Use of the EM algorithm to detect QTL affecting multiple-traits in an across half-sib family analysis

R.J. Kerr, G.M. Mclachlan, J.M. Henshall

Genet. Sel. Evol. 37 (2005) 83–103

DOI: 10.1051/gse:2004037

Original article

Use of the EM algorithm to detect QTL affecting multiple-traits in an across half-sib family analysis

R.J. KERR[a*], G.M. MCLACHLAN[b], J.M. HENSHALL[c]

[a] Animal Genetics and Breeding Unit, University of New England, Armidale 2351, Australia
[b] Department of Mathematics, University of Queensland, Brisbane 4072, Australia
[c] CSIRO Division of Livestock Industries, Australia

(Received 30 October 2003; accepted 17 August 2004)

Abstract – QTL detection experiments in livestock species commonly use the half-sib design. Each male is mated to a number of females, each female producing a limited number of progeny. Analysis consists of attempting to detect associations between phenotype and genotype measured on the progeny. When family sizes are limiting experimenters may wish to incorporate as much information as possible into a single analysis. However, combining information across sires is problematic because of incomplete linkage disequilibrium between the markers and the QTL in the population. This study describes formulæ for obtaining MLEs *via* the expectation maximization (EM) algorithm for use in a multiple-trait, multiple-family analysis. A model specifying a QTL with only two alleles, and a common within sire error variance is assumed. Compared to single-family analyses, power can be improved up to fourfold with multi-family analyses. The accuracy and precision of QTL location estimates are also substantially improved. With small family sizes, the multi-family, multi-trait analyses reduce substantially, but not totally remove, biases in QTL effect estimates. In situations where multiple QTL alleles are segregating the multi-family analysis will average out the effects of the different QTL alleles.

QTL / EM algorithm / interval mapping / half-sib families

1. INTRODUCTION

The paternal half-sib design remains a popular design for the mapping of quantitative trait loci (QTL) in livestock. Families can be easily generated from existing out-crossed populations and the results of the mapping experiment are likely to be more applicable to commercial populations. Designs that use crosses between divergent lines or breeds offer greater chance of detecting QTL. However there is the risk that favourable alleles are already at a high frequency in the target population.

* Corresponding author: richard.kerr@mmigenomics.com
Current address: MMI Genomics, 1756 Picasso Ave, Davis, CA 95616, USA.

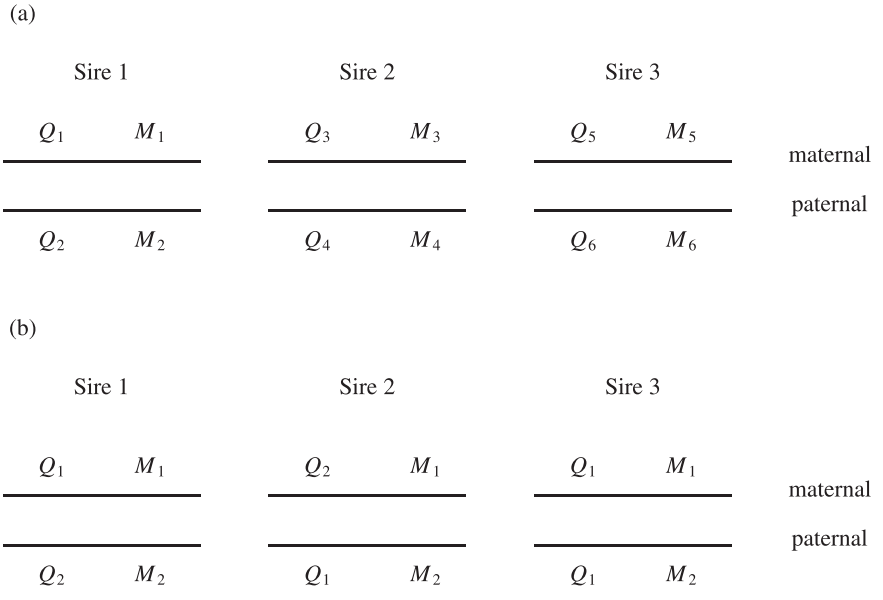

Figure 1. Region of chromosome pair (maternally and paternally inherited) with QTL and marker for 3 sires. (a) Different QTL alleles are assumed linked to different marker alleles in each sire family. (b) Biallelic QTL is assumed and marker alleles are now labelled as being of maternal (1) or paternal (2) origin.

The basis for using a half-sib design can be summarised in Figure 1a, which shows the same region of a chromosome pair in three sires. The region contains a QTL and a genetic marker and represents only a small section of the genome that is to be scanned. A genetic map of moderate resolution is assumed, for example, a 10–20 cM map. Polymorphic markers such as microsatellites span each chromosome of interest. Markers are screened for their informativeness, *i.e.*, there is a high fraction of progeny for which the allele inherited from the sire can be deemed as having derived from one paternal grandparent as opposed to the other. Up to six or more alleles can segregate in a population for a typical microsatellite, hence the marker in Figure 1a is denoted M_i with $i = 1 \ldots 6$. A difference between the phenotypic means of the two offspring groups inheriting the alternative marker alleles indicates that the marker alleles are linked to QTL alleles, and one QTL allele has an effect on phenotype clearly distinquishable from the effect of the other. QTL genotypes are unknown and the exact number of alleles cannot be ascertained. Hence in Figure 1a we have the situation of different QTL alleles linked to different marker alleles in each family. This presents no problem if QTL analyses are completed separately within each half-sib family.

For a fixed experimental resource, it is often preferable to test more small half-sib families than fewer large half-sib families. One reason is that there is greater chance of detecting a rare allele. In the case of a validation study significance tests should not be as stringent as for the initial genome scan. Hence smaller sample sizes suffice, allowing many more families to be screened. However because within sire analysis of small half-sib families provides limited power to detect QTL, experimenters may desire to incorporate information from all sires. Many statistical techniques are available to perform a joint analysis of multiple half-sib families. Linear regression described by Knott *et al.* [12] uses a model that fits a separate QTL effect for each sire, and a common error variance. This model allows multiple QTL alleles to be segregating in the population. Knott *et al.* [12] compared linear regression with maximum likelihood. The likelihood function assumed a biallelic QTL and two possible linkage phases in the population. Numerical methods were used to obtain maximum likelihood estimates (MLEs) of the QTL parameters. They found the two approaches gave similar power and estimates for QTL location. While linear regression [11] and maximum likelihood [9] have been extended to the analysis of multiple traits, there has been no extension of these techniques to the analysis of both multiple traits and multiple half-sib families.

Recently maximum likelihood (ML) has been extended to the analysis of more complex models. For example, Jansen *et al.* [8] describe a ML approach with the potential to analyse complex pedigrees such as multiple half-sib families with genetic ties among families. Likelihoods are optimised using a Monte Carlo expectation-maximization algorithm. Farnir *et al.* [3] describe likelihood functions for analysis of multiple half-sib families assuming linkage disequilibrium between markers and QTL at the population level. Likelihoods are optimised using quasi-Newton techniques. It is unknown how these optimisation techniques are equipped to handle the addition of multiple-traits to the models.

In this paper we describe an exact expectation-maximisation (EM) algorithm suitable for the maximum likelihood analysis of multiple-traits and multiple-families. Further, the benefits of a multi-trait, multi-family analysis, relative to simpler types of analyses are illustrated using computer simulation.

2. MODELS AND ASSUMPTIONS

This algorithm was designed for use in initial genome scans and follow up validation studies. In these situations genetic maps of sparse to moderate resolution are the norm and often mothers are not genotyped. Hence we do not consider likelihood functions that model linkage disequilibrium between the

markers and the QTL at the population level. Such likelihood functions are more applicable in fine mapping studies. The half-sib families are assumed independent. To provide a workable model a biallelic QTL and two possible linkage phases are assumed. Given a fixed map position, markers are used to provide prior probabilities that offspring inherit one of two alternative QTL alleles, labelled Q_1 and Q_2. Under these assumptions it becomes simpler to consider meiosis switches or inheritance states of the marker loci, rather than work with the marker genotypes. The meiosis switch states that the allele transmitted to the offspring due to a meoisis in the sire is either the sire's maternal or paternal allele. Ascertaining the meiosis switches in offspring for all marker loci on the paternal chromosome can be achieved using an algorithm such as the Lander-Green algorithm [13]. In two generational pedigrees such as the half-sib design, the actual phase of the alleles in progeny chromosomes (grandpaternal or grandmaternal origin) remain unknown. However, paternal haplotypes with phase choice having the highest likelihood can still be computed using the Lander-Green algorithm, which is limited to small pedigrees such as independent half-sib families. The chromosomal region in Figure 1b is now represented using different nomenclature to that in Figure 1a. Meiosis switches on marker alleles are represented by 1 (2) for maternal (paternal) origin. The three sires represent three of four possible scenarios: the linkage phase between the QTL and the marker in sire 1 is of the first type, arbitrarily denoted phase 1; the linkage phase in sire 2 is the reverse of that in sire 1 (phase 2); sire 3 is homozygous for the Q_1 allele. A sire can also be homozygous for the Q_2 allele, but this is not shown.

For the development of the following algorithm we will assume two flanking markers rather than a single marker. Flanking markers allow location along a chromosome to be inferred. In half-sib designs it is almost certain that for every individual not every marker can be scored for the allele inherited from the sire. If the flanking marker scores are incomplete for a given individual the typical approach is to use "offsets". That is, the nearest informative marker in one or both directions is taken.

3. MAXIMUM LIKELIHOOD *VIA* THE EM ALGORITHM

The following linear model is used to test for a QTL affecting n_t traits, and located on an interval of markers m and $m + 1$

$$y_{ij1} = \mu_{i1} + (z_{111,ij} + z_{122,ij})b_1 - (z_{112,ij} + z_{121,ij})b_1 + e_{ij1}, \qquad (1)$$

$$y_{ij2} = \mu_{i2} + (z_{111,ij} + z_{122,ij})b_2 - (z_{112,ij} + z_{121,ij})b_2 + e_{ij2}, \quad (2)$$

$$\vdots$$

$$y_{ijn_t} = \mu_{in_t} + (z_{111,ij} + z_{122,ij})b_{n_t} - (z_{112,ij} + z_{121,ij})b_{n_t} + e_{ijn_t} \quad (3)$$

where y_{ijk} is the observation for the kth trait on the jth progeny of the ith sire, μ_{ik} is the mean of the i half-sib family for trait k and b_k is the magnitude of the effect of the QTL on trait k. The b_k are parameterised using the restriction

$$b_k = 0.5\left(b^*_{k1} - b^*_{k2}\right)$$

where b^*_{kl} is the additive effect of allele l (l = 1,2) on the kth trait. The random error terms (e_{ijk}) include environmental variance and a genetic component due to the different polygenic contributions. With a half-sib design there is generally insufficient data to estimate the effects of the QTL allele inherited from the dam, hence this effect is also contained in the random error term. Random error terms are multivariate, normally distributed with a mean that is zero and are uncorrelated between individuals if full-sibs are not included, and correlated between traits recorded on the same individual. A common error variance-covariance matrix, $\mathbf{\Sigma}$, is assumed for all half-sib groups and is defined as

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma^2_1 & \sigma_{12} & \cdots & \sigma_{1n_t} \\ \sigma_{21} & \sigma^2_2 & \cdots & \sigma_{2n_t} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n_t1} & \sigma_{n_t2} & \cdots & \sigma^2_{n_t} \end{pmatrix}.$$

The variables $z_{111,ij}$, $z_{112,ij}$, etc., are indicator variables taking the value zero or one, where:

$z_{111,ij}$ = 1, if the ith sire is heterozygous, has phase 1 and its jth progeny has inherited the Q_1 allele;

$z_{112,ij}$ = 1, if the ith sire is heterozygous, has phase 1 and its jth progeny has inherited the Q_2 allele;

$z_{121,ij}$ = 1, if the ith sire is heterozygous, has phase 2 and its jth progeny has inherited the Q_1 allele;

$z_{122,ij}$ = 1, if the ith sire is heterozygous, has phase 2 and its jth progeny has inherited the Q_2 allele.

The variables $z_{111,ij}$, $z_{121,ij}$, $z_{112,ij}$ and $z_{122,ij}$ all have 1 as the first sub-index, indicating the sire is heterozygous. It takes the value 2 when the sire is homozygous and is necessary in the development of the algorithm in the next

section. The probabilities that the variables $z_{111,ij}$, $z_{121,ij}$, $z_{112,ij}$ and $z_{122,ij}$ are zero or one depend on: the unknown prior probability h that the sire is heterozygous; the probability that the sire is one of two equally likely possible linkage phases; and the p_{ij} denoting the specified prior probabilities that the progeny has inherited the Q_1 allele, conditional on the genotypes of markers m and $m + 1$ and the position being tested ($1 - p_{ij}$ is the prior probability the progeny has inherited the Q_2 allele). For example, the prior probability that $z_{111,ij} = 1$ is equal to $.5hp_{ij}$.

Denoting the number of sires by n_s, the number of progeny within in each half-sib group by n_i, the observations for the ith half-sib group by a $n_i \times n_t$ matrix $\mathbf{Y}_i$, all μ_{ik} within the ith half-sib group by the vector $\boldsymbol{\mu}_i$, and finally all b_k by the vector $\mathbf{b}$, the likelihood function is then given by

$$L(h, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_{n_s}, \mathbf{b}, \boldsymbol{\Sigma}) =$$

$$\prod_{i=1}^{n_s} \left\{ \begin{array}{l} .5h \prod_{j=1}^{n_i} \left(p_{ij}\, \phi(\mathbf{y}_{ij}; \boldsymbol{\mu}_i + \mathbf{b}, \boldsymbol{\Sigma}) + (1 - p_{ij})\, \phi(\mathbf{y}_{ij}; \boldsymbol{\mu}_i - \mathbf{b}, \boldsymbol{\Sigma})\right) \\ + .5h \prod_{j=1}^{n_i} \left(p_{ij}\, \phi(\mathbf{y}_{ij}; \boldsymbol{\mu}_i - \mathbf{b}, \boldsymbol{\Sigma}) + (1 - p_{ij})\, \phi(\mathbf{y}_{ij}; \boldsymbol{\mu}_i + \mathbf{b}, \boldsymbol{\Sigma})\right) \\ + (1 - h) \prod_{j=1}^{n_i} \phi(\mathbf{y}_{ij}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}) \end{array} \right\} \quad (4)$$

where $\phi(\mathbf{y}_{ij}; \boldsymbol{\mu}_i + \mathbf{b}, \boldsymbol{\Sigma})$, $\phi(\mathbf{y}_{ij}; \boldsymbol{\mu}_i - \mathbf{b}, \boldsymbol{\Sigma})$ and $\phi(\mathbf{y}_{ij}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ represent the multivariate normal density functions of the vector variable $\mathbf{y}_{ij}$ (the jth row of the matrix $\mathbf{Y}_i$) with means $\boldsymbol{\mu}_i + \mathbf{b}$, $\boldsymbol{\mu}_i - \mathbf{b}$ and $\boldsymbol{\mu}_i$, respectively, and covariance matrix $\boldsymbol{\Sigma}$.

The univariate representation of this likelihood function has been compared to more complicated likelihood functions by Goffinet *et al.* [5]. Other likelihood functions that were considered modeled different QTL substitution effects and/or different residual variances for each sire. Generally they found that in terms of power there were no appreciable differences between the alternative formulations.

General formulæ for obtaining the MLEs

The maximum likelihood estimate (MLE) of the vector $\boldsymbol{\theta}' = (h, \boldsymbol{\mu}, \mathbf{b}, \boldsymbol{\Sigma})$ of unknown parameters is obtained as an appropriate solution of the likelihood equation

$$\partial \log L(h, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_{n_s}, \mathbf{b}, \boldsymbol{\Sigma}) / \partial \boldsymbol{\theta} = \mathbf{0}. \quad (5)$$

Rather than working with (5) directly, we shall apply the EM algorithm, which is a general method of finding MLEs from a given data set when the data is

incomplete or missing: see, for example, McLachlan and Krishnan [14]. The incomplete data are declared to be the unobservable indicator variables, $z_{111,ij}$, $z_{112,ij}$, $z_{121,ij}$, and $z_{122,ij}$ as defined above. The following indicator variables are also declared to be incomplete data:

$$
\begin{aligned}
z_{1,i} &= 1, \text{ if the } i\text{th sire is heterozygous;} \\
z_{2,i} &= 1, \text{ if the } i\text{th sire is homozygous;} \\
z_{11,i} &= 1, \text{ if the } i\text{th sire is heterozygous and has phase 1;} \\
z_{12,i} &= 1, \text{ if the } i\text{th sire is heterozygous and has phase 2.}
\end{aligned}
$$

Assuming these indicator variables to be incomplete data is equivalent to assuming QTL transmission probabilities, heterozygosity and phase of sires to be missing.

In this framework, the complete-data log likelihood that can be formed on the basis of the observable data and these incomplete data is

$$
\begin{aligned}
\log L_c(\boldsymbol{\theta}) = \sum_{i=1}^{n_s} \sum_{j=1}^{n_i} \Big[& z_{111,ij} \log \phi(\mathbf{y}_{ij};\ \boldsymbol{\mu}_i + \mathbf{b},\ \boldsymbol{\Sigma}) \\
& + z_{112,ij} \log \phi(\mathbf{y}_{ij};\ \boldsymbol{\mu}_i - \mathbf{b},\ \boldsymbol{\Sigma}) \\
& + z_{121,ij} \log \phi(\mathbf{y}_{ij};\ \boldsymbol{\mu}_i + \mathbf{b},\ \boldsymbol{\Sigma}) \\
& + z_{122,ij} \log \phi(\mathbf{y}_{ij};\ \boldsymbol{\mu}_i - \mathbf{b},\ \boldsymbol{\Sigma}) \\
& + (1 - z_{111,ij} - z_{112,ij} - z_{121,ij} - z_{122,ij}) \log \phi(\mathbf{y}_{ij};\ \boldsymbol{\mu}_i,\ \boldsymbol{\Sigma}) \Big] \\
& + \sum_{i=1}^{n_s} \sum_{j=1}^{n_i} \Big[(z_{111,ij} + z_{112,ij} + z_{121,ij} + z_{122,ij}) \log h \\
& + (1 - z_{111,ij} - z_{112,ij} - z_{121,ij} - z_{122,ij}) \log(1 - h) \Big] \\
& + \text{terms involving } p_{ij} \text{ which are known.}
\end{aligned}
$$

The events $[z_{111,ij} = 1]$, $[z_{121,ij} = 1]$, $[z_{112,ij} = 1]$, $[z_{122,ij} = 1]$ and $[1 - z_{111,ij} - z_{112,ij} - z_{121,ij} - z_{122,ij}]$ (the sire is homozygous) are mutually exclusive, therefore you have that the logarithm of the sum is the sum of the logarithms, when passing from the log likelihood of (4) to the complete-data log likelihood. This point illustrates an important difference between standard parametric models and finite mixture distributions. A parametric model is identifiable if distinct values of the parameters determine distinct members of the parametric family. Identifiability for mixture distributions is defined slightly different in that distinct values of the parameters determine distinct members

of the mixture family, allowing permutations of the component labels, *i.e.* the indicator variables; see, for example, McLachlan and Peel [15].

The algorithm proceeds by completing the maximization step or M-step first. At the tth iteration the estimate of the vector $\boldsymbol{\theta}$ of unknown parameters is updated by the global maximizer of $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t-1)})$, which is the conditional expectation of the complete-data log likelihood function given the observed data, using $\boldsymbol{\theta}^{(t-1)}$ for $\boldsymbol{\theta}$. The updated estimates of the unknown parameters $\mathbf{b}$, $\boldsymbol{\mu}_i$ $(i = 1,\ldots,n_t)$ and $\boldsymbol{\Sigma}$ so obtained can be given in closed form. Concerning the updating of the estimates of $\mathbf{b}$, we have that

$$b_k^{(t)} = A_k^{(t)} \left[\sum_{i=1}^{n_s} \sum_{j=1}^{n_i} \left\{ \tau_{111,ij}^{(t-1)} + \tau_{112,ij}^{(t-1)} + \tau_{121,ij}^{(t-1)} + \tau_{122,ij}^{(t-1)} \right\} \right]^{-1} \tag{6}$$

for all k and where

$$A_k^{(t)} = \sum_{i=1}^{n_s} \sum_{j=1}^{n_i} \left[\tau_{111,ij}^{(t-1)} \left(y_{ijk} - \mu_{ik}^{(t-1)} \right) - \tau_{112,ijk}^{(t-1)} \left(y_{ijk} - \mu_{ik}^{(t-1)} \right) - \tau_{121,ij}^{(t-1)} \left(y_{ijk} - \mu_{ik}^{(t-1)} \right) + \tau_{122,ij}^{(t-1)} \left(y_{ijk} - \mu_{ik}^{(t-1)} \right) \right].$$

In the first iteration the following starting values can be used:

$$\begin{aligned} \tau_{111,ij}^{(0)} &= .25 p_{ij}; \\ \tau_{112,ij}^{(0)} &= .25(1 - p_{ij}); \\ \tau_{121,ij}^{(0)} &= .25(1 - p_{ij}); \\ \tau_{122,ij}^{(0)} &= .25 p_{ij}. \end{aligned}$$

To test for the existence of a QTL it will be necessary use a model with no QTL fitted. If this model is completed prior to completing a model with a QTL fitted the resulting half-sib family means can be conveniently used for $\mu_{ik}^{(0)}$. At the tth iteration the means of each half-sib family are found using

$$\begin{aligned} \mu_{ik}^{(t)} = n_i^{-1} \sum_{j=1}^{n_i} \Big[& \tau_{111,ij}^{(t-1)} (y_{ijk} - b_k^{(t)}) \\ & + \tau_{112,ij}^{(t-1)} \left(y_{ijk} + b_k^{(t)} \right) \\ & + \tau_{121,ij}^{(t-1)} \left(y_{ijk} + b_k^{(t)} \right) \\ & + \tau_{122,ij}^{(t-1)} \left(y_{ijk} - b_k^{(t)} \right) \\ & + \left(1 - \tau_{112,ij}^{(t-1)} - \tau_{121,ij}^{(t-1)} - \tau_{122,ij}^{(t-1)} \right) y_{ijk} \Big]. \end{aligned}$$

Next a solution is found for the common variance-covariance matrix in all half-sib families. If an element in the kth row and lth column of $\mathbf{\Sigma}$ is represented as σ_{kl} then

$$\begin{aligned}\sigma_{kl}^{(t)} = N^{-1} \sum_{i=1}^{n_s} \sum_{j=1}^{n_i} \Big[& \tau_{111,ij}^{(t-1)} (y_{ijk} - \mu_{ik}^{(t)} - b_k^{(t)})(y_{ijk} - \mu_{il}^{(t)} - b_l^{(t)}) \\ & + \tau_{112,ij}^{(t-1)} (y_{ijk} - \mu_{ik}^{(t)} + b_k^{(t)})(y_{ijk} - \mu_{il}^{(t)} + b_l^{(t)}) \\ & + \tau_{121,ij}^{(t-1)} (y_{ijk} - \mu_{ik}^{(t)} + b_k^{(t)})(y_{ijk} - \mu_{il}^{(t)} + b_l^{(t)}) \\ & + \tau_{122,ij}^{(t-1)} (y_{ijk} - \mu_{ik}^{(t)} - b_k^{(t)})(y_{ijk} - \mu_{il}^{(t)} - b_l^{(t)}) \\ & + (1 - \tau_{112,ij}^{(t-1)} - \tau_{121,ij}^{(t-1)} - \tau_{122,ij}^{(t-1)})(y_{ijk} - \mu_{ik}^{(t)})(y_{ijk} - \mu_{il}^{(t)}) \Big]\end{aligned}$$

where $N = \sum_{i=1}^{n_s} n_i$.

The expectation or E-step requires taking the conditional expectation of the complete-data log likelihood $\log L_c(\boldsymbol{\theta})$ given the observed data, using the current fit for the vector of unknown parameters. As $\log L_c(\boldsymbol{\theta})$ is linear in the unobservable indicator variables, the E-step is simply effected by replacing them by their conditional expectations given the observed data. As these indicator variables are zero-one variables, their conditional expectations are the posterior probabilities that they are equal to one; that is at the tth iteration

$$\begin{aligned}\tau_{111,ij}^{(t)} &= \text{pr}\{z_{111,ij} = 1 \mid \mathbf{y}_{ij}, \theta^{(t)}\}; \\ \tau_{112,ij}^{(t)} &= \text{pr}\{z_{112,ij} = 1 \mid \mathbf{y}_{ij}, \theta^{(t)}\}; \\ \tau_{121,ij}^{(t)} &= \text{pr}\{z_{121,ij} = 1 \mid \mathbf{y}_{ij}, \theta^{(t)}\}; \\ \tau_{122,ij}^{(t)} &= \text{pr}\{z_{122,ij} = 1 \mid \mathbf{y}_{ij}, \theta^{(t)}\}.\end{aligned}$$

The first step in computing the above posterior probabilities is to compute the posterior probability the ith sire is heterozygous at the QTL, which is denoted $\tau_{1,i}^{(t)}$. To compute $\tau_{1,i}^{(t)}$ we require the probability that the ith sire is heterozygous, $f_{1,i}^{(t)}$ and the probability that the ith sire is homozygous, $f_{2,i}^{(t)}$, at the tth iteration.

That is,

$$
\begin{aligned}
f_{1,i}^{(t)} = \frac{1}{2}h^{(t-1)} \prod_{j=1}^{n_i} \Big[& p_{ij}\, \phi(\mathbf{y}_{ij};\ \boldsymbol{\mu}_i^{(t)} + \mathbf{b}^{(t)},\ \boldsymbol{\Sigma}^{(t)}) \\
& + (1 - p_{ij})\, \phi(\mathbf{y}_{ij};\ \boldsymbol{\mu}_i^{(t)} - \mathbf{b}^{(t)},\ \boldsymbol{\Sigma}^{(t)}) \Big] \\
+ \frac{1}{2}h^{(t-1)} \prod_{j=1}^{n_i} \Big[& p_{ij}\, \phi(\mathbf{y}_{ij};\ \boldsymbol{\mu}_i^{(t)} - \mathbf{b}^{(t)},\ \boldsymbol{\Sigma}^{(t)}) \\
& + (1 - p_{ij})\, \phi(\mathbf{y}_{ij};\ \boldsymbol{\mu}_i^{(t)} + \mathbf{b}^{(t)},\ \boldsymbol{\Sigma}^{(t)}) \Big] \\
f_{2,i}^{(t)} = (1 - h^{(t-1)}) & \prod_{j=1}^{n_i} \phi(\mathbf{y}_{ij};\ \boldsymbol{\mu}_i^{(t)},\ \boldsymbol{\Sigma}^{(t)}).
\end{aligned}
$$

In the first iteration a convenient starting value for $h^{(0)}$ is 0.5. A value for $\tau_{1,i}^{(t)}$ is then found by normalizing. That is,

$$\tau_{1,i}^{(t)} = \frac{f_{1,i}^{(t)}}{f_{1,i}^{(t)} + f_{2,i}^{(t)}}.$$

The posterior probability at the tth iteration that the ith sire is homozygous at the QTL is

$$\tau_{2,i}^{(t)} = 1 - \tau_{1,i}^{(t)}.$$

Once the $\tau_{1,i}^{(t)}$, for $i = 1, \ldots, n_s$, have been computed, the MLE of h at the tth iteration can be found at this point:

$$h^{(t)} = \frac{1}{n_s} \sum_{i=1}^{n_s} \tau_{1,i}^{(t)}.$$

The next step is to compute the posterior probabilities the ith sire is heterozygous, and either phase 1 or phase 2:

$$\tau_{11,i}^{(t)} = \frac{\frac{1}{2} \prod_{j=1}^{n_i} [p_{ij}\, \phi(\mathbf{y}_{ij};\ \boldsymbol{\mu}_i^{(t)} + \mathbf{b}^{(t)},\ \boldsymbol{\Sigma}^{(t)}) + (1 - p_{ij})\, \phi(\mathbf{y}_{ij};\ \boldsymbol{\mu}_i^{(t)} - \mathbf{b}^{(t)},\ \boldsymbol{\Sigma}^{(t)})]}{f_{1,i}^{(t)}}$$

$$\tau_{12,i}^{(t)} = 1 - \tau_{11,i}^{(t)}.$$

Finally, the posterior probabilities that the *j*th progeny received either Q_1 or Q_2 from the *i*th sire, who in turn is heterozygous and of phase 1, are then

$$\tau^{(t)}_{111,ij} = \tau^{(t)}_{1,i} \cdot \tau^{(t)}_{11,i} \cdot \left[\frac{p_{ij}\, \phi(\mathbf{y}_{ij};\ \boldsymbol{\mu}_i^{(t)} + \mathbf{b}^{(t)},\ \boldsymbol{\Sigma}^{(t)})}{p_{ij}\, \phi(\mathbf{y}_{ij};\ \boldsymbol{\mu}_i^{(t)} + \mathbf{b}^{(t)},\ \boldsymbol{\Sigma}^{(t)}) + (1 - p_{ij})\, \phi(\mathbf{y}_{ij};\ \boldsymbol{\mu}_i^{(t)} - \mathbf{b}^{(t)},\ \boldsymbol{\Sigma}^{(t)})} \right]$$

$$\tau^{(t)}_{112,ij} = \tau^{(t)}_{1,i} \cdot \tau^{(t)}_{11,i} \cdot \left[\frac{(1 - p_{ij})\, \phi(\mathbf{y}_{ij};\ \boldsymbol{\mu}_i^{(t)} - \mathbf{b}^{(t)},\ \boldsymbol{\Sigma}^{(t)})}{p_{ij}\, \phi(\mathbf{y}_{ij};\ \boldsymbol{\mu}_i^{(t)} + \mathbf{b}^{(t)},\ \boldsymbol{\Sigma}^{(t)}) + (1 - p_{ij})\, \phi(\mathbf{y}_{ij};\ \boldsymbol{\mu}_i^{(t)} - \mathbf{b}^{(t)},\ \boldsymbol{\Sigma}^{(t)})} \right].$$

Likewise, the posterior probabilities that the *j*th progeny received either Q_1 or Q_2 from the *i*th sire, who in turn is heterozygous and of phase 2, are then

$$\tau^{(t)}_{121,ij} = \tau^{(t)}_{1,i} \cdot \tau^{(t)}_{12,i} \cdot \left[\frac{p_{ij}\, \phi(\mathbf{y}_{ij};\ \boldsymbol{\mu}_i^{(t)} - \mathbf{b}^{(t)},\ \boldsymbol{\Sigma}^{(t)})}{p_{ij}\, \phi(\mathbf{y}_{ij};\ \boldsymbol{\mu}_i^{(t)} - \mathbf{b}^{(t)},\ \boldsymbol{\Sigma}^{(t)}) + (1 - p_{ij})\, \phi(\mathbf{y}_{ij};\ \boldsymbol{\mu}_i^{(t)} + \mathbf{b}^{(t)},\ \boldsymbol{\Sigma}^{(t)})} \right]$$

$$\tau^{(t)}_{122,ij} = \tau^{(t)}_{1,i} \cdot \tau^{(t)}_{12,i} \cdot \left[\frac{(1 - p_{ij})\, \phi(\mathbf{y}_{ij};\ \boldsymbol{\mu}_i^{(t)} + \mathbf{b}^{(t)},\ \boldsymbol{\Sigma}^{(t)})}{p_{ij}\, \phi(\mathbf{y}_{ij};\ \boldsymbol{\mu}_i^{(t)} - \mathbf{b}^{(t)},\ \boldsymbol{\Sigma}^{(t)}) + (1 - p_{ij})\, \phi(\mathbf{y}_{ij};\ \boldsymbol{\mu}_i^{(t)} + \mathbf{b}^{(t)},\ \boldsymbol{\Sigma}^{(t)})} \right].$$

This completes the E-step and a new iteration begins. Convergence is reached when the value of the difference in the log likelihood (4) between successive iterations is below a set threshold.

The EM equations have been described elsewhere for the single-trait, single-family maximum likelihood analysis [16] and the multiple-trait, single familiy maximum likelihood analysis [9]. The EM equations for the single-trait, multiple-family maximum likelihood analysis are simply the univariate representations of the formulae outlined above.

4. COMPARING ANALYSIS METHODS USING SIMULATED DATA

Two QTL detection experiments, denoted experiment A and experiment B were simulated in order to compare the multiple-trait, multiple-family analysis method with simpler methods. Experiment A was run under ideal conditions, in order to eliminate almost all sources of bias, thus demonstrating that the EM algorithm yields unbiased estimates under such conditions. Experiment B was run under less favourable, but more realistic conditions.

In both experiments six half-sib families were simulated with the same number of offspring. In experiments A and B the values of n_i used were 200 and 25, respectively, for $i = 1, \ldots, 6$. A single chromosome with six equally spaced markers and a single QTL was considered. The recombination fraction between each marker was 0.2. In experiment A, the single QTL was positioned in the middle of the third interval, in the middle of the chromosome. In experiment B, the single QTL was located in the last interval, and closest to the last marker. Using the Haldane mapping function these positions translated to distances of 63.9 and 122 cM from the first marker, for experiments A and B, respectively. Paternal allele statuses (1 or 2) for all markers and the QTL were assigned to each progeny. The status of the allele at the first marker was randomly assigned and the status of the allele at the second marker was the same as for the allele of the first marker with probability $1 - r$, with $r = 0.2$. The process was repeated for each successive marker. However r was equal to 0.1127 when sampling the allele at the QTL and the fourth marker in experiment A and was equal to 0.167 and 0.05 when sampling the allele at the QTL and sixth marker, respectively, in experiment B.

In experiment A all progeny were informative at each marker. That is, the paternal allele statuses (1 or 2) of the marker alleles were revealed to the analysis. In actual half-sib experiments the determination of paternal allele status is rarely possible for all progeny, for any one marker. Often, one or more sires are homozygous for a particular marker, or, if a sire is heterozygous, the dam is also segregating for the same two alleles. Thus in experiment B paternal allele statuses (1 and 2) for 75% of the progeny were revealed to the analysis. These progeny were selected at random. The paternal allele statuses of the remaining 25% of the progeny were assumed unknown in the analysis. For each marker a different subset of progeny was selected to have their paternal allele statuses assumed known. The nearest informative markers were used to infer prior probability of inheriting the Q_1 allele in cases when a marker was non-informative.

Table I. Values used in each sire family for the size of the QTL effect (**b**), within sire variance (**Σ**). The linkage phase in each sire family is also shown.

	Sire					
	A	B	C	D	E	**F**
b	[1.58,–1.58][1]	[1.58,–1.58]	[1.58,–1.58]	[1.1,–1.1]	[1.58,–1.58]	[1.58,–1.58]
phase	1	2	1	1	2	not segregating
Σ	[10,5,10][1]	[10,5,10]	[10,5,10]	[10,5,10]	[22.5,11.25,22.5]	[10,5,10]

[1] First and second values are the effects of the QTL on traits 1 and 2, respectively.
[2] First value is the within sire error variance for trait 1, the second value is the covariance between traits and the third value is the variance for trait 2.

Two within sire error terms (e_{ij1} and e_{ij2}), for traits 1 and 2, were assigned to each progeny of each sire and were drawn at random from a bivariate normal distribution with zero means and variances and correlation equal to the values in Table I. Sires A, B, C, D and E were segregating for the QTL. Sire F was not segregating. Sire families segregating for the QTL were assigned a phase of 1 or 2 as shown in Table I. For example, sire A has a phase of 1 which implies $z_{11,A} = 1$ and $z_{12,A} = 0$, whereas sire B has a phase of 2 which implies $z_{11,B} = 0$ and $z_{12,B} = 1$. Sires B and E had opposite linkage phase to sires A, C and D. Paternal allele status at the QTL for the *j*th progeny of the *i*th sire implies that only one of four possible indicator variables ($z_{111,ij}$, $z_{112,ij}$, $z_{121,ij}$, $z_{122,ij}$) is non-zero and trait phenotypes for all progeny can be constructed using equations (1) and (2). The values for b_1 and b_2 are shown in Table I. The values are the same for each sire except sire D, which implies a third QTL allele is segregating in this particular sire. The proportion of the within sire variance due to the QTL was 20% for both traits, except for sire D where it was 10%. One hundred replicate populations containing six half-sib families as described in Table I were generated and analysed with the following interval mapping analysis methods:

– single-family, single trait analyses for each half-sib family;
– single-family, multi-trait analyses for each half-sib family;
– multi-family, single trait analyses for the following combinations of families: ABC (one family has opposite phase); ABD (three alleles are actually segregating); ABE (one family has a larger variance); ABF (one family is not segregating for the QTL);
– multi-family, multi-trait analyses for the above combinations of families.

Chromosomes transmitted to progeny from sires were tested at 22 analysis positions, spaced approximately 6 cM apart. Standard hypotheses were used for testing the existence of a QTL at each tested position. For single-trait analyses the hypotheses tested are

$$\begin{aligned} H_O: &\quad b = 0 \qquad \text{(there is no QTL at that position)} \\ H_A: &\quad b \neq 0 \qquad \text{(there is a QTL at that position).} \end{aligned}$$

For multiple-trait analyses the hypotheses tested are the same as that outlined in Jiang and Zeng [9], that is

$$\begin{aligned} H_O: &\quad \mathbf{b} = 0 \qquad \text{(there is no QTL at that position)} \\ H_A: &\quad \text{at least one element of the vector } \mathbf{b} \text{ is not zero.} \end{aligned}$$

To reject the null hypothesis a likelihood ratio test (LRT) statistic $-2\log[\sup_{\Theta_O} L(\theta)/\sup_{\Theta} L(\theta)]$ was calculated, where Θ_O and Θ are the restricted and unrestricted parameter spaces, respectively. Because classical distribution theory of the LRT statistic does not hold for testing homogeneity against mixture alternatives, the threshold value to reject the null hypothesis cannot be chosen from a χ^2 distribution. Instead empirical threshold values were obtained by permutation testing. Permutation testing was carried out using a variation on the method of Churchill and Doerge [2]. Rather than shuffle phenotypes between progeny, the approach used in the present study changed the identity of a progeny's paternally inherited QTL allele to its alternative, for a random selection of the progeny. For example, if p_{ij} is the probability that progeny j inherited sire i's Q_1 allele, and m_{ij} is a permuted value of p_{ij}, then in each permutation, $m_{ij} = p_{ij}$ with 50% probability, and $m_{ij} = 1 - p_{ij}$ with 50% probability. In previous testing this method produced similar significance levels to the method of Churchill and Doerge [2].

A chromosome wide significance threshold value was obtained for each analysed chromosome by storing the maximum LRT statistic across all the analysis points, for each of 1000 permutations. Once the chromosome is completed, the statistics can then be ranked across permutations to derive an appropriate chromosome wide significance threshold.

In every replicate the maximum LRT statistic on the chromosome was retained for each analysis method. The parameter estimates (position, effects of the QTL, the within-sire variance and the probability of sires being heterozygous) associated with the maximum LRT statistic was also retained. In the results section the means and standard errors of the parameter estimates computed over 100 replicates are presented. This is because means and standard

errors should be reported for estimates computed over the entire parameter space. However, the power of each analysis method was defined as the percentage of replicates in which the maximum LRT statistic exceeded the chromosome wide significance threshold at the 5% level.

5. RESULTS

The results are presented in separate tables for each analysis method: single-trait, single-family (Tab. II); multi-trait, single-family (Tab. III); single-trait, multi-family (Tab. IV); and multi-trait, multi-family (Tab. V). With large half-sib family sizes ($n_i = 200$) all analysis methods resulted in unbiased estimates for all parameters. The power to detect QTL is 100% in all methods. In the single-trait, single-family analysis method, the power is slightly less than 100% only for sires D and E. When analysing sire F, which does not segregate for the QTL, there were no significant results, which is not consistent with the expected false positive rate (5%) under the null hypothesis of no QTL. It is possible that 100 replicates are too few to assess empirically the false positive rate associated with our chromosome wide test. Hence a further 900 replicates were run for this particular sire, using the single-trait, single-family analysis method. Out of 1000 replicates the power to detect QTL was estimated as 5%. The mean position of the maximum LRT statistic when analysing sire F was 62 cM. This is expected when no QTL exists on a chromosome because peaks in the LRT statistic profile should be uniformly distributed across the length of the chromosome. The mean of a uniform distribution which is bounded by 0 and 127.7 is 63.9.

The results when $n_i = 200$ verify the correctness of the maximum likelihood techniques outlined in the methods section. However, it is not realistic to expect family sizes of this magnitude in actual planned experiments. The comparison of analysis methods when $n_i = 25$, and when markers are not completely informative, is of more practical value. Under these conditions the single-trait, single-family analyses have low power to detect QTL. In addition, parameter estimates are biased. Generally, the effect of QTL allele substitution is overestimated and the within sire variance is underestimated. The use of additional information, such as information on correlated traits, and/or by combining information from different sires, will increase power to detect QTL and decrease bias in estimation of QTL parameters. For example, the power of single-trait, single-family analyses to detect QTL ranges from 11 to 34%. The power of multi-trait, multi-family analyses ranges from 92 to 100%. In addition, the estimates of QTL parameters from the multi-trait, multi-family analyses appear

Table II. Results of single-trait, single-family analyses: percentage of replicates in which chromosome wide significance level of 5% was achieved (Power), mean estimated position measured in cM (Position); mean estimated QTL allele effect on trait 1 ($\hat{b}_1$); mean estimated within sire variance for trait 1 ($\hat{\sigma}_1^2$); and for two values of half-sib family size (n_i). Standard error of means in parentheses. IC is information content.

Sire	Power	Position	$\hat{b}_1^{\dagger}$	$\hat{\sigma}_1^{2\ddagger}$	
n_i = 200; true position = 63.9 cM; IC content of markers = 100%					
A	100	64 (7)	1.51 (0.19)	10.07	(0.80)
B	100	61 (10)	1.43 (0.27)	9.90	(0.84)
C	100	62 (13)	1.51 (0.31)	10.02	(0.86)
D	97	65 (16)	1.11 (0.24)	10.01	(0.85)
E	95	62 (12)	1.67 (0.26)	22.80	(2.28)
F	0	62 (39)	0.57 (0.22)	9.83	(0.75)
n_i = 25; true position = 122 cM; IC of markers = 75%					
A	34	100 (33)	1.88 (0.74)	7.77	(2.54)
B	16	93 (44)	1.84 (0.56)	8.50	(2.74)
C	29	92 (45)	1.87 (0.65)	8.26	(2.44)
D	4	85 (45)	1.49 (0.54)	7.67	(2.65)
E	11	80 (45)	2.22 (0.79)	18.93	(6.11)
F	4	56 (49)	1.81 (0.98)	7.21	(2.38)

† True effect: sires A, B, C and E $b_1 = 1.58$; sire D $b_1 = 1.1$.
‡ True variance: sires A, B, C and D $\sigma_1^2 = 10$; sire E $\sigma_1^2 = 22.5$.

to have only slight bias. Importantly, the precision of estimating QTL position has also been vastly improved. Using single-trait, single-family analyses the standard error of mean estimated position ranges from 33 to 49 cM. Using multi-trait, multi-family analyses the standard error ranges from 5 to 12 cM.

The use of additional information on correlated traits alone helps substantially when progeny size is limiting. The power of single-family, multi-trait analyses ranges from 22 to 72%. This represents a 100% improvement in power over single-family, single-trait analyses. Similar improvement over single-trait methods was found in the study of Gilbert and Le Roy [4].

The results show that when analysing a combination of sires, when in truth the sires have different QTL alleles segregating, or, there are different within sire error variances, the bi-allelic, common variance model will tend to average out the effects. For example in the single-trait, multi-family analysis of the

Table III. Results of multi-trait, single-family analyses: percentage of replicates in which chromosome wide significance level of 5% was achieved (Power), mean estimated position measured in cM (Position); mean estimated QTL allele effect on traits 1 and 2 ($\hat{b}_1$, $\hat{b}_2$); mean estimated within sire variances for traits 1 and 2 ($\hat{\sigma}^2_1$, $\hat{\sigma}^2_2$); and for two values of half-sib family size (n_i). Standard error of means in parentheses. IC is information content.

Sire	Power	Position	$\hat{b}_1^{\dagger}$	$\hat{\sigma}^{2\ddagger}_1$	$\hat{b}_2^{\dagger}$	$\hat{\sigma}^{2\ddagger}_2$
n_i = 200; true position 63.9 cM; IC content of markers = 100%						
A	100	64 (2)	1.55 (0.17)	9.95 (0.73)	1.59 (0.26)	9.76 (0.95)
B	100	63 (3)	1.43 (0.26)	9.90 (0.87)	1.71 (0.36)	9.99 (0.91)
C	100	65 (3)	1.56 (0.25)	9.99 (0.85)	1.58 (0.27)	9.63 (1.01)
D	100	63 (5)	1.08 (0.26)	10.05 (1.06)	1.07 (0.25)	9.75 (1.05)
E	100	63 (5)	1.60 (0.33)	22.97 (2.20)	1.59 (0.28)	22.57 (2.70)
F	0	60 (39)	0.50 (0.29)	9.43 (2.04)	0.44 (0.29)	9.70 (1.12)
n_i = 25; true position 122 cM; IC of markers = 75%						
A	72	118 (19)	1.67 (0.74)	8.50 (2.99)	1.54 (0.77)	9.84 (3.00)
B	64	116 (22)	1.64 (0.63)	9.08 (2.73)	1.54 (0.71)	9.85 (3.36)
C	70	118 (22)	1.64 (0.70)	8.97 (2.78)	1.45 (0.71)	9.43 (3.19)
D	29	112 (28)	1.29 (0.63)	8.13 (3.02)	1.21 (0.66)	9.05 (2.69)
E	22	99 (36)	1.79 (1.08)	20.10 (6.84)	1.95 (1.08)	18.89 (6.50)
F	5	67 (45)	1.55 (1.02)	7.53 (2.74)	1.33 (0.85)	7.66 (2.16)

† True effect: sires A, B, C and E b_1, b_2 = 1.58; sire D b_1, b_2 = 1.1.
† True variance: sires A to D σ^2_1, σ^2_2 = 10; sire E σ^2_1, σ^2_2 = 22.5.

sire combination A, B and D, the estimated QTL effect in the larger experiment (n_i = 200) is 1.36. The average of the true QTL effects is (1.58+1.1)/2 = 1.34. In the analysis of sire combination A, B and E, the estimated within sire error variance is 14.34. The average of the true within sire error variances is (10 + 22.2)/2 = 16.1. This averaging of effect estimates seems only a minor disadvantage when considering the substantial improvement in power, and accuracy and precision of QTL location estimates.

Computing time

Table VI presents the computing times for the various types of analyses. Computing time appears to increase linearly with the number of families in the analysis. By far the largest effect on computing time is the number of half-sibs

Table IV. Results of single-trait, multiple-family analyses: percentage of replicates in which chromosome wide significance level of 5% was achieved (Power); mean estimated position measured in cM (Position); mean estimated QTL allele effect on trait 1 ($\hat{b}_1$); mean estimated within sire variance for trait 1 ($\hat{\sigma}_1^2$); mean estimated heterozygosity parameter (h); and for two values of half-sib family size (n_i). Standard errors of means in parentheses. IC is information content.

Sire	Power	Position	$\hat{b}_1^{\dagger}$	$\hat{\sigma}_1^{2\ddagger}$	h
n_i = 200; true position 63.9 cM; IC content of markers = 100%					
ABC	100	64 (3)	1.50 (0.15)	10.03 (0.47)	1.00 (0.00)
ABD	100	64 (4)	1.36 (0.15)	10.08 (0.37)	0.99 (0.04)
ABE	100	63 (7)	1.53 (0.19)	14.34 (0.87)	1.00 (0.02)
ABF	100	63 (6)	1.46 (0.17)	10.02 (0.49)	0.67 (0.04)
n_i = 25; true position 122 cM; IC of markers = 75%					
ABC	56	120 (13)	1.77 (0.47)	9.00 (1.69)	0.94 (0.16)
ABD	39	113 (27)	1.73 (0.62)	8.91 (1.70)	0.87 (0.24)
ABE	42	104 (39)	1.94 (0.76)	12.71 (2.29)	0.88 (0.21)
ABF	29	112 (25)	1.85 (0.48)	8.73 (1.85)	0.80 (0.25)

† True effect: sires A, B, C and E $b_1 = 1.58$; sire D $b_1 = 1.1$.
‡ True variance: sires A to D $\sigma_1^2 = 10$; sire E $\sigma_1^2 = 22.5$.

per sire. The longest analysis (a multi-trait, six-family analysis with large half-sib family sizes) took just under 3 h to complete. The same analysis with the smaller family size took only 17 min to complete.

6. DISCUSSION AND CONCLUSIONS

The analysis of small half-sib families is not uncommon in many Australian QTL studies. A series of DNA marker validation experiments have been performed in the Australian Beef Quality Cooperative Research Centre program to confirm the locations of QTL for beef tenderness, marbling and yield. These are confirmations of linkages initially detected in a large experimental pedigree, known as the CBX cattle [6]. The validations were performed on 45 sires of tropical and temperate origin. The progeny number per sire ranged from 28 to 78. In swine, a QTL mapping project has been directed at providing a resource for evaluating QTL, either discovered in a previous linkage experiment or reported in the literature (Moran, personal communication). The resource consisted of eight sire families comprised of 38 to 65 progeny. In all these

Table V. Results of multi-trait, multiple-family analyses: percentage of replicates in which chromosome wide significance level of 5% was achieved (Power); mean estimated position measured in cM (Position); mean estimated QTL allele effect on traits 1 and 2 ($\hat{b}_1$, $\hat{b}_2$); mean estimated within sire variances for traits 1 and 2 ($\hat{\sigma}^2_1$, $\hat{\sigma}^2_2$); mean estimated heterozygosity parameter (h); and for two values of half-sib family size (n_i). Standard errors of means in parentheses. IC is information content.

Sire	Power	Position	$\hat{b}_1^{\dagger}$	$\hat{\sigma}_1^{2\ddagger}$	$\hat{b}_2^{\dagger}$	$\hat{\sigma}_2^{2\ddagger}$	h
n_i = 200; true position 63.9 cM; IC content of markers = 100%							
ABC	100	64 (0)	1.51 (0.16)	9.97 (0.45)	1.63 (0.19)	9.83 (0.62)	1.00 (0.00)
ABD	100	64 (1)	1.35 (0.13)	10.05 (0.39)	1.46 (0.19)	9.95 (0.51)	1.00 (0.00)
ABE	100	64 (1)	1.54 (0.14)	14.29 (0.83)	1.65 (0.17)	14.11 (1.11)	1.00 (0.00)
ABF	100	64 (1)	1.49 (0.17)	9.97 (0.46)	1.65 (0.24)	9.90 (0.50)	0.67 (0.00)
n_i = 25; true position 122 cM; IC of markers = 75%							
ABC	100	123 (4)	1.57 (0.40)	9.40 (1.68)	1.52 (0.42)	10.05 (1.77)	0.99 (0.05)
ABD	97	124 (5)	1.46 (0.45)	9.18 (1.68)	1.40 (0.45)	10.09 (1.48)	0.95 (0.12)
ABE	92	121 (12)	1.68 (0.54)	13.16 (2.41)	1.60 (0.58)	13.63 (2.27)	0.97 (0.10)
ABF	97	122 (8)	1.60 (0.50)	9.09 (1.78)	1.45 (0.54)	9.83 (1.52)	0.75 (0.15)

† True effect: sires A, B, C and E b_1, b_2 = 1.58; sire D b_1, b_2 = 1.1.
‡ True variance: sires A to D σ^2_1, σ^2_2 = 10; sire E σ^2_1, σ^2_2 = 22.5.

Table VI. Computing times depending on number of families for both single- and multiple-trait analyses. Time is elapsed time in hours:minutes:seconds. Twenty-two positions across the chromosome were tested, with 1000 permutations at each position.

Number of families analysed jointly	1	2	3	4	5	6
n_i = 200						
Single-trait	0:0:45	0:7:30	0:13:42	0:23:15	0:29:40	0:34:11
Multi-trait (2 traits)	0:2:21	0:28:21	0:58:13	1:35:53	2:05:53	2:44:23
n_i = 25						
Single-trait	0:0:8	0:0:53	0:1:47	0:2:42	0:3:27	0:3:36
Multi-trait (2 traits)	0:0:26	0:3:17	0:6:42	0:11:10	0:14:02	0:17:34

studies the EM algorithm was used to estimate the effects of a bi-allelic QTL in an across half-sib family analysis. The use of this particular model in actual datasets has demonstrated several practical advantages over simpler methods of analysis. The increase in power, as demonstrated by the simulations, is apparent. Often QTL are detected in multi-family analyses, but not in single-family analyses. Conversely, there are significant single-family results that are

not significant in multi-family analyses. There is a real chance that the significant single-family results may be false positives due to multiple-testing. The multi-family analysis is one way to partly reduce the number of tests performed on the data. The goal of these validation studies is to demonstrate a consistent QTL effect across a wide cross-section of the population. It is also important to ascertain the effects of the QTL on as many traits as possible. If the model fitted a separate QTL effect for each sire, it becomes increasingly difficult to interprete and summarise the results. The simulations have shown that the bi-allelic QTL model will average out the effects of the different alleles in cases where multiple alleles are segregating.

There have been numerous publications that address the EM algorithm in the context of QTL mapping [1,7,10]. While MLEs have been derived for a variety of genetic models and for a variety of population structures, no publication has yet dealt with the type of finite mixture problem discussed in this study. Other publications have used similar likelihood functions, but used quasi-Newton methods to derive MLEs [3, 11]. We have compared the EM algorithm to the quasi-Newton routine (E04JAF) from the NAG library (Numerical Algorithms Group 1990) and found the NAG routine to have less accuracy and power. Warnings of local maxima found were often given and the routine was especially unstable when used in a multi-family, multi-trait analysis. The advantage of the EM algorithm, apart from its numerical stability and greater accuracy, is the facility to provide the posterior probabilities that the *i*th sire is heterozygous, and either phase 1 or phase 2, which are the $\tau_{11,i}$ and $\tau_{12,i}$ described earlier. These indicators will aid any subsequent marker assisted selection of progeny.

In conclusion the EM algorithm described in this study provides the experimenter with a stable and reliable method for combining information across sires and traits in QTL mapping. The strategy of combining information is more critical when faced with small family sizes.

ACKNOWLEDGEMENTS

The financial support of Australian Pork Limited is gratefully acknowledged. We thank two anonymous referees for comments on the submitted version of the manuscript.

REFERENCES

[1] Carbonell E.A., Gerig T.M., Balansard E., Asins M.J., Interval mapping in the analysis of nonadditive quantitative trait loci, Biometrics 48 (1992) 305–315.

[2] Churchill G.A., Doerge R.W., Empirical threshold values for quantitative trait mapping, Genetics 138 (1994) 963–971.
[3] Farnir F., Grisart B., Coppieters W., Riquet J., Berzi P., Cambisano N., Karim L., Mni M., Moisio S., Simon P., Wagenaar D., Vilkki J., Georges M., Simultaneous mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees: revisiting the location of a quantitative trait locus with major effect on milk production on bovine chromosome 14, Genetics 161 (2002) 275–287.
[4] Gilbert H., Le Roy P., Comparison of three multitrait methods for QTL detection, Genet. Sel. Evol. 35 (2003) 281–304.
[5] Goffinet B., Roy P.L., Boichard D., Elsen J., Mangin B., Alternative models for QTL detection in livestock. III. Heteroskedastic model and models corresponding to several distributions of the QTL effect, Genet. Sel. Evol. 31 (1999) 341–350.
[6] Hetzel D.J.S., Davis G.P., Corbet N.J., Shorthose W.R., Stark J., Kuypers R., Scacheri S., Mayne C., Stevenson R., Moore S.S., Byrne K., Detection of gene markers linked to carcass and meat quality traits in a tropical beef herd, in: Proceedings of the Twelth Conference of the Association for the Advancement of Animal Breeding and Genetics, Dubbo, 1997, pp. 442–446.
[7] Jansen R., Maximum likelihood in a generalised linear finite mixture model by using the EM algorithm, Biometrics 49 (1993) 227–231.
[8] Jansen R.C., Johnson D.L., Arendonk J.A.M.V., A mixture model approach to the mapping of quantitative trait loci in complex populations with an application to multiple cattle families, Genetics 148 (1998) 391–399.
[9] Jiang C., Zeng Z.B., Multiple trait analysis of genetic mapping for quantitative trait loci, Genetics 140 (1995) 1111–1127.
[10] Kao C.H., Zeng Z.B., General formulas for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm, Biometrics 53 (1997) 653–665.
[11] Knott S., Haley C., Multitrait least squares for quantitative trait loci detection, Genetics 156 (2000) 899–911.
[12] Knott S.A., Elsen J.M., Haley C.S., Methods for multiple-marker mapping of quantitative trait loci in half-sib populations, Theor. Appl. Genet. 93 (1996) 71–80.
[13] Lander E.S., Green P., Construction of multilocus genetic linkage maps in humans, Proc. Natl. Acad. Sci. USA 84 (1987) 2363–2367.
[14] McLachlan G.J., Krishnan T., The EM algorithm and Extensions, Wiley, New York, 1st edn., 2000.
[15] McLachlan G.J., Peel D., Finite Mixture Models, Wiley, New York, 1st edn., 2000.
[16] Zeng Z.B., Precision mapping of quantitative trait loci, Genetics 136 (1994) 1457–1468.