# QTL MAPPING IN MULTIPLE FAMILIES USING LOGISTIC REGRESSION

**Y. D. Zhang and B. Tier**

Animal Genetics and Breeding Unit*, University of New England, Armidale NSW 2351

## SUMMARY

This study compares logistic regression (LR) with maximum likelihood (ML) methods for mapping quantitative trait loci (QTL) in multiple half-sib families under selective or full genotyping strategies, with various levels of marker informativeness and marker interval. In ideal conditions involving evenly located polymorphic markers and all individuals genotyped, both LR and ML methods showed a high power of detecting QTL and produced accurate estimates of QTL locations and effects. Under selective genotyping strategy, the power of ML is limited by regions with low information content. The LR method performed better than ML and is a straight-forward and robust method for this case.

## INTRODUCTION

There are many methods available for detecting QTL in half-sib populations when all individuals are genotyped (eg Zeng 1993, 1994; Kao *et al.* 1999; Kerr *et al.* 2005). These methods generally rely on good genetic information, that is, polymorphic markers spaced evenly at close intervals along the chromosome. This ideal case is uncommon in real genome scans, because of a lack of informative markers in some chromosomal regions. Selective genotyping is used to reduce the cost of genotyping and achieve a reasonable power of detecting QTL (Lebowitz *et al.* 1987). Low amounts of genetic information, selective genotyping or a spike in the phenotypic distribution can generate spurious QTL peaks (eg Broman 2003; Feenstra and Skovgaard 2004). In a previous study, the logistic regression (LR) method effectively detected QTL in single half-sib families and avoided the spurious QTL peaks (Zhang and Tier 2005). Using simulated data, this study compares LR with the maximum likelihood (ML) methods for detecting QTL in multiple families with various levels of marker informativeness, marker interval and under two types of genotyping schemes (full or selective).

## MATERIALS AND METHODS

**Data.** Ten datasets were simulated and analysed to examine the performance of LR and ML methods (five for each genotyping scheme). For each dataset one chromosome and 4 half-sib families (two families without QTL, the other two with QTL segregating but their effects in opposite directions) were simulated. Each progeny had a different dam. Genetic information for full genotyping datasets (1 to 5) and selective genotyping datasets (S1 to S5) is shown in Table 1. Residuals were drawn from a normal distribution with a mean of zero and variance of 10. The QTL effect (difference between two homozygous genotypes) was assumed to be half of the residual standard deviation (1.6). For selective genotyping, 100 individuals in each single family were simulated and the 25 individuals with highest and 25 with lowest phenotypic values were genotyped. Fifty individuals per family were simulated and genotyped in the full genotyping cases. As a result, 200 genotyped individuals in each dataset were analyzed. For each case, 100 replicates were generated and analyzed using both LR and ML methods. The chromosome was searched in steps of 5 cM. A chromosome wide significant threshold value was obtained using 300 or 500 permutations across each test point using LR or ML, respectively. Genotypes in each family were shuffled. The power was defined as a percentage of replicates in which a chromosome wide

* AGBU is a joint venture of NSW Department of Primary Industries and University of New England.

significance level of 5% was achieved. The QTL position was identified by the test location with the highest test statistic. The average QTL position and effect were calculated using all replicates where a 5% chromosome wide significance was achieved.

**Table 1. Genetic information used in each multiple family dataset, number (N), position (cM) and information content range (MIC) of markers, minimum MIC with position in parenthesis. Simulated QTL position is also shown for full and selective genotyping cases**

| Dataset | N | Marker position (cM) | QTL position | MIC range | Min MIC (cM) |
|---|---|---|---|---|---|
| 1, S1 | 6 | 0, 26, 51, 77, 102, 128 | 64 | 0.90-1.0 | 0.75 |
| 2, S2 | 6 | 0, 18, 30, 76, 102, 127 | 64 | 0.90-1.0 | 0.65(54) |
| 3, S3 | 6 | 0, 26, 51, 97, 109, 127 | 64 | 0.90-1.0 | 0.65(74) |
| 4, S4 | 5 | 0, 81, 106, 132, 157 | 93 | 0.90-1.0 | 0.33(40) |
| 5, S5 | 5 | 0, 81, 106, 132, 157 | 93 | 0.75-0.85 | 0.29(38) |

**Marker information content.** Paternal haplotypes were estimated using the Lander-Green algorithm (Lander and Green 1987). Prior QTL transmission probabilities (TP) for each individual at test points were estimated using genotypes and the recombinant fragment of flanking markers, as described by Kerr *et al*. (2005). The variation of TP is used as an indicator of the marker information content (MIC), which is 4 times the TP variance, ranging from 0.0 to 1.0.

**Statistical model.** The multiple-family approach integrates the single half-sib family analyses. The model for the single family analysis is the same as described in the previous study (Zhang and Tier 2005). Assuming that a bi-allelic QTL with alleles Q and q contributing to the variance of a quantitative trait. The quantitative trait value $y_i$ for individual *i*, being pre-adjusted for fixed effects and polygenic effects, can be related to the QTL by the model $y_i = \mu + \alpha x_i + e$, where $\alpha$ is the effect of the putative QTL, $x_i$ is a probability that a progeny inherits allele Q from the sire, and e is the residual. The logistic regression model is fitted and implemented as described by Dobson (2002). For the multiple-family analysis, different QTL phases were considered for different families. For a family the phase was determined by the largest likelihood. The log likelihood of the multiple-family estimation was the sum of log likelihoods from each family with respect to its most likely phase. The estimated LR coefficient and the total variance were used to calculate the QTL effect $\alpha$, as described by Henshall and Goddard (1999). The ML method was implemented as described by Kerr *et al.* (2005), the $\alpha$ value is estimated as described by Zeng (1994). In the case of selective genotyping, the $\alpha$ values from ML were adjusted for the consequences of selective genotyping, using the method of Darvasi and Soller (1992). Empirical threshold values were determined using permutation tests (Churchill and Doerge 1994).

## RESULTS AND DISCUSSION

Results are presented for the two genotyping strategies separately to compare performance of LR and ML methods. Differences between the two strategies are not compared here, because the underlying simulated datasets, for example, 1 and S1, are not the same.

**Full Genotyping.** As shown in Table 2, under the ideal conditions involving evenly located polymorphic markers and all individuals genotyped (dataset 1), both LR and ML showed a high power of detecting QTL and produced accurate estimates of QTL locations and effects, although QTL effects estimated using LR were slightly lower than the simulated value. The results for ML

methods are consistent with findings by Kerr *et al.* (2005), who also reported that a 100% detection power was achieved when sib-family size was 100. Although all animals were genotyped, when the marker genetic information was low (at various levels in datasets 2 to 5), power decreased dramatically, particularly for ML. Though the estimates of QTL effects were very conservative for both LR and ML methods the estimated QTL positions were close to that simulated.

**Table 2. Results of QTL analyses using Logistic Regression and Maximum Likelihood methods in the full and selective genotyping schemes. The power of each analysis, QTL position (cM) and effect (simulated at 1.6) are presented (standard error in parenthesis)**

| Dataset | Logistic regression | | | Maximum likelihood | | |
|---|---|---|---|---|---|---|
| | Power[1] | Position | Effect | Power | Position | Effect |
| | | | Full genotyping | | | |
| 1 | 84 | 62 | 1.39(0.04) | 87 | 63 | 1.64(0.06) |
| 2 | 31 | 71 | 1.21(0.10) | 27 | 70 | 1.36(0.11) |
| 3 | 27 | 57 | 1.17(0.10) | 18 | 56 | 1.21(0.13) |
| 4 | 28 | 106 | 1.10(0.08) | 24 | 95 | 1.47(0.13) |
| 5 | 31 | 101 | 1.17(0.09) | 21 | 94 | 1.42(0.13) |
| | | | Selective genotyping | | | |
| S1 | 44 | 59 | 1.64(0.08) | 42 | 59 | 1.49(0.11) |
| S2 | 36 | 68 | 1.60(0.10) | 39 | 65 | 1.63(0.10) |
| S3 | 38 | 55 | 1.54(0.09) | 37 | 63 | 1.68(0.12) |
| S4 | 50 | 96 | 1.66(0.08) | 37 | 88 | 1.85(0.14) |
| S5 | 51 | 97 | 1.48(0.08) | 34 | 53 | 2.73(0.14) |

[1] percentage of 100 replicates with a 5% chromosome wide significance level.

**Selective Genotyping.** Both LR and ML methods showed a similar power, QTL position and effects when markers were highly informative and evenly spaced (S1). When an interval containing the QTL was extended to 57 cM, QTL positions assessed by both methods also did not deviate very much from the true position when marker genetic information was not very low (for example, minimum MIC was 0.65 in datasets S2, S3). When a large marker interval (up to 80 cM) occurred next to the interval containing the QTL, LR performed better than ML in the positioning QTL and the power to detect QTL (S4). When marker information content was ever lower (0.6 – 0.8, 0.29 at the lowest point) and with a large marker interval neighboring the QTL region, LR also produced better estimates of QTL location and effect than those from ML (S5) and displayed high power. In this adverse case, ML is likely to produce spurious QTL peaks, far away from the true position and towards the region with low genetic information, and with an inflated estimate of the QTL effect (2.73 for dataset S5).

Chromosomal regions of low MIC frequently occur in QTL projects. Those regions are attributable to a lack of informative markers. ML can produce spurious peaks of test statistic in low information regions under selective genotyping. Feenstra and Skovgaard (2004) developed a

2-component mixture model to avoid such spurious peaks and applied to a back-cross design. However, to apply this method to F2 or half-sib design, a 3-component model is required and is yet to be developed. The LR method can overcome such problems and facilitate the efficient application of selective genotyping. In the previous study on a single family (Zhang and Tier 2005), both LR and ML methods displayed higher power than in this multiple family approach, suggesting this approach could effectively reduce the positive rate.

## CONCLUSION

In the ideal conditions involving evenly located polymorphic markers and all individuals genotyped, both LR and ML methods showed a high power of detecting QTL and produced accurate estimates of QTL locations and effects. Under selective genotyping, the power of the ML method is limited in low information regions, being very likely to produce spurious QTL peaks. In this case, LR performed better than ML in detecting QTL. The LR method provides a straightforward and robust solution for this situation.

## REFERENCES

Broman, K. W. (2003) *Genetics* **163**:1169
Churchill, G. A. and Doerge, R. W. (1994) *Genetics* **138**:963.
Darvasi, A. and Soller, M. (1992) *Theor. Appl. Genet.* **85**:353.
Dobson, A. J. (2002) "An Introduction to Generalized Linear Models" 2nd ed. Chapman and Hall, London.
Feenstra, B. and Skovgaard, I. M. (2004) *Genetics* **167**:959.
Henshall, J. M. and Goddard, M. E. (1999) *Genetics* **151**:885.
Kao, C. H., Zeng, Z. B. and Teasdale, R. D. (1999) *Genetics* **152**:1203.
Kerr, R. J., McLachlan, G.M. and Henshall J.M. (2005) *Genet. Sel. Evol*. **37:**83.
Lander, E. S. and Green, P. (1987) *Proc. Natl. Acad. Sci. USA* **84**:2363.
Lebowitz, R. J., Soller, M. and Beckmann, J. S. (1987) *Theor. Appl. Genet.* **73**:556.
Zeng, Z. B. (1993) *Proc. Natl. Acad. Sci. USA* **90**:10972.
Zeng, Z. B. (1994) *Genetics* **136**:1457.
Zhang, Y. D. and Tier, B. (2005) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **16**:354.