Rajshahi University Journal of Science & Engineering **ISSN 2309-0952**
Vol. 44: 95-99, 2016

Regression Based Robust QTL Analysis for F_2 Population

Md. Jahangir Alam[1*], Md. Alamin[2], Most. Humaira Sultana[1], Md. Amanullah[1] and Md. Nurul Haque Mollah[1]

*[1]*Bioinformatics Lab, Department of Statistics, University of Rajshahi, Rajshahi 6205, Bangladesh.*
[2]Department of Agronomy, Zhejiang University, Hangzhou, China.

*Corresponding author: *jahangir_statru63@yahoo.com*

Abstract

This Quantitative trait locus (QTL) analysis is a widely used statistical approach for the detection of important genes in the chromosomes. Maximum likelihood (ML) based interval mapping (IM) is one of the most popular approaches for QTL analysis. However, it is relatively complex and computationally slower than regression based IM. Haley-Knott (HK) and extended Haley-Knott (eHK) regression based IM save computation time and produce similar results as ML-IM. However, these approaches are not robust against phenotypic outliers. In this research, we have developed a robust regression based IM approach by maximizing beta-likelihood function for intercross (F_2) population. The proposed method reduces to the HK-IM method when beta $\rightarrow 0$. The tuning parameter beta controls the performance of the proposed method. The simulation results show that the proposed method improves performance over the existing IM approaches in the case of data contaminations; otherwise, it shows almost the same results as the classical IM approaches.

Keywords—QTL analysis; F_2 population; robust regression; maximum beta-likelihood estimation; beta-LRT criterion; robustness.

INTRODUCTION

The rapid increase in availability of fine-scale genetic markers due to the rapid advancement in molecular biology has led to the intensive use of QTL mapping in the genetic study of quantitative traits in bioinformatics. Reference [1] first proposed the idea of using two markers to bracket a region for testing QTLs. Reference [2] proposed a similar, but much improved, method which is known as interval mapping (IM) approach. This method uses two adjacent markers to test the existence of a QTL within the interval by performing a likelihood ratio test (LRT) at every position in the interval. Maximum likelihood (ML) based IM [2] and regression based IM [3] are two most popular and widely used interval mapping approaches.

In practice, QTL effects are treated as either fixed or random [4]. In fixed effects QTL model, allelic substitution effects are usually estimated and tested, and QTL variance is calculated from estimated allelic effects. In random effects QTL model, the QTL effects and QTL variance are directly estimated and tested. Since the conditional expectations of the QTL genotype given the flanking marker genotype are unknown in MLE based IM model [2], this QTL effect model can be treated as a random effects model (REM). On the other hands, in the HK regression based IM model the conditional expectation of the QTL genotype given the flanking marker genotype is considered as fixed [5] and this model can be treated as a fixed effect model (FEM).

The existing interval mapping based on REM [2] and FEM [3] are two most popular and widely used methods for QTL analysis. But these methods are not robust against phenotypic contaminations. In this work, we propose a robust method with FEM to perform QTL analysis for F_2 population. We also show a simulation study to investigate the performance of the proposed method with the existing random effect QTL model and fixed effect QTL model for F_2 population.

A QTL MAPPING FOR F_2 POPULATION USING REGRESSION APPROACH

Let us consider that there is no epistatis between two QTLs, no interference in crossing over, and there is only one QTL in the testing interval. The fixed effect model for F_2 population, for testing a QTL within a marker interval, is defined as

$$y_j = \mu + ax^*_{j|i} + dz^*_{j|i} + u_j,\ i = 1, 2, 3 \text{ and } j = 1, 2, \ldots, n \quad (1)$$

where y_j is the phenotypic value of the j-th individual, $x^*_{j|i} = p_{j|1} - p_{j|3}$, $z^*_{j|i} = p_{j|2}$, μ is the general mean effect, a is the QTL additive effect, d is the QTL dominance effect and $u_j \sim NID(0,\sigma^2)$ is a random error. Here, $x^*_{j|i}$ and $z^*_{j|i}$ are the probabilities for QTL genotypes conditional the flanking marker genotypes. Since conditional expectation is equivalent to conditional probabilities of QTL genotypes [5], $x^*_{j|i}$ and $z^*_{j|i}$ are fixed. Since $x^*_{j|i}$ and $z^*_{j|i}$ are fixed, so this model is called fixed effect model.

The conditional probabilities for QTL genotypes *QQ*, *Qq* and *qq* given the flanking marker genotypes are denoted by $p_{j|1}$, $p_{j|2}$ and $p_{j|3}$ respectively. The conditional probabilities $p_{j|1}$, $p_{j|2}$ and $p_{j|3}$ are shown in TABLE I for F_2 population. In TABLE I, p is defined as $p = r_{MQ}/r_{MN}$ where r_{MQ} is the recombination fraction between the left marker M and the putative QTL and r_{MN} is the recombination fraction between two flanking markers M and N. Also c is defined as $c = r^2_{MN}/[\ r^2_{MN} + (1 - r^2_{MN})]$. The possibility of a double recombination event in the interval is ignored.

To investigate the existence of a QTL at a given position within a marker interval, we want to test the hypothesis H_0: $a = 0$ and $d = 0$ (i.e., there is no QTL) versus H_1: H_0 is not true.

Under the normality assumption of error, the probability density function of the trait value (y) within each QTL genotype class is $N(\mu + ax^*_{j|i} + dz^*_{j|i},\ \sigma^2)$. Then the likelihood function for the parameters $\boldsymbol{\theta} = (\mu, a, d, \sigma^2)$ can be written as follows

$$L(\boldsymbol{\theta}\,|\,Y) = \prod_{i=1}^{n}\left[\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{y_j - \mu - ax^*_{j|i} - dz^*_{j|i}}{\sigma}\right)^2\right]\right] \quad (2)$$

To test H_0 against H_1, the likelihood ratio test (LRT) statistic is defined as

$$LRT = 2\left[\log\sup_{\boldsymbol{\theta}} L(\boldsymbol{\theta}\,|\,Y) - \log\sup_{\boldsymbol{\theta}_0} L(\boldsymbol{\theta}\,|\,Y)\right] = 4.608295 * LOD \quad (3)$$

where, Θ_0 and Θ are the restricted and unrestricted parameter spaces.

The threshold value to reject the null hypothesis can't be simply chosen from a chi-square distribution because of the violation of regularity conditions of asymptotic theory under H_0. The number and size of intervals should be considered in determining the threshold value. Since multiple tests are performed in mapping, the hypotheses are usually tested at every position of an interval and for all intervals of the genome to produce a continuous

LRT statistic profile. At every position, the position parameter p is predetermined and only μ, a, d and σ^2 are involved in estimation and testing. If the tests are significant in a chromosomal region, the position with the largest LRT statistic is inferred as the estimate of the QTL position and the maximum likelihood estimates (MLEs) at this position are the estimates of μ, a, d and σ^2 obtained by iterative way.

The MLEs of the parameters $\gamma = [\mu \;\; a \;\; d]^T$ and σ^2 are obtained as follows

$$\hat{\gamma} = (X^T X)^{-1}(X^T Y) \;\text{ and }\; \hat{\sigma}^2 = \frac{1}{n}(Y - X\hat{\gamma})^T (Y - X\hat{\gamma}) \tag{4}$$

Obviously these estimates are very much sensitive to outliers. Therefore, regression analysis by MLE produces misleading results in presence of outliers.

TABLE I. CONDITIONAL PROBABILITIES OF A PUTATIVE QTL GENOTYPE GIVEN THE FLANKING MARKER GENOTYPES FOR AN F_2 POPULATION

Marker Genotypes	**Expected Frequency**	**QTL Genotypes**		
		***QQ*(p_{ji1})**	***Qq*(p_{ji2})**	***qq*(p_{ji3})**
MN/MN	$(1-r)^2/4$	1	0	0
MN/Mn	$r(1-r)/2$	$(1-p)$	p	0
Mn/Mn	$r^2/4$	$(1-p)^2$	$2p(1-p)$	p^2
MN/mN	$r(1-r)/2$	p	$(1-p)$	0
MN/mn	$[(1-r)^2 + r^2]/2$	$cp(1-p)$	$1-2cp(1-p)$	$cp(1-p)$
Mn/mn	$r(1-r)/2$	0	$(1-p)$	p
mN/mN	$r^2/4$	p^2	$2p(1-p)$	$(1-p)^2$
mN/mn	$r(1-r)/2$	0	p	$(1-p)$
mn/mn	$(1-r)^2/4$	0	0	1

ROBUST QTL MAPPING FOR F_2 POPULATION USING REGRESSION APPROACH

The β-likelihood function (for details about β-likelihood, see [6]) for $\boldsymbol{\theta}$ is given by

$$L_\beta(\boldsymbol{\theta} \mid Y) = \frac{1}{\beta}\left[\frac{1}{n l_\beta(\boldsymbol{\theta})}\sum_{t=1}^{n} f_{\boldsymbol{\theta}}^{\beta}(y_t) - 1\right] \tag{5}$$

The β-likelihood equation is obtained as

$$\sum_{j=1}^{n}(y_j - \mu - a x^*_{j|i} - d z^*_{j|i}) w_j x_{kj} = 0;\; k = 0, 1, 2 \tag{6}$$

where $x_{0j} = 1$ for all $j = 1, 2, \ldots, n$ and $w_j = \exp[-(\beta/2\sigma^2)(y_j - \mu - a x^*_{j|i} - d z^*_{j|i})^2]$ for $i = 1, 2$. The function $w_j = w(y_j|\theta, x_{ij})$ is the weight function which produces almost zero weight for the outlying observations.

Solving (6), we get the proposed estimates of the parameters $\boldsymbol{\theta}$ as

$$\hat{\gamma} = (X^T X_W)^{-1}(X_W^T Y) \;\text{ and }\; \hat{\sigma}^2 = \frac{1}{n}(Y - X\hat{\gamma})^T (Y - X\hat{\gamma}) \tag{7}$$

where $X_W = X_{n\times 3} \otimes (W_{n\times 1}\, \mathbf{1}_{1\times 3})^{-1}(X_W^T Y)$. The notation $\otimes$ denotes the Hademerds product.

To test H_0: $a = 0$ and $d = 0$ against H_1: H_0 is not true, the proposed test criterion is defined as

$$\lambda_\beta = 2n[L_\beta(\hat{\boldsymbol{\theta}}_1 \mid Y) - L_\beta(\hat{\boldsymbol{\theta}}_0 \mid Y)], \text{ where } \hat{\boldsymbol{\theta}}_0 = (\hat{\mu}, \hat{\sigma}^2) \text{ and } \hat{\boldsymbol{\theta}}_1 = (\hat{\mu}, \hat{a}, \hat{d}, \hat{\sigma}^2) \quad (8)$$

By permutation test, we compute the p-value for testing H_0 vs H_1 using the following formula

$$p = \sum_{k=1}^{N_p} I_{[\hat{\lambda}_{\beta(k)} \le \hat{\lambda}_\beta]} / N_p \quad (9)$$

where N_p is the number of permutation under H_0 and $\hat{\lambda}_\beta$ is the estimate of λ_β for the original dataset and $\hat{\lambda}_\beta(k)$ is the estimate of λ_β for the k-th permutation of the values of the response variable. Note that, for $\beta \to 0$, $\hat{\lambda}_\beta$ reduces to the approximate χ^2 distribution.

SIMULATION RESULTS

To illustrate the performance of the proposed method in comparison of random effect and fixed effect model for QTL mapping with F_2 population, we have considered two unlinked QTLs with total 7 chromosomes and 13 equally spaced markers in each of chromosomes, where any two successive marker interval size is 5cM. The true QTL position is located in chromosome 1 and 3 with marker 7. The true values for the parameters in the fixed effect model are assumed as $\mu = 0.05$, $a = 0.8$, $d = 0.4$ and $\sigma^2 = 0.5$. We have generated 250 trait values with heritability $h^2 = 0.20$ which means that 20% of the trait variation is controlled by QTL and the remaining 80% is subject to the environmental effects (random error). To investigate the robustness of the proposed method in a comparison of the REM and FEM methods, we contaminated 12% trait values in this dataset by outliers. To perform the simulation study we have used R/qtl software [8].

Fig. 1(a) and Fig. 1(b) are representing the scatter plots of 250 trait values in presence and absence of outliers, respectively. Then we computed LOD scores by REM, FEM and the proposed methods for both types of data sets. Fig. 1(c) and Fig. 1(d) are showing the LOD scores profile plots for the uncontaminated and contaminated datasets, respectively.

In the LOD scores profile plots the dotted, two dash and solid lines represent the LOD scores at every 1cM position in the chromosomes for REM, FEM and the proposed method with $\beta = 0.2$, respectively. It is seen that the highest LOD score peak occurs in the true QTL position of the true chromosome 1 and 3 with marker 7 by all three methods for the uncontaminated dataset. However, in presence of outliers, the highest LOD score peak occurs in the true QTL position by the proposed method only [see Fig. 1(d)].

CONCLUSION

In this paper, a new robust regression based interval mapping approach has been discussed for QTL analysis by maximum β-likelihood estimation with F_2 population. The value of the tuning parameter β plays a key role on the performance of the proposed method. An appropriate value for the tuning parameter β can be selected by cross validation. The proposed method with tuning parameter $\beta = 0$ reduces to the traditional interval mapping approach. Simulation results show that the proposed method significantly improves the performance over the classical interval mapping approaches in presence of phenotypic outliers.

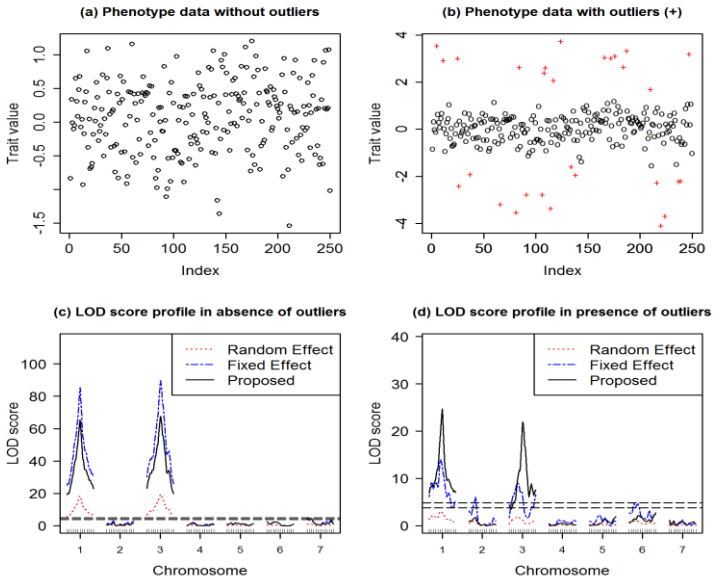

Fig. 1. Simulated phenotypic observations in (a) absence and (b) presence of 12% outliers, and LOD score profile in (c) absence and (d) in presence of 12% outliers.

Acknowledgment

We would like to thank the learned reviewers whose valuable comments helped to strengthen this paper. Also we would like to acknowledge HEQEP sub-project (CP-3603, W2, R3) for its financial support during this research work.

References

[1] J. M. Thoday, "Location of polygenes", Nature 191, pp. 368–370, 1960.

[2] E. S. Lander and D. Botstein, "Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps", Genetics, 121, pp. 185–199, 1989.

[3] C. S. Haley and S. A. Knott, "A simple regression method for mapping quantitative trait in line crosses using flanking markers", Heredity, 69, pp.315–324, 1992.

[4] S. Xu, "Mapping Quantitative Trait Loci Using Multiple Families of Line Crosses", Genetics, 148, pp.517–524, 1998.

[5] C. H. Kao, "On the differences between maximum likelihood and regression interval mapping in the analysis of quantitative trait loci", Genetics, 156, pp.855–865, 2000.

[6] M.N.H. Mollah, M. Minami and S. Eguchi, "Robust prewhitening for ICA by minimizing beta-divergence and its application to FastICA", Neural Processing Letters, 2007.

[7] M. Soller, T. Brody and A. Genizi, "On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbredlines, Theoret. Appl. Genetics, 47, pp. 35–39, 1976.

[8] Karl W. Broman, Hao Wu, Saunak Sen and Gary A. Churchill, "R/qtl: QTL mapping in experimental crosses", Bioinformatics, Vol. 19, pp. 889–890, 2003.