Multiple-Trait Mapping of Quantitative Trait Loci After Selective Genotyping Using Logistic Regression

John M. Henshall and Michael E. Goddard[1]

Animal Genetics and Breeding Unit,[2] University of New England, Armidale, New South Wales 2351, Australia

Manuscript received April 14, 1998
Accepted for publication October 30, 1998

ABSTRACT

Experiments to map QTL usually measure several traits, and not uncommonly genotype only those animals that are extreme for some trait(s). Analysis of selectively genotyped, multiple-trait data presents special problems, and most simple methods lead to biased estimates of the QTL effects. The use of logistic regression to estimate QTL effects is described, where the genotype is treated as the dependent variable and the phenotype as the independent variable. In this way selection on phenotype does not bias the results. If normally distributed errors are assumed, the logistic-regression analysis is almost equivalent to a maximum-likelihood analysis, but can be carried out with standard statistical packages. Analysis of a simulated half-sib experiment shows that logistic regression can estimate the effect and position of a QTL without bias and confirms the increased power achieved by multiple-trait analysis.

EXPERIMENTS to detect and locate quantitative trait loci (QTL) in livestock species are becoming more common, driven by the potential for increased genetic gain in traits of economic importance (*e.g.*, Soller 1978; Meuwissen and Goddard 1996). A variety of experimental designs have been used, and various statistical methods have been applied to the resulting data. With commercial livestock, designs commonly measure phenotype and genotype on half-sib families, because of the higher reproductive capacity of males. The phenotype records might be measured directly on the half-sibs, as in a daughter design, or on the progeny of the half-sibs, as in a granddaughter design (Weller *et al.* 1990). The data are commonly analyzed using maximum-likelihood or regression interval mapping (Lander and Botstein 1989; Haley and Knott 1992; Martinez and Curnow 1992). The regression methods are computationally less demanding, which may be relevant if techniques such as permutation testing (Churchill and Doerge 1994) are used to determine significance thresholds. These methods were initially developed for detecting QTL that affect a single trait. Even when phenotype measurements have been available on multiple traits, the results of single-trait analyses have generally been presented in the literature.

In recent years, the possibility of multiple-trait QTL detection has been considered by a number of researchers (*e.g.*, Jiang and Zeng 1995; Korol *et al.* 1995; Weller *et al.* 1997). There are several reasons why multiple-trait QTL mapping is of interest. Not only will an understanding of a QTL's part in the genetic covariance structure of economically important traits be important if selection decisions are to be based on QTL genotype, but the statistical power to detect QTL is potentially higher in multiple-trait analysis than in single-trait analysis. Both Korol *et al.* (1995) and Jiang and Zeng (1995) demonstrate such increased power on simulated datasets while maximizing the likelihood to obtain multiple-trait parameter estimates.

Another approach often applied in livestock QTL detection experiments is selective genotyping, in which only animals with extreme phenotypes are genotyped (see Lebowitz *et al.* 1987; Lander and Botstein 1989; Darvasi and Soller 1992; Muranty and Goffinet 1997; Bovenhuis and Spelman 1998). For a given number of animals genotyped, the power to detect QTL is increased with this approach. However, although simple regression methods can be used to estimate parameters with selective genotyping, the estimates will be biased. To obtain unbiased estimates, maximum likelihood can be applied to the full dataset, including the ungenotyped animals (Lander and Botstein 1989), or approximations to the parameters can be made (Darvasi and Soller 1992; Muranty and Goffinet 1997). Markov chain Monte Carlo methods, which sample missing data, are also appropriate for selectively genotyped data (see Bink *et al.* 1998; Jansen *et al.* 1998). The estimation of the effects of QTL in traits correlated to the trait in which selective genotyping occurred is also problematic (see Weller *et al.* 1997). Unless the selective genotyping and the correlation are taken into account, parameter estimates in the correlated trait will be biased. Maxi-

Corresponding author: John Henshall, Animal Genetics and Breeding Unit, The University of New England, Armidale, NSW 2351, Australia. E-mail: jhenshal@metz.une.edu.au

[1]*Present address:* Institute of Land and Food Resources, University of Melbourne, Parkville, Victoria, 3052 Australia.

[2]Animal Genetics and Breeding Unit (AGBU) is a joint institute of New South Wales Agriculture and The University of New England.

mum-likelihood methods or the less computationally demanding approximation methods of Muranty and Goffinet (1997) and Bovenhuis and Spelman (1998) can be applied. These authors state that these approximations are suitable when QTL effects are small.

Most statistical methods currently used for QTL detection, including those mentioned above, make comparisons between the phenotypes of alternate marker genotypes. Alternatively, the marker genotypes of differing phenotypes can be compared (Stuber *et al.* 1980, 1982). Lebowitz *et al.* (1987) called this approach trait based, as opposed to marker based, and presented methods to compare the marker allele frequencies in divergent selection lines or selectively genotyped individuals.

In this article, a more general trait-based method is presented. The method is suited to half-sib data and addresses the problems that arise with multiple-trait QTL detection on selectively genotyped data. Results comparable with those obtained with maximum likelihood on the full dataset can be achieved, using software in standard statistical packages. The method is regression based, but instead of regressing phenotype on genotype, the regression is genotype on phenotype. This replaces the assumption that the phenotypes are unselected with the assumption that there was no selection based on genotype. This assumption is easily satisfied by including all genotyped animals in the analysis. With half-sib experiments, in treating genotype as the response, the response variable is binary. Methods for analyzing binary data are well understood (*e.g.*, Cox and Snell 1989; Dobson 1990), and suitable subroutines are included in the major statistical computing packages.

STATISTICAL METHODS

Single trait, no recombination model: In the first case considered, it is assumed that there is no recombination between the genotyped locus and the locus affecting the quantitative trait. This would apply when testing for an effect from a candidate gene. A single-sire, half-sib design will be assumed, with no genotypes available on the dams. Let the QTL have two alleles, Q and q, and the genotypic marker have two alleles, M and m. The model can be written as

$$y_i = \mu + s_i\alpha + e_i,$$

where y_i is the phenotypic value for offspring i (adjusted for any contemporary group effects), μ is a sire mean, s_i is an indicator variable taking values of minus one if QTL allele q was inherited from the sire or the value of one if QTL allele Q was inherited from the sire, α is the allele substitution effect, and e_i is a random error term, which includes environmental variance, a genetic effect due to the QTL allele inherited from the dam, and a polygenic effect. The polygenic effect consists of deviation from the sire mean due to Mendelian sampling (0.25 additive genetic variance) and an effect due to the dam (0.5 additive genetic variance). In this model, the indicator variable s_i can be observed.

If we assume the e_i has a normal distribution with variance σ_e^2, then the distribution of phenotypes is a mixture of two normal distributions, with means $(\mu - \alpha)$ and $(\mu + \alpha)$, and common variance σ_e^2. Let these distributions be labeled f^- and f^+. Let Z_i be a random variable that takes the value one if allele M $(= Q$, because no recombination) was inherited, and zero if allele m $(= q)$ was inherited by offspring i, and let the probability that $(Z_i = 1) = p_i$ and the probability that $(Z_i = 0) = (1 - p_i)$. Then

$$p = \frac{f^+}{f^+ + f^-} \tag{1}$$

$$= \frac{\exp(-((Y - \mu) - \alpha)^2/(2\sigma_e^2))}{\exp(-((Y - \mu) - \alpha)^2/(2\sigma_e^2)) + \exp(-((Y - \mu) + \alpha)^2/(2\sigma_e^2))}$$

$$= \frac{\exp(2(Y - \mu)\alpha/\sigma_e^2)}{\exp(2(Y - \mu)\alpha/\sigma_e^2) + 1}. \tag{2}$$

This is the logistic model, with asymptotes zero and one. We can write

$$\log\left(\frac{p}{1 - p}\right) = 2(Y - \mu)\alpha/\sigma_e^2$$

or

$$\log\left(\frac{p}{1 - p}\right) = a + bY,$$

where $a = -2\mu\alpha/\sigma_e^2$ and

$$b = \frac{2\alpha}{\sigma_e^2}. \tag{3}$$

We can estimate a and b with standard logistic-regression software using Z as the response variable and Y as the explanatory variable. Here, for each animal, Z_i is the number of "successes" from one trial if success is inheriting allele M. A software package that allows single observations and a continuous explanatory variable should be chosen, to avoid having to group observations into classes based on phenotype.

The total variance, σ^2, is composed of σ_e^2 and the variance due to the sire QTL allele α^2. We have

$$\sigma^2 = \sigma_e^2 + \alpha^2. \tag{4}$$

Given an estimate of b and an estimate of σ^2 from all of the data, we can solve (3) and (4) for α and σ_e^2 to obtain

$$\alpha = \frac{-1 + \sqrt{1 + b^2\sigma^2}}{b}.$$

It is important that the estimate of σ^2 be from all of the data, not from a selected sample.

Figure 1 contains an example of the function p, where the sire QTL allele accounts for 20% of the total variance. The underlying normal distributions are also shown. The parameter b is related to the "slope," and it is our estimate of b that allows us to estimate the magnitude of the QTL. For the mixture distribution in Figure 1, selecting the upper and lower 5% of observations would exclude records with phenotypes between -2 and 2. For phenotypes of ± 2, the ratio of allele frequencies is around 0.1:0.9, and the shape of the curve p between -2 and 2 can be interpolated reasonably well even without observations in this region. If more extreme selection were applied, or if the QTL effect were larger, then the ratio of allele frequencies at the truncation points might approach 0.0:1.0. Then the shape of the curve between the truncation points could not be reliably interpolated. Selection on the basis of phenotype, as in selective genotyping, will reduce the precision with which we estimate b (and therefore α) compared with genotyping the whole sample, the degree by which the precision is reduced being a function of the percentage of records genotyped and the size of the QTL effect.

To be useful in QTL detection, a measure of how well the model fits the data as well as estimates of the parameters is

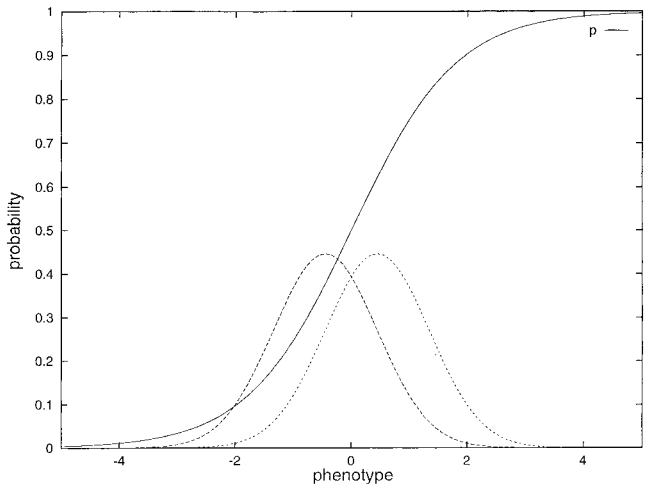

Figure 1.—An example of the function p, the probability that allele Q was inherited from the sire given the phenotype. Here, the sire QTL allele accounts for 20% of the total variance. Also shown are the two components of the underlying mixture distribution.

required. Standard logistic regression software generally provides a log-likelihood ratio test for whether $\hat{b}$ is significantly different from zero. As $b = 2\alpha/\sigma_\varepsilon^2$, this is equivalent to testing the hypothesis that $\alpha = 0$, so these tests may be used to draw conclusions about the significance of α.

Multiple-trait, no recombination model: The methods described above are easily extended to the situation where phenotypes are available on more than one trait. Let f_1 and f_2 in Equation 1 be multivariate normal distributions. If the sire mean is a vector μ, and the vector of half the average sire allele effects A, then the mean of animals in distribution f_1 will be $\mu_1 = \mu + A$, and the mean of animals in distribution f_2 will be $\mu_2 = \mu - A$. If the covariance matrix estimated from the data is Σ, and the covariance matrix within sire QTL genotype is V, then $\Sigma = V + AA'$. As in the single-trait model, V will contain both genetic and nongenetic components. Then,

$$p = \frac{f^+}{f^+ + f^-}$$

$$= \frac{\exp(-\tfrac{1}{2}(Y-\mu_1)'V^{-1}(Y-\mu_1))}{\exp(-\tfrac{1}{2}(Y-\mu_1)'V^{-1}(Y-\mu_1)) + \exp(-\tfrac{1}{2}(Y-\mu_2)'V^{-1}(Y-\mu_2))}$$

$$= \frac{\exp(-\tfrac{1}{2}((Y-\mu_1)'V^{-1}(Y-\mu_1) - (Y-\mu_2)'V^{-1}(Y-\mu_2)))}{\exp(-\tfrac{1}{2}((Y-\mu_1)'V^{-1}(Y-\mu_1) - (Y-\mu_2)'V^{-1}(Y-\mu_2))) + 1}$$

so

$$\log\left(\frac{p}{1-p}\right) = -\frac{1}{2}((Y-\mu_1)'V^{-1}(Y-\mu_1) - (Y-\mu_2)'V^{-1}(Y-\mu_2))$$

$$= -\frac{1}{2}(\mu_1'V^{-1}\mu_1' - \mu_2'V^{-1}\mu_2' - 2Y'V^{-1}(\mu_1-\mu_2)).$$

Multivariate logistic regression packages estimate β, where

$$\log\left(\frac{p}{1-p}\right) = Y'\beta;$$

so, ignoring the intercept in $Y'\beta$,

$$\beta = V^{-1}(\mu_1 - \mu_2)$$
$$= 2V^{-1}A$$

or $V\beta = 2A$. Combining this with $\Sigma = V + AA'$ we get

$$AA'\beta + 2A = \Sigma\beta \quad (5)$$

$$A(A'\beta + 2) = \Sigma\beta$$

$$A = \frac{\Sigma\beta}{A'\beta + 2}; \quad (6)$$

also from (5),

$$\beta'AA'\beta + 2\beta'A + 1 = \beta'\Sigma\beta + 1$$

$$(A'\beta + 1)^2 = \beta'\Sigma\beta + 1$$

$$A'\beta = \sqrt{\beta'\Sigma\beta + 1} - 1; \quad (7)$$

and combining (6) and (7),

$$A = \frac{\Sigma\beta}{1 + \sqrt{\beta'\Sigma\beta + 1}}.$$

So, using multivariate logistic-regression software, we can estimate the vector β, and using an estimate of Σ from the complete experimental data, we can simultaneously estimate multiple-trait QTL effects. Again, any selective genotyping carried out is selection on the explanatory variable(s), and, while it may reduce the precision of the estimates, it should not cause systematic bias.

Recombination model: Both the single- and multiple-trait methods described above assume that there is no recombination between the genotyped locus and the locus affecting the quantitative trait. In QTL detection experiments using markers, this assumption cannot be made. The indicator variable s_i is now unobservable. If we consider a single marker with a recombination rate r with the QTL, the multivariate logistic model becomes

$$p = r + \frac{(1-2r)\exp(Y'\beta)}{\exp(Y'\beta) + 1}, \quad (8)$$

where p is the probability that $Z = 1$, where Z relates to the marker allele inherited. Again this is the logistic equation, with additional parameters r and $(1 - 2r)$. For a single marker, these represent the horizontal asymptotes, with r the lower asymptote and $r + (1 - 2r)$ the upper asymptote. If r is unknown, then in theory, the estimates of the asymptotes provided by standard logistic-regression software could be used to estimate r. However, for QTL of moderate size, there will be little information about the asymptotes in the data. Therefore it is not recommended that this approach be used.

A more common use of this model is where an estimate of α is required, given a map position relative to a number of markers, as in interval mapping. A method is therefore required to summarize the information from multiple markers into a form that can be used in Equation 8, in which we require a vector p and an associated vector r. As for maximum-likelihood methods, we can calculate the probability that animal i inherited the sire allele Q, on the basis of the observed marker transmission, the recombination rates between the postulated locus and the markers, and the assumed mapping function. Let this probability be q_i. Although q_i was estimated from multiple markers, we can proceed as if it had been estimated from a single marker and recover a value p_i, which will be either zero or one, and a value r_i, which will be <0.5. As

$$q = \begin{cases} 1 - r & \text{if } p = 1 \\ r & \text{if } p = 0 \end{cases}$$

and as $r < 0.5$, we can write

$$\begin{cases} p = 0, r = q & \text{if } q < 0.5 \\ p = 1, r = 1 - q & \text{if } q > 0.5. \end{cases}$$

There are several numerical methods that can be used to estimate β given vectors *p* and *r*. In testing, both fitting Equation 8 using nonlinear least squares and iteratively maximizing the likelihood of Equation 8 appeared to work well with a single trait. However, as the focus of this article is on using standard statistical software for multiple-trait analysis, an approximate method of interval mapping is described. There are two parts to the problem, the estimation of β and the evaluation of the log-likelihood ratio statistic. We have already shown how to estimate β at the markers, and provided that the markers are not too far apart, simple interpolation will provide sufficiently accurate estimates between the markers. Given an estimate of β, the evaluation of the log-likelihood is straightforward. For the logistic model, the log-likelihood takes the form

$$D = 2\sum o \log\frac{o}{e}, \tag{9}$$

where *e* is the expected frequency and *o* is the observed frequency (Dobson 1990). The estimate of β can be substituted into (8) along with the vector of recombination rates *r* to estimate $\hat{p}$. Then, (9) can be evaluated using *p* and $\hat{p}$ as the observed and expected frequencies. If the interpolated estimate of β is to be used to estimate the QTL effect, then an adjustment for recombination between the markers and the QTL will be required.

RESULTS OF SIMULATION STUDIES

To compare the logistic-regression method to alternative methods and to generally examine its performance, various simulation studies were carried out. Phenotypic values were generated by adding randomly generated error terms to genotypic values. In all cases, a half-sib design was simulated with phenotypes available on 1000 half-sibs. All simulations were repeated 100 times, and mean estimates and significance levels were calculated.

Single trait, no recombination: A sire QTL effect and an error term were simulated, with the sire allele accounting for 0, 1, 4, and 25% of the total phenotypic variance. As the total phenotypic variance was 1.0, the magnitudes of the sire allele effects were 0.0, 0.1, 0.2, and 0.5. Three levels of selective genotyping were tested, with genotypes available on all animals (*i.e.*, no selective genotyping), on 50% of animals, and on 10% of animals. Where selective genotyping was applied, genotypes for the animals with the highest and lowest phenotypes were made available.

The allele effect (α) was estimated with logistic regression (LR), with maximum likelihood (ML), and with the methods of Darvasi and Soller (1992; DS) and Muranty and Goffinet (1997; MG). Estimates of *b* for the LR method were obtained using SAS procedure LOGISTIC (SAS 1990). The response variable, or marker genotype, was coded 0.0 or 1.0, the independent variable was the phenotype, and a dummy variable *n*, for number of trials, was set to 1. Where markers were uninformative, or the animal was not genotyped, then that record was not included. However, all of the records were used to estimate the variance (σ^2), which is required to estimate α. For the ML analysis, the log-likelihood was maximized numerically using NAG subroutine E04JAF (NAG 1991). The likelihood used was

$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \left(q_i \exp\left(\frac{-1}{2\sigma^2}(y_i - \mu - \alpha)^2\right) + (1 - q_i) \exp\left(\frac{-1}{2\sigma^2}(y_i - \mu + \alpha)^2\right)\right), \tag{10}$$

where q_i is either 1 or 0 if the marker is informative, and 0.5 if the marker is uninformative, or if the animal was not genotyped. The DS and MG methods require no special software.

Three models were used to simulate the data. In the first model, the error term was normally distributed, and the selective genotyping was the only selection applied to the data. This would be the case if all markers were fully informative. This resulted in truncation selection, with equal proportions informative in each tail. Table 1 summarizes the results obtained. All four methods provide similar estimates and standard errors when QTL effects are small. It is only when the QTL effect is large that differences are observed between the methods. For a QTL effect of 0.5, with selective genotyping, methods DS and MG underestimate the QTL effect, and methods LR and ML overestimate the QTL effect.

To test the performance of the methods when the error is not normally distributed, errors were simulated from other distributions in the second model. Two distributions were used: a mixture distribution, composed of two normal distributions, with the difference in means responsible for a variance of 0.5, and a χ^2-distribution with 4 d.f., scaled to produce a total variance of 1.0. These are similar to the distributions used by other researchers (*e.g.*, Muranty and Goffinet 1997). The mixture distribution could occur because of the effect of other QTL segregating or because of failure to correctly account for fixed effects. The χ^2-error produces a skewed distribution of phenotypes.

Table 2 contains the results for sire QTL allele effects of 0.1 and 0.2 and one result for a sire QTL allele effect of 0.5. When genotype records were available on all animals, the nonnormal error had little effect on the estimates of the QTL effect. With selective genotyping however, all methods had difficulties in estimating the QTL effect. When the error was from the mixture distribution, method DS produced estimates of QTL effect closest to those simulated, with little difference between the mean estimates produced by methods MG, LG, and ML. When the error term was simulated from the χ^2-distribution, the effect of reducing the number of genotype records available appeared to be nonlinear. Genotyping most animals, or small QTL effects, resulted in overestimation of the QTL effect, while genotyping less animals, or large QTL effects, resulted in underestimation of the QTL effect. This pattern was consistent for all of the methods, with methods DS and LR requiring

TABLE 1

Estimates of QTL allele effects, normal error, truncation selection

	Estimate of α			
Gen (%)	DS	MG	LR	ML
		Simulated value = 0.0		
100	0.001 ± 0.003	0.001 ± 0.003	0.001 ± 0.003	0.001 ± 0.003
50	0.000 ± 0.003	0.000 ± 0.003	0.000 ± 0.003	0.000 ± 0.003
10	−0.005 ± 0.004	−0.005 ± 0.004	−0.005 ± 0.004	−0.005 ± 0.004
		Simulated value = 0.1		
100	0.101 ± 0.003	0.101 ± 0.003	0.101 ± 0.003	0.101 ± 0.003
50	0.100 ± 0.003	0.100 ± 0.003	0.100 ± 0.003	0.100 ± 0.003
10	0.097 ± 0.004	0.096 ± 0.004	0.099 ± 0.005	0.097 ± 0.005
		Simulated value = 0.2		
100	0.201 ± 0.003	0.201 ± 0.003	0.201 ± 0.003	0.201 ± 0.003
50	0.200 ± 0.003	0.200 ± 0.003	0.201 ± 0.003	0.201 ± 0.003
10	0.197 ± 0.004	0.196 ± 0.004	0.206 ± 0.005	0.204 ± 0.005
		Simulated value = 0.5		
100	0.501 ± 0.002	0.501 ± 0.002	0.501 ± 0.002	0.501 ± 0.002
50	0.489 ± 0.002	0.490 ± 0.002	0.502 ± 0.003	0.502 ± 0.003
10	0.408 ± 0.002	0.420 ± 0.002	0.526 ± 0.006	0.511 ± 0.005

Estimates are means and standard errors of 100 replicates, where Gen refers to the percentage of animals genotyped. Estimation methods were those of Darvasi and Soller (1992; DS) and Muranty and Goffinet (1997; MG). LR, logistic regression; ML, maximum likelihood.

TABLE 2

Estimates of QTL allele effects, nonnormal error

	Estimate of α			
Gen (%)	DS	MG	LR	ML
		Simulated value = 0.1, mixture error		
100	0.101 ± 0.003	0.101 ± 0.003	0.101 ± 0.003	0.101 ± 0.003
50	0.106 ± 0.003	0.108 ± 0.003	0.108 ± 0.003	0.108 ± 0.003
10	0.114 ± 0.004	0.128 ± 0.005	0.132 ± 0.005	0.131 ± 0.005
		Simulated value = 0.2, mixture error		
100	0.201 ± 0.003	0.201 ± 0.003	0.200 ± 0.003	0.201 ± 0.003
50	0.210 ± 0.003	0.214 ± 0.003	0.214 ± 0.003	0.215 ± 0.003
10	0.221 ± 0.004	0.250 ± 0.005	0.263 ± 0.005	0.270 ± 0.006
		Simulated value = 0.1, χ^2-error		
100	0.100 ± 0.003	0.100 ± 0.003	0.101 ± 0.003	0.100 ± 0.003
50	0.126 ± 0.003	0.120 ± 0.003	0.132 ± 0.003	0.121 ± 0.003
10	0.210 ± 0.005	0.131 ± 0.005	0.228 ± 0.006	0.133 ± 0.005
		Simulated value = 0.2, χ^2-error		
100	0.200 ± 0.003	0.200 ± 0.003	0.205 ± 0.003	0.200 ± 0.003
50	0.241 ± 0.003	0.228 ± 0.003	0.253 ± 0.003	0.232 ± 0.003
10	0.335 ± 0.003	0.207 ± 0.004	0.422 ± 0.005	0.216 ± 0.004
		Simulated value = 0.5, χ^2-error		
10	0.382 ± 0.002	0.327 ± 0.003	0.526 ± 0.006	0.359 ± 0.004

Estimates are means and standard errors of 100 replicates, where Gen refers to the percentage of animals genotyped. Estimation methods were those of Darvasi and Soller (1992; DS) and Muranty and Goffinet (1997; MG). LR, logistic regression; ML, maximum likelihood. Error terms simulated were a mixture of normal distributions and χ^2 with 4 d.f.

TABLE 3

Estimates of QTL allele effects, nontruncation selection

Gen (%)	Estimate of α		
	MG	LR	ML
	Simulated value = 0.0		
100	0.001 ± 0.002	0.002 ± 0.007	0.002 ± 0.004
50	0.000 ± 0.003	−0.000 ± 0.007	0.000 ± 0.004
10	−0.006 ± 0.005	−0.008 ± 0.014	−0.006 ± 0.006
	Simulated value = 0.1		
100	0.080 ± 0.002	0.101 ± 0.007	0.146 ± 0.004
50	0.099 ± 0.003	0.098 ± 0.007	0.108 ± 0.004
10	0.095 ± 0.006	0.103 ± 0.018	0.098 ± 0.006
	Simulated value = 0.2		
100	0.160 ± 0.003	0.200 ± 0.006	0.279 ± 0.004
50	0.199 ± 0.003	0.201 ± 0.007	0.215 ± 0.003
10	0.197 ± 0.006	0.247 ± 0.019	0.206 ± 0.006
	Simulated value = 0.5		
100	0.398 ± 0.007	0.501 ± 0.004	0.585 ± 0.003
50	0.489 ± 0.003	0.506 ± 0.005	0.512 ± 0.003
10	0.421 ± 0.003	0.631 ± 0.019	0.516 ± 0.006

Estimates are means and standard errors of 100 replicates, where Gen refers to the percentage of animals genotyped. Estimation methods were those of Muranty and Goffinet (1997; MG). LR, logistic regression; ML, maximum likelihood. Fifty percent of markers were informative, with the ratio of identifiable alleles being 10:90%.

greater selection, or larger QTL effects, before the underestimation occurred.

Another departure from the model assumed by the estimation methods is when the selection is not truncation selection with fixed proportions of animals with genotype records in each tail. This might occur when the marker allele inherited from the sire cannot be determined for some animals, as occurs when the marker genotype of the animal is the same as that of the sire, with no genotype available for the dam. If one of the sire's marker alleles is at a high frequency in the dam population, then one of the sire's marker alleles will be identified more often in the offspring than the other. The effect of this was tested in the third model, with 50% of genotyped markers assumed to be uninformative, but with one sire allele informative 90% of the time and the other sire allele informative only 10% of the time.

The DS method was not applied to this model because it requires the assumption of known, equal, selected proportions. Results from application of the MG method to data in which all animals were genotyped, are presented for the equation that assumed selection, as it performed better than the equation that did not assume selection. Table 3 contains the results obtained for methods MG, LR, and ML. All of the methods tested have problems with this model. Method MG underestimates the QTL effect, except when the proportion selected is ~50%. For simulated values of α of 0.2 and 0.5, method LR overestimates the QTL effect when only 10% of animals are genotyped. These overestimates are accompanied by relatively high standard errors. Method ML overestimated the allele effect. This was less apparent when selective genotyping was applied.

Multiple trait, no recombination: A bivariate analysis was performed, where the two traits had covariance matrix

$$V = \begin{bmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{bmatrix}$$

within QTL genotype. Vectors of QTL effects $A = [0.3, 0.3]'$, $A = [0.3, 0.0]'$, and $A = [0.0, 0.3]'$ were simulated, with either all animals genotyped, or 10% selective genotyping on the first trait. It was assumed that there was no recombination between the QTL and a marker, with all markers informative. Both single-trait and multiple-trait logistic-regression analyses were carried out, using SAS procedure LOGISTIC. The multiple-trait MG methods and methods of Bovenhuis and Spelman (1998; BS) were also applied to the simulated data. The results are summarized in Table 4.

When all animals are genotyped, there is no difference between the allele effect estimates produced using single- or multiple-trait analysis, and the MG and LR multiple-trait methods produce identical results. When selective genotyping is applied, the single-trait analysis produces good estimates of α_1, the effect on the trait used to select animals for genotyping, but biased estimates of α_2, the effect on the correlated trait. The esti-

TABLE 4

Estimates of QTL allele effects

Method (%)	α_1	α_2	α_1	α_2	α_1	α_2
	Values simulated					
	0.30	0.30	0.30	0.00	0.00	0.30
	Single-trait analysis					
LR100	0.30 (0.003)	0.30 (0.003)	0.30 (0.003)	0.00 (0.003)	0.00 (0.003)	0.30 (0.003)
LR10	0.31 (0.005)	0.43 (0.007)	0.31 (0.005)	0.28 (0.008)	0.00 (0.004)	0.17 (0.007)
	Multiple-trait analysis					
LR100	0.30 (0.003)	0.30 (0.003)	0.30 (0.003)	0.00 (0.003)	0.00 (0.003)	0.30 (0.003)
MG100	0.30 (0.003)	0.30 (0.003)	0.30 (0.003)	−0.00 (0.003)	−0.00 (0.003)	0.30 (0.003)
LR10	0.30 (0.005)	0.29 (0.009)	0.30 (0.005)	0.00 (0.009)	−0.01 (0.005)	0.30 (0.008)
MG10	0.28 (0.004)	0.25 (0.007)	0.28 (0.004)	0.02 (0.007)	−0.01 (0.005)	0.29 (0.008)
BS10		0.25 (0.008)		0.01 (0.008)		0.29 (0.009)

Estimates are means and standard errors of 100 replicates, where estimation is by either logistic regression (LR), the method of Muranty and Goffinet (1997; MG), or the method of Bovenhuis and Spelman (1998; BS), with either all animals genotyped (LR100 and MG100) or 10% of animals genotyped (LR10, MG10, and BS10). α_1 is the effect of the first trait, and α_2 is the effect of the second trait. Where applied, selective genotyping was on the phenotype of the first trait. The within-QTL genotype variance was 1.0, and the within-QTL genotype covariance between the traits was 0.5. Correlations used in the analyses were estimated from the complete data in each replicate.

mates of QTL effect from the multiple-trait methods are much less biased than the single-trait analysis estimates, with the estimates from the LR method less biased than the estimates from the MG and BS methods. For smaller QTL effects there were no differences between the results produced by the multiple-trait methods (results not shown).

Single trait, recombination: Markers and the QTL were simulated in the order

$$M_1 - r_1 - M_2 - r_2 - Q - r_3 - M_3,$$

where M_1, M_2, and M_3 are markers, and Q is the QTL. r_1, r_2, and r_3 are recombination rates, taking the values 0.1, 0.03, and 0.07, respectively. All markers were fully informative. The total phenotypic variance was 1.0, and α was 0.1, so 1% of the total variance was explained by the sire QTL allele.

Interval mapping was carried out, using both ML and LR. For the ML analysis, the likelihood was maximized numerically using NAG subroutine E04JAF. The likelihood used was again (10), but with q_i calculated from the observed marker transmission for the nearest flanking markers and the locus being mapped. Haldane's mapping function was assumed. SAS procedure LOGISTIC was used to obtain the LR estimates of a and b at the markers and the log-likelihood ratio statistic between the markers calculated from interpolated values of a and b.

Figure 2 contains the log-likelihood ratio profiles obtained. When 100% of animals were genotyped, at the markers, there is almost no difference between the log-likelihood statistics produced by the two methods. Between the markers, the LR profile is slightly lower than that produced by ML. In addition to the ML and LR analyses, the data were analyzed using the regression method of Haley and Knott (1992). The resulting profile (not shown) was more similar to the ML profile than to the LR profile. When only 10% of animals were genotyped the log-likelihood statistics were less than when all animals were genotyped, but there was little difference between the ML and LR profiles. Regardless of whether selective genotyping was applied, it appears that provided that the distance between markers is not too great, and provided that most markers are informa-

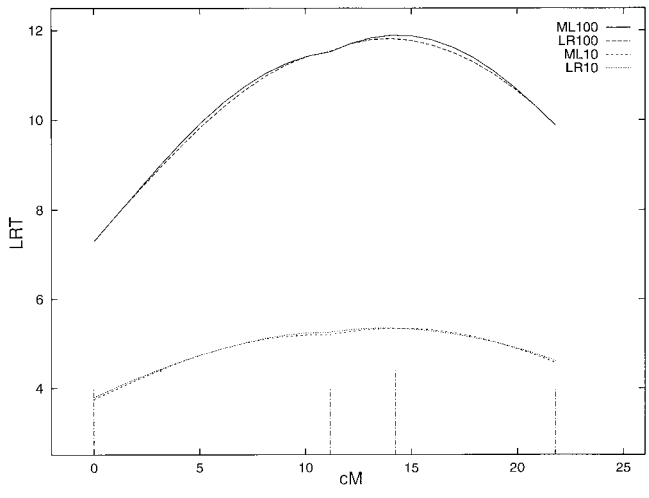

Figure 2.—Log-likelihood ratio test profiles for logistic regression with 100% (LR100) and 10% (LR10) of animals genotyped, and maximum likelihood with 100% (ML100) and 10% (ML10) of animals genotyped. The marker locations are displayed as short impulses and the simulated QTL position is displayed as a long impulse. The sire QTL allele accounts for 1% of the phenotypic variance.

tive, the profiles are for practical purposes equivalent. Having some markers not informative is equivalent to increasing the distance between markers for some animals, and the profiles produced by the two methods may differ. However, if significance thresholds are determined through permutation testing (Churchill and Doerge 1994), similar conclusions should be drawn despite the differences between the profiles.

Multiple trait, recombination: The marker and QTL locations simulated in the single-trait analysis were used for a bivariate analysis. The two traits had covariance matrix

$$V = \begin{bmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{bmatrix}$$

within QTL genotype. A vector of QTL effects $A = [0.1, 0.1]'$ was simulated, with 10% selective genotyping on the first trait. LR analyses were carried out, using SAS procedure LOGISTIC to estimate β at the markers, and with estimates between the markers interpolated. These estimates were then used to calculate the log-likelihood ratio statistic. Single-trait analyses on the two traits were carried out for comparison. Figure 3 plots the mean log-likelihood ratio profiles obtained. The log-likelihood ratio for the second trait is lower than for the first trait, on which the selective genotyping was based, reflecting the loss of power because of selective genotyping on a correlated trait. As in other results reported in the literature, the log-likelihood profile for the multiple-trait analysis is higher than for the single-trait analysis. It must be noted that more degrees of freedom are used in fitting the multiple-trait model.

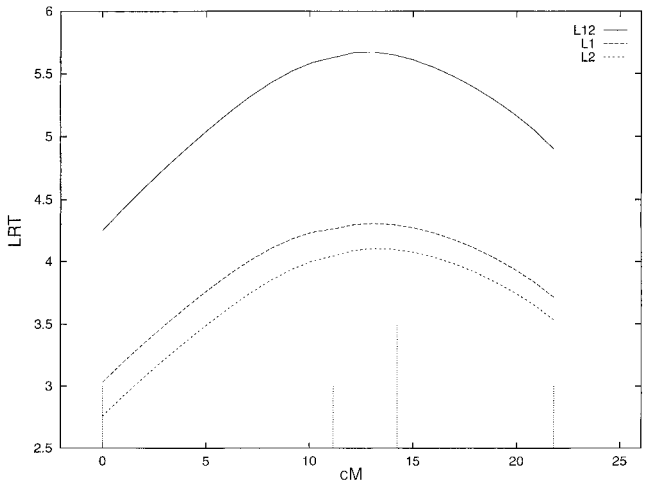

Figure 3.—Log-likelihood ratio test profiles from logistic-regression analysis of multiple-trait data, with single-trait analysis of trait 1 (L1), trait 2 (L2), and a multiple-trait analysis (L12). Selective genotyping was applied to trait 1, with 10% of animals genotyped. The within-QTL genotype variance was 1.0, and the within-QTL genotype covariance between the traits was 0.5. The effect of the QTL on both traits was 0.1.

DISCUSSION

It has been shown in this article that in half-sib families, multiple-trait QTL detection is a simple matter of performing multivariate logistic regression. As no assumptions are made regarding selection among phenotypes, the method is useful for selectively genotyped data as well as for experiments where genotypes are known for all animals.

All methods for estimating QTL effects with missing genotype information require assumptions regarding the distribution of phenotypes within QTL genotype classes. The methods considered in this study all assumed normal distributions for phenotypes, and when the data were simulated according to this assumption, with equal proportions genotyped in each phenotypic tail, differences between the results produced by the methods were minor. The exception to this was when a large QTL was segregating and a small percentage of animals were genotyped. It is under these conditions that the assumptions regarding the distributions of phenotype within unknown genotype become most critical. However, although there is little power for any method to accurately estimate such a QTL effect, in practice this is not a major problem. A more important problem is that most QTL effects are too small to be significant. The situation with very large QTL is that we can tell that they are big, but we can't tell how big.

The robustness of the methods to departure from normal error is of interest. The performance of the methods appears to depend heavily on the distribution of the error, the percentage of animals genotyped, and the relative size of the QTL. When error was generated from a mixture distribution, the DS method produced less bias due to selective genotyping, but with a χ^2-error the MG method and the ML method produced less bias. However, this was only the case for small QTL effects; for larger QTL effects the LR method produced less bias with high levels of selection. It appears that no method is "best" for all circumstances, but all perform reasonably well provided that at least 50% of animals are genotyped. If the data suggest that the distribution of phenotypes is not normal, then the use of data transformations might be considered.

If the selection is not simple truncation selection based on phenotype, then methods such as DS become difficult to implement, because of their use of the expected mean of a truncated distribution. The MG method can be applied to this type of data, but the estimates produced appear to be biased, even when all animals are genotyped. The LR method produces acceptable estimates of QTL effect except when QTL effects are large or selection extreme. The performance of the ML method for this type of data was unsatisfactory. With all animals genotyped the QTL effect was overestimated even for small QTL, but with less genotypic information available the estimates improved. This may be

due to the ratio of QTL genotypes among the animals of unknown genotype. In the likelihood function used in the simulations (Equation 10), when genotype was unknown, a value of 0.5 was coded for the probability of inheriting each of the sire's QTL alleles, implying equal probability of either genotype. With all animals genotyped, almost all of the animals with unknown genotype carry the same QTL allele, but, when only some animals are genotyped, the distribution of QTL alleles in the animals of unknown genotype is more balanced. Probabilities for noninformative markers based on the allele frequencies in the dam population could be assigned to alleviate this problem; the other methods might also benefit from such data preparation.

The DS and MG methods estimate QTL effects as functions of means and variances of the data. As such they are quickly and easily computed. However, the cost of this computational simplicity is reduced generality. The DS method requires simple truncation selection. Two single-trait equations are provided to estimate QTL effects by the MG method, one applicable when all genotypes are available and one applicable when selection has taken place. When the selection is based on other than phenotype, as with noninformative marker information, it is not obvious which equation should be applied.

The LR method has more in common with the ML method computationally. In practice, the LR procedures in statistical packages may use ML to fit the logistic curve to the data. Alternatively another iterative method, such as weighted least squares, may be used. Therefore there may not be much advantage in computing time to using LR. However, LR is a common statistical procedure, and the algorithms for LR in the major statistical packages should be highly optimized. The major advantage of the LR method over ML is the availability and ease of use of appropriate software. For example, the commands required to estimate β and the log-likelihood for a two-trait model with SAS procedure LOGISTIC (SAS 1990) are

```
proc logistic;
  model Q/n = Y Z;
run;
```

where Q is the marker genotype, coded 0 or 1, *n* is the number of trials for each observation (= 1), and Y and Z are the phenotype records. Additional traits are easily added. This is in contrast to the complexities of maximizing multiple-trait likelihood functions.

LR can be used for any experimental design producing markers relating to a mixture of two normal distributions. This would include backcross designs, where the difference between animals heterozygous for the QTL and animals homozygous for the QTL can be estimated, and granddaughter designs, which are essentially half-sib designs with repeated records. Where the genotyped animals are the F_2 generation resulting from a cross between inbred lines, ML methods estimate effects for the two classes of homozygous animals and a dominance effect. As LR is a method for binary data, it is not possible to fit the full model for this type of data directly. However, it should be possible to estimate the differences between two QTL classes, for example, the difference between the homozygous classes or the difference between heterozygous animals and one homozygous class. In this case weights should be used, proportional to the probability that the animal has one of the QTL genotypes under consideration.

As with ML interval mapping, it is desirable to account for both linked and unlinked QTL in estimating QTL effects and locations. Iterative methods, such as that of Zeng (1994) should be adaptable to the LR method. If multiple-trait LR is being performed, then the method is comparable with that of Jiang and Zeng (1995). Also, as for ML or regression-interval mapping, permutation testing (Churchill and Doerge 1994) will provide significance thresholds that should account for any peculiarities in the data.

The results of the simulation studies presented here are no different from those of earlier studies. Jiang and Zeng (1995) and Korol *et al.* (1995) provide convincing arguments for multiple-trait interval mapping. Muranty and Goffinet (1997) use a bivariate ML analysis as a benchmark against which to compare their approximation methods for multiple-trait estimation under selective genotyping. What we have demonstrated here is that using LR, multiple-trait QTL analysis becomes a straightforward application of standard statistical software. The method is applicable regardless of whether selective genotyping was applied. The achievable results are comparable to those obtained from the closely related ML methods, but without the complexity of multiple-trait ML.

J. Henshall was in receipt of a supplementary stipend from the Cooperative Research Centre for Cattle and Beef Industry (Meat Quality) while undertaking this work.

LITERATURE CITED

Bink, M. C. A. M., J. A. M. Van Arendonk and R. L. Quaas, 1998 Breeding value estimation with incomplete marker data. Genet. Sel. Evol. **30**(1): 45–58.

Bovenhuis, H., and R. J. Spelman, 1998 Selective genotyping to detect QTL for multiple traits in outbred populations, pp. 241–244 in *Proceedings of the 6th World Congress on Genetics Applied to Livestock Production 26*, Armidale, New South Wales, Australia.

Churchill, G. A., and R. W. Doerge, 1994 Empirical threshold values for quantitative trait mapping. Genetics **138:** 963–971.

Cox, D. R., and E. J. Snell, 1989 *Analysis of Binary Data*, Ed. 2. Chapman and Hall, London.

Darvasi, A., and M. Soller, 1992 Selective genotyping for determination of linkage between a marker locus and a quantitative locus. Theor. Appl. Genet. **85:** 353–359.

Dobson, A. J., 1990 *An Introduction to Generalized Linear Models*. Chapman and Hall, London.

Haley, C. S., and S. A. Knott, 1992 A simple regression method

for mapping quantitative trait loci in line crosses using flanking markers. Heredity **69:** 315–324.
Jansen, R. C., D. L. Johnson and J. A. M. van Arendonk, 1998 A mixture model approach to the mapping of quantitative trait loci in complex populations with an application to multiple cattle families. Genetics **148:** 391–399.
Jiang, C., and Z.-B. Zeng, 1995 Multiple trait analysis of genetic mapping for quantitative trait loci. Genetics **140:** 1111–1127.
Korol, A. B., Y. I. Ronin and V. M. Kirzhner, 1995 Interval mapping of quantitative trait loci employing correlated trait complexes. Genetics **140:** 1137–1147.
Lander, E. S., and D. Botstein, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121:** 185–199.
Lebowitz, R. J., M. Soller and J. S. Beckman, 1987 Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines. Theor. Appl. Genet. **73:** 556–562.
Martinez, O., and R. N. Curnow, 1992 Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theor. Appl. Genet. **85:** 480–488.
Meuwissen, T. H. E., and M. E. Goddard, 1996 The use of genetic markers in animal breeding schemes. Genet. Sel. Evol. **28:** 161–176.
Muranty, H., and B. Goffinet, 1997 Selective genotyping for location and estimation of the effect of a quantitative trait locus. Biometrics **53**(2): 629–643.
NAG, 1991 *The NAG Fortran Library Manual, Mark 15.* The Numerical Algorithms Group Limited, Oxford.
SAS, 1990 *SAS/STAT User's Guide,* Version 6, Ed. 4. SAS Institute Inc., Cary, NC.
Soller, M., 1978 The use of loci associated with quantitative traits in dairy cattle improvement. Anim. Prod. **27:** 133–139.
Stuber, C. W., R. H. Moll, M. M. Goodman, H. E. Schaffer and B. S. Weir, 1980 Allozyme frequency changes associated with selection for increased grain yield in maize (*Zea mays* L.). Genetics **95:** 225–236.
Stuber, C. W., M. M. Goodman and R. H. Moll. 1982 Improvement of yield and ear number resulting from selection at allozyme loci in a maize population. Crop Sci. **22:** 737–740.
Weller, J. I., Y. Kashi and M. Soller, 1990 Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle. J. Dairy Sci. **73:** 2525–2537.
Weller, J. I., J. Z. Song, Y. I. Ronin and A. B. Korol, 1997 Designs and solutions to multiple trait comparisons. Anim. Biotechnol. **8**(1): 107–122.
Zeng, Z.-B., 1994 Precision mapping of quantitative trait loci. Genetics **136:** 1457–1468.

Communicating editor: C. Haley