International Journal of Statistical Sciences *ISSN 1683-5603*
Vol. 17, 2019, pp 97-114
© 2019 Dept. of Statistics, Univ. of Rajshahi, Bangladesh

SNP Based Robust Fast-eQTL Mapping for Identification of Important Genes

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

[1]Institutue of Crop Science & Institute of Bioinformatics, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310058, China

[2]Laboratory of Bioinformatics, Department of Statistics, University of Rajshahi, Rajshahi-6205, Bangladesh
[*]Correspondence should be addressed to Mollah M.N.H. (mollah.stat.bio@ru.ac.bd)

[Received June 10, 2018; Revised April 15, 2019; Accepted May 20, 2019]

Abstract

Gene expression quantitative trait locus (eQTL) mapping measures the relationship between gene expressions and genotypic variations in a certain chromosomal location. Most of the existing eQTL mapping algorithms are not suitable for the identification of important genes related to human phenotypic problems like a disease. However, SNP based fast eQTL mapping is suitable for the identification of important genes related to human phenotypic problems. But, the existing SNP based eQTL mapping is sensitive to outliers. In this paper, we robustify the SNP based Fast eQTL mapping using outlier modification rule. For outlier modification, we use the minimum β-divergence method. First, we detect the outlier by β-weight function then we replace outlier data points by its respective robust mean produced by the minimum β-divergence method. Then we investigate the performance of the proposed method in a comparison of the existing method using simulation study. Simulation results show that the proposed

method performs better than traditional method in presence of outliers; otherwise it keeps almost equal performance.

Keywords: Gene Expression, SNP, eQTL, β-weight, Outlier detection and modification, Robustness.

AMS Classification: Primary 62G35; Secondary 92-08.

1. Introduction

The central biological attention is finding the casual connection of polymorphism with phenotypes (Korte and Farlow, 2013). The capability to calculate genetic risk factors for human sickness and agronomically important traits like development rate, yield in plants need an understanding of both the exact loci that trigger a phenotype and the genetic architecture of a trait (Korte and Farlow, 2013). The connection between phenotype and genotype has been of main attention since Mendel hypothesized the presence of 'internal factors' that are passed on to the following generation. An extensively use method to recognize the genomic loci linked with phenotypic variation in a genetically segregating population is a QTL mapping which has been greatly successful in determining causative loci underlying numerous disease phenotypes (Cervino et al. 2005; Hillebrandt et al. 2005; Wang et al. 2004) and can be broadly partitioned into two classes namely linkage mapping and association mapping. Likelihood and regression approaches are used to map QTL for ordinary and linkage mapping to experimental crosses, where flanking markers used to infer genotypes in the interval between widely spaced markers (Haley and Knott, 1992; Lander and Botstein, 1989). Linkage statistics might be calculated at single marker loci with minimal loss in precision or power as marker density increases (Kong and Wright, 1994). On the other

hand, simple association mapping calculated only at the marker loci instead of considering the linkage disequilibrium structure between marker loci.

QTL mapping has proved, and remains, an influential method to recognize regions of the genome that co-segregate with a specified trait either in F_2 populations or Recombinant Inbred Line (RIL) families (Korte and Farlow, 2013). To overcome the limitations of the traditional QTL mapping methods, we can use the most popular and widely used eQTL analysis. eQTL mapping, an innovative combination of traditional quantitative trait mapping and microarray technology, is a genetic mapping of genome-wide gene expression or transcriptome. It aims to identify genomic locations to which expression traits are linked. eQTL mapping studies have been applied in several model organisms and humans recently (Brem et al. 2002; Schadt et al. 2003; Morley et al. 2004; Chesler et al. 2005; Stranger et al. 2005; WANG et al. 2006). These studies thus far have demonstrated several advantages of this line of research from identifying candidate genes (Schadt et al. 2003) to elucidating regulatory networks (Brem et al. 2002; Schadt et al. 2003; Yvert et al. 2003). Some statistical methods for eQTL mapping have been discussed by C. M. Kendziorski et al. (2006) and Wang et al. (2011). Fast eQTL mapping (FastMap) in homozygous populations has been discussed by Gatti et al. (2009). The cost of collecting gene expression and high-density genotype data on the identical population have lowered due to current developments of gene expression and single nucleotide polymorphism (SNP) microarray machinery (Gatti et al. 2009).

Gatti et al. (2009) reported that recent advances technologies were used to produced high-density SNP datasets with thousands of transcripts and millions of allele calls in both mice (Frazer et al. 2007b; Szatkiewicz et al. 2008) and humans

(Frazer et al. 2007a). The computational challenges for current tools are the calculation of the relationship between tens of thousands of transcripts and thousands to millions of SNPs. A large number of available SNPs and transcripts made these challenges more complicated. Numerous methods have been used to overcome these issues including a mutual way of addressing several comparisons among markers is a resampling method (Carlborg et al. 2005; Churchill and Doerge 1994; Peirce et al. 2006) which is also used by numerous existing QTL mapping tools (Broman et al. 2003; Manly et al. 2001; Wang et al. 2003). Likelihood ratio statistic (LRS) (Chesler et al. 2005) or the mixture over markers method (Kendziorski et al. 2006) used for transcript-specific testing of association with SNPs has been beforehand addressed by thresholding transcripts using q-values (Storey and Tibshirani, 2003) for multiple comparisons among transcripts.

The challenge of eQTL mapping lies in the fact that there are an enormous amount of hereditary variants and gene expression characters, and hence the exploration space for potential eQTLs is vast. A possible explanation to the computational challenges associated with the eQTL analysis is the parallel computation (Carlborg et al. 2005). Gatti et al. (2009) developed the FastMap algorithm and implemented it as a Java-based desktop software package that performs eQTL analysis using association mapping, where they addressed the growing need for eQTL mapping in high-density SNP datasets and the pitiable scalability of the current computational tools. They attained computational efficiency over the use of a data structure named as Subset Summation Tree (SST). Either single marker mapping (SMM) or haplotype association mapping (HAM) by sliding an m-SNP window across the genome was performed by FastMap (Pletcher et al. 2004). However, existing eQTL mapping algorithms are very much sensitive to outliers. Therefore, it is crucial to develop new robust methods which are not affected by

outliers. In this paper, we propose a robust fast eQTL mapping algorithm for the identification of important genes. We investigate the performance of the proposed method in a comparison of the existing method using simulation study. Simulation results show that the proposed method performs better than the traditional method in presence of outliers; otherwise, it keeps almost equal performance.

2. Methods and Materials

The FastMap algorithm used association mapping for eQTL analysis. The input data consists of two matrices in association mapping for homozygous inbred strains: real-valued transcript expression measurements contained in the first and SNP allele calls, coded as 0 for the major allele and 1 for minor allele contain in the second matrix. The same number of strains contains in each matrix. SMM or HAM by sliding an *m*-SNP window across the genome was performed by FastMap.

2.1 1-SNP sliding window

The calculation of test statistics (correlation) for SMM in a 1-SNP sliding window in association mapping with homozygous inbred strains for a given transcript g and SNP s is given below (Gatti et al. 2009):

$$\text{cor}(g,s) = \frac{\frac{1}{n}\sum_{i=1}^{n} g_i s_i - \frac{1}{n^2}\sum_{i=1}^{n} g_i \sum_{i=1}^{n} s_i}{\sqrt{Var(g)\,Var(s)}} \quad (1)$$

Simplifying the formula, authors assumed without loss of generality that each transcript expression vector g is centered and standardized such that,

$$\sum_{i=1}^{n} g_i = 0 \text{ and } \sum_{i=1}^{n} {g_i}^2 = 1 \quad (2)$$

In this case, the reduces correlation expression is

$$\text{cor}(g,s) = \frac{\sum_{i=1}^{n} g_i s_i}{\sqrt{n}\sqrt{\frac{1}{n}\sum_{i=1}^{n} {s_i}^2 - \left(\frac{1}{n}\sum_{i=1}^{n} s_i\right)^2}} \quad (3)$$

The denominator of (3) depends upon the Hamming weight of *s*, which can be calculated once for each but the numerator must be calculated for every SNP-transcript pair.

To speed up the calculation, the numerator of (3) is denoted by $M_g(s)$:

$$M_g(s) = cov(g,s) = \sum_{i=1}^{n} g_i s_i = \sum_{i:s_i=1} g_i$$

$M_g(s)$ is simply the sum of transcript expression values over a subset of samples defined by the minor allele of the SNP as the SNPs are binary. To explain how the calculation of the $M_g(s)$ can be simplified, consider two SNPs *s* and s'that vary only at the i-th position (thus s and s' have Hamming distance of 1):

$$\text{s} = (s_1, s_2, \ldots, s_{i-1}, s_i = 0, s_{i+1}, \ldots, s_n)$$

$$s' = (s_1, s_2, \ldots, s_{i-1}, {s_i}' = 1, s_{i+1,\ldots,} s_n)$$

In this case, the $M_g(s')$ can be calculated quickly from $M_g(s)$as follows:

$$M_g(s') = \sum g_i {s_i}' = \sum g_i s_i + g_i({s_i}' - s_i) = M_g(s) + g_i \quad (4)$$

The association statistic is the same for SNPs with the same strain distribution pattern (SDP) for any given transcript. Hence, the authors calculated the association statistic once for each unique SDP.

2.2 Permutation-based significance thresholds

For a single transcript, the association statistic is calculated between the observed values of that transcript and all SNPs. The transcript data are then permuted while the SNP data are held fixed. Association statistics are calculated between the

permuted transcript values and all SNPs and the maximum association statistic is stored. The distribution of the maximum association statistics obtained from 1000 permutations of the transcript's values is used to define significance thresholds for individual pairs, and to assign a percentile-based *p*-value to the observed maximum association of the transcript across SNPs. Permutation-based maximum association test used to get p values to test the multiple comparisons across transcripts.

2.3 SNP Based Robust Fast eQTL Mapping (Proposed)

In this paper, we would like to apply the outlier modification rule for SNP Based Robust Fast eQTL mapping as follows:

2.3.1 Robust 1-SNP sliding window

From equation (3), it is obvious that cor(g, s), the existing Fast-eQTL mapping (Gatti et al. 2009) correlation estimator is very much sensitive to outliers. Therefore, an attempt is made to propose a new robust FastMap approach by the minimum β-divergence method as follows.

Outlier detection and modification for gene expressions:

We used the following equation to detect the outliers

$$W_\beta(x_{ij}|\theta_j) = \exp\{-\frac{\beta}{2\sigma_j^2}(x_{ij} - \mu_i)^2\}, \quad (5)$$

where, $\theta = (\mu_j, \sigma_j), x_{ij}$ are the i-th observation (strain) of j-th SNPs and μ_i is the mean of the i-th strains.

The equation (5) is called the β-weight function (Mollah et al. 2010).

The value of the tuning parameter β plays a key role in the performance of the proposed method. It controls the trade-off between robustness and efficiency of estimators. We consider the β-weight function with β = 0.2 as a measure for outlier detection.

where $0 \leq W_\beta \leq 1$. If $W_\beta(g_{ij}) < 0.2$, then we consider g_{ij} as an outlier.

Let us consider g'_{ij} as follow:

$$g'_{ij} = [g_{1j,}\, g_{2j}, \dot{g}_{3j}\,,\ \dot{g}_{4j},\ \ldots,\ g_{nj}] \tag{6}$$

Suppose, $\dot{g}_{3j}$ and $\dot{g}_{4j}$ genes are affected by outliers. To overcome this problem, these genes values must calculate from the robust mean. After processing by our proposed method, we get our desire gene data which is not affected by outliers.

2.3.2 Robust Prewhitening for Gene Expressions

After outlier detection, we proposed robust pre-processing by β–divergence method as follow:

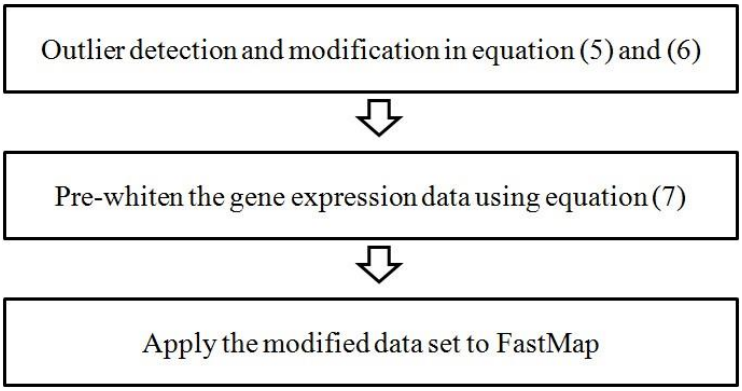
$$g_\beta{}' = \frac{x-\hat{\mu}_{j,\beta}}{\hat{\sigma}_{j,\beta}{}^2} \tag{7}$$

where, $\hat{\mu}_{j,\beta}$ and $\hat{\sigma}_{j,\beta}{}^2$ are the robust estimator of mean and variance, respectively. We robustify (3) by our proposed minimum β- divergence methods. The proposed minimum β-divergence estimators $\hat{\theta}_{j,\beta} = (\hat{\mu}_{j,\beta}\ , \hat{\sigma}_{j,\beta}{}^2)$ of the parameters $\theta_j = (\mu_j, \sigma_j{}^2)$ are computed iteratively as follows,

$$\mu_{j,t+1} = \frac{\sum_{i=1}^{n} W_\beta(x_{ij}|\theta_{j,t}) x_{ij}}{\sum_{i=1}^{n} w_\beta(x_{ij}|\theta_{j,t})} \text{ and } \sigma_{j,t+1}{}^2 = \frac{\sum_{i=1}^{n} W_\beta(x_{ij}|\theta_{j,t})(x_{j,t}-\mu_{j,t})^2}{(\beta+1)^{-1} \sum_{i=1}^{n} w_\beta(x_{ij}|\theta_{j,t})} \tag{8}$$

2.3.3 Flow-Chart for Robust 1-SNP sliding window

For easily understanding our proposed methods working pathway, we proposed the following flowchart:



3. Results and Discussion

We have performed a simulation study to compare the performance of our proposed method with the existing FastMap method for identifying important genes.

3.1 Single SNP association

We have considered that gene expression is controlled by a single SNP. That is, particular gene expression is associated or correlated with a single SNP. To illustrate the performance of the proposed method in a comparison of the FastMap method, in our simulation study we have considered 50 individuals (or strains) and generate 100 SNPs where each containing 50 observations (strains). We have considered 5 SNPs per each of the 20 chromosomes. We have considered the gene expression controlled from the association between 25^{th} gene and at 21^{st} SNP. We have performed the simulation 100000 times.

We have calculated correlations between SNPs and gene expressions. Then we have plotted the SNPs along the X-axis and the corresponding absolute correlations along the Y-axis. The black color with the solid line indicated classical method, whereas dotted line with the red color indicated by the proposed method. Figure 1(a) shows that the association is detected correctly by the FastMap and our proposed method at 21st SNP in the absence of outliers. To investigate the robustness of the proposed method in a comparison of the FastMap, we contaminated dataset by 20% outliers. Figure 1(b) shows the plot of SNPs along the X-axis and the corresponding correlation along the Y-axis in the presence of contaminated dataset. However, Figure 1(b) shows that the classical FastMap method fails to detect the association correctly at 21st SNP in the presence of outliers. But our proposed method can detect the association correctly at the 21st SNP in the presence of outliers.

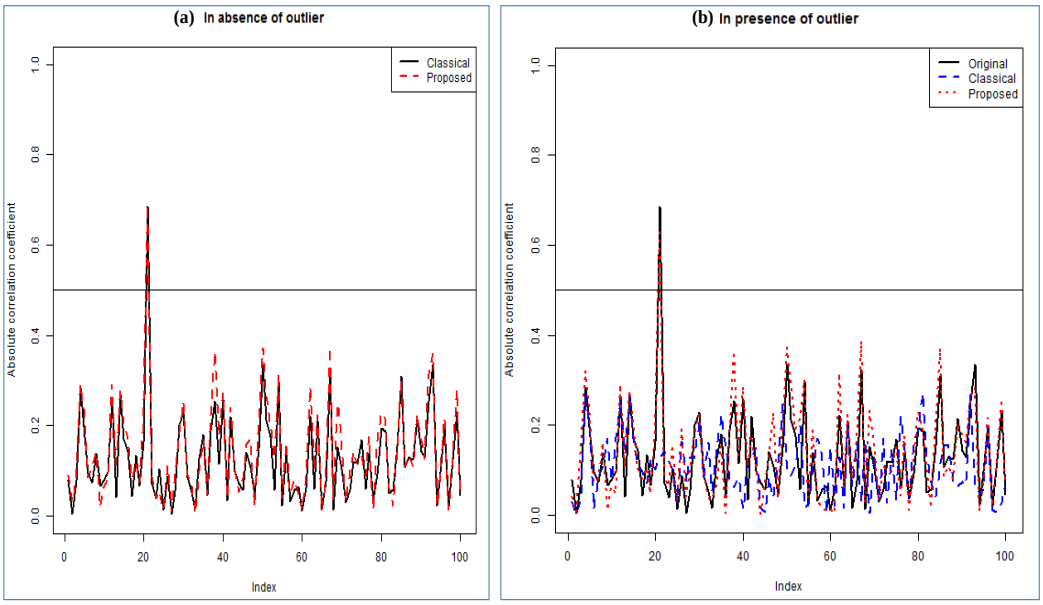


Figure 1: Single SNP association. (a) In absence of outliers
(b) In presence of outliers

3.2 Two SNPs association

To illustrate the performance of the proposed method in a comparison of FastMap method for gene expression dataset, we considered that a particular gene expression is controlled by two SNPs in absence and presence of outliers in the datasets. That is, a particular gene expression is associated or correlated with two SNPs. In our simulation study, we have considered 50 individuals (or strains) and generate 100 SNPs where each containing 50 observations (strains). We have considered 5 SNPs per each of the 20 chromosomes. We have considered the gene expression controlled from the association between 25^{th} gene and at 55^{th} and 95^{th} SNP. We have performed the simulation 100000 times.

We have calculated correlations between SNPs and gene expressions. Then we have plotted the SNPs along the X-axis and the corresponding correlations along the Y-axis. Figure 2(a) shows that the association is detected by the FastMap and our proposed method correctly at 55^{th} and 95^{th} SNP in the absence of outliers. To investigate the robustness of the proposed method in a comparison of the FastMap, we contaminated dataset by 20% outliers. Figure 2(b) shows that the FastMap method fails to detect the association correctly at 55^{th} and 95^{th} SNP in the presence of outliers. But our proposed method can detect the association correctly at the 55^{th} and 95^{th} SNP in the presence of outliers in case of two SNPs association.

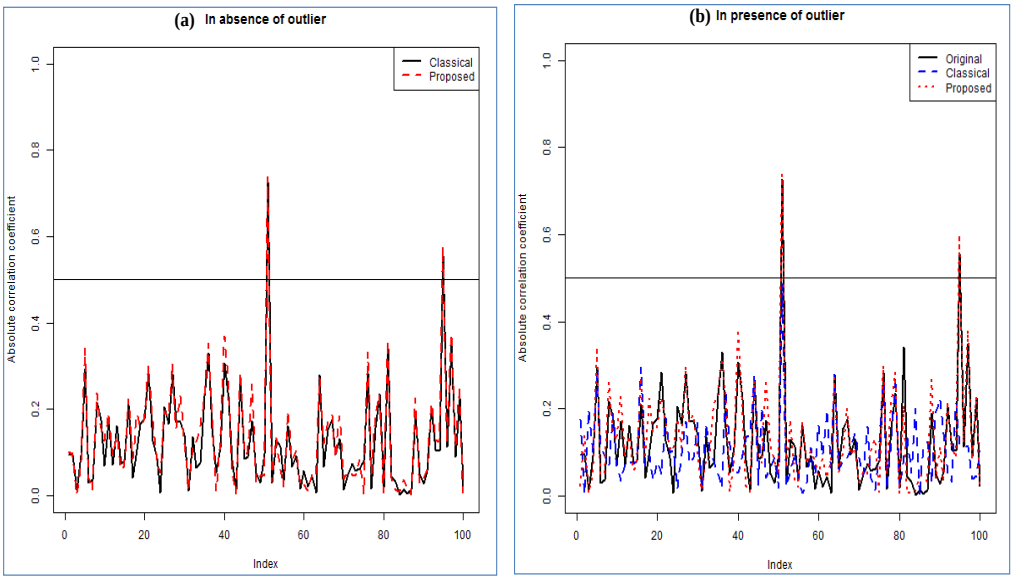


Figure 2: Two SNPs association. (a) In absence of outliers

(b) In presence of outliers

3.3 Correlation plot of genes with single and multiple SNPs

Correlogram is a graph of the correlation matrix. It is very useful to highlight the most correlated variables in a data table. We generated 20 genes and 20 SNPs. Here we considered single, two and three SNPs on different genes to see their respective correlation. We considered that the first gene expression depends on first SNP, 5^{th} gene expression depends on 3^{rd} and 10^{th} SNPs, and the 10^{th} gene expression depends on 6^{th}, 8^{th} and 10^{th} SNPs in the absence of outliers. The color intensity and the size of the circle are proportional to the correlation coefficients. From Figure 3(a), we see that the color intensity and the size of the circle corresponding to the first gene and first SNP is high. So, we can say that 1^{st} gene expression controlled by 1^{st} SNP. From Figure 3(a), we see that gene color

intensity and the size of the circle corresponding to the 5th gene with 3rd and 10th SNPs are high. Form this figure; we can say that 5th gene influence by 3rd and 10th SNPs. Similarly, from Figure 3(a), we see that gene color intensity and the size of the circle corresponding to the 10th gene corresponding to 6th, 8th and 10th SNPs are high. Form this figure we can say that 10th gene expression controlled by 6th, 8th and 10th SNPs. So, we can conclude that in the absence of outliers, FastMap detects the true expression. We also generated a correlation plot in case of contaminated data. Here we also considered the same number of genes and SNPs position as mentioned above to see the genes and SNPs correlation. We contaminated dataset by 20% outliers. In the case of outliers, from Figure 3(b), we see that the color intensity and the size of the respective genes corresponding to SNPs reduced and some cases give misleading results. So, form Figure 3(b), we can conclude that the correlation between genes and SNPs are not correctly detected in the case of outliers by FastMap.

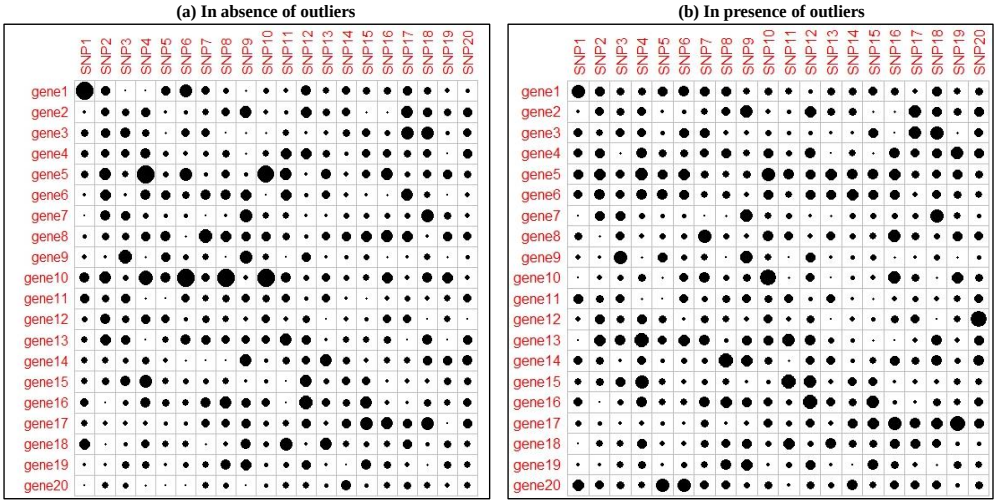


Figure 3: Correlogram for Genes and SNPs (a) In absence of outliers

(b) In presence of outliers

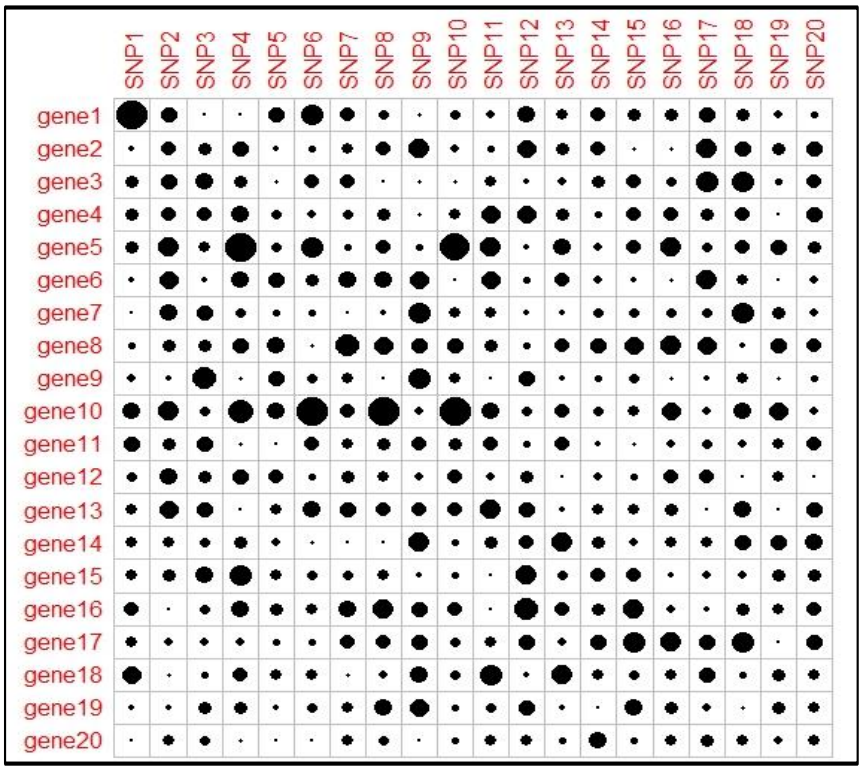


Figure 4: Correlogram for Genes and SNPs after modification of outliers.

Again we generated the data after modification of outliers with the same genes and SNPs position as mentioned above to see the genes and SNPs correlation. We modified our dataset by our proposed method. After modification of dataset, we apply FastMap and see the result, which is given in Figure 4. In the case of outliers modification, we see that the correlation results between genes and corresponding SNPs are similar to the original dataset (Figure 4). From Figure 4, we see that the correlation between genes and SNPs are like as an original dataset correlation. So, we conclude that our proposed method performing well after modification dataset.

4. Conclusions

In this paper, we discuss the proposed SNP based Robust Fast eQTL mapping for the identification of essential genes. We discuss classical SNP based FastMap approach and our proposed methods. For outlier modification, we use the minimum β-divergence method. First, we detect the outlier by β-weight function then we replace outlier data points by its respective robust mean produced by the minimum β-divergence method. The value of the tuning parameter β plays a key role in the performance of the proposed method. We discuss the simulation study of the proposed method in comparison with FastMap-eQTL mapping method. The parameters of this model are estimated by maximizing β-likelihood function. The value of the tuning parameter β plays a key role in the performance of the proposed method. An appropriate value for the tuning parameter β is selected by cross-validation. Simulation studies show that the proposed method significantly improves the performance over the FastMap-eQTL in the absence of outlier. However, in the presence of outlier, our proposed method is performing better than the FastMap method to detect the important gene correctly in case of single and two SNPs. From correlation results between genes and SNPs, we conclude that our proposed method performs better than the classical FastMap method.

References

[1] Brem, R.B., Yvert, G., Clinton, R. and Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast, Science, 296, 752–755.

[2] Broman, K. et. al. (2003). R/qtl: QTL mapping in experimental crosses, Bioinformatics, 19, 889–890.

[3] Carlborg, O., De koning, D.J., Manly, K.F., Chesler, E., Williams, R.W., Haley, C.S. (2004). Methodological aspects of the genetic dissection of gene expression, Bioinformatics, 21, 2383–2393.

[4] Cervino, A.C., Li, G., Edwards, S., Zhu, J., Lauric, C., Tokiwa, G., Lum, P.Y., Wang, S., Castellani, L.W., Lusis, A.J., Carison, S., Sachs, A.B. and Schadt, E.E. (2005). Integrating QTL and high-density SNP analyses in mice to identify Insig2 as a susceptibility gene for plasma cholesterol levels, Genomics, 86(5), 505–517.

[5] Chesler, .EJ., Lu, L., Shou, S., Qu, Y., Gu, J., Wang, J., Hsu, H.C., Mountz, J.D., Baldwin, N.E., Langston, M.A., Threadgill, D.W., Manly, K.F. and Williams, R.W. (2005). Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function, Nature Genetics, 37(3), 233–242.

[6] Churchill, G. and Doerge, R. (1994). Empirical threshold values for quantitative trait mapping, Genetics, 138, 963–971.

[7] Frazer, K.A. et al. (2007a). A second generation human haplotype map of over 3.1 million SNPs, Nature, 449, 851–862.

[8] Frazer, K.A. et al. (2007b). A sequence-based variation map of 8.27 million SNPs in inbred mouse strains, Nature, 448, 1050–1053.

[9] Gatti, D.M., Shabalin, A.A., Lam, T.C., Wright, F.A., Rusyn, I. and Nobel, A.B. (2009). FastMap: Fast eQTL mapping in homozygous populations, Bioinformatics, 25(4), 482–489.

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

[11] Hillebrandt, S., Wasmuth, H.E., Weiskirchen, R., Hellerbrand, C., Keppeler, H., Werth, A., Schirin-Sokhan, R., Wilkens, G., Geier, A., Lorenzen, J., Köh, J., Gressner, A.M., Matern, S. and Lammert, F. (2005). Complement

factor 5 is a quantitative trait gene that modifies liver fibrogenesis in mice and humans, Nature Genetics, 37, 835–843.

[12] Kendziorski, C.M,, Chen, M., Yuan, M., Lan, H. and Attie, A.D. (2006). Statistical methods for expression quantitative trait loci (eQTL) mapping, Biometrics, 62,19-27.

[13] Kong, A. and Wright, F. (1994). Asymptotic theory for gene mapping, Proc. Natl Acad. Sci, 91, 9705–9709.

[14] Korte, A. and Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review, Plant Methods, 9:29.

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

[16] Manly, K.F., Cudmore, R.H., Jr., Meer, J.M. (2001). Map Manager QTX, cross-platform software for genetic mapping, Mammalian Genome, 12, 930–932.

[17] Mollah, M.N.H., Sultana, N., Minami, M. and Eguchi, S. (2010). Robust extraction of local structures by the minimum beta-divergence method, Neural Network, 23(2), 226-238.

[18] Morly, M., Molony, C.M., Weber, T.M., Devlin, J.L., Ewens, K.G., Spielman, R.S. and Cheung, V.G. (2004). Genetic analysis of genome-wide variation in human gene expression, Nature, 430, 743–747.

[19] Pletcher, M.T., McClurg, P., Batalov, S., Su, A.I., Barnes, S.W., Lagler, E., Korstanje, R., Wang, X., Nusskern, D., Bogue, M. A., Mural, R.J., Paigen, B. and Wiltshire, T. (2004). Use of a dense single nucleotide polymorphism map for in silico mapping in the mouse, Plos Biology, 2, e393.

[20] Schadt, E.E., Monks, S.A., Drake, T., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G., Linsley, P.S., Mao, M.,

Stoughton, R.B.and Friend, S.H. (2003). Genetics of gene expression surveyed in maize, mouse and man, Nature, 422, 297–302.

[21] Storey, .J and Tibshirani, R. (2003). Statistical significance for genomewide studies, Proc. Natl Acad. Sci., 100, 9440–9445.

[22] Stranger, B.E., Forrest, M.S., Clark, A.G., Minichiello, M.J., Deutsch, S., Lyle, R., Hunt, S., Kahl, B., Antonarakis, S.E., Tavare, S., Deloukas, P. and Dermitzakis, E.T. (2005). Genome-wide associations of gene expression variation in humans, Plos Genetics, 1(6),e78.

[23] Szatkiewicz, J.P., Beane, G.L., Din, Y., Hutchins, L., De Villena, F.P. and Churchill, G.A. (2008). An imputed genotype resource for the laboratory mouse, Mammalian Genome, 19(3), 199–208.

[24] Wang, P., Dawson, J.A., Keller, M.P., Yandell, B.S., Thronberry, N.A., Zhang, M.B., Wan I-M, Schadt, E.E., Attie, A.D. and Kendziorski, C. (2011). A Model Selection Approach for Expression Quantitative Trait Loci (eQTL) Mapping, Genetics, 187(2), 611–621.

[25] Wang, S., Yehya, N., Schadt, E.E., Wang, H., Drake, T.A. and Lusis, A.J. (2006). Genetic and genomic analysis of a fat mass trait with complex inheritance reveals marked sex specificity, Plos Genetics, 2(2), e15.

[26] Wang, X., Korstanje, R., Higgins, D. and Paigen, B. (2004). Haplotype analysis in multiple crosses to identify a QTL gene, Genome Research, 14, 1767-1772.

[27] Wang, J., Williams, R.W. and Manly, K.F. (2003). Web QTL web-based complex trait analysis, Neuroinformatics, 1(4), 299–308.

[28] Yvert, G., Brem, R.B., Whittle, J., Akey, J.M., Foss, E., Smith, E.N., Mackelprang, R. and Kruglyak, L. (2003). Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors, Nature Genetics, 35, 57–64.