MedCrave
Step into the World of Research

Journal of Cancer Prevention & Current Research

Research Article

Open Access

CrossMark

An application of matrix eQTL to billions hypothesis testing to identify expression quantitative trait loci in genome wide association studies of inflammatory bowel disease

Abstract

Introduction: Genome wide association studies (GWAS) have been widely used in recent years to identify new information on genetic variants which are associated with complex trait in many diseases. Advances in identifying the Single Nucleotide Polymorphisms (SNPs) facilitate the study of etiologies of common disorders including cancers, inflammatory bowel diseases (IBD) and colorectal cancer. Variations in gene expression demonstrate that transcript levels of many RNAs behave as heritable quantitative traits. Studying the genetics of gene expression can provide additional power to the roles of GWAS variants. Expression quantitative trait loci (eQTL) mapping links the genome-wide SNPs with RNA expression.

Methods: In this study, we performed expression quantitative trait loci (eQTL) analysis using the Matrix eQTL R package. This technique implements matrix covariance calculation and efficiently runs ANOVA and linear regression analysis for eQTL studies. The statistical test determines the association between SNP and gene expression, where the null hypothesis is no association between genotype and phenotypes. False Discovery Rate (FDR) is used to identify significant cis and trans eQTL and adjust for multiple hypothesis testing.

Results: We applied matrix eQTL to a real data set consisting of 730,256 SNP and 33,298 RNA for 173 samples. SNPs with minor allele frequency (MAF) less than 0.05 and those violating the Hardy_Weinberg equilibrium (HWE), were excluded from the study. In this study, 15,408 cis eQTL and 27,562 trans eQTL are identified at a FDR less than 0.05, corresponding to p value thresholds of 8e-5 and 1e-8, respectively.

Conclusion: We found out that matrix eQTL is a computationally efficient and user friendly method for analysis of eQTL studies. Our application provides insight into the genomic architecture of gene regulation in inflammatory bowel disease (IBD).

Keywords: genome wide, single nucleotide polymorphisms, inflammatory bowel diseases, expression quantitative trait loci

Volume 11 Issue 2 - 2020

Fahimeh Moradi,[1] Morteza Hajihosseini,[1] Elham Khodayari-Moez, Irina Dinu
School of Public Health, University of Alberta, Canada

Correspondence: Irina Dinu, Associate Professor Biostatistics, School of Public Health, University of Alberta, Canada , Tel 7802000631, Email idinu@ualberta.ca

Received: April 04, 2020 | **Published:** April 16, 2020

Introduction

One of the main goals for human genetics is to understand the inherited balance of common, complex diseases and help improve treatment or diagnosis.[1] Development of a complex disease begins with a genetic event in a normal cell. Genome wide association studies (GWAS) have been widely used in recent years to identify the new information on genetic variants which are associated with a complex trait in many diseases. In GWAS, hundreds of thousands of Single Nucleotide Polymorphisms, called SNPs, across the genome are genotyped to investigate the association between SNPs marker and trait. In the past decade GWAS have identified genetic loci, using SNPs that are associated with trait or risk factors. Advances in identifying the SNPs and their utilization as heritable markers facilitate the understanding of genetic basis of disease susceptibility in polygenic disorders. A SNP occurs when a very small minority of a population does not carry the same nucleotide in a specific position in the DNA sequence. Within populations, SNPs can allocate a minor allele frequency (MAF) which is the ratio of chromosomes in the population carrying the less common variant to the one with more common variants. It is noteworthy that SNPs allele can be common in one geographical area or ethnic group and rare in other ones.[2,3] Although it is possible that a specific SNP does not cause a disorder, SNPs act as biological markers and can be associated with some diseases. This association helps scientists discover an individual's genetic predisposition towards developing a disease. When SNPs happen within or near a region, they can play an important role in the disease process by affecting the gene function. Also, if a certain SNP has an association with a trait, then scientists can try to find out the stretches of DNA around the SNP and determine a gene or genes that are responsible for the trait. With the advancement in biotechnologies, scientists can study associations between SNPs and gene expressions, explain heritability in a population and improve understanding of disease mechanisms. The association between SNPs and genes is referred as expression quantitative trait loci.[4]

Expression quantitative trait loci (eQTLs) are the genomic loci that

[1]authors equally contributed to the manuscript

©2020 Moradi et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and build upon your work non-commercially.

influence genomic regions of a gene expression in the sample that was taken from a population of individuals. The concept of genome-wide eQTL was proposed by Jansen. This method is similar to traditional QTL mapping since both determine genomic location.[5] So the gene is mapped to a related SNP region.[6] The eQTL studies are looking to test the association between a locus in genetic variation and expression variation of genes.[7] They are also classified as cis-acting or trans-acting depending on the location of eQTL and distance to the gene. The cis and trans terms are introduced by Haldane for the first time.[8]

The DNA sequence polymorphism causes variation in expression level within or in the gene. The cis acting eQTL considers DNA in a specific location of the gene. So the DNA variation of a gene affects the transcript level of the gene.[9] Another type of eQTL is trans acting or distal, when the variation acts further from a regulated gene. Depending on the regulation of a gene, the trans eQTL can be anywhere in the genome.[10] So there is no physical linkage to a transcript-encoding gene in trans eQTL. Trans eQTLs are the result of polymorphism which changes regulation in the gene.[11] Cis corresponds to the distance between eQTL and a target gene of less than 1Mb. Trans corresponds to a distance larger than 100 kb.[9] In practice, the variation of a gene can be regulated by a mixture of cis and trans eQTL.

One of the challenges in Genome-wide association studies is the large number of hypothesis tests for finding an association between markers with thousands of transcripts.[12] Most of the eQTL studies run a test statistic for each pair consisting of a transcript and aSNP.[13] This is time consuming and needs lots of calculations in large sample size studies. In genome-wide eQTL, the number of transcripts is in the order of tens of thousands and the number of SNPs is in the order of hundreds of thousands. An important challenge with this analysis is computational in nature. Most of the existing methods take days to complete the association between SNPs and gene expressions. Since the dimension of data for recent eQTL studies has increased, there is a need for efficient eQTL analysis methods.[13] Matrix eQTL was designed to address this need. It is two to three times faster than other eQTL/QTL methods and provides efficient eQTL mapping. The method is implemented as a package in statistical software R. It takes advantage of innovative ways to handle large matrix operations built in R software. In this study, we focused on the ability of Matrix eQTL to identify gene-SNP associations in inflammatory bowel disease. In spite of its computational advantages, matrix eQTL is not used much in the study of complex diseases, since its publication as a novel fast eQTL method in 2012. A quick search on PubMed reveals only two applications of this method to studies of regulatory architecture of gene expression variation in liver[14] and colon.[15]

Methods

Matrix eQTL

Here we explain the matrix structure in matrix eQTL. S denotes the genotype matrix and G denotes the gene expression matrix. Each row of these matrices holds different measurements for a single SNP among samples and a single gene across samples, respectively. Also, the samples (columns of S and G matrices) should match. Data matrices are sliced in blocks of up to 10,000 variables. Then gene expression and genotype matrices are standardized. For each pair of blocks, the correlation matrix for a relevant block is calculated. A quick check to see if the absolute value of any correlation exceeds a predefined threshold or not is performed. After the last step, matrix eQTL calculates and reports the test statistic and p-value, only for those gene-SNP associations selected based on their correlations being larger than the threshold.[13]

The Matrix eQTL method can model the influence of genotype when added as categorical (ANOVA model) or linear (least square model) and test association between each SNP and transcript. Also, it is able to test the interaction between genotype and covariate and check for significant associations. Matrix eQTL has the ability to include covariates factors such as gender, clinical variables, population structure and surrogate variables. The user can chose between two different options: simple regression and analysis of variance (ANOVA) model.

Simple linear regression

A simple linear regression is performed for each pair of SNP and gene. We assume a linear relationship between gene expression G and SNPS (coded as 0, 1 and 2 according to the frequency of the minor allele):

$$G = \alpha + \beta S + \varepsilon, \varepsilon \sim iid.\ Nor(0, \sigma^2)$$

The statistical test determines the association between SNP and gene expression. In matrix eQTL method, p-value is not calculated for each pair of SNP-transcript. A threshold is defined first, based on the test statistics, and then p-value is calculated only for those pairs whose test statistics exceed the threshold, leading to faster computational time. In this method the test statistics t, F, R2 and Likelihood Ratio Test (LRT) are equivalent. All of these test statistics can be defined as functions of Pearson correlation r:

$$t = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}}, F = t^2, R^2 = r^2, LRT = -n\log(1-r^2)$$

$$r_{gs} = Corr\left(S, G\right) = \frac{\sum(S_i - \bar{S})(G_i - \bar{G})}{\sqrt{\sum(S_i - \bar{S})^2 (G_i - \bar{G})^2}}$$

Matrix eQTL considers the absolute value of Pearson correlation r as the test statistic, identifies a threshold and looks for significant SNP-transcript associations. Several factors should be considered to define a statistically significant threshold. Sample size and type I error are key factors. Furthermore, the size of dataset under investigation is important in setting a threshold, i.e. for larger datasets, lower threshold values are used. We note that Pearson correlation is scale invariant. To save computation time, genotype and gene expression variables are standardized prior to running eQTL.

ANOVA Model

The eQTL analysis can also be done through ANOVA model. In this approach we considered both additive and dominant effect of the genotype. The genotype variable should be treated as categorical. ANOVA model demonstrates the genotype effect on gene expression and can be written as a linear regression model (S_1 and S_2 are dummy variables for SNPs):

$$G = \alpha + \beta_1 S_1 + \beta_2 S_2 + \varepsilon, \varepsilon \sim iid.\ Nor(0, \sigma^2)$$

The F or LRT statistics can be used for testing joint significance of S_1 and S_2. Similar to the linear regression model, in ANOVA model,

the F and Likelihood Ratio statistics are monotone functions of the Pearson correlation coefficient, and equivalent test statistics. Here are the steps to perform matrix eQTL using ANOVA option:

a. Center variables G, S_1 and S_2

b. S_2 is orthogonalized with respect to S1 for every gene expression:

$$\tilde{S}_2 = S_1 - corr(S_1, S_2)S_1$$

c. Variables S_1 and $S_2$are standardized.

d. The following test statistic is calculated using large matrix R functions:

$$R^2 = corr\left(G, S_1\right)^2 + corr\left(G, S_2\right)^2$$

e. The threshold for R^2 and the p-value is calculated based on the F statistic:

$$F = \frac{(n-k-1)R^2}{k(1-R^2)}$$ where k=2 corresponds to the two regressors S1 and S2.

Advantages of matrix eQTL

Compared to other techniques, Matrix eQTL is faster and can handle huge datasets. Shabalin[13] measured the performance of Matrix eQTL and 6 other tools: Fastmap, Merlin, snpMatrix, R/qtl, PLINK and eMap. The computational time for Matrix eQTL is two to three times faster than these methods and remains unchanged when the covariates are added to the model. Moreover, Matrix eQTL implements FDR to account for multiple comparisons which is separately estimated for both cis and trans eQTLs. To the best of our knowledge, matrix eQTL is the only tool that implements ANOVA model to find associations between gene expression and SNPs.

Multiple hypothesis testing

One of the main concerns in eQTL analysis is multiple hypothesis testing. For a single testing hypothesis, when we reject the null hypothesis because the p-value is less than our threshold, there is a chance that we reject our null hypothesis incorrectly and a false positive error occurs. Various methods have been developed to estimate an overall measure of error for multiple hypotheses such as family-wise error rate (FWER), Bonferroni, and False Discovery Rate (FDR). Bonferroni and FWER have some limitations for gene expression studies. Matrix eQTL uses FDR to control for false positive errors. FDR is defined as the expected proportion of false positives among all the significant tests.[16]

Data description and result

Data description

We applied matrix eQTL to a study of inflammatory bowel disease. The data consist of SNPs and RNA for 173 samples. These samples are obtained from individuals who enrolled at Mount Sinai Hospital in Toronto, Ontario. The cohort contains patients with a diagnosis of ulcerative colitis or familial adenomatous polyposis. A board of clinical information and biospecimens was collected on enrollment, including whole blood for DNA extraction and tissue biopsy specimens for RNA analysis. We downloaded the data from Gene Expression Omnibus (GEO), with accession ID GSE40292. Endoscopical and histological normal tissue biopsies from every eligible subject were obtained. RNA was extracted with the QIAGEN miRNeasy Kit, and mRNA analysis was performed on Affymetrix Human Gene 1.0 ST arrays. Affymetrix GeneChip Command Console produced summarized probe cell intensity data. RNA was background-adjusted, normalized and log transformed with the robust multiarray average algorithm in Affymetrix (17). Also, the genomic DNA was extracted from whole blood samples from the same individuals. Then samples were hybridized to HumanOmniExpress or HumanOmni2.5 Beadchips (Illumina). Arrays werescanned using iSCan system. Data were genotyped with Illumina beadchips.[17]

Processing data for matrix eQTL

We used PLINK to recode the SNP data. PLINK selects the minor and major allele for each SNP data, then recodes SNP values into 0, 1 and 2. According to data quality standard approaches, SNPs were removed if their Minor allele frequency (MAF) was less than 0.05. Also, the SNPs were removed from further analysis if they were not in Hardy-Weinberg equilibrium (P > 10-6). There are 733,202 SNPs and 33,298 RNA for 173 samples. In total, there are 592,645 SNPs remaining in data sets with MAF<0.05 and Hardy_Weinberg equilibrium (P>10-6). We included all of the 173 samples in the study.

Results

After performing quality control and filtering, we examined how SNPs regulate RNA expression. First, we performed the eQTL analysis without considering the gene/SNP location. We used matrix eQTL with ANOVA option. We found 552,011 significant eQTL. Figure 1 shows the eQTL identified from RNA data in all the subjects, using FDR thresholds of 0.01, 0.05, 0.10 and 0.25.

We restricted cis to within 1Mb of transition starting site. The p-value thresholds for cis and trans were set at 8×10^{-5} and 1×10^{-8}, respectively. We found 15,408 cis eQTL and 27,562 trans eQTL. Then we matched the chromosome on RNA and SNPs. The majority of eQTL were identified from distal effect (trans eQTL). Table 1 presented the number of cis and trans eQTL under different FDR values, after matching by chromosome.

Table 1 Cis and trans eQTL for FDR thresholds of 0.01 and 0.05, after matching by chromosome

FDR	Cis/Trans	eQTL	Genes	SNPs
<0.01	Cis	9247	981	7088
	trans	27562	1427	4672
<0.05	Cis	14227	1707	10552
	trans	27562	1427	4672

QQ plots for all p-values, and local and distant p-values are presented in Figure 1. We focused our study on cis eQTL, since they are more biologically interesting. The significant eQTL are matched by chromosome. Figure 4 shows the boxplots for nine significant cis eQTL.

The nine significant cis eQTL are among the top twelve significant eQTL reported by Kabachiev et al.[17] employing the Kruskal Wallis test, a nonparametric alternative to ANOVA to analyze the same dataset. We note that Kruskal Wallis can be conservative, giving p-values considerably higher than ANOVA, especially for larger sample sizes.

This may be the reason the p-values thresholds for cis and trans were set at higher thresholds of $1x10^{-3}$ and $2x10^{-7}$ in,[17] compared to our analysis performed by matrix eQTL, with thresholds of 8x10-5 and 1x10-8, for cis and trans, respectively.

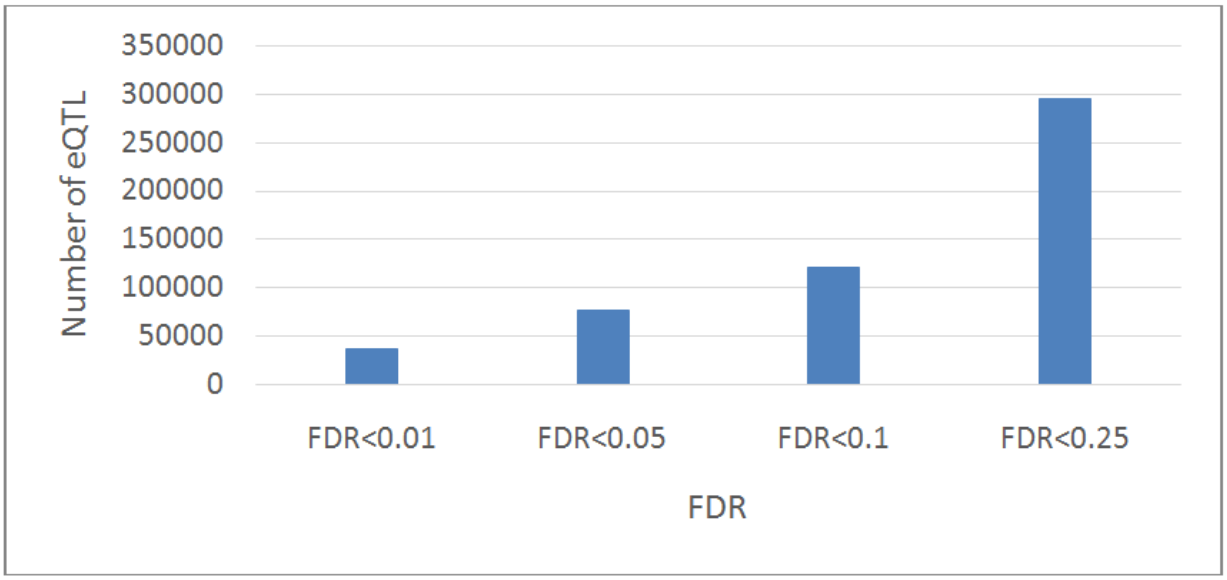
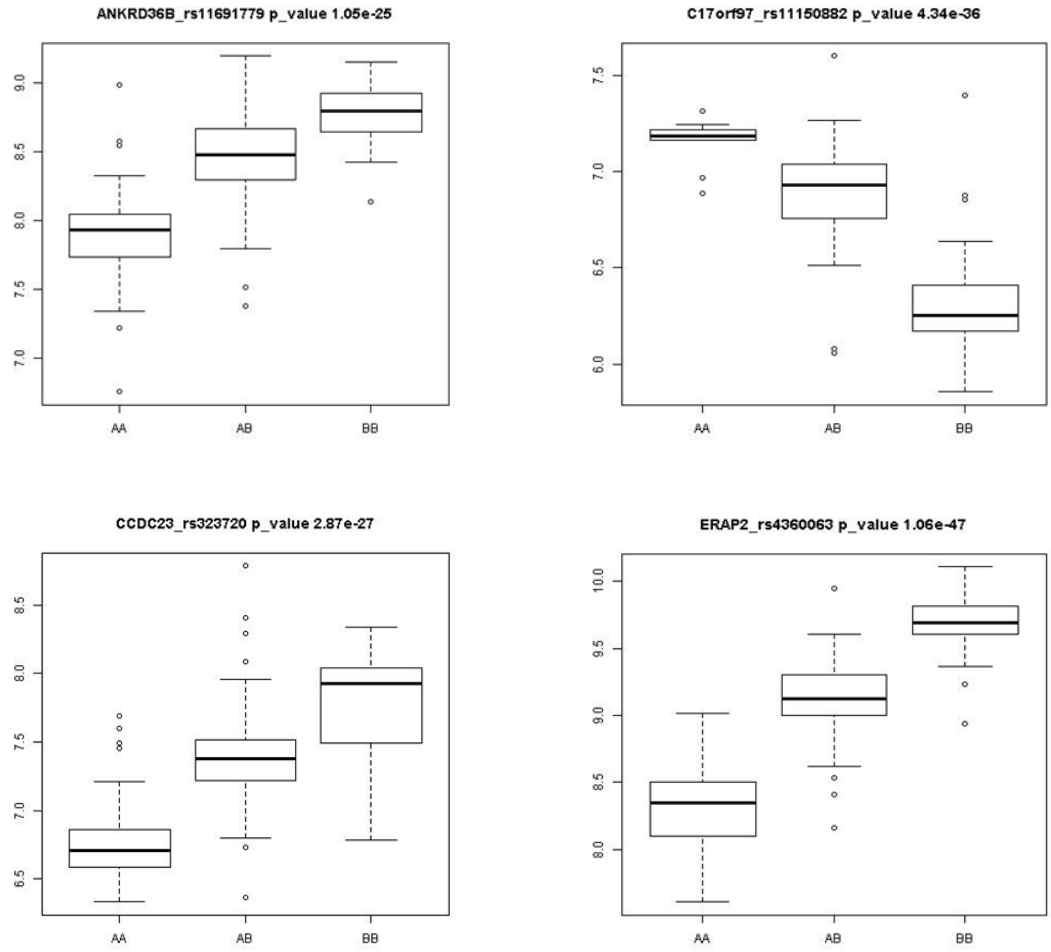


Figure 1 Number of eQTL mapping for various FDR thresholds.



Citation: Moradi F, Hajihosseini M, Khodayari-Moez E, et al. An application of matrix eQTL to billions hypothesis testing to identify expression quantitative trait loci in genome wide association studies of inflammatory bowel disease. *J Cancer Prev Curr Res.* 2020;11(2):41–46. DOI: 10.15406/jcpcr.2020.11.00424

Copyright: ©2020 Moradi et al.

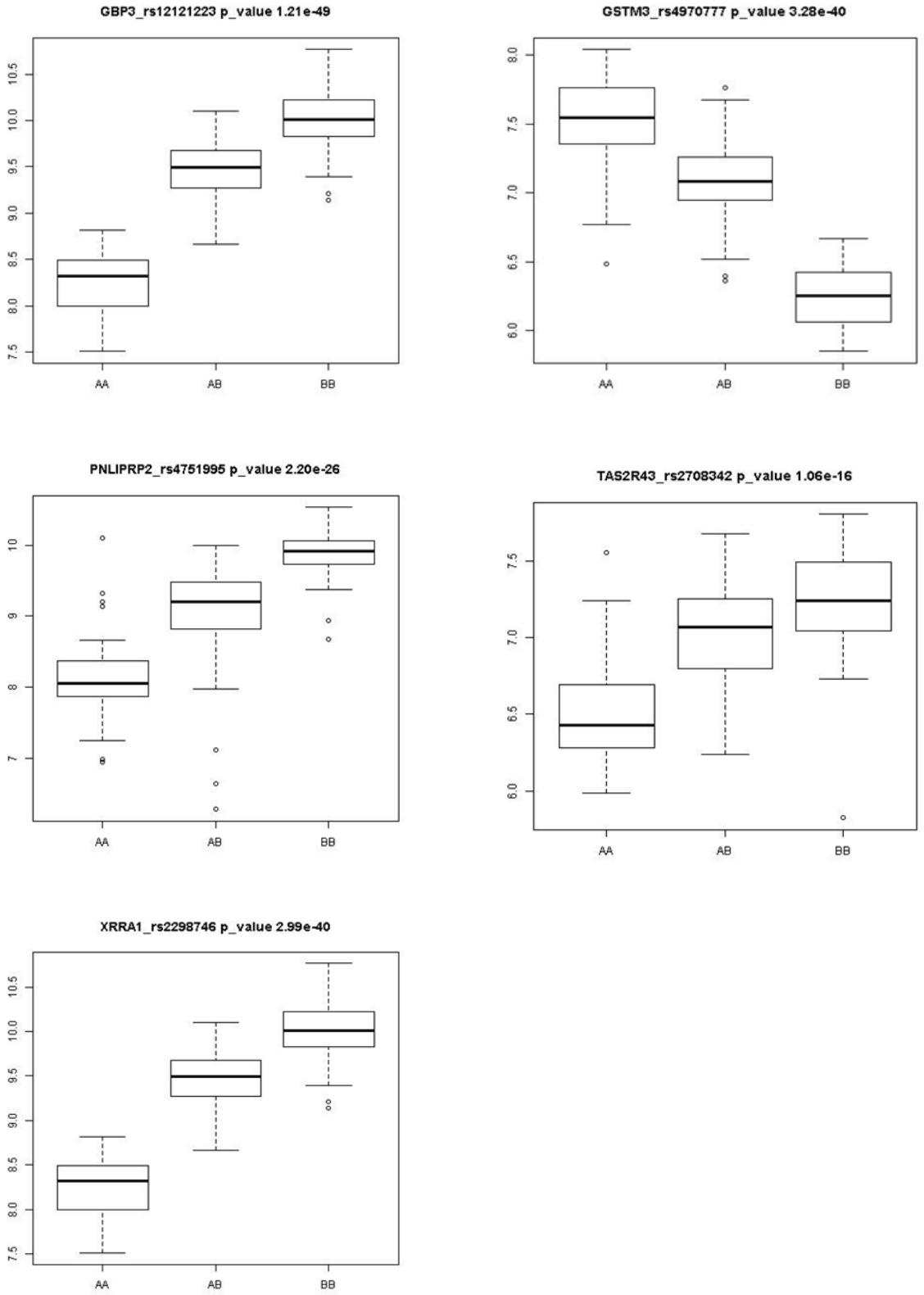


Figure 2 Boxplots of gene- SNP associations for 9 significant cis eQTL.

Citation: Moradi F, Hajihosseini M, Khodayari-Moez E, et al. An application of matrix eQTL to billions hypothesis testing to identify expression quantitative trait loci in genome wide association studies of inflammatory bowel disease. *J Cancer Prev Curr Res.* 2020;11(2):41–46. DOI: 10.15406/jcpcr.2020.11.00424

The top gene identified was endoplasmic reticulum aminopeptidase 2, *ERAP2*. This gene has been previously identified as a potential candidate gene for IBD.[18] Both *ERAP1* and *ERAP2* have been previously associated with human leucocyte antigen diseases, such as ankylosing spondylitis and Behcet's disease, in GWAS studies. The second top gene was X-ray radiation resistance associated 1, *XRRA1*. This gene has been linked to colorectal cancer (CRC).[19] Blocked *XRRA1* expression can lead to cell cycle arrest in CRC. Expressed *XRRA1* can reduce cell cycle arrest and increase cell proliferation in CRC.

Discussion and conclusion

Despite its computational advantages compared to other methods, a search on PubMed indicates only two applications of Matrix eQTL method, in the past six years, after publication. We present here an application of Matrix eQTL to investigate the association between genotype and gene expression in a study of inflammatory bowel disease. In this study, a large dataset is implemented which consists of RNA data for intestinal tissue and SNPs from blood sample in same cohort of individuals. There are various statistical methods and tools to identify eQTL. Most of the eQTL methods take days to complete the analysis. Matrix eQTL takes only a few minutes to run the analysis of SNP and gene expression associations in large datasets. An important step in eQTL analysis is data cleaning and filtering. For instance gene expression data should be normalized. Also, the marker should pass filtering missing value, MAF, and HWE. SNPs were filtered by PLINK if their MAF were less than 0.05 and HWE ($P > 10-6$). After data processing, the gene expressionsassociated with SNPs can be identify using ANOVA model option in matrix eQTL. The ANOVA model is more flexible and uses more slopes compared to simple linear regression. Since the factors are orthogonal, the interaction terms can be analyzed in a timely manner.

Acknowledgments

The authors thank the reviewers and editor for their useful comments and suggestions to improve the manuscript.

Author contributions

I.D. identified the computational challenge in eQTL analysis; F.M. and E.K. performed literature search on eQTL methods. M.J., F.M. and E.K. performed statistical analysis, interpretation and manuscript writing. All authors approved the final draft.

Conflicts of interest

The authors have no conflicts of interest to declare for this study.

Funding

None.

References

1. Hirschhorn JN. Genome wide association studies--illuminating biologic pathways. *N Engl J Med*. 2009;360(17):1699–1701.
2. (NLM) NLoM. What are single nucleotide polymorphisms (SNPs)? 2016.
3. Assembler DS. Single nucleotide polymorphism analysis and mutation detection. 2018.
4. Petretto E, Mangion J, Dickens NJ, et al. Heritability and tissue specificity of expression quantitative trait loci. *PLoS Genet*. 2006;2(10):e172.
5. Kendziorski C, Chen M, Yuan M, et al. Statistical methods for expression quantitative trait loci (eQTL) mapping. *Biometrics*. 2006;62(1):19–27.
6. Zhang W, Liu JS. From QTL Mapping to eQTL Analysis. Frontiers in Computational and Systems Biology: Springer; 2010:301–329.
7. Gilad Y, Rifkin SA, Pritchard JK. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet*. 2008;24(8):408–415.
8. Haldane JBS. New paths in genetics. George allen & Unwin; 1941.
9. Sieberts SK, Schadt EE. Moving toward a system genetics view of disease. *Mamm Genome*. 2007;18(6-7):389–401.
10. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. *Nat Rev Genet*. 2015;16(4):197–212.
11. Kliebenstein D. Quantitative genomics: analyzing intraspecific variation using global gene expression polymorphisms or eQTLs. *Annu Rev Plant Biol*. 2009;60:93–114.
12. Mackay TF, Stone EA, Ayroles JF. The genetics of quantitative traits: challenges and prospects. *Nat Rev Genet*. 2009;10(8):565–577.
13. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*. 2012;28(10):1353–1358.
14. Strunz T, Grassmann F, Gayán J, et al. A mega-analysis of expression quantitative trait loci (eQTL) provides insight into the regulatory architecture of gene expression variation in liver. *Sci Rep*. 2018;8(1):5865.
15. Guo CC, Wei N, Liang SH, et al. Population-specific genome-wide mapping of expression quantitative trait loci in the colon of Han Chinese. *J Dig Dis*. 2016;17(9):600–609.
16. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society Series B (Methodological)*. 1995;57(1):289–300.
17. Kabakchiev B, Silverberg MS. Expression quantitative trait loci analysis identifies associations between genotype and gene expression in human intestine. *Gastroenterology*. 2013;144(7):1488–1496.
18. Castro-Santos P, Moro-García MA, Marcos-Fernández R, et al. ERAP1 and HLA-C interaction in inflammatory bowel disease in the Spanish population. *Innate Immun*. 2017;23(5):476–481.
19. Wang W, Guo M, Xia X, et al. XRRA1 Targets ATM/CHK1/2-Mediated DNA Repair in Colorectal Cancer. *Biomed Res Int*. 2017;2017:5718968.