International Journal of Statistical Sciences *ISSN 1683-5603*
Vol. 16, 2017, pp 115-126

Composite Interval Mapping for Identifying QTLs in Presence and Absence of Phenotypic Outliers

Md. Mamun Monir[1,2*], Mita Khatun[2], Md. Asif Ahsan[1,3], Md. Matiur Rahaman[1,3], S M Sayadat Amin[1,4], Munirul Alam[5], and Md. Nurul Haque Mollah[1*]

[1]Department of Statistics, Rajshahi University, Rajshahi, Bangladesh.
[2]Institute of Bioinformatics, College of Agriculture and Biotechnology, Zhejiang University, China.
[3]Department of Bioinformatics, College of Life Science, Zhejiang University, China.
[4]Bangladesh Medical Research Council (BMRC), Dhaka, Bangladesh.
[5]International Centre for Diarrhoeal Disease Research, Bangladesh (ICDDR, B), Dhaka, Bangladesh.

[*]Correspondence should be addressed to Md. Nurul Haque Mollah
(Email: mollah.stat.bio@ru.ac.bd)

[Received March 11, 2017; Accepted October 10, 2017]

Abstract

Simple interval mapping approaches have been playing the significant role for QTL (quantitative trait loci) analysis. However, these approaches cannot detect the multiple linked QTLs. Composite Interval Mapping (CIM) is one of the most popular approaches for identifying multiple linked QTLs even if they are located nearby at the same chromosome. However, classical CIM approach fails to identify QTLs in presence of phenotypic outliers. Therefore, in this study, an attempt was made to investigate performance of robust CIM approach, which was developed by maximizing β-likelihood function for robust QTL analysis. The performance of the robust CIM method depends on the value of tuning parameter β. It reduces to the traditional CIM when $\beta \to 0$. We compared the robust CIM method with classical CIM and interval mapping approaches for identifying multiple linked QTLs by extensive simulation study. It was observed that simple interval mapping approaches cannot detect the multiple linked QTLs, whereas CIM can detect multiple linked QTLs correctly in absence of phenotypic outliers. However, the robust CIM method can detect the multiple linked QTLs in absence and presence of phenotypic outliers. Therefore, the robust CIM method improves the

performance over the existing CIM approach for QTL analysis. In this study, we developed a pipeline for imputing missing genotypes by using Bayesian classification approach. Simulation result shows that, if we impute 25% missing genotypes than maximum 5% genotypes for the markers could be incorrectly imputed, and we can identify most of the causal QTLs using the imputed genotypes.

Keywords: Composite interval mapping, link genes, robust QTL analysis, gaussian mixture distribution, maximum β-likelihood estimation, genotype Imputation.

AMS Classification: 62P10.

1. Introduction

Lander and Botstein (1989) first developed an idea of using flanking markers for detecting putative QTL positions in experimental crosses, which improved significantly in the past decade in identifying causal genes associated with traits of interest. They proposed Interval Mapping (IM) approach, which have been widely used for identifying putative QTL positions for different organisms (Islam, et al., 2011; Tan, et al., 1998). Estimation procedure of genetic parameters of IM approach is based on maximum likelihood method, which is relatively complex and computationally slow. Haley and Knott (1992) proposed a QTL mapping approach based on multiple regression, which is relatively simple and can be applied using any general statistical package. This approach produce very similar results to those obtained by using maximum likelihood-based IM approach. For further improvement of the regression-based approach, Feenstra, et al. (2006) proposed extension of the Haley–Knott regression method by using estimating equations. Both approaches are approximations of the maximum likelihood-based IM method. In this study, we conducted simulation to observe performance of these methods. In experimental cross populations, the genetic effects associated with marker genotypes are confounded by the position of a functional QTL and its actual effect (Doerge, 2002). Linked QTLs are situated close to each other in the same chromosome. In testing the putative QTLs, it is necessary to fully isolate the effects of multiple possible linked QTLs on chromosomes. However, simple interval mapping approaches cannot separate effects of linked QTLs. Therefore, these approaches cannot precisely detect multiple linked QTLs, and provide flatted confidence interval for QTL positions. CIM approach is a powerful analytical technique, which can separate the effects of QTLs from its position along with increase the reliability and accuracy of QTL mapping (Zeng, 1994).

CIM approach has several advantages than other mapping methods. The basis of this method is an interval test in which the test statistic is constructed to be unaffected by QTLs located outside a defined interval. This is achieved by using the properties of multiple regression analysis (Zeng, 1994). A natural way to eliminate the influence of genetic background is to attempt to remove this confounding information using covariates or cofactor. CIM constructs test statistics by combining interval mapping on two flanking markers and multiple regression analysis on the other markers. For this property it can identify linked genes which are located near at the same chromosome. However, CIM approach is not robust against phenotypic outlier because of using classical maximum likelihood approach in estimating genetic parameters. Classical CIM approach can be robustified by replacing its parameter estimation procedure by maximum beta-likelihood approach (Mihoko and Eguchi, 2002; Mollah and Eguchi, 2008). In this study, we investigated performance of the robust CIM approach for identifying causal QTL positions in presence and absence of phenotypic outliers. We compared the robust CIM approach with classical interval mapping approaches.

Missing genotypes of multiple markers is very common in real experimental cross genotype data. Therefore, before performing QTL analysis it is needed to efficiently impute the missing genotypes of multiple markers over whole genome. In this study, an attempt was made to develop pipeline for imputing missing genotypes by using Bayesian classification approach. We investigated performance of the developed pipeline for imputing missing genotypes and impact of the genotype imputation on QTL analysis.

2. Materials and Methodology

2.1 Statistical Genetic Model for CIM Approach

The CIM is a powerful interval mapping approach for detecting QTLs positions. This approach modifies the standard interval mapping approach by including background markers as cofactors, described independently by Zeng (1993), and Jansen and Stam (1994). For Backcross population, suppose that we want to test for a QTL on a marker interval (i, i+1). If we use markers i and i+1as an indicator for the genotype of the putative QTL within the interval, then the statistical model suggested by Zeng as –

$$y_j = b_0 + b^* x_j^* + \sum_{k \neq i, i+1} b_k x_{jk} + e_j \quad j = 1, 2, ..., n \tag{1}$$

where, y_j is the trait value of the j^{th} individual, b_0 is the mean of the model, b^* is the effect of the putative QTL expressed as a difference in effects between homozygote and heterozygote mode of QTL, x_j^* is an indicator variable, taking a value 1 or 0 with probability depending on the genotypes of the markers i and i+1 for j^{th} individual and the position of the QTL, b_k is the effect of k^{th} marker cofactor, x_{jk} is genotype of k^{th} marker for j^{th} individual, taking a value 1 or 0 depending on whether the marker type is homozygote or heterozygote and e_j is a random variable.Assuming e_j's are identically and independently normally distributed with mean zero and variance σ^2, the likelihood function under H_1 is given by–

$$L_1 = \prod_{j=1}^{n} \left[p_j(1) f_j(1) + p_j(0) f_j(0) \right] \tag{2}$$

where $p_j(1)$ gives a prior probability of x_j^* =1, $p_j(0)$ = 1− $p_j(1)$,$f_j(1)$ and $f_j(0)$specify a normal density function for the random variable y_j with mean $b_0 + b^* + \sum_{k \neq i, i+1} b_k x_{jk}$ and $b_0 + \sum_{k \neq i, i+1} b_k x_{jk}$, respectively, and a variance σ^2. By differentiating the likelihood function (2) with respect to individual parameters, setting the derivatives equal to zero, and then solving the equations, the maximum likelihood (ML) estimates of the parameters b^*, b_k's and σ^2 can found as follows-

$$\hat{b}^* = (Y - X\hat{B})' \hat{P} / \hat{c} \tag{3}$$

$$\hat{B} = (X'X)^{-1} X'(Y - \hat{P}\hat{b}^*) \tag{4}$$

$$\hat{\sigma}^2 = [(Y - X\hat{B})'(Y - X\hat{B}) - \hat{c}\hat{b}^{*2}] / n \tag{5}$$

where, Y is a (n × 1)vector of phenotypes y_j's, $\hat{B}$ is a ((t − 1) × 1) vector of the ML estimates of b_k's, X is an (n × (t − 1)) matrix of x_{jk}'s, $\hat{P}$ is a (n × 1) vector with elements $\hat{P}_j$ specifying the ML estimate of the posterior probability of x_j^* = 1:

$$\hat{P}_j = p_j(1) \hat{f}_j(1) / [p_j(1) \hat{f}_j(1) + p_j(2) \hat{f}_j(2)] \tag{6}$$

and, $\hat{c} = \sum_{j=1}^{n} \hat{P}_j$ (7)

Estimates of the parameters can be found by iterative procedure of the above equations via the expectation conditional maximization (ECM) algorithm.

The hypotheses are H_0: $b^* = 0$ and H_1: $b^* \neq 0$. The likelihood function under the null hypothesis is

$$L_0 = \prod_{j=1}^{n} f_j(0) \tag{8}$$

with the ML estimates

$$\hat{B} = (X'X)^{-1}X'Y \tag{9}$$

$$\hat{\sigma}^2 = (Y - X\hat{B})'(Y - X\hat{B}) / n \tag{10}$$

The likelihood ratio (LR) test statistic is defined by-

$$LR = -2\ln\left(\frac{L_0}{L_1}\right) \tag{11}$$

Estimated effects of QTLs by using this approach are unaffected by other linked QTLs.

2.2 Robust CIM by Maximizing β-Likelihood Function Using EM Algorithm

The β-likelihood function was developed by minimizing β-divergence (BASU, et al., 1998; Mihoko and Eguchi, 2002; Mollah and Eguchi, 2008). It is defined as-

$$L_\beta(\theta \mid Y, X) = \frac{1}{n\beta}\sum_{j=1}^{n} f_\theta^{\beta}(y_j \mid X_j) - b_\beta(\theta, X) \tag{12}$$

In our current context

$$f_\theta(y_j \mid X_j) = \sum_{i=1}^{2} p_{ij}\phi\left(\frac{y_j - \mu_{ji}}{\sigma}\right)$$

is the normal mixture model. Then the estimators of the parameters are obtained by maximizing β-likelihood function using EM like algorithm by treating the normal mixture model as an incomplete data density. And the estimators of the parameter under alternative hypothesis are given as–

$$b^{*(t+1)} = (Y - XB^{(t)})^T P_\beta^{(t)} D\left[1^T P_\beta^{(t)}(D \# D)\right]^{-1} \tag{13}$$

$$B^{(t+1)} = \left[X^T\{X \#(P_\beta^{(t)}1)\}\right]^{-1}\left[X^T\{Y \#(P_\beta^{(t)}1) - P_\beta^{(t)} Db^{*(t+1)}\}\right] \tag{14}$$

And

$$\begin{aligned}\sigma^{2(t+1)} = {} & [(Y - XB^{(t+1)})^T\{(Y - XB^{(t+1)}) \#(P_\beta^{(t)}1)\} - 2(Y - XB^{(t+1)})^T P_\beta^{(t)} Db^{*(t+1)} \\ & + V^{(t)} b^{*2(t+1)}][1^T P_\beta^{(t)} 1 - \beta(1+\beta)^{-\frac{3}{2}}]^{-1}\end{aligned} \tag{15}$$

where, $P_\beta = \left\{ \left[\exp\left\{ -\frac{1}{2}\left(\frac{y_j - \mu_{ji}}{\sigma} \right) \right\} p_{ji} \right]^\beta P_{ji} \right\}_{n\times 2}$ (16)

$V = 1^T P_\beta (D\#D)$ and the notation # denotes Hadamard product and D is the design matrix, D^T = [1, 0]. Under null hypothesis the minimum β-divergence estimators for the parameters are obtain iteratively as follows-

$$B^{(t+1)} = \left[X^T \left\{ X \#\left(W_\beta^{(t)} 1 \right) \right\} \right]^{-1} \left\{ X \#(W_\beta^{(t)} 1 \right\}^T Y \quad (17)$$

$$\sigma^{2(t+1)} = (Y - XB^{(t+1)})^T [(Y - XB^{(t+1)}) \# W_\beta^{(t)}] \quad (18)$$

$$[1^T W_\beta^{(t)} - \beta(1+\beta)^{-\frac{3}{2}}]^{-1}$$

where, $W_\beta = \left[\exp\left\{ -\frac{\beta}{2}\left(\frac{y_j - X_j B}{\sigma} \right)^2 \right\} \right]_{n\times 1}$ (19)

which is the vector of the β-weight under H_0. Thus β-*LOD* score for the evidence of a QTL is given by-

$$LOD_\beta = 0.434n \left\{ \sup_\theta L_\beta(\theta \mid Y, X) - \sup_{\theta_0} L_\beta(\theta \mid Y, X) \right\} \quad (20)$$

For $\beta \rightarrow 0$, the LOD_β reduce to the classical *LOD* score.

2.2.1 Initialization Parameters

We used initial value for the parameter *B*as-

$$B_0 = \left(X^T X \right)^{-1} X^T Y$$

where, X is $(n \times (t-1))$ incidence matrix of the $(t-2)$ markers; b_0 is the model mean; and Y is the refined phenotypic value of the individuals. In this case, phenotypic data must need to between $0.5Q_1$ and $2Q_3$, where Q_1 is first quartile and Q_3 is third quartile of the phenotypic observations. If y_j is greater than $2Q_3$ or less than $0.5Q_1$ than it was replaced by the median of the phenotype value. We used $b^* = 0$ as the initial value suggested by Zeng (1994).

2.3 Imputing Missing Marker Genotype

Missing of genotypes of the genetic markers is very common in QTL analysis. If markers contain missing genotypes, estimating genetic parameters using the CIM method is not feasible. For estimating the genetic parameters of the CIM method, it is needed to impute missing genotypes of the markers. In this study, we developed a pipeline of imputing missing genotypes based on naïve bayes classification approach. We imputed missing marker-genotypes using Bayesian classification based on conditional genotype probabilities, which can be calculated

by different map functions (e.g., Haldane, Kosambi, c-f and Morgan). We used haldane map function to calculate conditional genotype probabilities. If $p_{ij}(1)$ and $p_{ij}(2)$ are the probabilities for *QQ* and *Qq*, respectively, and if the value of the genotype *QQ* is 1 and the value of genotype *Qq* is 2, then we calculated P_{ij} as- $P_{ij} = 1\times\ p_{ij}(1) + 2 \times\ p_{ij}(2)$. Based on known genotype of the i^{th} marker and corresponding P_i, we trained the naïve bayes model and based on P_i of the unknown marker genotype the missing genotypes were imputed. Two types of genotypes were treated as two different populations for employing naïve bayes classification approach.

3. Result

3.1 Performance of Mapping Approaches in Absence of Outliers

Main advantage of CIM over simple interval mapping approaches is identifying linked QTLs. In this study, we investigated the performance of linked QTLs identification of the CIM, IM, Hally-Knott (HK), extended Hally-Knott (eHK), and beta-likelihood based robust CIM method. We simulated genotype data by considering 4 chromosomes and each with 15 markers with maker interval 10cM. For phenotypic data, we considered total 10 QTLs are associated with simulated phenotypic traits, which are contributed to 81% phenotypic variations. Out of 10 QTLs, one is located at the 3rd marker of the 1st chromosome, three are located at the 2nd, 5th and 8th markers positions of the 2nd chromosome, another three are located at the 4th, 7th and 13th markers positions of the 3rd chromosome, and last three are located at the 2nd, 5th and 10thmarkers positions of the 4th chromosome. Data for mapping QTLs consisted of markers information and phenotypic values for 300 individuals. Genotypes for each of the markers can be recorded in digital form, such as 1 and 0 for distinguishing the two marker types (homozygote and heterozygote). We conducted 100 simulations and calculated average LOD scores for each of the genome positions. From figure-1a, we observed that only CIM and robust CIM methods calculated high LOD score at the causal QTL position at 3rd marker of the 1st chromosome. For both methods, LOD scores exceeded the predefined threshold for 5% level of significance, whereas average LOD scores calculated using simple interval mapping methods were below the threshold level. Form figure 1b, it was observed that calculated LOD scores using simple interval mapping methods exceeded predefined threshold for a large region. In this region, there were two different QTLs, however the methods failed to distinguished them. Similar problems for simple interval mapping was also observed for chromosome-3 and chromosome-4. However, in all chromosomes CIM and robust CIM successfully identified causal QTLs.

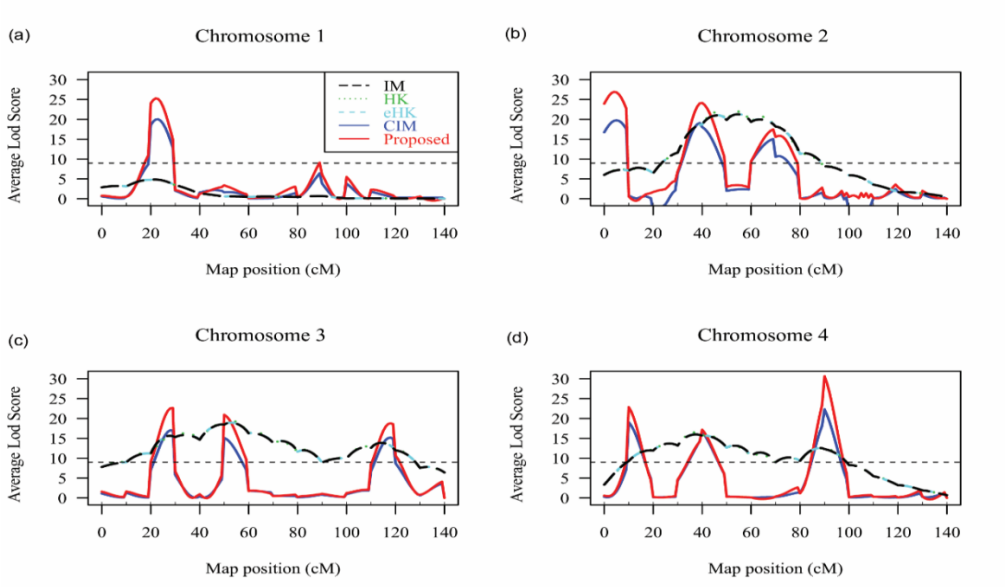

Figure 1: The performance of different methods used to identify multiple linked QTLs in 100 simulations. Threshold values for two different methods were calculated by using permutation test with 5% level of significance. Threshold line was based on the average of threshold values for the methods. In the absence of phenotypic outliers, the LOD score of the genome locations were calculated by using different mapping methods. Different colors and line types were used for plotting LOD scores for different methods, e.g., blue solid line for classical CIM, red solid line for robust CIM, black dash line for IM, green dash line for HK, and sky blue for eHK. LOD scores for each of the four different chromosomes were plotted separately in fig1a, fig1b, fig1c and fig1d for observing clear view of the differences among the approaches.

3.2 Performance of Robust and Classical CIM Approaches in Presence of Outliers

We further investigated performance of CIM and robust CIM methods in presence of phenotypic outliers. We randomly allocated 5% outlying observations to the phenotypic data and calculated LOD scores using CIM and robust CIM approaches. It was observed that classical CIM is failed to detect QTLs in presence of phenotypic outlying observations (Figure 2). Moreover, LOD scores calculated using classical CIM approach exceeded predefined threshold in several wrong positions, which are false discovery of this method. Therefore, classical CIM approach could perform poorly for discovering QTL positions in presence of

phenotypic outliers. However, robust CIM approach identified all the QTLs positions correctly in presence of outlying observations. Moreover, no false discovery was observed in four different chromosomes. Therefore, robust CIM method provide good results in absence and presence of phenotypic outlying observation and solved the main drawback of the classical CIM approach.

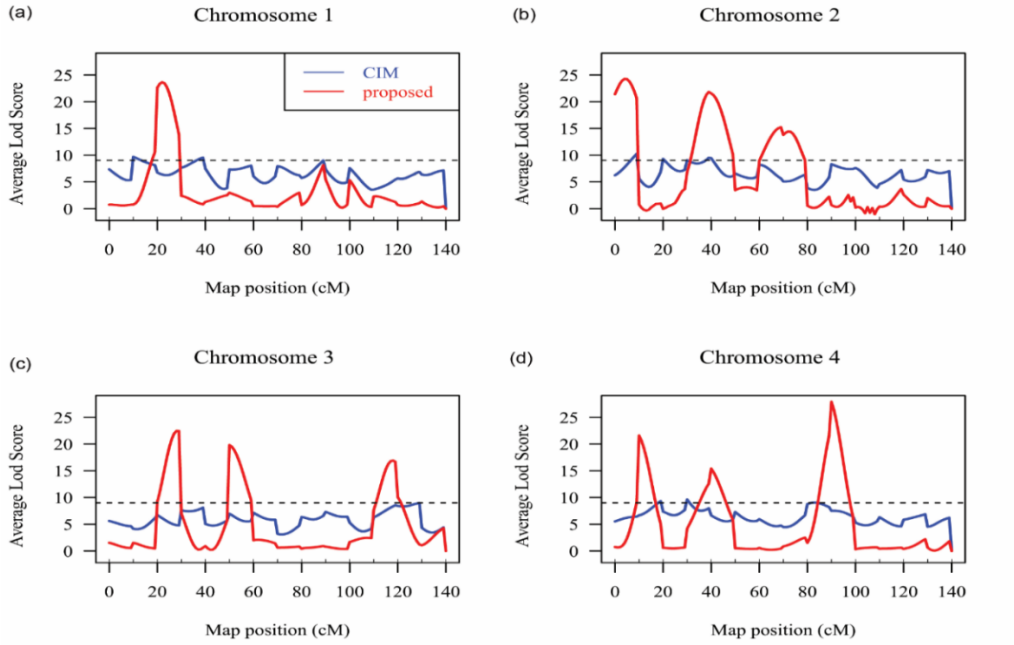

Figure 2: Performance of the classical CIM and robust CIM approaches for linked QTL identification in presence of outlier. Threshold values for two different methods with 5% level of significance were calculated by using permutation test. Threshold line was based on the average of the threshold values for the methods. LOD scores genome locations were calculated by using CIM and proposed robust CIM methods in presence of phenotypic outliers. Then average LOD scores for 100 simulations for each of the genome locations were plotted. Similar to figure-1, LOD scores for the four different chromosomes were plotted separately in fig2a, fig2b, fig2c, and fig2d.

3.2 Missing Genotype Imputation and QTL Analysis

We estimated missing genotypes of the markers using Bayesian classification and conditional genotype probabilities. We randomly generated 25% missing genotypes of the markers and predicted missing genotypes through classification. In this case, we considered 10 QTLs for generating phenotypic data. For convenience of graphical presentation we located 10 QTLs as follows- three QTLs

at the 2nd, 5th and 10th markers positions of the 1st chromosome, three at the 2nd, 5th and 8th markers positions of the 2nd chromosome, three at the 4th, 7th and 13th markers positions of the 3rd chromosome, and last one at the 3rd marker position of the 4th chromosome. In figure- 3(a), we plotted missing genotypes of the markers. It was observed that missing genotypes were randomly distributed. After employing developed pipeline for predicting missing genotypes, it was observed that maximum 5% imputed genotypes could be incorrect. For most of the markers, misclassification rate is below of 3%. Therefore, developed pipeline can significantly correctly impute the missing genotypes. And this pipeline can be used for missing genotype imputation rather than arbitrarily guessing the missing genotypes. In simulation study, we performed QTL analysis with imputed and original genotypic data. It was observed that with imputed genotype data it is possible to identify most of the causal QTLs (Figure 3c-d).

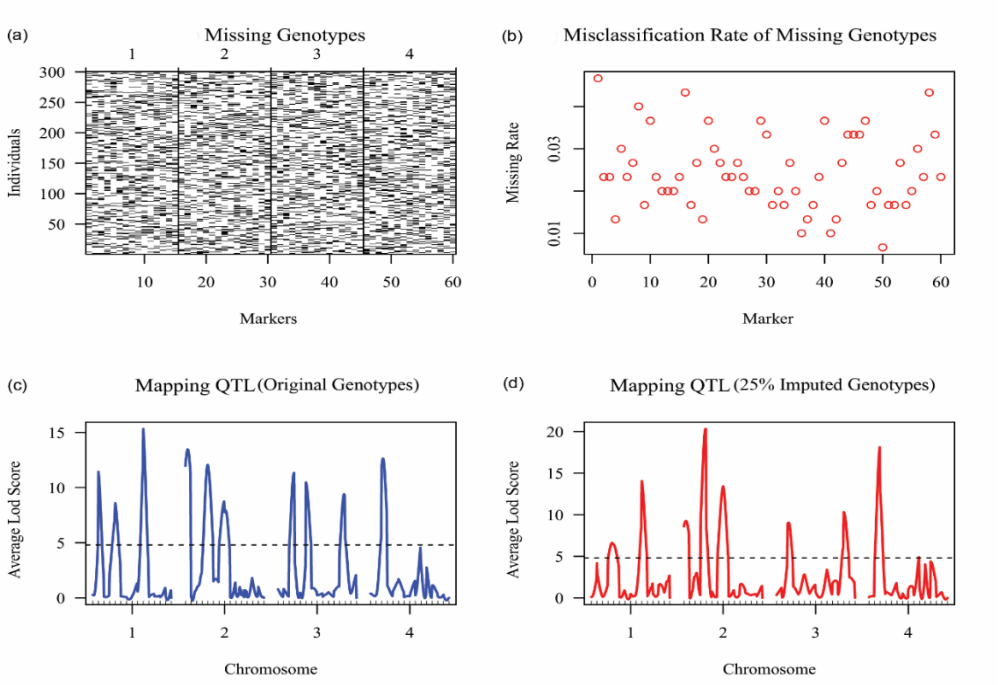

Figure 3: Missing genotypes imputation using Bayesian classification approach and QTL analysis using robust CIM approach. Here,(a) missing genotypes for 60 markers with respect to 300 individual lines; (b) mis-classification rate in genotype imputation for each of the 60 markers; (c) original genotypes of the markers were analyzed using robust CIM approach; and (d) Genotype data with 25% imputed genotypes using Bayesian classification was analyzed using robust CIM approach.

4. Discussion

QTL analysis is a powerful approach for discovering genes associated with phenotypic traits of different experimental crosses of animal and plants, largely being used since last decade (Barchi, et al., 2009; Quarrie, et al., 1997; van der Schaar, et al., 1997). There are many methods proposed for QTL analysis, such as, maximum likelihood-based interval mapping, regression-based interval mapping, composite interval mapping etc. Composite interval mapping is one of the popular QTL mapping approaches, which is used for precisely discovering linked and unlinked QTLs (Rodriguez-Zas, et al., 2002; Zeng, 1994). However, in presence of phenotypic outliers this approach could fail to identify causal QTLs and by the same time could detect wrong genome locations as QTLs. To solve this problem, robust CIM approach was developed based on beta-likelihood method for QTL analysis (Mollah and Eguchi, 2008). In this study, we discussed about the robustification of CIM algorithm for identification of both linked and unlinked genes by maximizing β-likelihood function using EM algorithm. Adjusting the value of parameter β plays a key role in the performance of the robust method. An appropriate value for the tuning parameter β can be selected by cross validation. However, in this study we used $\beta = 0.2$ heuristically. Simulation studies show that the robust method significantly improves the performance over the traditional IM and CIM methods in presence of phenotypic outliers; otherwise, it keeps equal performance to CIM. Missing genotypes of genetic markers very common in the real genotype data sets. Therefore, in this study an attempt was made to develop a pipeline for predicting missing genotypic data. We developed a procedure for predicting the missing genotype using Bayesian classification. Simulation study shows that, if we impute 25% missing genotypes by the developed pipeline than maximum 5% genotypes might be wrongly imputed. And, we can identify actual QTL positions using imputed genotypes. Our study could significantly improve QTL analysis results for different experimental crosses.

Reference

[1] Barchi, L., et al. (2009). QTL analysis of plant development and fruit traits in pepper and performance of selective phenotyping. Theoretical and Applied Genetics, 118(6), 1157-1171.

[2] BASU, A., et al. (1998). Robust and efficient estimation by minimising a density power divergence. Biometrika, 85(3), 549-559.

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

[4] Feenstra, B., Skovgaard, I.M. and Broman, K.W. (2006). Mapping quantitative trait loci by an extension of the Haley-Knott regression method using estimating equations. Genetics, 173(4), 2269-2282.

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

[6] Jansen, R.C. and Stam, P. (1994). High-Resolution of Quantitative Traits into Multiple Loci Via Interval Mapping. Genetics, 136(4), 1447-1455.

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

[8] Mihoko, M. and Eguchi, S. (2002). Robust blind source separation by beta divergence. Neural computation, 14(8), 1859-1886.

[9] Mollah, M.N.H. and Eguchi, S. (2008). Robust Composite Interval Mapping for QTL Analysis by Minimum beta-Divergence Method. In, 2008 IEEE International Conference on Bioinformatics and Biomedicine. IEEE:115-120.

[10] Quarrie, S.A., et al. (1997). QTL analysis to study the association between leaf size and abscisic acid accumulation in droughted rice leaves and comparisons across cereals. Oryza: From Molecule to Plant:155-165.

[11] Rodriguez-Zas, S., et al. (2002). Interval and composite interval mapping of somatic cell score, yield, and components of milk in dairy cattle. Journal of Dairy Science, 85(11), 3081-3091.

[12] Tan, X., et al. (1998). Genetic analysis of rice CMS-WA fertility restoration based on QTL mapping. Theoretical and Applied Genetics, 97(5-6), 994-999.

[13] van der Schaar, W., et al. (1997). QTL analysis of seed dormancy in Arabidopsis using recombinant inbred lines and MQM mapping. Heredity, 79(2), 190-200.

[14] Zeng, Z.B. (1993). Theoretical Basis for Separation of Multiple Linked Gene Effects in Mapping Quantitative Trait Loci. P Natl AcadSci USA, 90(23), 10972-10976.

[15] Zeng, Z.B. (1994). Precision Mapping of Quantitative Trait Loci. Genetics, 136(4), 1457-1468.