of log L indicating that the approximation did not bias the results presented in this study.