A MIXTURE MODEL APPROACH FOR QTL MAPPING IN PRODUCTION TRAITS OF DAIRY CATTLE

Y. Liu[1], G.B. Jansen[1] and C.Y. Lin[1,2]

[1] CGIL, Dept. of Animal & Poultry Science, University of Guelph, Ontario, Canada N1G 2W1
[2] Dairy and Swine Research and Development Centre, Agriculture and Agri-Food Canada

INTRODUCTION

The QTL mapping studies for production traits in dairy cattle have often used the regression method of interval mapping (Knott *et al*., 1996) or single marker analysis. Although the mixture model maximum likelihood (Lander and Botstein, 1989) has been widely used in plants or laboratory animals, it has seldom been used for QTL mapping in dairy cattle. The mixture model maximum likelihood and the regression method (Haley and Knott, 1992) have been compared for QTL mapping through simulation of line crossing designs. The former was found to have a smaller residual variance (Xu, 1995) and a larger likelihood ratio statistic (Kao, 2000) than the latter. The mixture model method uses not only marker information for inferring the probability of QTL genotypes but also phenotypic observations, resulting in a more accurate estimation of the probability of QTL genotypes than the regression method. Kao (2000) reported that the mixture model maximum likelihood method was more effective in detecting closely linked QTL than the regression method. There have been a few studies on QTL mapping by mixture model maximum likelihood in livestock (George *et al*., 1995; Knott *et al*., 1996, Song and Weller 1998). However, the method still needs to be refined for practical application to livestock population because these methods either analyzed sire families separately or were restricted to bi-allelic QTL. The purpose of this study is to develop a mixture model maximum likelihood method for half-sib designs and to compare it with regression method on the basis of both actual and simulated data in dairy cattle.

MATERIALS AND METHODS

Mixture model method. To detect QTL, phenotypic observations need to be described in terms of QTL alleles inherited from parents. In a half-sib design, a phenotypic observation, y_{ij} ($i = 1,2,...k$ and $j = 1,2,...n_i$) of offspring j of sire i can be expressed as,

$$y_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \xi_{ij}\boldsymbol{z}_{ij}'\boldsymbol{a} + e_{ij} \quad (1)$$

where $\boldsymbol{\beta}$ is a vector of fixed effects and may include an overall mean μ, half-sib family means, *etc;* $\boldsymbol{x}_{ij}'$ is the row of design matrix $\boldsymbol{X}$ corresponding to y_{ij}; $\boldsymbol{a} = (a_1, a_2, ..., a_k)'$ is a vector of substitution effects; $\boldsymbol{z}_{ij}'$ is the row of design matrix $\boldsymbol{Z}$ corresponding to y_{ij}; ξ_{ij} is an indicator variable and equal to 1 if son ij inherits Q_{i1} from sire i and equal to 0 if son ij inherits Q_{i2} from sire i; and e_{ij} is residual error with $e_{ij} \sim N(0, \sigma_{ij}^2)$ or $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{R}\sigma^2)$. For dairy cattle, y_{ij} can be estimated breeding values (EBV) with reliability r_{ij}. The variance of residual

error e_{ij} is taken as σ^2 / r_{ij} and $\boldsymbol{R} = diag\{1/r_{ij}\}_{N\,xN}$ where $N = \sum n_i$. Given phenotypic observations and marker data, the likelihood of the parameters ($\boldsymbol{\beta}, \boldsymbol{a}$ and σ_{ij}^2) is,

$$f(\boldsymbol{\beta},\boldsymbol{a},\sigma^2 \mid \boldsymbol{y},M) = \prod_{i=1}^{k}\prod_{j=1}^{n_i}(2\pi\sigma_{ij}^2)^{-1/2}\{p_{ij}\exp[-\frac{1}{2\sigma_{ij}^2}(y_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\beta} - a_i)^2] + (1-p_{ij})\exp[-\frac{1}{2\sigma_{ij}^2}(y_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\beta})^2]\} \quad (2)$$

where p_{ij} is the conditional probability that offspring ij inherits the first QTL allele of sire i , inferred from marker information. The EM algorithm was used to maximize the likelihood by taking the inherited QTL allele as missing data. This leads to the following equations:

$$\begin{pmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{U} \\ \boldsymbol{U}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{U} \end{pmatrix}\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{a}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{U}'\boldsymbol{R}^{-1}\boldsymbol{y} \end{pmatrix} \quad (3)$$

$$\hat{\sigma}^2 = \frac{1}{N}[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{a}'\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{U}\,\boldsymbol{a} - 2(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{R}^{-1}\boldsymbol{U}\,\boldsymbol{a} \quad (4)$$

$\boldsymbol{U} = \{\pi_{ij}\boldsymbol{z}_{ij}'\}_{N\,xk}$ where $\pi_{ij} = \dfrac{p_{ij}\phi(y_{ij} \mid \boldsymbol{\beta},\boldsymbol{a},\sigma_{ij}^2)}{p_{ij}\phi(y_{ij} \mid \boldsymbol{\beta},\boldsymbol{a},\sigma_{ij}^2) + (1-p_{ij})\phi(y_{ij} \mid \boldsymbol{\beta},\boldsymbol{a},\sigma_{ij}^2)}$

The posterior probability π_{ij} is calculated in the E-step while the parameters are estimated in the M-step based on equations (3) and (4). The likelihood ratio statistic is used to test QTL effect under the null hypothesis that $a_i = 0$ for $i = 1,2,...k$. The alternative hypothesis is that at least one of the sires has significant QTL effect, suggesting that the test position may contain a QTL. Chromosomewise, single trait critical values were determined by permutation test.

Mapping QTL of production traits. Daughter yield deviations (DYD) for milk, fat and protein yields and fat and protein percentage were analyzed with interval mapping by both mixture model and regression methods. The data (Nadesalingam *et al*., 2001; Plante *et al*., 2001) included 6 Holstein grandsires with 71 to 74 sons each (N=433). Grandsires and sons were genotyped for 64 microsatellite markers distributed over chromosomes 1, 3, 5, 6, 9, 10, 13, 15, 17, 20, 23 and 26. The markers covered 867.4 cM of the genome with two to nine markers per chromosome. Sons' dams were not genotyped. The MARC linkage map was applied. The genome was scanned by 1 cM steps. The test statistic used was the likelihood ratio for the mixture model and the F-value for the regression method. The empirical distribution of the test statistics were determined from 1000 permutation replicates.

Data simulation. Nine cases were simulated with different number of grandsires and sons per grandsire family, variance proportion explained by the simulated QTL (the ratio of QTL variance to phenotypic variance), and interval size in cM (Table 1). A chromosome of 100 cM with evenly spaced markers, ten alleles each, was simulated. Sires were heterozygous at all marker loci. A single QTL was set at 34 cM for all cases. Two different QTL alleles were simulated for each sire and dam. Each case consists of 100 replicates. The critical values

($\alpha = 0.05$ or 0.01) were determined by ranking the test statistic of 1000 simulated replicates without QTL effects included in the phenotypes.

Table 1. Parameter combinations of the nine simulated cases

Simulated cases	1	2	3	4	5	6	7	8	9
No. of sires	5	10	15	10	10	10	10	10	10
No. of sons per sire	150	75	50	50	150	75	75	75	75
QTL variance ratio	0.1	0.1	0.1	0.1	0.1	0.1	0.05	0.15	0.1
Interval size (cM)	10	10	10	10	10	10	10	5	20

RESULTS AND DISCUSSION

Analysis of actual data. The detected QTL positions and p values are listed in Table 2 for *p<0.05* by either method. Based on the critical values of $\alpha = 0.05$ from permutation tests, five QTL were detected by the regression method as compared to nine by the mixture model method. This result indicates that the mixture model method offers higher power for QTL detection than does the regression method. The mixture model analysis also showed that the QTL for protein yield on chromosome 20 has correlated effects on milk yield.

Table 2. QTL positions (cM) detected by mixture model (Mix) and regression model (Reg) and their p values (in parenthesis)

Chr.	Method	Milk	Fat	Protein	Fat%	Protein%
1	Mix		28 (0.005)	47 (0.004)	16 (0.009)	
	Reg		27 (0.056)	46 (0.042)	17 (0.054)	
3	Mix	40 (0.010)			35 (0.032)	20 (0.001)
	Reg	39 (0.043)			34 (0.099)	27 (0.041)
6	Mix	42 (0.010)				15 (0.024)
	Reg	40 (0.035)				14 (0.067)
20	Mix	21 (0.023)		20 (0.006)		
	Reg	21 (0.109)		19 (0.034)		

Simulation results. The number of significant tests among 100 simulated replicates of each case (Table 3) shows that the statistical power of the two methods does not differ in cases 1, 2, 5, 7 and 8. Notably, these cases have favourable combinations of parameters for QTL mapping because they have a greater number of sons per sire and/or a smaller interval size. In the remaining cases, less favourable for QTL mapping, the mixture model shows higher power than the regression method. The dispersion of estimated QTL positions (Table 4) was slightly smaller in the mixture model than in the regression method in 6 of 9 cases simulated. The parameters used in the simulation have clear effects on the results with expected patterns in all cases. Interestingly, given a fixed sample size, the family size (number of sons per sire) has a very large influence on QTL mapping. The estimated dispersion of QTL positions was much smaller with 150 sons per sire than with 75 or 50 sons per sire.

Table 3. Number of significant tests among 100 simulated replicates

	Simulated cases								
Method	1	2	3	4	5	6	7	8	9
Mix α =0.05	93	83	76	64	95	43	95	88	74
Mix α =0.01	91	67	39	38	94	20	90	76	61
Reg α =0.05	93	83	74	62	95	41	95	88	74
Reg α =0.01	91	67	36	39	94	21	90	76	59

Table 4. Means and S. D. (in parenthesis) of estimated QTL positions (cM) over 100 replicates

Meth.	Simulated cases								
	1	2	3	4	5	6	7	8	9
Mix	33.70 (6.22)	34.49 (11.13)	34.94 (14.56)	39.19 (20.69)	34.35 (3.88)	38.04 (21.89)	34.27 (6.21)	34.43 (11.51)	35.27 (11.42)
Reg	33.21 (6.72)	33.03 (10.99)	35.19 (15.21)	39.39 (20.92)	33.48 (3.85)	38.05 (22.01)	33.36 (6.56)	33.36 (11.16)	35.37 (12.78)

CONCLUSION

A mixture model maximum likelihood method based on the EM algorithm was developed and compared to the regression method for analysis of both actual and simulated data sets. The five QTLs detected by the regression method are quite compatible with previous analyses of the data (Nadesalingam *et al*. 2001, Plante *et al*. 2001). Additional QTLs detected by the mixture model analysis are: QTLs for fat percentage between TGLA49 and RM095 and for fat yield between RM095 and ILSTS004 on chromosome 1, a QTL around INRA023 for fat percentage on chromosome 3, and a QTL for protein percentage near BM143 on chromosome 6. The mixture model analysis also detected that the QTL for protein yield on chromosome 20 also has a correlated effects on milk yield. Results from actual data showed that the mixture model method has higher statistical power than the regression method. The simulation study supports the advantage of the mixture model, however, the evidence is not as strong as that based on actual data analysis. The superiority of the mixture model method was greatest in scenarios with limited information for QTL detection.

REFERENCES

George, M., Nelson, D., Mackinnon M., *et al.* (1995) *Genetics* **139**: 907-20.
Kao, C. H. (2001) *Genetics* **156**: 855-865.
Haley, C. S. and Knott, S. A. (1992) *Genetics* **69**: 315-324.
Knott, S. A., Elsen, J. M. and Haley, C. S. (1996) *Theor. Appl. Genet*. **93**: 71-80.
Lander, E. S. and Botstein, D. (1989) *Genetics* **121**: 185-199.
Nadesalingam, J., Plante, Y. and Gibson, J. P. (2001) *Mammalian Genome* **12**: 27-31.
Plante, Y., Gibson, J. P., Nadesalingam, J., *et al*. (2001) *J. Dairy Sci*. **84**: 1516-1524.
Song, J. Z. and Weller, J. I. (1998) *Proc. 6th WCGALP* **26**: 341-344.
Xu, S. C. (1995) *Genetics* **141:** 1657-1659.