J. Bio-Sci. 24: 75-81, 2016 ISSN 1023-8654
http://www.banglajol.info/index.php/JBS/index

ROBUST LINEAR REGRESSION BASED SIMPLE INTERVAL MAPPING FOR QTL ANALYSIS WITH BACKCROSS POPULATION

MJ Alam[1*], M Alamin[2], MR Hossain[1], SMS Islam[3] and MNH Mollah[1]

[1]Bioinformatics Laboratory, Department of Statistics, University of Rajshahi, Bangladesh; [2]Department of Agronomy, Zhejiang University, Hangzhou, China; [3]Institute of Biological Sciences, University of Rajshahi, Rajshahi 6205, Bangladesh

Abstract

Simple interval mapping (SIM) is one of the most important techniques for the identification of quantitative trait locus (QTL). Most of the approaches of SIM are very sensitive to phenotypic outliers and produce misleading results. There is a robust approach of SIM only for F_2 population. However, there is no robust SIM method for Backcross population. The objective was to develop a new approach of SIM with Backcross population which is robust against phenotypic outliers and performs almost the same as existing classical methods in absence of outliers. Maximum likelihood (ML) and linear regression (LR) based approaches of SIM are not robust against phenotypic outliers. In this research, we have developed a robust regression based SIM approach by maximizing β-likelihood function for Backcross population. The proposed method reduces to the LR-based SIM method when $\beta = 0$. To measure the performance of the proposed method in comparison of ML and LR based SIM with backcross population; we have generated phenotypic and genotypic data for Backcross population using simulation technique. LOD score profile plot shows that the highest peaks of LOD scores occur in the true QTL positions of the true chromosomes at true markers by all three methods for the uncontaminated dataset. However, in presence of outliers, only the proposed method gives the highest LOD score peaks at the true QTL positions on the true chromosomes. The simulation results showed that the proposed method improves performance over the existing SIM methods in presence of phenotypic contaminations.

Key words: Backcross population, beta-LRT criterion, maximum beta-likelihood estimation, QTL analysis, robustness, robust linear regression

Introduction

The rapid advancement in molecular biology has increased the availability of fine scale genetic markers which facilitate the wide use of QTL analysis in the genetic study of quantitative traits in bioinformatics. Liu (1997) and Wu et al. (2007) discussed various techniques of QTL mapping in their texts. Thoday (1961) first proposed the idea of using two markers to bracket a region for testing QTLs. Soller et al. (1976) examined the power of experiments at detecting linkage between a quantitative locus and a marker locus. Similar to Thoday's (1961), but much improved, method called interval mapping (IM) approach was proposed by Lander and Botstein (1989) which is based on linkage relationships between a QTL and flanking markers. Maximum likelihood (ML) based IM (Lander and Botstein 1989) and regression based IM (Haley and Knott 1992) are two most popular and widely used interval mapping approaches.

In practice, QTL effects are treated as either fixed or random (Xu 1998). In fixed effects QTL model, allelic substitution effects are usually estimated and tested, and QTL variance is calculated from estimated allelic

*Author for correspondence: jahangir_statru63@yahoo.com

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 (Lander and Botstein 1989), 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 (Kao 2000) and this model can be treated as a fixed effect model (FEM).

The existing interval mapping based on REM (Lander and Botstein 1989) and FEM (Haley and Knott 1992) are two most popular and widely used methods for QTL analysis. But these methods are not robust against phenotypic contaminations. There is a regression based robust approach of SIM for QTL mapping only with F_2 population (Alam et al. 2015). In this work, we propose a robust method with FEM to perform QTL analysis for Backcross population. Also we have investigated the performance of the proposed method with the existing random effect QTL model and fixed effect QTL model for Backcross population by simulation study.

Materials and Methods

Linear regression based SIM approach for QTL detection with backcross population

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

$$y_j = \mu + ax_{j|i} + u_j,\ i = 1, 2 \text{ and } j = 1, 2, \ldots, n \qquad (1)$$

where y_j is the phenotypic value of the j-th individual, μ is the general mean effect, $x_{j|i} = p_{j|1}$, a is the QTL additive effect and $u_j \sim NID(0,\sigma^2)$ is a random error. Here, $x_{j|i}$ is the conditional probability for QTL genotypes given the flanking marker genotypes. Since conditional expectation is equivalent to conditional probabilities of QTL genotypes (Kao 2000), $x_{j|i}$ is fixed for QTL genotypes given flanking marker genotypes. Since $x_{j|i}$ is fixed, so this model is called fixed effect model.

The conditional probabilities for QTL genotypes *QQ* and *Qq* given the flanking marker genotypes are denoted by $p_{j|1}$ and $p_{j|2}$, respectively. The conditional probabilities $p_{j|1}$ and $p_{j|2}$ are shown in Table 1 for Backcross population. In Table 1, 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*. The possibility of a double recombination event in the interval is ignored.

Table 1. Conditional Probabilities of a putative QTL genotype given the flanking marker genotypes for a backcross population.

Marker genotypes	Expected frequency	QTL genotypes	
		QQ ($p_{j\|1}$)	*Qq* ($p_{j\|2}$)
MN/MN	$(1-r)/2$	1	0
MN/Mn	$r/2$	$(1-p)$	p
MN/mN	$r/2$	p	$(1-p)$
MN/mn	$(1-r)/2$	0	1

To investigate the existence of a QTL at a given position within a marker interval, we want to test the hypothesis H_0: $a = 0$ (i.e., there is no QTL at a given position) 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 $N(\mu + ax_{jli}, \sigma^2)$.

Then the likelihood function for the parameters $\boldsymbol{\theta} = (\mu, a, \sigma^2)$ can be written as follows

$$L(\boldsymbol{\theta} \mid Y) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y_j - \mu - ax_{jli}}{\sigma} \right)^2 \right] \tag{2}$$

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

$$LRT = -2\ln\left[\frac{\sup_{\Theta_0} L(\boldsymbol{\theta} \mid Y)}{\sup_{\Theta} L(\boldsymbol{\theta} \mid Y)} \right] = 4.608295 * \mathrm{LOD} \tag{3}$$

where, Θ_0 and Θ are the restricted (H_0) and unrestricted (H_1) parameter spaces.

The threshold value to reject the null hypothesis cannot 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 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 and σ^2 obtained by iterative way.

The MLEs of the parameters μ, a and σ^2 are as follows

$$\hat{\mu} = \bar{y} - a\bar{x}, \hat{a} = \frac{\sum_{j=1}^{n} (x_{jli} - \bar{x})(y_j - \bar{y})}{\sum_{j=1}^{n} (x_{jli} - \bar{x})^2} \text{ and } \hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n} (y_j - \hat{\mu} - \hat{a}x_{jli})^2 \tag{4}$$

where $\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$ and $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_{jli}, i = 1, 2.$

Obviously these ML estimates of μ, a and σ^2 are very much sensitive to outliers. Therefore, regression analysis by MLE produces misleading results in presence of contaminated data.

Robust linear regression based SIM for QTL detection with backcross population

The β-likelihood function (for details about β-likelihood (Mollah et al. 2007)) for $\boldsymbol{\theta}$ is given by

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

The β-likelihood equation is obtained as

$$\sum_{j=1}^{n}(y_j-\mu-ax_{j|i})w(y_j\,|\,\boldsymbol{\theta},x_{j|i})x_{kj}=0;\,k=0,1,2 \tag{6}$$

where $x_{0j}=1$ for all $j=1, 2, \ldots, n$ and $w(y_j\,|\,\boldsymbol{\theta},x_{j|i})=\exp\left[-\frac{\beta}{2\sigma^2}(y_j-\mu-ax_{j|i})^2\right]$ for $i=1, 2$. The function $w(y_j|\boldsymbol{\theta}, x_{j|i})$ is the weight function which produces almost zero weight for the outlying observations. Solving equation (6), we get the proposed estimates of the parameters $\boldsymbol{\theta}$ as

$$\hat{\mu}=\bar{y}_w-a\bar{x}_w,\ \hat{a}=\frac{\sum_{j=1}^{n}w(y_j\,|\,\boldsymbol{\theta},x_{j|i})(x_{j|i}-\bar{x}_w)(y_j-\bar{y}_w)}{\sum_{j=1}^{n}w(y_j\,|\,\boldsymbol{\theta},x_{j|i})(x_{j|i}-\bar{x}_w)^2}\ \text{and}$$

$$\hat{\sigma}^2=\frac{\sum_{j=1}^{n}w(y_j\,|\,\boldsymbol{\theta},x_{j|i})(y_j-\hat{\mu}-\hat{a}x_{j|i})^2}{\sum_{j=1}^{n}w(y_j\,|\,\boldsymbol{\theta},x_{j|i})} \tag{7}$$

where $\bar{y}_w=\frac{\sum_{j=1}^{n}w(y_j\,|\,\boldsymbol{\theta},x_{j|i})y_j}{\sum_{j=1}^{n}w(y_j\,|\,\boldsymbol{\theta},x_{j|i})}$ and $\bar{x}_w=\frac{\sum_{j=1}^{n}w(y_j\,|\,\boldsymbol{\theta},x_{j|i})x_{j|i}}{\sum_{j=1}^{n}w(y_j\,|\,\boldsymbol{\theta},x_{j|i})},\ i=1,2.$

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

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

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

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

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.

Simulated data

To measure the performance of the proposed method in comparison of the fixed effect and random effect models for QTL mapping with Backcross population, we have generated phenotypic and genotypic data for Backcross population using simulation technique. We have considered two unlinked QTLs, total 10 chromosomes and 11 equally spaced markers in each of the 10 chromosomes, where any two successive marker interval size is 5 cM. The true QTL position is located in chromosome 2, 3and 5 at marker 5 (locus position 20 cM). The true values for the parameters in the fixed effect model are assumed as $\mu = 0.5$, $a = 0.8$, $d = 0.4$ 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 (Broman et al. 2003, homepage: http://www.rqtl.org/).

Results and Discussion

Table 2 shows QTL positions (i.e., chromosome, marker and locus position) identified by REM, FEM and the proposed method. Fig. 1a and 1b are representing the scatter plots of 250 trait values in presence and absence of outliers, respectively. Then we computed LOD scores based on REM, FEM and the proposed methods for both types of data sets. Fig. 1c and 1d 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 2, 3 and 5 at marker 5 (locus position 20 cM) 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 (Fig. 1d).

Table 2. QTL positions identified by each method in absence and presence of outliers.

Methods	True QTL positions	Identified QTL positions	
		In absence of outliers	In presence of outliers
REM	On chromosomes 2, 3 and 5 at marker 5 (locus position 20 cM) for each chromosome.	On chromosomes 2, 3 and 5 at marker 5 (locus position 20 cM) for each chromosome.	REM fails identify any QTL on any chromosome.
FEM	On chromosomes 2, 3 and 5 at marker 5 (locus position 20 cM) for each chromosome.	On chromosomes 2, 3 and 5 at marker 5 (locus position 20 cM) for each chromosome.	(i) On chromosome 1 at marker 10 (locus position 45 cM). (ii) On chromosome 2 at marker 5 (locus position 20 cM). (iii) On chromosome 5 at marker 1 (locus position 0 cM). (iv) On chromosome 10 at marker 11 (locus position 50 cM)
Proposed model	On chromosomes 2, 3 and 5 at marker 5 (locus position 20 cM) for each chromosome.	On chromosomes 2, 3 and 5 at marker 5 (locus position 20 cM) for each chromosome.	On chromosomes 2, 3 and 5 at marker 5 (locus position 20 cM) for each chromosome.

From Table 2 and Fig. 1 we observe that all of the 3 methods (REM, FEM and proposed method) identify the true QTL positions correctly in absence of outliers. But in presence of outliers the REM fails to identify any significant QTL position and the FEM identify QTLs on chromosomes 1 at marker 10 (locus position 45 cM), on chromosome 2 at marker 5 (locus position 20 cM), on chromosome 5 at marker 1 (locus position 0 cM) and on chromosome 10 at marker 11 (locus position 50 cM). The positions on chromosome 5 at marker 1 and on chromosome 10 at marker 11, identified by FEM, are not the true position of QTLs. However, in presence of outliers, the proposed method identify the QTLs on chromosome 2, 3 and 5 at marker 5 (locus position 20 cM) which are the true QTL positions.

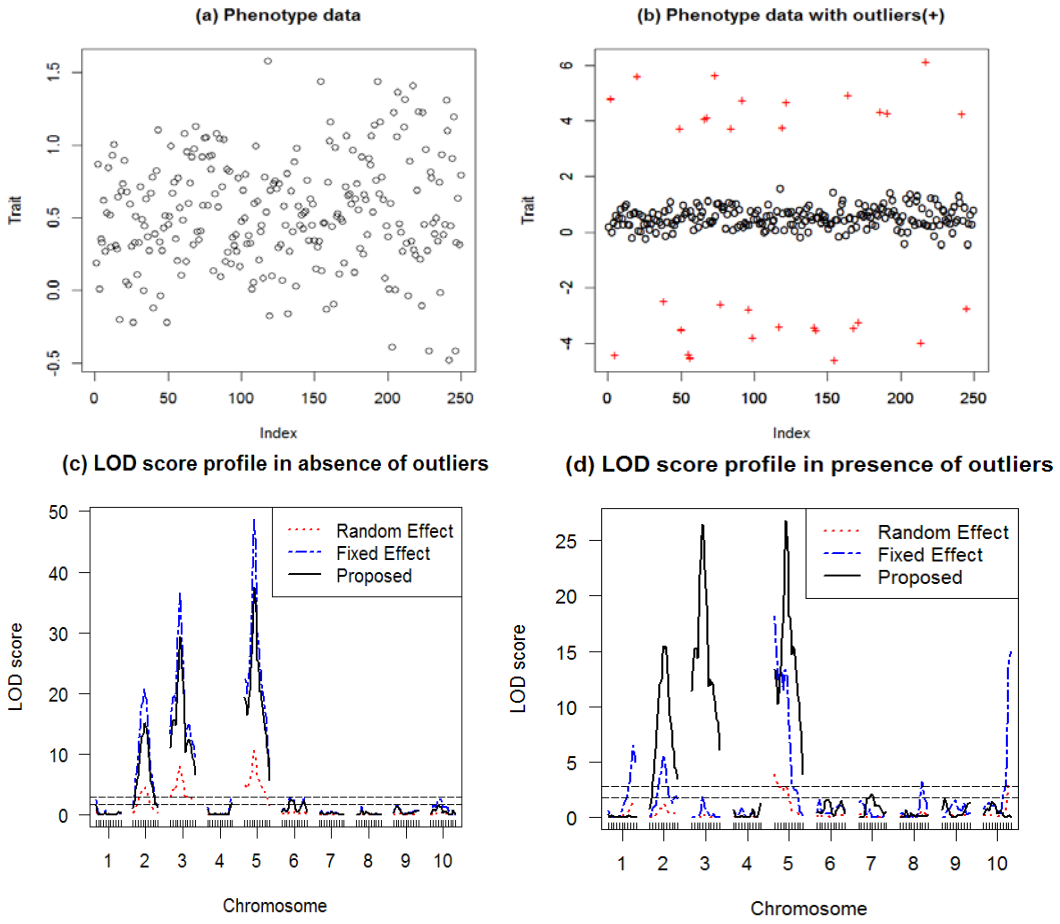

Fig. 1. Simulated phenotypic observations in (a) absence and (b) presence of 12% outliers, and LOD score profile in (c) absence and (d) in presence of 12% outliers.

Hence, in presence of outliers, the classical methods of SIM (REM and FEM) fail to identify the all the true QTL positions whereas the proposed method successfully identifies all the true QTL positions. Also in absence of outliers the proposed method is working as the classical methods.

Conclusion

Under this study we have proposed a new robust regression based simple interval mapping approach for QTL analysis by maximum β-likelihood estimation with Backcross population. The performance of the proposed method is controlled by the tuning parameter β. An appropriate value for the tuning parameter β can be selected by cross validation. The proposed method reduces to the traditional interval mapping approach when the tuning parameter $\beta = 0$. Simulation results show that the proposed method significantly improves the performance over the classical simple interval mapping approaches in presence of phenotypic outliers. Also in absence of outliers it shows similar performance to the classical methods of SIM.

Acknowledgement

We would like to acknowledge to the HEQEP Sub-Project (CP-3603, W2, R3) for providing financial support for this study.

References

Alam MJ, Alamin M, Humaira SM, Amanullah M and Mollah MNH (2015). Regression Based Robust QTL Analysis using Flanking Marker with Intercross (F_2) Population, Proceedings of the International Conference on Materials, Electronics and Information Engineering, Faculty of Engineering, University of Rajshahi, Rajshahi, Bangladesh, Paper ID - 125. http://180.211.185.216/icmeie2015/proceedings/pdfs/125.pdf.

Liu BH (1997). Statistical Genomics: Linkage, Mapping, and QTL Analysis. CRC Press LLC, 648 pp.

Broman KW and Sen S (2009). A Guide to QTL Mapping with R/qtl. Springer Science + Business Media, New York, 396 pp.

Broman KW, Wu H, Sen S and Churchill GA (2003). R/qtl: QTL mapping in experimental crosses, Bioinformatics 19(7): 889-890.

Haley CS and Knott SA (1992). A simple regression method for mapping quantitative trait in line crosses using flanking markers, Heredity 69(4): 315-324.

Kao CH (2000). On the differences between maximum likelihood and regression interval mapping in the analysis of quantitative trait loci, Genetics 156(2): 855-865.

Lander ES and Botstein D (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps, Genetics 121(1): 185-199.

Mollah MNH, Minami M and Eguchi S (2007). Robust prewhitening for ICA by minimizing beta-divergence and its application to FastICA, Neural Processing Letters 25(2): 91-110.

Wu R, Ma CX and Casella G (2007). Statistical genetics of quantitative traits: linkage, maps and QTL. Springer-Verlag, New York, 368 pp.

Soller M, Brody T and Genizi A (1976). On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines, Theoretical and Applied Genetics 47(1): 35-39.

Thoday JM (1961). Location of polygenes, Nature 191: 368-370.

Xu S (1998). Mapping quantitative trait loci using multiple families of line crosses, Genetics 148(1): 517-524.