International Journal of Statistical Sciences *ISSN 1683-5603*
Vol. 20(1), 2020, pp 61-78
© 2020 Dept. of Statistics, Univ. of Rajshahi, Bangladesh

Robust QTL Analysis Based on Robust Estimation of Bivariate Normal Distribution with Backcross Population

Md. Jahangir Alam[1*], Md. Ripter Hossain[1], S. M. Shahinul Islam[2] and Md. Nurul Haque Mollah[1*]

[1]Bioinformatics Laboratory, Department of Statistics,
University of Rajshahi, Rajshahi-6205, Bangladesh

[2]Institute of Biological Science, University of Rajshahi,
Rajshahi-6205, Bangladesh

[*]Correspondence should be addressed to Md. Jahangir Alam and
Md. Nurul Haque Mollah
(jahangir_statru63@yahoo.com) and (mollah.stat.bio@ru.ac.bd)

[Received Jan. 8, 2020; Revised February 10, 2020; Accepted March 3, 2020]

Abstract

Simple interval mapping (SIM) is one of the most popular approaches for genome-wide single quantitative trait locus (QTL) analysis. Maximum likelihood (ML) and least squares (LS) regressions are widely used methods for SIM. However, these approaches are very complex and time-consuming in terms of statistical computation. In this study, we have introduced a new approach for single-trait QTL analysis using the properties of bivariate normal distribution (BND) with the backcross population. In this approach, statistical calculations are very straight forward because the calculations depend on only the sample means, sample variances and sample covariances. In spite of computational simplicity, our proposed classical method is very sensitive to phenotypic outliers like other existing methods and it provides misleading results in presence of phenotypic contaminations. To overcome this problem, we have developed a new robust approach of SIM for single-trait QTL analysis by robustifying our proposed classical BND based SIM approach using the minimum β–divergence method. The proposed robust method reduces to the proposed classical SIM approach when the tuning parameter $\beta = 0$. Simulation study and real data analysis show that our proposed classical method shows almost the same performance as the existing classical methods of SIM in all cases and our proposed robust approach outperforms over the classical SIM approaches in presence of outliers.

Also, in absence of outliers, the proposed robust approach shows almost the same performance as the classical SIM approaches.

Keywords: Simple interval mapping, maximum likelihood, β-likelihood function, bivariate normal distribution, LOD statistic.

AMS Classification: 62F35.

1. Introduction

Recent advancements in biotechnology, particularly in molecular marker technology, have expedited the availability of large fine-scaled genetic markers data which facilitate the genome-wide quantitative trait locus (QTL) analysis in the genetic study for identifying the important genes which control specific quantitative trait. The idea of using two flanking markers bracketing a region for testing QTLs was first proposed by Thoday (1961). Lander and Botstein (1989) proposed a much improved approach based on the maximum likelihood (ML) estimation method, which uses the linkage relationship between the flanking markers and a QTL for identifying important QTLs. This method is called simple interval mapping (SIM) for QTL mapping. Similar to Lander and Botstein (1989), a linear regression based SIM approach was proposed byHaley and Knott (1992), which uses the least squares(LS) method for estimation of the model parameters. This LS regression-based SIM is also well known as HK regression-based interval mapping to the biologists. Kao (2000)investigated the differences between ML and (LS) based simple interval mapping for QTL analysis analytically and numerically.Liu (1997),Wu et al. (2007),Weller (2009),Rifkin (2012),Xu (2013),Chen (2016) and Caballero (2020) discussed different methods for single-trait QTL analysis in their textbooks.

The existing SIM approaches based on the ML method(Lander and Botstein, 1989) and the LS method (Haley and Knott, 1992) are the two most popular and widely used methods for single-trait QTL analysis. Substantial work has been done in single-trait QTL analysis using ML and LS based SIM methods (Boopathi, 2020; Broman, 2001; Broman and Sen, 2009; Churchill and Doerge, 1994; Doerge, 2002; Huang et al., 2020; Jansen, 1993; Knott, 2005; Kwak et al., 2014; Moser et al., 1998; Ngwako, 2008; Nobari et al., 2012; Sharma et al., 2019; Singh et al., 2018). The main limitation of ML based SIM is that its calculations are very complex and it is very time-consuming because it uses the expectation-maximization (EM) algorithm. Although LS based SIM takes less time than ML based SIM, its computations are also complex because parameter estimation depends on the least squares method, and the calculation of test statistic needs calculation of residuals and residual variance. In this study, we have developed a new approach of single-trait QTL analysis using the properties of bivariate normal distribution (BND) with the backcross (BC) population. In this approach, the parameter estimation and calculation of test statistic are very straight forward because the calculations depend only on the sample means, sample variances and sample covariances of phenotype and the conditional probability of QTL genotype given the flanking marker genotypes. Although our proposed BND based SIM are very useful methods for QTL analysis, it is very sensitive to phenotypic contaminations and provides misleading results when the phenotypic data are contaminated by outliers.

To overcome this problem of phenotypic contaminations, we have also developed a new robust approach of SIM for single-trait QTL analysis with BC population by robustifying our proposed classical BND based SIM using minimum *β*–divergence method. We have performed a simulation study to investigate the performance of the proposed methods in comparison with the existing methods of SIM for QTL analysis with BC population. Although we have developed our proposed methods for BC population, these methods can easily be extended for other populations, such as double haploid (DH) and intercross (F_2) population, with some simple modification.

2. Materials and Methodology

2.1 Regression based SIM for single-trait QTL analysis using the properties of MND (Proposed1)

Let us consider no epistasis between two QTLs, no interference in crossing over, and only one QTL in the testing interval for a BC population. Then for testing a QTL within a marker interval, the linear regression model for BC population is as follows:

$$y_j = \alpha + \gamma x_{j|i} + \varepsilon_j, i = 1, 2 \text{ and } j = 1, 2, \ldots, n \quad (1)$$

where y_j is the phenotypic value of the j^{th} individual, α is the general mean effect, $\gamma$is the QTL effect, $x_{j|i} = p_{j|i} = x_j$ is the conditional probability of the putative QTL genotypes given the flanking marker genotypes for the j^{th} individual (see Table 1) and ε_j~$NID(0, \sigma^2)$ is a random error.

Let the flanking markers of the QTL testing interval are denoted by $\mathbf{M}_{\mathrm{L}}$ (left marker)with alleles M_L and m_L, and $\mathbf{M}_{\mathrm{R}}$ (right marker) with alleles $M_R$andm_R. Suppose that the locus of the unobserved putative QTL located within the testing interval bracketed by the flanking marker $\mathbf{M}_{\mathrm{L}}$ and $\mathbf{M}_{\mathrm{R}}$ is denoted by **Q** with alleles Q and q. 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 the BC population. The recombination fraction between the two markers is denoted by r. The possibility of the event of double recombination within the interval of two flanking markers is ignored.

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

Marker Genotypes	Expected Frequency	QTL Genotypes	
		$QQ(p_{j|1})$	$Qq(p_{j|2})$
M_LM_R/M_LM_R	$(1-r)/2$	1	0
M_LM_R/M_Lm_R	$r/2$	$(1-p^*)$	p
M_LM_R/m_LM_R	$r/2$	p	$(1-p)$
M_LM_R/m_Ln_R	$(1-r)/2$	0	1

$^*p = r_{\mathbf{M_LQ}}/r_{\mathbf{M_LM_R}}$, where $r_{\mathbf{M_LQ}}$ is the recombination fraction between the left marker $\mathbf{M}_{\mathrm{L}}$ and the putative QTL **Q**, and $r_{\mathbf{M_LM_R}}$ is the recombination fraction between two flanking markers $\mathbf{M}_{\mathrm{L}}$ and $\mathbf{M}_{\mathrm{R}}$.

We want to test the null hypothesis $H_0: \gamma = 0$ (i.e., there is no QTL at a given position within a marker interval) against $H_1: H_0$ is not true.Under the null hypothesis (H_0), the model (1) reduces to the following model

$$y_j = \alpha + \varepsilon_j,\ j = 1, 2,\ \dots, n \tag{2}$$

In order to estimate the model parameters and the variance of the random error, let us consider that $\boldsymbol{Z} = (Y, X)$ follows a bivariate normal distribution

$N\left(\underset{(2\times1)}{\boldsymbol{\mu}_{\boldsymbol{Z}}}, \underset{(2\times2)}{\boldsymbol{\Sigma}_{\boldsymbol{Z}}}\right)$ with mean vector $\boldsymbol{\mu}_{\boldsymbol{Z}}$and covariance matrix$\boldsymbol{\Sigma}_{\boldsymbol{Z}}$, where Y and X are introduced in (1).Then the probability density function for Z can be written as

$$f(\boldsymbol{Z}) = \frac{1}{(2\pi)|\boldsymbol{\Sigma}_{\boldsymbol{Z}}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{Z} - \boldsymbol{\mu}_{\boldsymbol{Z}})^T \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}(\boldsymbol{Z} - \boldsymbol{\mu}_{\boldsymbol{Z}})\right] \tag{3}$$

We can partition the mean vector $\boldsymbol{\mu}_{\boldsymbol{Z}}$ as $\boldsymbol{\mu}_{\boldsymbol{Z}} = [\mu_Y \quad \mu_X]^{\boldsymbol{T}}$ and the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{Z}}$ as $\boldsymbol{\Sigma}_{\boldsymbol{Z}} = \begin{bmatrix} \sigma_Y^2 & \sigma_{XY} \\ \sigma_{YX} & \sigma_X^2 \end{bmatrix}$, where μ_Y = population mean of Y , μ_X = population mean of X , $\sigma_X^2 = E[(X - \mu_X)^2]$ = population variance of X , $\sigma_Y^2 = E[(Y - \mu_Y)^2]$ = population variance of and $\sigma_{XY} = \sigma_{YX} = E[(X - \mu_X)(Y - \mu_Y)]$ = population covariance between X and Y.

Then the conditional mean of Y given X is obtained as

$$E(Y|X = x) = \mu_Y + \sigma_{YX}\sigma_X^{-2}(X - \mu_X) \tag{4}$$

Equation (4) can be expressed as

$$\begin{aligned} (Y|X = x) &= \mu_Y + \sigma_{YX}\sigma_X^{-2}X - \sigma_{YX}\sigma_X^{-2}\mu_X \\ &= (\mu_Y - \sigma_{YX}\sigma_X^{-2}\mu_X) + (\sigma_{YX}\sigma_X^{-2})X \\ &= \alpha + \gamma X \end{aligned} \tag{5}$$

which is known as simple linear regression surface of Y on X, where $\alpha = (\mu_Y - \sigma_{YX}\sigma_X^{-2}\mu_X)$ is the general mean effect and the (m×1) vector$\gamma = (\sigma_{YX}\sigma_X^{-2})$ is called the regression coefficient. For BC population γ is the additive QTL effects.

Using (4), the prediction error can be written as

$$\varepsilon = Y - E(Y|X) = Y - \mu_Y - \sigma_{YX}\sigma_X^{-2}(X - \mu_X) \tag{6}$$

Now, the variance of the prediction error is

$$\sigma^2 = V(\varepsilon) = E[\{\varepsilon - E(\varepsilon)\}^2] = E[\varepsilon^2], \text{ since } E(\varepsilon) = 0 \tag{7}$$

Using (6) in (7), we can write

$$\begin{aligned}\sigma^2 &= E[\{Y - \mu_Y - \sigma_{YX}\sigma_X^{-2}(X - \mu_X)\}^2] \\ &= E[(Y - \mu_Y)^2 - 2(Y - \mu_Y)\{\sigma_{YX}\sigma_X^{-2}(X - \mu_X)\} \\ &\qquad + \{\sigma_{YX}\sigma_X^{-2}(X - \mu_X)\}^2] \\ &= E[(Y - \mu_Y)^2] - 2\sigma_{YX}\sigma_X^{-2}E[(Y - \mu_Y)(X - \mu_X)] \\ &\qquad + \sigma_{YX}^2\sigma_X^{-4}E[(X - \mu_X)^2] \\ &= \sigma_Y^2 - 2\sigma_{YX}\sigma_X^{-2}\sigma_{YX} + \sigma_{YX}^2\sigma_X^{-4}\sigma_X^2 \\ &= \sigma_Y^2 - 2\sigma_{YX}^2\sigma_X^{-2} + \sigma_{YX}^2\sigma_X^{-2} \\ &= \sigma_Y^2 - \sigma_{YX}^2\sigma_X^{-2} \end{aligned} \tag{8}$$

Because $\boldsymbol{\mu_Z}$and $\boldsymbol{\Sigma_Z}$are typically unknown, they must be estimated from a random sample in order to construct the multivariate linear predictor and determine expected prediction errors.

Based on a random sample of size *n*, the maximum likelihood estimator of the $\boldsymbol{\mu_Z}$and $\boldsymbol{\Sigma_Z}$ are given by

$$\boldsymbol{\hat{\mu}_Z} = \begin{bmatrix} \hat{\mu}_Y \\ \hat{\mu}_X \end{bmatrix} = \begin{bmatrix} \bar{Y} \\ \bar{X} \end{bmatrix} \text{and} \boldsymbol{\hat{\Sigma}_Z} = \begin{bmatrix} \hat{\sigma}_Y^2 & \hat{\sigma}_{YX} \\ \hat{\sigma}_{XY} & \hat{\sigma}_X^2 \end{bmatrix} = \left(\frac{n-1}{n}\right)\begin{bmatrix} S_Y^2 & S_{YX} \\ S_{XY} & S_X^2 \end{bmatrix} \tag{9}$$

where $\bar{X} = \frac{1}{n}\sum_{j=1}^{n} x_j$, $\bar{Y} = \frac{1}{n}\sum_{j=1}^{n} y_j$, $S_Y^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(y_j - \bar{Y}\right)^2$, $S_{XY} = S_{YX} = \frac{1}{n-1}\sum_{j=1}^{n}\left(y_j - \bar{Y}\right)\left(x_j - \bar{X}\right)andS_X^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_j - \bar{X}\right)^2$.

Hence, based on a random sample of size *n*, we can get the maximum likelihood estimators of the regression parameters α and γ, and the error variance σ^2.

Using (9) into (5), we can write

$$\hat{\alpha} = (\hat{\mu}_Y - \hat{\sigma}_{YX}\hat{\sigma}_X^{-2}\hat{\mu}_X) = \bar{Y} - S_{YX}S_X^{-2}\bar{X} \tag{10}$$

and

$$\hat{\gamma} = (\hat{\sigma}_{YX}\hat{\sigma}_X^{-2}) = S_{YX}S_X^{-2} \tag{11}$$

Therefore, using (9) in (4), the maximum likelihood estimator of the regression function is

$$\hat{Y} = \hat{\alpha} + \hat{\gamma}X = \bar{Y} - S_{YX}S_X^{-2}\bar{X} + S_{YX}S_X^{-2}X = \bar{Y} + S_{YX}S_X^{-2}(X - \bar{X}) \tag{12}$$

Based on a random sample of size *n*, using (9) in (8), the maximum likelihood estimators of σ^2 under the full model and the reduced model are, respectively,

$$\hat{\sigma}^2 = \hat{\sigma}_Y^2 - \hat{\sigma}_{YX}^2\hat{\sigma}_X^{-2} = \left(\frac{n-1}{n}\right)(S_Y^2 - S_{YX}^2S_X^{-2}) \tag{13}$$

and

$$\hat{\sigma}_0^2 = \hat{\sigma}_Y^2 = \left(\frac{n-1}{n}\right)S_Y^2 \tag{14}$$

Let $L_1(\alpha, \gamma, \sigma^2)$ is the likelihood function under the full model (1) and $L_0(\alpha, \sigma^2)$ is the likelihood function under the reduced model (2). To test H_0againstH_1, the likelihood ratio test (LRT) statistic is defined as

$$\begin{aligned} \text{LRT} &= -2\ln\left[\frac{\max\limits_{\alpha,\sigma^2} L_0(\alpha, \sigma^2)}{\max\limits_{\alpha,\gamma,\sigma^2} L_1(\alpha, \gamma, \sigma^2)}\right] \\ &= -2\ln\left[\frac{L_0(\hat{\alpha}_0, \hat{\sigma}_0^2)}{L_1(\hat{\alpha}, \hat{\gamma}, \hat{\sigma}^2)}\right] = -n\ln\left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right) \end{aligned} \tag{15}$$

where $\hat{\alpha}$, $\hat{\gamma}$ and $\hat{\sigma}^2$ are the maximum likelihood (ML) estimates of the parameters α, γ and σ^2 under the full model (1), and $\hat{\alpha}_0$ and $\hat{\sigma}_0^2$ are the ML estimates of the parameters α and σ^2 under the reduced model (i.e., under H_0).

Under the null hypothesis (H_0), the LRT statistic in (15) is expected to have an approximate chi-square distribution with 1 degree of freedom for a given QTL position in the genome. However, the threshold value to reject the null hypothesis (H_0) cannot be simply chosen from the χ^2 distribution because of the violation of regularity conditions of the asymptotic theory under H_0.An alternative way is to use the log of odds (LOD) score (Lander and Botstein, 1989; Ott, 1999; Terwilliger and Ott, 1994; Wu et al., 2007; Xu, 2013) as a test statistic to test the null hypothesis of no QTL (H_0). The LOD score is the transformation of the LRT statistic, defined as

$$\text{LOD} = \frac{\text{LRT}}{2\times \log(10)} = \frac{\text{LRT}}{4.605} = 0.217\ \text{LRT} \tag{16}$$

According toLander and Botstein (1989), the typical threshold of LOD score should be between 2 and 3 to ensure a 5% overall false positive error for identifying a QTL. Terwilliger and Ott (1994), Ott (1999); Wu et al. (2007), and Xu (2013)suggested a value of LOD = 3 as the critical threshold for declaring the existence of QTL. Thus, the LOD > 3 can be used as a criterion to declare a significant QTL.

2.2 Robust SIM for single-trait QTL analysis using robust bivariate normal distribution (Proposed2)

All the approaches discussed in previous sections are very sensitive to phenotypic outliers and produce misleading results in presence of outliers. So, we need some robust approach that produces similar results in absence of outliers and performs better in presence of outliers being less sensitive to outliers. We observe that the estimates in (9) – (15) are very sensitive to outliers and give misleading results in presence of outliers. In this section, we have discussed the robustification of the estimates in (9) – (15) using β–divergence method (Mihoko and Eguchi, 2002; Mollah et al., 2007) to obtain the robust estimates of model parameters and the robust test statistics (LRT and LOD). From (9) – (15) we observe that if we can robustify the sample means, sample variances and sample covariance, then we can obtain the robust estimates of the model parameters and the test statistics (LRT and LOD).

According to (Mihoko and Eguchi, 2002; Mollah et al., 2007), the β-divergence between two probability density functions $p(\boldsymbol{z})$ and $q(\boldsymbol{z})$ is defined by

$$D_\beta(p,q) = \int \left[\frac{1}{\beta}\{p^\beta(z) - q^\beta(\boldsymbol{z})\}p(\boldsymbol{z}) - \frac{1}{\beta+1}\left\{p^{\beta+1}(\boldsymbol{z}) - q^{\beta+1}(\boldsymbol{z})\right\}\right] d\boldsymbol{z},\ for\ \beta > 0 \qquad (17)$$

which is non-negative, that is $D_\beta(p,q) \geq 0$, equality holds iff $p = q$.

The minimum β-divergence estimators of the parameters $\boldsymbol{\theta} = (\boldsymbol{\mu_Z}, \boldsymbol{\Sigma_Z})$ can be obtained by the iterative solution of the following equations:

$$\boldsymbol{\mu}_{\boldsymbol{z},t+1} = \frac{\sum_{j=1}^{n} w_\beta(\boldsymbol{z}_j|\boldsymbol{\theta}_t)\boldsymbol{z}_j}{\sum_{j=1}^{n} w_\beta(\boldsymbol{z}_j|\boldsymbol{\theta}_t)} \qquad (18)$$

and

$$\boldsymbol{\Sigma}_{\boldsymbol{Z},t+1} = (1+\beta)\frac{\sum_{j=1}^{n} w_\beta(\boldsymbol{z}_j|\boldsymbol{\theta}_t)(\boldsymbol{z}_j - \boldsymbol{\mu}_{Z,t})(\boldsymbol{z}_j - \boldsymbol{\mu}_{Z,t})^T}{\sum_{j=1}^{n} w_\beta(\boldsymbol{z}_j|\boldsymbol{\theta}_t)} \qquad (19)$$

where $w_\beta\left(\boldsymbol{z}_j\middle|\boldsymbol{\theta}_t\right), j = 1, 2, \dots, n$, is called the β-weight function and defined as

$$w_\beta\left(\boldsymbol{z}_j\middle|\boldsymbol{\theta}_t\right) = \exp\left[-\frac{\beta}{2}(\boldsymbol{z}_j - \boldsymbol{\mu}_{Z,t})^T \boldsymbol{\Sigma}_{Z,t}^{-1}(\boldsymbol{z}_j - \boldsymbol{\mu}_{Z,t})\right].$$

If β→0, then (18) and (19) reduces to the classical non-iterative solution.

Let the robust estimates (i.e., β-estimates) of $\boldsymbol{\mu_Z}$ and $\boldsymbol{\Sigma_Z}$ are denote by $\widehat{\boldsymbol{\mu}}_{\boldsymbol{Z}(\beta)}$and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}(\beta)}$. Then we can write

$$\widehat{\boldsymbol{\mu}}_{\boldsymbol{Z}(\beta)} = \begin{bmatrix} \hat{\mu}_{Y(\beta)} \\ \hat{\mu}_{X(\beta)} \end{bmatrix} \text{and } \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}(\beta)} = \begin{bmatrix} \hat{\sigma}_{Y(\beta)}^2 & \hat{\sigma}_{YX(\beta)} \\ \hat{\sigma}_{XY(\beta)} & \hat{\sigma}_{X(\beta)}^2 \end{bmatrix} \qquad (20)$$

Then the robust estimates of the regression parameters can be written as

$$\hat{\alpha}_{(\beta)} = \left(\hat{\mu}_{Y(\beta)} - \hat{\sigma}_{YX(\beta)}\hat{\sigma}^{-2}_{X(\beta)}\hat{\mu}_{X(\beta)}\right) \tag{21}$$

and

$$\hat{\gamma}_{(\beta)} = \left(\hat{\sigma}_{YX(\beta)}\hat{\sigma}^{-2}_{X(\beta)}\right) \tag{22}$$

Now, the robust estimates of σ^2 under the full model and the reduced model are, respectively,

$$\hat{\sigma}^2_{(\beta)} = \hat{\sigma}^2_{Y(\beta)} - \hat{\sigma}^2_{YX(\beta)}\hat{\sigma}^{-2}_{X(\beta)} \tag{23}$$

and

$$\hat{\sigma}^2_{0(\beta)} = \hat{\sigma}^2_{Y(\beta)} \tag{24}$$

Then we get the robust LRT statistic as follows:

$$\mathrm{LRT}_{(\beta)} = -n \ln\left(\frac{\hat{\sigma}^2_{(\beta)}}{\hat{\sigma}^2_{0(\beta)}}\right) \tag{25}$$

The modified LRT statistic has an approximate χ^2-distribution with 1 degree of freedom. Then the robust LOD statistic can be written as

$$\mathrm{LOD}_{(\beta)} = \frac{\mathrm{LRT}_{(\beta)}}{2\times\log(10)} = \frac{\mathrm{LRT}_{(\beta)}}{4.605} = 0.217\ \mathrm{LRT}_{(\beta)} \tag{26}$$

We have developed the proposed method for BC population. However, methods for other mapping populations, such as F_2 and double haploid (DH), are the simple extension of that for the BC population with some modifications.

3. Result and Discussions

3.1 Simulation Results

To measure the performance of the proposed methods (Proposed1: Classical BND and Proposed2: Robust BND) in comparison of the maximum likelihood (ML) and least squares (LS) methods of SIM for QTL mapping with BC population, we have generated phenotypic and genotypic data with BC population using simulation technique. We have considered three unlinked QTL sacross ten chromosomes and 11 equally spaced markers in each of the ten chromosomes, where any two successive marker interval size is 5 cM. The true QTL positions are located on chromosomes 2, 3 and 5 at marker 5 (locus position 20 cM). The true values of the parameters in the model are assumed as $\alpha = 0.5, \gamma = 0.8$ and $\sigma^2 = 0.25$. We have generated 300 trait values with heritability $h^2 = 0.39$ which means that 39% of the trait variation is controlled by QTL and the remaining 61% is subject to the environmental effects (random error). To investigate the robustness of the Proposed2 (robust BND) method in a comparison of the ML, LS and Proposed1 (classical BND) methods, we have contaminated 12% of the trait values (i.e., phenotypic 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/).

Table 2 shows QTL positions (i.e., chromosome, marker and locus position) identified by the ML, LS, Proposed1 and Proposed2 methods in presence and absence of outliers. Figure 1(a) and Figure 1(b) are representing the scatter plots of 300 trait values in presence and absence of outliers, respectively. Then we computed LOD scores based on the ML, LS, Proposed1 and Proposed2 methods for both types of data sets (uncontaminated and contaminated). Figure 1(c) and Figure 1(d) are showing the LOD scores profile plots for the uncontaminated and contaminated datasets, respectively. In the LOD scores profile plots, the dotted (red colour), two dash (green colour), dot dash (blue colour) and solid (black colour) lines represent the LOD scores at every 1cM position in the chromosomes for the ML, LS, classical BND (Proposed1) and robust BND (Proposed2) method of SIM, respectively, with $\beta = 0.2$. We have selected the best value of the tuning parameter β by cross-validation.

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

Method	**True QTL position**	**Identified QTL position**	
		In absence of outliers	**In presence of outliers**
ML	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.	ML method fails to identify any QTL on any chromosome.
LS	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.	LS method fails to identify any QTL on any chromosome.
Proposed1 (Classical BND)	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 3 at marker 8 (locus position 35 cM) (ii) On chromosome 5 at marker 5 (locus position 20 cM) (iii) On chromosome 6 at marker 2 (locus position 5 cM) (iv) On chromosome 8 at marker 3 (locus position 10 cM)
Proposed2 (Robust BND)	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 Figure 1 it is seen that the highest LOD score peak occurs at the true QTL position on the true chromosome 2, 3 and 5 at marker 5 (locus position 20 cM) for all four methods for the uncontaminated dataset (Figure 1(c)). However, in presence of outliers, the highest LOD score peak occurs at the true QTL positions on true chromosomes for the Proposed2 (robust BND) method only (Figure 1(d)). That is, from Table 2 and Figure 1 we observe that all of the four methods (ML, LS, Proposed1 and Proposed2) identify the true QTL positions correctly in absence of outliers. But in presence of outliers, the ML and LS fail to identify any significant QTL position, and the Proposed1 (classical BND) method identify QTLs on chromosomes 3 at marker 8 (locus position 35 cM), on chromosome 5 at marker 5 (locus position 20 cM), on chromosome 6 at marker 2 (locus position 5 cM) and on chromosome 8 at marker 3 (locus position 10 cM). In presence of outliers, only the position on chromosome 5 at marker 5 identified by the classical BND method is the true QTL position, and all other positions identified by the classical BND are not the true QTL positions. However, in presence of outliers, the Proposed2 (robust BND) method has identified the QTLs on chromosome 2, 3 and 5 at marker 5 (locus position 20 cM) which are the true QTL positions.

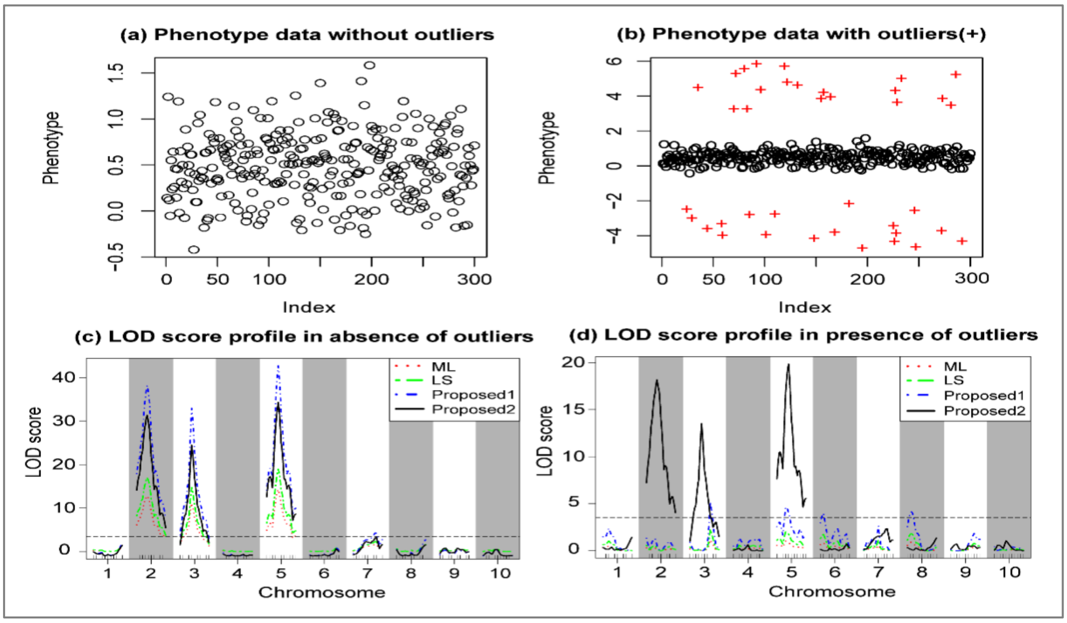

Figure 1: Simulated phenotypic observations in (a) absence and (b) presence of 12% outliers, and LOD score profile in (c) absence and (d) in presence of 12% outliers.**Proposed1:** Classical BND based method of SIM. **Proposed2:** Robust BND based method of SIM.

Hence, in presence of outliers, the classical methods of SIM (ML, LS and classical BND) fail to identify all the true QTL positions whereas the Proposed2 (robust BND) method successfully identifies all the true QTL positions. Also in absence of outliers, the Proposed2 method is working as the classical methods of SIM for single-trait QTL analysis.

3.2 Real Data Analysis Results

To investigate the performance of the proposed methods for real data analysis in a comparison of traditional ML and LS methods, we have considered the hypertension dataset of Sugiyama et al. (2001) which is available in R/qtl package (Broman et al., 2003), homepage: http://www.rqtl.org. This dataset was analyzed to investigate the genetic control of salt-induced hypertension on male mice from a reciprocal backcross between the salt-sensitive c57BL/6J and the non-salt-sensitive A/J (A) inbred mouse strains.

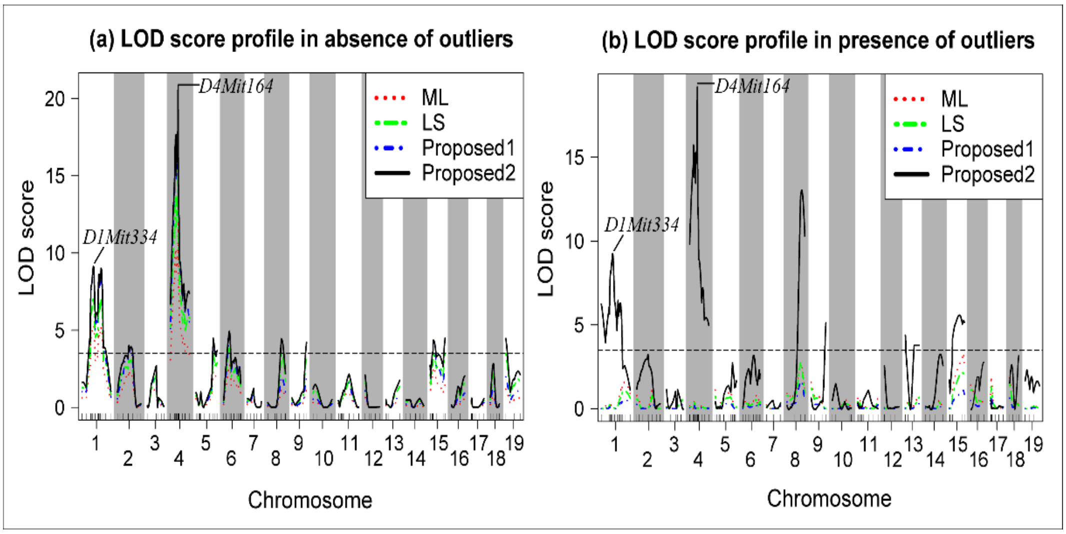

Figure 2: LOD score profile plots with hypertension using ML, LS, Proposed1 (classical BND) and Proposed2 (robust BND). (a) LOD score plot in absence of phenotypic outliers and (b) in presence of 12% phenotypic outliers.

Figure 2 represents the LOD score profile plots in absence and presence of outliers. Figure 2(a) shows the LOD scores profile in absence of outliers, where dotted (red colour), two dash (green colour), dot dash (blue colour) and solid (black colour) lines represents the LOD scores at every 1cM position on the chromosomes for the ML, LS, Proposed1 (classical BND) and Proposed2 (robust BND) method, respectively. Figure 2(b) shows the LOD scores profile for the contaminated dataset, where the LOD scores at every 1cM position on the chromosomes for the ML, LS, Proposed1 (classical BND) and Proposed2 (robust BND) methods are presented by the same line styles and colours as Figure 2(a).For the Proposed2 method with contaminated data, we have selected the best value of the tuning parameter β as β = 0.2by cross-validation (for details see Mollah et al., 2007).

Figure 2(a) shows that two QTLs on chromosome 1 (QTL/marker: *D1Mit334*) and chromosome 4 (QTL/marker: *D4Mit164*) are statistically significant genome-wide, and one QTL on each of chromosomes 2 (QTL/marker: *D2Mit62*), 6 (QTL/marker: *D6Mit8*), 8 (QTL/marker: *D8Mit271*) and 15 (QTL/marker: *D15Mit152*) are suggestive to be important for controlling blood pressure genome-wide by all four methods for the uncontaminated real dataset. In presence of outliers, almost similar results are obtained by the robust BND (Proposed2) method only as shown in Figure 2(b) whereas all the classical methods fail to identify the same QTL positions as identified in absence of outliers. Therefore, the Proposed2 method (robust BND) significantly outperforms over the traditional ML and LS method as well as the Proposed1 (classical BND) method in presence of outliers. Otherwise, it shows equal performance.

Sugiyama et al. (2001) found that the QTL *D1Mit334* on chromosome 1 and the QTL *D4Mit164* on chromosome 4 were significantly associated with hypertension in mouse which supports our findings by the proposed methods (Proposed1 and Proposed2) in absence of outliers and by the Proposed2 method in presence of outliers. They also suggested the QTLs *D6Mit15* and *D15Mit152* on chromosomes 6 and 15, respectively, as important QTLs for affecting blood pressure which are similar to our suggestive QTLs responsible for hypertension based on our proposed methods.

4. Conclusion

In this paper, first, we have introduced a new approach of SIM (Proposed1) using the properties of BND. Then a new robust approach of SIM (Proposed2) for QTL analysis has been developed by robustifying the classical BND based SIM approach (Proposed1) using maximum β-likelihood estimation with BC population. The value of the tuning parameter β plays a key role in the performance of the Proposed2 method. An appropriate value for the tuning parameter β can be selected by cross-validation. The Proposed2 method with tuning parameter β = 0 reduces to the traditional interval mapping approach. Simulation and real data analysis results show that the Proposed1 (classical BND) method exhibits almost the same performance as the traditional ML and LS methods of SIM in all cases (presence and absence of outliers).However, simulation and real data analysis results reveal that the Proposed2 (robust BND) method significantly improves the performance over the classical interval mapping approaches in presence of phenotypic outliers.

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

Reference

[1] Boopathi, N. M. (2020). QTL Analysis. In Genetic Mapping and Marker Assisted Selection: Basics, Practice and Benefits (pp. 253-326). Singapore: Springer Singapore.

[2] Broman, K. (2001). Review of statistical methods for QTL mapping in experimental crosses. Lab animal, 30(7), 44-52.

[3] Broman, K. W., and Sen, S. (2009). A Guide to QTL Mapping with R/qtl (Vol. 46): Springer.

[4] Broman, K. W., Wu, H., Sen, Ś., and Churchill, G. A. (2003). R/qtl: QTL mapping in experimental crosses. Bioinformatics, 19(7), 889-890.

[5] Caballero, A. (2020). Quantitative Genetics: Cambridge University Press.

[6] Chen, Z. (2016). Statistical methods for QTL mapping: Chapman and Hall/CRC.

[7] Churchill, G. A., and Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. Genetics, 138(3), 963-971.

[8] Doerge, R. W. (2002). Mapping and analysis of quantitative trait loci in experimental populations. Nature Reviews Genetics, 3(1), 43-52.

[9] Haley, C. S., and Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity, 69(4), 315-324.

[10] Huang, F., Chen, Z., Du, D., Guan, P., Chai, L., Guo, W., . . . Ni, Z. (2020). Genome-wide linkage mapping of QTL for root hair length in a Chinese common wheat population. The Crop Journal. doi:https://doi.org/10.1016/j.cj.2020.02.007

[11] Jansen, R. C. (1993). Interval mapping of multiple quantitative trait loci. Genetics, 135(1), 205-211.

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

[13] Knott, S. A. (2005). Regression-based quantitative trait loci mapping: robust, efficient and effective. Philosophical Transactions of the Royal Society B: Biological Sciences, 360(1459), 1435-1442. doi:doi:10.1098/rstb.2005.1671

[14] Kwak, I.-Y., Moore, C. R., Spalding, E. P., and Broman, K. W. (2014). A simple regression-based method to map quantitative trait loci underlying function-valued phenotypes. Genetics, 197(4), 1409-1416. doi:10.1534/genetics.114.166306

[15] Lander, E. S., and Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics, 121(1), 185-199.

[16] Liu, B. H. (1997). Statistical Genomics: Linkage, Mapping, and QTL Analysis: CRC Press.

[17] Mihoko, M., and Eguchi, S. (2002). Robust blind source separation by beta divergence. Neural computation, 14(8), 1859-1886. doi:https://doi.org/10.1162/089976602760128045

[18] Mollah, M. N. H., Eguchi, S., and Minami, M. (2007). Robust prewhitening for ICA by minimizing β-divergence and its application to FastICA. Neural Processing Letters, 25(2), 91-110.

[19] Moser, G., Mueller, E., Beeckmann, P., Yue, G., and Geldermann, H. (1998). Mapping of QTLs in F2 generations of Wild Boar, Pietrain and Meishan pigs.Paper presented at the Proceedings of the 6th World Congress on Genetics Applied to Livestock Production.

[20] Ngwako, S. (2008). Mapping quantitative trait loci using marker regression and interval mapping methods. Pakistan J Biol Sci, 11, 553-558.

[21] Nobari, K., Nassiry, M. R., Aslaminejad, A. A., Tahmoorespur, M., and Esmailizadeh, A. K. (2012). Effects of QTL parameters and marker density on efficiency of Haley–Knott regression interval mapping of QTL with complex traits and use of artificial neural network for prediction of the efficiency of HK method in livestock. Journal of Applied Animal Research, 40(3), 247-255. doi:10.1080/09712119.2012.667647

[22] Ott, J. (1999). Analysis of human genetic linkage (3rd ed.). Baltimore, Maryland: Johns Hopkins University Press.

[23] Rifkin, S. A. (2012). Quantitative Trait Loci (QTL): Methods and Protocols: Humana Press.

[24] Sharma, U., Banerjee, P., Joshi, J., Kapoor, P., and Vijh, R. K. (2019). Identification of QTLs for low somatic cell count in Murrah buffaloes. Indian Journal of Animal Sciences, 89(7), 54-63.

[25] Singh, G., Kuzniar, A., van Mulligen, E. M., Gavai, A., Bachem, C. W., Visser, R. G. F., and Finkers, R. (2018). QTLTableMiner++: semantic mining of QTL tables in scientific articles. BMC Bioinformatics, 19(1), 183. doi:10.1186/s12859-018-2165-7

[26] Sugiyama, F., Churchill, G. A., Higgins, D. C., Johns, C., Makaritsis, K. P., Gavras, H., and Paigen, B. (2001). Concordance of murine quantitative trait loci for salt-induced hypertension with rat and human loci. Genomics, 71(1), 70-77.

[27] Terwilliger, J. D., and Ott, J. (1994). Handbook of human genetic linkage: JHU Press.

[28] Thoday, J. (1961). Location of polygenes. Nature, 191, 368-370.

[29] Weller, J. I. (2009). Quantitative Trait Loci Analysis in Animals (2nd ed.): CABI.

[30] Wu, R., Ma, C., and Casella, G. (2007). Statistical genetics of quantitative traits: linkage, maps and QTL: Springer Science & Business Media.

[31] Xu, S. (2013). Principles of statistical genomics: Springer.