QTL-MLE: A MAXIMUM LIKELIHOOD QTL MAPPING PROGRAM FOR FLEXIBLE MODELLING USING THE R COMPUTING ENVIRONMENT

P.C. Thomson[1,2], S.C. Brown[3] and H.W. Raadsma[1,2]

[1] Centre for Advanced Technologies in Animal Genetics and Reproduction (ReproGen) and
[2] CRC for Innovative Dairy Products, University of Sydney, NSW 2570
[3] Department of Statistics, Macquarie University, NSW 2141

SUMMARY

In this paper, a maximum likelihood-based procedure is described for mapping QTL in non-inbred line crosses and specifically half-sib families such as commonly found in sheep. The algorithms developed here address limitations of currently available procedures, to ensure that the most reliable information is extracted. A key feature is the use of the E-M algorithm for likelihood maximisation, as this allows standard algorithms (e.g. linear and generalised linear models) to be used, in an iterative process. A suite of programs written in R (QTL-MLE) has been developed, making use of the flexible and integrated modelling environment of that package.

INTRODUCTION

A large resource flock has been developed at The University of Sydney for mapping of QTL in sheep, derived between two breeds, Merino and Awassi. This is an extreme cross for many traits, including milk production, wool characteristics, and parasite resistance (Raadsma *et al*. 2007). Initial QTL mapping efforts were hampered by several limitations, including: (1) inability to make use of the known family structure – and resultant linkage phase – which can be correctly inferred from the backcross nature of the experiment; (2) inability to make effective use of "semi-informative" heterozygous markers of backcross progeny; (3) inability to build flexible modelling strategies for analysing the QTL effects e.g. cannot fit separate QTL effects for males and females; and (4) incorrect modelling of error variance structure when regression-based approaches are used. No analytical software was ready available to address all these limitations. To ensure that the best quality information is extracted from this resource flock, a novel QTL mapping method was derived to overcome these and other limitations. In the following, some details of the algorithm are derived and some implementations issues in the computing environment R are discussed. The method is illustrated using body weight QTL, and also a (derived) binary trait analysis.

METHODS

Backcross design and genotyping. The program has been designed for a backcross design between Awassi × Merino F_1 sires whereby the F_1 sire is backcrossed to Merino ewes ((Aw × Me) × Me). For the current description, the offspring from a single sire will be considered. Informative markers have been screened for the F_1 sire and given a genotype label of '12' with allele '1' being of Awassi origin and allele '2' being of Merino origin. Backcross progeny were subsequently genotyped as either '1', '2', or '12', with '1' corresponding to '11' or '1*x*', '2' corresponding to '22' or '2*x*', with *x* indicating an unscored allele, but in both these cases the allele transmitted from the F_1 sire is determined with certainty. However, a backcross genotype of '12' is not fully informative, as there in uncertainty as to whether allele '1' or '2' was transmitted by the F_1 sire.

Phenotype model. For a normally distributed trait, a linear model may be appropriate, of the form $y_i = \boldsymbol{\beta}'\mathbf{x}_i + \gamma q_i + \varepsilon_i$, where y_i = observed trait value of animal i, $i = 1, \ldots n$; $\mathbf{x}_i$ = set of covariates and fixed effects for animal i; $\boldsymbol{\beta}$ = corresponding set of regression parameters; γ = sire family allelic QTL effect (Q relative to q); q_i = unobserved QTL genotype of animal i, = 1 if Q, 0 if q; and ε_i = random error, assumed $N(0,\sigma^2)$. Note the Merino dam effects will be absorbed in to this last term. The genotype of the F_1 sire is assumed to be Qq, with Q originating from the Awassi line and q from the Merino line.

Genotype model. QTL genotypes are not observed, so the phenotype is a mixture of two distributions, so we calculate the QTL transmission probability (π_i), i.e. the probability of the sire transmitting QTL genotype Q is $\pi_i = p(q_i = 1 \mid \mathbf{m}_i)$ while the probability of transmitting the other allele q is $1 - \pi_i = p(q_i = 0 \mid \mathbf{m}_i)$, where $\mathbf{m}_i$ is the "flanking" marker genotype information. The probabilities depend on the distance from the putative QTL to the marker(s) via Haldane's mapping function. For "informative" markers (genotyped as '1' or '2'), the immediate flanking markers provide all the information. However for the "semi-informative" markers ('12'), the minimal set of markers that contains all the information for a QTL at d are those up to the unambiguous genotypes of '1' or '2', as is seen in the following:

QTL

Animal	M1	M2	M3	M4	M5	M6	M7
1	1	**2**	**12**	**12**	**1**	2	12
2	**12**	**12**	**12**	**1**	12	1	1
3	12	**1**	**1**	1	12	1	2
…							

d

To calculate the transmission probabilities, we determine all the possible "pathways" between unambiguous markers (starting at maker $j = 0$ ending at marker $j = k + 1$). (So for example for the sequence '2–12–12–1', consider all the possible F_1 sire and Merino dam gametes that could have produced this.) It can be shown that the resultant transmission probability is

$$\pi = p(q = 1 \mid \mathbf{m}) = \frac{p(\mathbf{m}, q = 1)}{p(\mathbf{m})} = \frac{\sum\limits_{i \in \{i:\, t_{i,k+1} = g_{k+1}\}} \left[\prod\limits_{j=0}^{k} r'^{\,s_{ij}}_{j,j+1} n'^{\,1-s_{ij}}_{j,j+1} \prod\limits_{j=1}^{k} p_{j1}^{1-t_{ij}} p_{j2}^{t_{ij}} \right]}{\sum\limits_{i \in \{i:\, t_{i,k+1} = g_{k+1}\}} \left[\prod\limits_{j=0}^{k} r^{s_{ij}}_{j,j+1} n^{1-s_{ij}}_{j,j+1} \prod\limits_{j=1}^{k} p_{j1}^{1-t_{ij}} p_{j2}^{t_{ij}} \right]}$$

where $r_{j,j+1}$ is the probability of a recombination between markers j and $j + 1$ ($n_{j,j+1} = 1 - r_{j,j+1}$) in the sire gamete; $r'_{j,j+1} = r_{j,j+1}$ if the putative QTL is not flanked by markers j and $j + 1$, otherwise it is the probability of recombination between the markers *and* of transmitting QTL allele Q (i.e. $q = 1$). To cater for the different pathways, the following indicator variables have been introduced,

$$g_j = \begin{cases} 1 & \text{if marker } j \text{ has genotype '1'} \\ 0 & \text{otherwise (genotype '2')} \end{cases} \qquad j = 0, k + 1;$$

$$s_{ij} = \begin{cases} 1 & \text{if recombination between marker } j \text{ and } j + 1 \text{ for pathway } i \\ 0 & \text{otherwise;} \end{cases}$$

and

$$t_{ij}=\begin{cases}1 & \text{if ram transmits allele '1' at locus } j\text{, for pathway } i\\ 0 & \text{otherwise,}\end{cases}$$

which may be calculated recursively as $t_{ij}=t_{i,j-1}(1-s_{i,j-1})+(1-t_{i,j-1})s_{i,j-1}$, setting $t_{i0}=g_0$ as required, depending on the genotype of the first informative marker of the genotype sequence. The terms p_{j1} and p_{j2} are the frequencies of alleles '1' and '2' at locus j in the Merino population, and maximum likelihood estimates for these have been derived.

Interval mapping procedure: At regular distances (typically 1 cM) along the length of the chromosome, the log-likelihood is constructed assuming a QTL at that position (d), i.e.

$$\log_e L(d)=\sum_{i=1}^{n}\log_e\left[\pi_i f(y_i \mid q_i=1)+(1-\pi_i)f(y_i \mid q_i=0)\right]$$

where $f(\cdot)$ is the probability density function (PDF) for a normal distribution (assuming that is the appropriate model for the data type). The log-likelihood is maximised using the E-M algorithm (e.g. Jansen 1992), which allows standard linear model software to be used, in a weighted iterative manner. This requires computation at each iteration of the posterior probabilities (τ_i) that the sire transmits allele Q, conditional on its phenotype,

$$\tau_i=p(q_i=1 \mid y_i,\mathbf{m}_i)=\frac{\pi_i f(y_i \mid q_i=1)}{\pi_i f(y_i \mid q_i=1)+(1-\pi_i)f(y_i \mid q_i=0)}.$$

The τ_i and $1-\tau_i$ values are used as weights for $q_i = 1$ and $q_i = 0$ respectively in the weighted linear model. At the peak log-likelihood position (i.e. estimated QTL location), these τ_i values can be used to classify backcross animals with high probability of having received the Q (or q) allele.

Model flexibility: the R computing environment: This QTL mapping procedure (termed QTL-MLE) has been developed using R (http://www.r-project.org/). This has the advantage that the QTL mapping procedure can be extended within other modelling and graphical capabilities of this package. For normally distributed traits, the linear model function `lm()` is used, and this easily allows model extension to include interactions between the QTL and other fixed effects, such as sex-specific QTL effects: most other QTL analysis programs do not allow such extensions. Another advantage of the R system is the relative ease to model traits of different types. This is achieved by only a few lines of code change, primarily (1) replace the `lm()` call by another function call, and (2) replace the normal PDF in the τ_i calculation (`dnorm()`) by the appropriate PDF (or discrete probability function) for the required distribution. Thus QTL mapping for binary (binomial distribution), count (Poisson distribution), ordinal (multinomial distribution), and survival time (Weibull distribution) data can be readily accomplished.

RESULTS AND DISCUSSION

Application: As an illustration, bodyweight data from the backcross progeny were mapped for QTL on OAR16 using QTL-MLE. There were two age cohorts in the data set. Three different models were fitted, (1) BW ~ QTL, (2) BW ~ QTL + Cohort, and (3) BW ~ QTL * Cohort, and the resultant interval maps are shown below (Figure 1). There is a significant QTL×Cohort interaction, and this is supported by a likelihood ratio test (QTL * Cohort vs QTL + Cohort: $\chi^2 = 4.32$, df = 1, $P = 0.038$), evaluated at the peak QTL position (29 cM). But more importantly, the QTL would not been detected without the interaction, assuming a LOD 2 threshold was adopted. (As in other QTL mapping, permutation thresholds can also be evaluated, avoiding the need to rely on theoretical asymptotic results of likelihood ratio tests.)

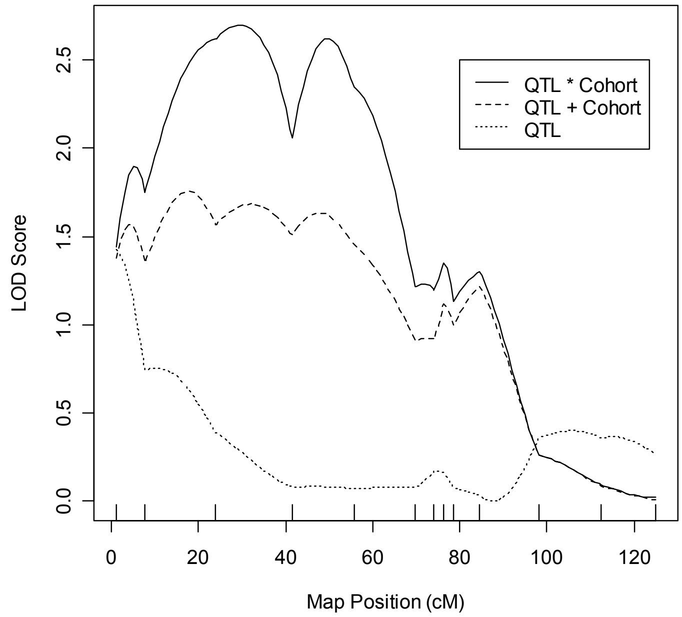

Figure 1. QTL interval maps for bodyweight on OAR16.

CONCLUSION

The QTL-MLE suite of programs has been developed to provide a means of QTL mapping in half-sib designs from non-inbred lines in such a way as to include all available information. Because of the flexible modelling environment of R, it is easily extendable for the analysis of different types of traits and other model types including QTL* fixed effect interactions and different underlying distributions for a complex suite of traits. However, additional developments are being planned, including mufti-sire analysis and multi-trait analysis. The R code can be obtained from the first author (PeterT@camden.usyd.edu.au).

ACKNOWLEDGEMENTS

The research is partly funded by the CRC for Innovative Dairy Products.

REFERENCES

Jansen, R. C. (1992). A general mixture model for mapping quantitative trait loci by using molecular markers. *Theoretical and Applied Genetics* **85**, 252.

Raadsma, H. W., *et al.* (2007). Building an integrated QTL map for all major production traits in sheep from two QTL mapping experiments. *Proc. Assoc. Advmt. Anim. Breed. Genet.* **17**: 127.