A POISSON GLM APPLICATION TO QTL ANALYSIS OF DISCRETE F_2 GENERATION DATA USING THE E-M ALGORITHM

S. A. Kayis[1], P. C. Thomson[1] and F. W. Nicholas[2]

[1]Biometry Unit, Department of Crop Sciences, University of Sydney, NSW 2006
[2]Department of Animal Science, University of Sydney, NSW 2006

SUMMARY

In this study a statistical model for quantitative trait loci (QTL) analysis of count data, such as litter size, for an F_2 design is proposed. A Poisson model is assumed for data and parameters are estimated in the framework of Generalised Linear Models (GLM) using the E-M algorithm. The model allows interval mapping as well. Statistical properties of the model are determined by a Monte Carlo (MC) simulation study.

Keywords: QTL, GLM, E-M algorithm, F_2 generation, discrete character

INTRODUCTION

Many methods have been developed to detect QTL. However, they are almost always based on a normal, continuos, distribution of the trait, and therefore may not be suitable for non-normal data such as litter size, calving ease etc.

Kayis *et al* (1998a) proposed a single-marker-single-QTL method, for QTL analyses of backcross mice litter size data - a discrete variable. Litter size was assumed to have a Poisson distribution, and parameters were estimated using the E-M algorithm. Kayis *et al* (1998b) extended this method in the framework of GLM that allows interval mapping and incorporates fixed effects easily.

In this study a method is presented for analysing QTL data in an F_2 design for discrete data. It is analogous to the method presented by Kayis *et al* (1998b).

METHODS

Genetic design. Two highly inbred lines (P_1 and P_2) are considered here. It is assumed that P_1 is homozygous for the alleles Q and M_k for the QTL and genetic markers, respectively, while P_2 is homozygous for the alleles q and m_k (ie single QTL flanking markers model with k genetic markers). The F_1, obtained from mating of P_1 and P_2, is heterozygous for QTL and marker genotypes. F_2 offspring are obtained from $F_1 \times F_1$ matings.

A Poisson GLM Model. Countable phenotypes (y), such as litter size, are assumed to have a Poisson distribution with parameter $\lambda = \exp(\mu + \beta_i + \mathbf{q}'_j\gamma)$, where β = fixed effect, such as parity; γ = QTL effects; $\mathbf{q}_j$ is the indicator random vector indexing the QTL genotypes (unobservable) QQ, Qq, qq as in McLachlan and Basford (1988).

The complete log likelihood is $l(\theta;\mathbf{y}_j,\mathbf{Q}) = \sum_j \log P(\mathbf{y}_j,\mathbf{q}_j;\theta)$ where $\mathbf{y}_j$ = vector of litter size for animal j; $\mathbf{Q}$ is a matrix of $\mathbf{q}_j$ and $\theta = (\mu, \beta_1, \beta_2, \beta_3, \beta_4, \gamma_1, \gamma_2, \gamma_3)$ is the set of parameters to be estimated.

Since the indicator variables are not observed, they can be treated as missing data and the E-M algorithm (Dempster *et al.* 1977) can be applied to estimate parameters. In the E step (expectation), the expected log likelihood is calculated by replacing $\mathbf{q}_j$ with their conditional expectations (posterior probabilities of QTL genotype group membership, given the phenotype data). Then in the M step (maximisation), the expected log likelihood is maximised over θ. The E-M sequence is continued until $\left\|\hat{\theta}_{k+1} - \hat{\theta}_k\right\|$ is sufficiently small (10^{-6}).

Monte Carlo simulation. To asses the performance of the model, three series of 1200 MC simulations were performed, in accordance with the proposed model, with n_{F2}=100 per simulation. Each series is made up of a combination of two different QTL locations; see Figure 1. Genetic markers are equally spaced. Haldane's mapping function is used as a distance measurement. The parameters assigned to the simulation were μ=1, β= (0.0, 0.25, 0.50, 0.75), γ=(0.0, 0.40, 0.80).

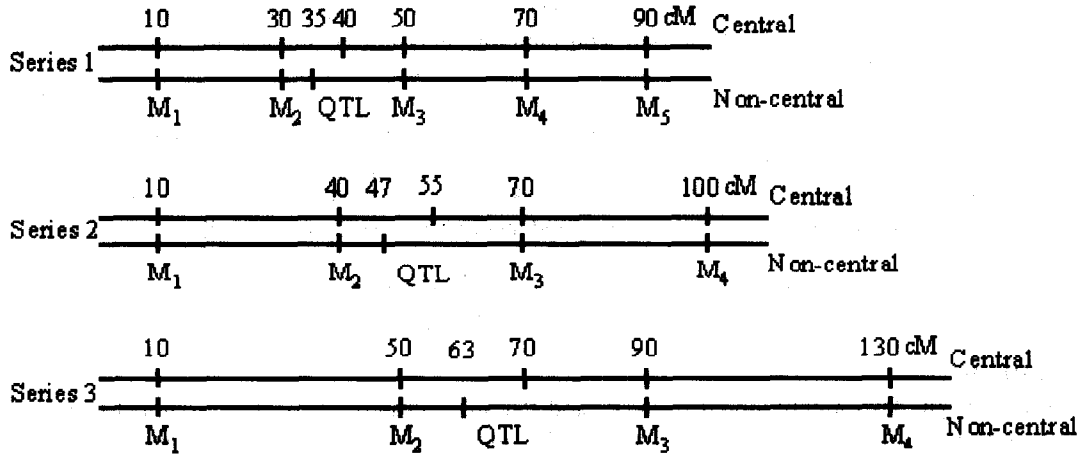

Figure 1. Marker and QTL locations for three series (3 different marker spacings) of MC simulations. In each series, a QTL is located centrally or non-centrally between the markers.

RESULTS

Results of MC study is summarised in Table 1. Another MC simulation has been performed for the series 1, adding random animal effects, u, (ie $\lambda = \exp(\mu + \beta_i + u_j + \mathbf{q}'_j\gamma)$) with $u_j \sim N(0, 0.1)$. Parameters have been estimated via the same model, thus allowing the robustness of the procedure to an incorrectly specified model to be assessed. Results of second MC study for the series 1 is given in Table 2.

Table 1. Results of MC simulations for different marker space and QTL position combinations

		Central			Non-central		
Series	Parameter	Value	Mean	Std.err*.	Value	Mean	Std.err*.
	QTL location	0.40	0.399	0.0288	0.35	0.345	0.0318
	μ	1.00	0.999	0.0733	1.00	0.997	0.0686
	β_1	(0)			(0)		
	β_2	0.25	0.250	0.0667	0.25	0.253	0.0625
1	β_3	0.50	0.499	0.0617	0.50	0.501	0.0623
	β_4	0.75	0.750	0.0594	0.75	0.751	0.0602
	γ_1	(0)			(0)		
	γ_2	0.40	0.399	0.0680	0.40	0.402	0.0631
	γ_3	0.80	0.799	0.0683	0.80	0.797	0.0657
	QTL location	0.55	0.548	0.0397	0.47	0.463	0.0398
	μ	1.00	1.004	0.0750	1.00	1.003	0.0728
	β_1	(0)			(0)		
	β_2	0.25	0.248	0.0644	0.25	0.254	0.0651
2	β_3	0.50	0.500	0.0601	0.50	0.502	0.0626
	β_4	0.75	0.750	0.0580	0.75	0.752	0.0588
	γ_1	(0)			(0)		
	γ_2	0.40	0.394	0.0709	0.40	0.394	0.0654
	γ_3	0.80	0.792	0.0736	0.80	0.788	0.0699
	QTL location	0.70	0.698	0.0474	0.60	0.583	0.0532
	μ	1.00	1.000	0.0766	1.00	1.015	0.0807
	β_1	(0)			(0)		
	β_2	0.25	0.252	0.0670	0.25	0.244	0.0656
3	β_3	0.50	0.500	0.0632	0.50	0.501	0.0609
	β_4	0.75	0.752	0.0590	0.75	0.751	0.0589
	γ_1	(0)			(0)		
	γ_2	0.40	0.397	0.0755	0.40	0.380	0.0751
	γ_3	0.80	0.795	0.0739	0.80	0.778	0.0792

*Empirical standard deviation of the parameter estimates from the MC simulation.

CONCLUSION AND FUTURE DEVELOPMENT

Correctly specified model. In general, MC simulation results show that parameter estimates are mostly unbiased. When markers are spaced 20 cM and 30 cM aparts, parameter estimates are very close to true values irrespective of central or non-central QTL placement between markers. However, when the markers are located 40 cM apart, the parameter estimates in the central QTL placement is less biased then non-central for detection of QTL location, and QTL effects.

Incorrectly specified model. Parameter estimates for detection of QTL location and parity effects are unbiased. However, the model overestimates the QTL effects (ie. biased) when the model is incorrectly specified while generating data (ie. adding random animal effects, u).

Relative performance of GLM to a model that assumes normal distribution. A comparison between the GLM and a model that assumes normal distribution of phenotype for backcross design is given by Kayis *et. al.* (1999). Phenotype was mice litter size.

Future development. A study on including random animal effects to the model is under investigation. Another consideration is extracting information from the backcross and F_2 generations simultaneously.

Table 2. Results of second MC study for series 1

		Central			Non-central		
Series	Parameter	Value	Mean	Std.err*.	Value	Mean	Std.err*.
	QTL location	0.40	0.408	0.0912	0.35	0.353	0.1107
	μ	1.00	0.935	0.1371	1.00	0.943	0.1262
	β_1	(0)			(0)		
	β_2	0.25	0.254	0.0614	0.25	0.251	0.0642
1	β_3	0.50	0.503	0.0610	0.50	0.496	0.0579
	β_4	0.75	0.755	0.0558	0.75	0.748	0.0568
	γ_1	(0)			(0)		
	γ_2	0.40	0.467	0.1674	0.40	0.471	0.1597
	γ_3	0.80	1.013	0.1827	0.80	1.018	0.1707

*Empirical standard deviation of the parameter estimates from the MC simulation.

REFERENCES

Kayis, S.A., Thomson, P.C., and Nicholas, F.W. (1998a) *Proceedings of the 6th Wold Congress on Genetics Applied to Livestock Production.* **26:** 237

Kayis, S.A., Thomson, P.C., and Nicholas, F.W. (1998b) *45th Annual Conference Genetics Society of Australia*, Sydney, Australia. 67

Kayis, S.A., Maqbool, N.J., Thomson, P.C., Moran, C., and Nicholas, F.W. (1999) *Proc. Aust. Assoc. Advmt. Anim. Breed. And Genet.* **13:**321

Dempster, A. P., Laird, N. M, and Rubin, D. B. (1977) *J. Royal Statist. Soc. B* **39** 1

McLachlan, G. J., and Basford, K. E. (1988) Mixture models: inference and applications to clustering. M. Dekker, New York