Detection of Quantitative Trait Loci in Presence of Phenotypic Contamination

Md. Nurul Haque Mollah mnhmollah@yahoo.co.in
Department of Statistics,
Faculty of Science,
University of Rajshahi,
Rajshahi-6205, Bangladesh

Abstract

Genes controlling a certain trait of organism is known as quantitative trait loci (QTL). The standard Interval mapping [8] is a popular way to scan the whole genome for the evidence of QTLs. It searches a QTL within each interval between two adjacent markers by performing likelihood ratio test (LRT). However, the standard Interval mapping (SIM) approach is not robust against outliers. An attempt is made to robustify SIM for QTL analysis by maximizing β-likelihood function using the EM like algorithm. We investigate the robustness performance of the proposed method in a comparison of SIM algorithm using synthetic datasets. Experimental results show that the proposed method significantly improves the performance over the SIM approach for QTL mapping in presence of outliers; otherwise, it keeps equal performance.

Keywords: Quantitative trait loci (QTL), Gaussian mixture distribution, Likelihood ratio test (LRT), Method of maximum likelihood, Method of maximum *β* -likelihood and Robustness.

1. INTRODUCTION

The basic methodology for mapping QTLs involves arranging a cross between two inbred strains differing substantially in a quantitative trait: segregating progeny are scored both for the trait and for a number of genetic markers. A cross between two parental inbred lines *A* and *B* is performed to produce an *F1* population. The *F1* progeny are all heterozygote's with the same genotype [4]. Typically, the segregating progeny are produced by a backcross *B1 = F1*×parent or an intercross *F2 = F1 × F1.*

Let (μ_A, σ_A^2), (μ_B, σ_B^2), $(\mu_{F1}, \sigma_{F1}^2)$ and $(\mu_{F2}, \sigma_{F2}^2)$

denote the set of mean and variance of a phenotype in the *A, B, F1* and *F2* population, respectively. Let $\mu_B - \mu_A > 0$ denote the phenotypic difference between the strains. The cross will be analyzed under the classical assumption that the phenotypic variations are influenced by the effects of both genetic and non-genetic (environmental) factors. In particular, we assume complete codominance and no epistasis. These implies that

$$\mu_{F1_1} = \tfrac{1}{2}(\mu_A + \mu_B),\ \mu_{F1} = \tfrac{1}{3}(\mu_A + \mu_{F1} + \mu_B) \text{ and } \sigma_A^2 = \sigma_B^2 = \sigma_{F1}^2 < \sigma_{F2}^2.$$

The variance within the *A, B* and *F1* populations equal the environmental variance, σ^2_E among genetically identical individuals, while the variance within the *F2* progeny also includes genetic variance, $\sigma^2_G = \sigma^2_{F2} - \sigma^2_E$. Frequently, phenotypic measurements must be mathematically transformed so that parental phenotypes are approximately normally distributed.

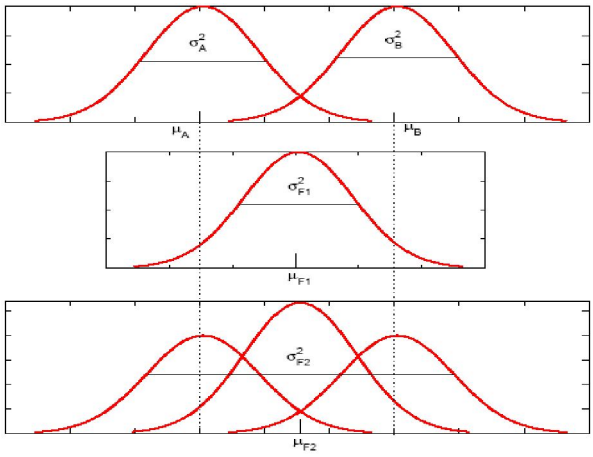

Figure-1: Phenotype distribution. Schematic drawing of phenotype distributions in the *A* and *B* parental, *F1* hybrid and *F2* intercross populations.

With the rapid advances in molecular biology, it has become possible to gain fine-scale genetic maps for various organisms by determining the genomic positions of a number of genetic markers (RFLP, isozymes, RAPDs, AFLP, VNTRs, etc.) and to obtain a complete classification of marker genotypes by using codominant markers. These advances greatly facilitate the mapping and analysis of quantitative trait loci(QTLs). Thoday [12] first proposed the idea of using two markers to bracket a region for testing QTLs. Lander and Botstein [8] implemented a similar, but much improved, method to use two adjacent markers to test the existence of a QTL in the interval by performing a likelihood ratio test (LRT) at every position in the interval. This is known as standard interval mapping (SIM) approach. It is a comparatively popular way to detect a QTL position in a chromosome [11]. However, It is not robust against outliers [5, 8, 9]. In this project, an attempt is made to robustify the SIM approach [8] by maximizing *β*-likelihood function [10] using EM like algorithm [3].

In section 2, we discuss the genetic model and its extension to statistical SIM model.
Section 3 introduce the robustification of SIM approach for QTL analysis. We demonstrate the performance of the proposed method using simulated datasets in section 4 and make a conclusion of our study in section 5.

2. GENETIC MODEL FOR SIM APPROACH

Let us consider a putative QTL locus Q of two alleles *Q* and *q* with allelic frequencies *p* and (1-*p*), respectively. Define an indicator variable for alleles by

$$u = \begin{cases} 1 & \text{for} \quad Q \\ 0 & \text{for} \quad q \end{cases}$$

and

$$v = u - E(u) = u - p = \begin{cases} 1-p & \text{for} \quad Q \\ -\text{p} & \text{for} \quad q, \end{cases}$$

where *v* is a standardized indicator variable with mean zero. Then the genetic-effect design variables for two alleles are defined as

$$x = v_1 + v_2 = \begin{cases} 2(1-p) & \text{for } QQ \\ 1-2p & \text{for } Qq \\ -2p & \text{for } qq \end{cases}$$

and

$$x = -2v_1v_2 = \begin{cases} -2(1-p)^2 & \text{for } QQ \\ 2p(1-p) & \text{for } Qq \\ -2p^2 & \text{for } qq, \end{cases}$$

where v_1 and v_2 are for the two alleles in an individual. Then the genetic model for a QTL in the *F2* population is defined as

$$G = \mu + ax^* + dz^*, \quad (1)$$

where *a* and *d* are additive and dominance effects of QTL, and $x^*=x$ and $z^*=z$ are the genetic-effect design variables with $p=1/2$. Therefore, in matrix notation, the genetic model for a QTL in the *F2* population can be written as

$$G = \begin{bmatrix} G_2 \\ G_1 \\ G_0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} 1 & \frac{-1}{2} \\ 0 & \frac{1}{2} \\ -1 & \frac{-1}{2} \end{bmatrix} \begin{bmatrix} a \\ d \end{bmatrix} = 1_{3\times 1}\mu + DE.$$

It was proposed to model the relation between a genotypic value *G* and the genetic parameters *μ, a* and *d.* Here G_2, G_1 and G_0 are the genotypic values of genotypes *QQ, Qq* and *qq*. We call *D* the genetic design matrix. The first and second columns of *D*, denoted by D_1 and D_2, represent the status of the additive and dominance parameters of the three different genotypes.

Let loci M, with allels *M* and *m*, and N with alleles *N* and *n*, denote two flanking markers for an interval where a putative QTL is being tested. Let the unobserved QTL locus Q with alleles *Q* and *q* be located in the interval flanked by markers M and N. The distribution of unobserved QTL genotypes can be inferred from the observed flanking marker genotypes according to the recombination frequencies between them. To infer the distribution of QTL genotype, we assume

Table 1: Conditional Probabilities of a putative QTL genotype given the flanking marker genotypes for an *F2* population.

Marker genotypes	Expected frequency	QTL genotypes		
		$QQ(p_{j1})$	$Qq(p_{j2})$	$qq(p_{j3})$
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 (or mN/Mn)	$(1-r)^2/2+ r^2/2$	$cp(1-p)$	$1-2cp(1-p)$	$cp(1-p)$
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

$p=r_{MQ}/r_{MN}$, 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. $c=r^2_{MN}/[r^2_{MN}+(1-r_{MN})2]$. The possibility of a double recombination event in the interval is ignored.

that there is no crossover interference and also that double recombination events within the interval are very rare and can be ignored to simplify the analysis. The conditional probabilities of the QTL genotypes given marker genotypes are given in Table 1 for the *F2* population. We extract the conditional probabilities from this table to form a matrix **Q** for *F2* population.

2.1 Statistical Model for SIM Approach

Let us assume no epistasis between QTLs, no interference in crossing over, and only one QTL in the testing interval. A statistical model for SIM approach based on the genetic model (1) for testing a QTL in a marker interval is defined as

$$y_j = \mu + ax_j^* + dz_j^* + e_j^*, \tag{2}$$

where

$$(x_j^*,\ z_j^*) = \begin{cases} (1,\ -1/2) & \text{for } QQ \\ (0,\ \ 1/2) & \text{for } Qq \\ (-1,\ -1/2) & \text{for } qq, \end{cases}$$

y_j is the phenotypic value of the jth individual, μ is the general mean effect; and ϵ_j is a random error. We assume $\epsilon_j \sim N(0, \sigma^2)$. To investigate the existence of a QTL at a given position in a marker interval, we want to test the following statistical hypothesis.

Null Hypothesis, H_0 : a=0 and d=0 (i.e. there is no QTL at a given position).
Alternative Hypothesis, H_1 : H_0 is not true (i.e. there is a QTL at a given position).

2.2 SIM Approach by Maximizing Likelihood Function

In the SIM model (2), each phenotypic observation (y_j) is assumed to follow a mixture of three possible Gaussian densities with different means and mixing proportion, since observation y_j's are influenced by three QTL genotypes *QQ, Qq* and *qq*. Therefore, the density of each phenotypic observation (y_j) is defined as

$$f(y_j \mid \theta) = \sum_{i=1}^{3} p_{ji}\, \phi\left(\frac{y_j - \mu_{ji}}{\sigma}\right), \tag{3}$$

where $\theta=(\mu,p,a,d,\sigma^2)$, ϕ is a standard normal probability density function, $\mu_{j1}=\mu+a-d/2$, $\mu_{j2} = \mu + d/2$ and $\mu_{j3} = \mu - a - d/2$. The mixing proportions p_{ji}'s which are functions of the QTL position parameter p, are conditional probabilities of QTL genotypes given marker genotypes. For n individuals, the likelihood function for $\theta=(\mu, p,a,d,\sigma^2)$ is given by

$$L(\theta \mid \mathbf{Y}) = \prod_{j=1}^{n} f(y_j \mid \theta). \tag{4}$$

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

$$\begin{aligned} \text{LRT} &= -2\log\left[\frac{\sup_{\theta_0} L(\theta \mid \mathbf{Y})}{\sup_{\theta} L(\theta \mid \mathbf{Y})}\right] \\ &= 2[\log \sup_{\theta} L(\theta \mid \mathbf{Y}) - \log \sup_{\theta_0} L(\theta \mid \mathbf{Y})], \end{aligned} \tag{5}$$

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 hypothesis 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 a, d, μ and σ^2 are involved in estimation and testing. If the tests are significant in a chromosome region, the position with the largest LRT statistic is inferred as the estimate of the QTL position p, and the MLEs at this position are the estimates of a, d, μ and σ^2 obtained by EM algorithm treating the normal mixture model as an incomplete-data problem [8]. Note that EM algorithm has been also used to obtain MLEs in several studies of QTL mapping analysis [6, 7].

3. ROBUSTIFICATION OF SIM APPROACH BY MAXIMIZING β-LIKELIHOOD FUNCTION

The β-likelihood function for $\theta = (\mu, p, a, d, \sigma^2)$ as defined in equation (3) is given by

$$L_\beta(\theta \mid \mathbf{Y}) = \frac{1}{\beta}\left[\frac{1}{\mathrm{nl}_\beta(\theta)}\sum_{j=1}^{n}\left\{f(y_j \mid \theta)\right\}^\beta - 1\right], \tag{6}$$

where

$$l_\beta(\theta) = \left[\int \left\{f(y \mid \theta)\right\}^\beta dy\right]^{\beta/(1+\beta)}.$$

Note that maximization of β-likelihood function is equivalent to the minimization of β-divergence for estimating θ [10]. The β-likelihood function reduces to the log-likelihood function for $\beta \to 0$. In this paper, our proposal is to use the β-LOD score for the evidence of a QTL in a marker interval from the robustness point of view. It is defined by

$$\mathrm{LOD}_\beta = 2n[\sup_\theta \mathrm{L}_\beta(\theta \mid \mathbf{Y}) - \sup_{\theta_0} \mathrm{L}_\beta(\theta \mid \mathbf{Y})], \tag{7}$$

where θ_0 and θ are the restricted and unrestricted parameter spaces as before. For $\beta \to 0$, the LOD_β reduces to the likelihood ratio test (LRT) criterion as defined in equation. The threshold value to reject the null hypothesis H_0 can be computed by permutation test following the work of Churchill and Doerge [2]. At every position, the position parameter p is predetermined and only a, d, μ and σ^2 are involved in estimation and testing as before. If the tests are significant in a chromosome region, the position with the largest LOD_β is inferred as the estimate of the QTL position p, and the β-estimators at this position are the estimates of a, d, μ and σ^2 obtained by EM algorithm treating the normal mixture model as an incomplete-data problem.

3.1 Maximization of β-Likelihood Function Using EM Algorithm

The EM algorithm can be used for obtaining the maximum β-likelihood estimators of a, d, μ and σ^2 treating the normal mixture model as an incomplete-data density. Let

$$(\mathrm{x}_\mathrm{j}^*, \mathrm{z}_\mathrm{j}^*) = \begin{cases} \mathrm{p}_{\mathrm{j}1} & \text{if } \mathrm{x}_\mathrm{j}^* = 1,\ \mathrm{z}_\mathrm{j}^* = -\frac{1}{2} \\ \mathrm{p}_{\mathrm{j}2} & \text{if } \mathrm{x}_\mathrm{j}^* = 0,\ \mathrm{z}_\mathrm{j}^* = \frac{1}{2} \\ \mathrm{p}_{\mathrm{j}3} & \text{if } \mathrm{x}_\mathrm{j}^* = -1,\ \mathrm{z}_\mathrm{j}^* = -\frac{1}{2} \end{cases}$$

be the distribution of QTL genotype specified by x_j^* and z_j^*. Let us treat the unobserved QTL genotypes (x_j^* and z_j^*) as missing data, denoted by $y_{j(mis),}$ and the trait y_j as observed data, denoted by $y_{j(obs).}$ Then, the combination of $y_{j(mis)}$ and $y_{j(obs)}$ is the complete data, denoted by $y_{j(com)}$. The conditional distribution of observed data, given missing data, is considered as an independent sample from a population such that $y_j|(\theta,x_j^*,z_j^*)\sim N(\mu+ax_j^*+dz_j, \sigma^2)$

The complete-data density model in this problem is regarded as a two-stage hierarchical model. First the values of random variables (x_j^*, z_j^*) are sampled by a trinomial experiment to decide QTL genotype, and then a normal variate for that genotype is generated. The values of random variables (x_j^*, z_j^*) of individual *j* are (1, -1/2), (0, 1/2) or (-1, -1/2) for QTL genotype *QQ, Qq,* or *qq* with probability p_{j1}, p_{j2} *or* p_{j3}, respectively. Thus the complete-data density function is given by

$$\varphi(y_{j(com)} \mid \theta) = \left\{ p_{j1}\phi\left(\frac{y_j-\mu_{j1}}{\sigma}\right)\right\}^{-\frac{1}{2}(x_j^*+1)(z_j^*-\frac{1}{2})} \times \left\{ p_{j2}\phi\left(\frac{y_j-\mu_{j2}}{\sigma}\right)\right\}^{(x_j^*+1)(z_j^*+\frac{1}{2})} \times \left\{ p_{j3}\phi\left(\frac{y_j-\mu_{j3}}{\sigma}\right)\right\}^{\frac{1}{2}(x_j^*-1)(z_j^*-\frac{1}{2})}. \quad (8)$$

At a given position, *p* is determined. The EM algorithm is used for obtaining the maximum *β*-likelihood estimators of *a, d, μ* and σ^2 for the complete-data density. The iteration of the (*t*+1) EM-step is as follows:

E-step: The conditional expected complete-data *β*-likelihood with respect to the conditional distribution of Y_{mis} given Y_{obs} and the current estimated parameter value $\theta^{(t)}$ is given by

$$Q_\beta(\theta \mid \theta^{(t)}) = \int L_\beta(\theta \mid Y_{com})\, h(Y_{mis} \mid Y_{obs},\ \theta=\theta^{(t)})\, dY_{mis}$$
$$= \frac{1}{n\beta b_\beta(\theta)} \sum_{j=1}^{n}\sum_{i=1}^{3} \left\{ p_{ji}\phi\left(\frac{y_j-\mu_{ji}}{\sigma}\right)\right\}^{\beta} \times \pi_{ji}^{(t)} - \frac{1}{\beta}, \quad (9)$$

where

$$b_\beta(\theta) = (1+\beta)^{-\beta/2(1+\beta)}(2\pi\sigma^2)^{-\beta^2/2(1+\beta)}$$

and

$$\pi_{ji} = \frac{p_{ji}\phi\left(\frac{y_j-\mu_{ji}}{\sigma}\right)}{\sum_{i=1}^{3} p_{ji}\phi\left(\frac{y_j-\mu_{ji}}{\sigma}\right)}$$

which is the posterior probability of *j*-th individual given the *i-th* QTL genotype, *i*=1, 2 and 3 for QTL genotypes *QQ, Qq* and *qq*, respectively.

M-step: Find $\theta^{(t+1)}$ to maximize the conditional expected *β*-likelihood by taking the derivatives of $Q_\beta(\theta|\ \theta^{(t)})$ with respect to each parameter. The solutions of parameters in closed form are as follows.

$$\mu^{(t+1)} = \left[\mathbf{1}_n^T \left\{Y \# (\Pi_\beta^{(t)}\mathbf{1}_3) - \Pi_\beta^{(t)} DE^{(t)}\right\}\right]\left[\mathbf{1}_n^T \Pi_\beta^{(t)}\mathbf{1}_3\right]^{-1}, \quad (10)$$

$$a^{(t+1)} = \frac{(Y - \mathbf{1}_n \mu^{(t+1)})^T \Pi_\beta^{(t)} D_1 - \mathbf{1}_n^T \Pi_\beta^{(t)}(D_1 \# D_2)d^{(t)}}{\mathbf{1}_n^T \Pi_\beta^{(t)}(D_1 \# D_1)}, \quad (11)$$

$$d^{(t+1)} = \frac{(Y - \mathbf{1}_n \mu^{(t+1)})^T \Pi_\beta^{(t)} D_2 - \mathbf{1}_n^T \Pi_\beta^{(t)} (D_1 \# D_2) a^{(t+1)}}{\mathbf{1}_n^T \Pi_\beta^{(t)} (D_2 \# D_2)} \quad (12)$$

and

$$\sigma^{2(t+1)} = (1+\beta)[(Y - \mathbf{1}_n \mu^{(t+1)})^T \left\{(Y - \mathbf{1}_n \mu^{(t+1)}) \# (\Pi_\beta \mathbf{1}_3)\right\} - 2(Y - \mathbf{1}_n \mu^{(t+1)})^T \Pi_\beta^{(t)} DE^{(t+1)}$$
$$- E^{(t+1)T} V^{(t)} E^{(t+1)}][\mathbf{1}_n^T \Pi_\beta^{(t)} \mathbf{1}_3]^{-1}, \quad (13)$$

where

$$V = \begin{bmatrix} \mathbf{1}_n^T \Pi_\beta (D_1 \# D_1) & \mathbf{1}_n^T \Pi_\beta (D_1 \# D_2) \\ \mathbf{1}_n^T \Pi_\beta (D_2 \# D_1) & \mathbf{1}_n^T \Pi_\beta (D_2 \# D_2) \end{bmatrix}$$

which is a symmetric matrix. Here # denotes Hadamards product, which is the element-by-element product of corresponding elements of two same-order matrices and

$$\Pi_\beta = \left\{ \left[\exp\left\{ -\frac{1}{2}\left(\frac{y_j - \mu_{ji}}{\sigma} \right)^2 \right\} \right]^\beta \pi_{ji} \right\}_{n \times 3}$$

which is called the matrix of β-weighted posterior probabilities. For β=0, the matrix Π_β reduces to the matrix of standard posterior probabilities. It should be noted here that each element of n-vector $\mathbf{1}_n$ is 1. The *E* and *M* steps are iterated until a convergent criterion is satisfied. The converged values of *a, d,* μ and σ^2 are the values of minimum β-divergence estimators. Note that minimum β-divergence estimators (10-13) with β=0 reduce to maximum likelihood estimators (MLE) of SIM for QTL analysis.

Under null hypothesis Ho: *a*=0, *d*=0, the minimum β-divergence estimators for the parameters μ and σ^2 are obtained iteratively as follows

$$\mu^{(t+1)} = [YW_\beta^{(t)}][\mathbf{1}_n^T w_\beta^{(t)}]^{-1} \quad (14)$$

and

$$\sigma^{2(t+1)} = (1+\beta)(Y - \mathbf{1}_n \mu^{(t+1)})^T [(Y - \mathbf{1}_n \mu^{(t+1)}) \# W_\beta^{(t)})][\mathbf{1}_n^T w_\beta^{(t)}]^{-1}, \quad (15)$$

where

$$W_\beta = \left[\exp\left\{ -\frac{\beta}{2}\left(\frac{y_j - \mu}{\sigma} \right)^2 \right\} \right]_{n \times 1}$$

which is called the β-weight vector.

4. SIMULATION RESULTS

To illustrate the performance of the proposed method in a comparison of SIM approach [8] for QTL analysis, we consider *F2* intercross population for simulation study. In this study, we assume only one QTL on a chromosome with 10 equally spaced markers, where any two successive marker interval size is 5 cM . Marker positions and their genotypes are generated using R/qtl

software [1], homepage: http://www.rqtl.org/). The successive marker interval size 5 is considered. The QTL position is located by the 5*th* marker of chromosome-10. The true values for the parameters in the SIM model are assumed as a=0.4, d=0.8, μ=0.05 and σ^2=1. We randomly generated 250 trait values with heritability h^2=0.2 using the SIM model (2). A trait with heritability h^2=0.2 means that 20% of the trait variation is controlled by QTL and the remaining 80% is subject to the environmental effects (random error).

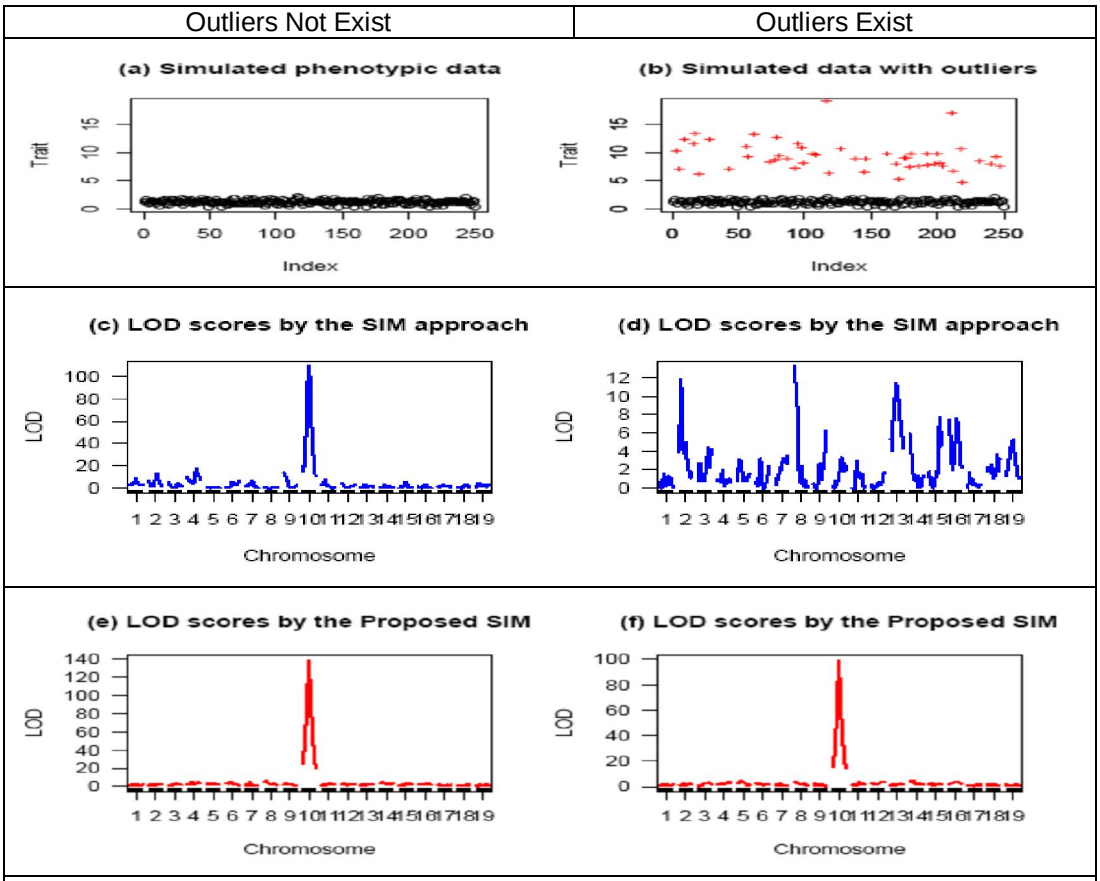

Figure-2: Simulation results based on two successive marker interval size 5 cM. (a-b) Simulated phenotypic observations in absence and presence of outliers, respectively. (c-d) LOD scores by the SIM approach in absence and presence of 15% outliers, respectively. (e-f) LOD scores by the proposed robustified SIM approach in absence and presence of 15% outliers, respectively.

Figure 2(a) represent the scatter plot of a sample of 250 trait values and a covariate. To investigate the robustness of the proposed method in a comparison of the SIM method for QTL mapping, we replaced 15% trait values randomly in each replication of the previous example by outliers (+) so that data contamination rate is 15% in each replication. Figures 2(b) represent the scatter plot of a sample of 250 trait values in presence of outliers.

To determine the QTL position, we compute LOD scores by both the SIM and the proposed robust SIM method. It should be noted here that the name 'LOD scores' is used for convenience of presentation instead of both LRT scores of SIM method and the β-LOD scores of the proposed method. According to the theoretical settings, the largest LOD score should be occurred with the true QTL positions in the entire genome. We computed LOD scores by both SIM and robust SIM approaches. Figures 2(c) and 2(e) represents the LOD scores at every 2 cM position in the chromosomes in absence of outliers by SIM and the proposed method, respectively. Similarly,

figures 2(d) and 2(f) represents the LOD scores at every 2 cM position in the chromosomes in presence of outliers by SIM and the proposed method, respectively. It is seen that the highest and significant LOD score peak occurs with the true QTL position in the chromosome 10 by both methods in absence of outliers, while in presence of phenotypic outliers, highest and significant LOD score peak occurs with the true QTL position by the proposed method only. Therefore, performance of both methods is almost same and good in absence of outliers, while the performance of the proposed method is better than the SIM approach in presence of outliers.

5. CONCLUSION

This paper proposes the robustification of the SIM algorithm for QTL mapping by maximizing β-likelihood function using EM like algorithm. The β-likelihood function reduces to the log-likelihood function for $\beta \rightarrow 0$. The proposed robust SIM algorithm with the tuning parameter $\beta=0$ reduces to the traditional SIM algorithm. The value of the tuning parameter β has a key role on the performance of the proposed method for QTL mapping. An appropriate value for the tuning parameter can be selected by cross-validation [10]. However, one can choose $\beta \in (0.1, 0.5)$, heuristically. Simulation results show that the robustified SIM algorithm with $\beta>0$ significantly improves the performance over the SIM approach in presence of phenotypic outliers. It keeps equal performance otherwise.

6. REFERENCES

1. K. W. Broman, H. Wu, S. Sen and G. A. Churchill. "*R/qtl: QTL mapping in experimental crosses*". Bioinformatics, Vol. **19**, pp. 889-890, 2003.
2. G. A. Churchill and R. W. Doerge,. "*Empirical Threshold Values for Quantitative Triat Mapping*". Genetics, Vol **138**, pp. 963-971, 1994.
3. A. P. Dempster, Laird, and Rubin, D. B.: "*Maximum likelihood from incomplete data via the EM algorithm".* J. Roy. Statist. Soc. B, **39**, pp. 1-38, 1977.
4. Elston, R. C., Stewart, J. "*The Analysis of Quantitative Traits for Simple Genetic Models from Parental, F1 and Backcross Data"*. Genetics, **73**, pp. 695-711, 1973.
5. 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.
6. R. C. Jansen, " *A general mixture model for mapping quantitative trait loci by using molecular markers*". Theor Appl Genet., **85**, 252-260, 1992.
7. 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.
8. E. S. Lander and D. Botstein. "*Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps"*. Genetics, **121**, pp. 185-199, 1989.
9. M. N. H. Mollah and S. Eguchi. "*Robust Composite interval Mapping for QTL Analysis by Minimum β-Divergence Method*". Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (IEEE BIBM08) , pp. 115-120, Philadelphia, USA, 2008.
10. M. N. H. Mollah, N. Sultana, M. Minami and S. Eguchi. "*Robust extraction of local structures by the minimum β-divergence method*". Neural Network, **23**, pp. 226-238, 2010.
11. A. H. Paterson, S. Damon, J. D. Hewitt, D. Zamir, H. D. Rabinowitch, S.E. Lincoln, E. S. Lander, S.D.Tanksley. "*Mendelian factors underlying quantitative traits in tomato: comparison across species, generations and environments"*. Genetics, **127**, pp. 181-197, 1991.
12. J. M. Thoday. "*Effects of disruptive selection. III. Coupling and repulsion"*. Heredity, **14**, pp. 35-49, 1960.