Identifying expression quantitative trait loci in genome wide association studies

by

Fahimeh Moradi

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Epidemiology

School of Public Health
University of Alberta

Abstract

Introduction: Genome wide association studies (GWAS) have been widely used in recent years to identify the 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. However, the known SNPs are not sufficient to explain the heritability associated with traits. Variations in gene expression demonstrate that transcript levels of many RNAs behave as heritable quantitative trait. 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.

Objectives: The objective of this thesis is to identify an efficient, statistically sound and user friendly method for analysis of eQTL studies.

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 linear regression analysis. The statistical test determines the association between SNP and gene expression, where the null hypothesis is no association between genotype and phenotypes. In eQTL mapping, the regulative variants are classified as cis and trans, the definition depending on the physical distance between a gene and transcript. A certain genomic distance (e.g. 1 Mb) is defined as the maximum distance at which cis or trans regulatory elements can be located from the gene they

regulate. False discovery rate (FDR) is used to identify significant cis and trans eQTL for multiple testing corrections.

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) $P > 10^{-6}$, 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. The results provide insight into the genomic architecture of gene regulation in inflammatory bowel disease (IBD).

Dedicated to my supportive husband and
my beloved parents

Acknowledgments

I am very thankful to my supervisor, Dr. Irina Dinu for her guidance and support throughout my MSc. program. Specially for having faith in me, giving me the opportunity to work on this project, and granting me freedom to explore my ideas. I really appreciate her constructive feedback and insight for conducting this thesis. I have been extremely lucky to have a supervisor who cared so much about my research work. Her assistance and guidance helped me improving my analysis and thesis writing skills.

I also want to thank Dr. Bei Jiang, Dr. William K. Midodzi and Dr. Devidas Menon for participating on my thesis committee.

I would like to express my sincerest love and gratitude to my husband, parents and my sister and brother, for their unflagging support and endless love, throughout my studies.

Finally, I would like to thank all those who helped and inspired with me throughout my research.

Table of Contents

List of Tables:

List of Figures:

Chapter 1: Introduction

Genome variability has been widely used in the past few years to study its association with different risk diseases. One of the important factors in explanation of the genome variants is to postulate the effects of the gene on various diseases (Nica, 2013). Genome wide association study (GWAS) have been used to identify genetic variants that are associated with complex diseases such as breast cancer, type 2 diabetes, Schizophrenia. However, GWAS explains only a small fraction of the heritability associated with traits/phenotypes. Since tissue specific gene expression is predominantly a heritable trait, associating gene expression with polymorphisms is believed to complement the search for missing heritability in complex disorders. New technology enables collecting gene expression data for disease related tissues. Gene expression analysis is implemented in the study of expression quantitative trait loci (eQTL). eQTLs are genomic regions which contain DNA sequence and influence the expression level of one or more genes. Standard eQTL explains association test between genetic variants and gene expression. Furthermore, new studies show that some of the single nucleotide polymorphisms (SNPs) discovered by GWAS can be an eQTL and can be helpful for the identification of the complex traits. Using GWAS in eQTL mapping is helpful to identify the new loci without any prior knowledge about regulatory regions (Nica, 2013).

1.1 Brief Overview of eQTL mapping

First step in eQTL mapping is the measurement of the gene expression in specific cells of multiple individuals. eQTL identification requires two different kind of data: firstly, individuals should be genotyped using SNPs microarray. Secondly, the expression of the gene in genome is measured using expression microarrays or RNA sequencing. A statistical test should be performed to test the effect of a specific SNP on gene expression. Over the past few years, a large number of studies are performed in eQTL to find the significant associations. Shabalin (2012) have studied the association between gene expression and SNPs using linear regression analysis, and ANOVA. Others have investigated the association using nonlinear models such as generalized linear model (Hernandez et al., 2012), mixed effects model (Kanget al., 2008) and Bayesian regression (Servin and Stephens, 2007; Bottolo et al., 2011; Chipman et al., 2011; Stegle et al., 2011). Several techniques have been developed to find the association between a group of SNPs and expression of each transcript (Zeng, 1993; Kaoet al., 1999; Hoggartet al., 2008, Leeet al., 2008; Michealson et al., 2009).

1.2 Challenges in Analytic Methods for eQTL

High dimensionality in eQTL data imposes significant computational issues in eQTL analysis. The modern eQTL studies contain over one million of SNPs and gene expression and the analysis includes over billions of tests. Shabalin (2012) has presented a fast eQTL analysis tool (Matrix eQTL) for high dimensional data that is 2 to 3 times faster than any existing QTL/eQTL algorithm.

1.3 Contributions

The existing eQTL methods take a lot of days to complete analysis. The modern eQTL data sets include over million SNPs and gene expression. So the number of association tests will exceed billions. It is very important to find an efficient method which can handle large data set and run billion association tests. Most of the proposed methods take a long time to complete the analysis. We focus on matrix eQTL, as it is computationally efficient and user friendly. In this study, Matrix eQTL has been used for the analysis of large data set of 173 samples. This study confirmed eQTL signals in inflammatory bowel diseases (IBD), while also identifying additional eQTLs unique to our study data. The identified significant cis and trans eQTLs enables us to specify genes that regulate IBD. The performance of Matrix eQTL tool was evaluated on a real SNPs and RNA dataset for patients with IBD.

1.4 Thesis Organization

This thesis is organized based on 6 chapters. The background on GWAS and eQTL is explained in chapter 2. Discussions regarding concepts of gene expression, RNA, SNP cis and trans eQTL as well as the statistical and computational challenges in eQTL analysis are presented here. Chapter 3 presents the matrix eQTL method and explains the details of this technique to determine significant eQTL. In chapter4, multiple comparisons, Bonferroni, and False discovery adjustment are discussed. In chapter 5, data preparation,

handling, and analysis are presented. Also, chapter 5 describes the results of the study. Chapter 6 explains the discussion and suggests future work.

Chapter 2: Background

In this chapter, we review the Genome wide association study. Then we describe the expression quantitative trait loci (eQTL) method. We review the cis and trans eQTL. Then we discuss statistical and computational challenges in eQTL methods. In the last section, we briefly compare the different tools for eQTL analysis.

2.1 Brief Overview of Genome_wide Association Studies

One of the main goals for human genetics is to understand the inherited balance of common, complex diseases and help to improve treatment or diagnosis (Hirschhorn, 2009).

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 the complex trait in many diseases. In genome_wide association studies hundreds of thousands of 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 Single Nucleotide Polymorphisms (SNPs) and their utilization as heritable markers facilitate the understanding of genetic basis of disease susceptibility in polygenic disorders.

GWAS include a good knowledge of common genetic variation as well as diagnosing the phenotype or its measurement. For these studies, a large number

of subjects are required, with cases and controls in the order of thousands. For this process, genotyping the SNPs is essential. The large SNPs set will be available through chip base microarray technology. Two companies, Affymetrix and Illumina provide chips with high accuracy and low cost (Bush et al., 2012). With the advances in quality control technology, data cleaning can be performed in an affordable and timely manner. Quality control helps to take away those samples such that the SNPs or DNA cases are not related (Psychiatric GWAS Consortium Coordinating Committee, 2009).

The selected well-defined phenotypes are used for statistical analysis and evaluation of the association with a disease or trait. According to study design, the p_value for the genome_wide significance threshold should be estimated to be less than 1×10^{-8} (Hardly et al., 2009).

Using the statistical analysis or biological credibility, significant SNPs or loci are replicated. This replication is essential for having a reliable association between SNPs and diseases in GWAS. In other words, GWAS is replicated for independent samples. The size of the sample for replication cohort can be the same or larger than the size of samples observed in the GWAS. Based on the statistical test results and imposing a threshold limit, the association between the loci and the disease is quantified. Data mining is further applied to investigate the true genetic association. Also, fine mapping of the genetic regions will find the new variants and identify those variants that contribute to the association with the disease (Manolio, 2010; Hardly et al., 2009).

2.2 SNP and gene expression

2.2.1 Brief overview of SNP

DNA sequence variation occurs when a single nucleotide (A, T, C, G) changes in the genome sequence. Everyone has many single nucleotide polymorphisms which together create a unique pattern of DNA for that person (Nature, 2016).

Single nucleotide polymorphisms, called (SNPs), are variation when a single nucleotide (A, T, C, and G) in the genome or other sequences differ between individuals. Figure 2-1 shows how DNA strand 1 differs from DNA strand 2 at a single polymorphism (International Society of Genetic Genealogy). SNPs occur throughout an individual‘s DNA. 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; this variation is called a SNP. 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 (National Library of Medicine, 2016; DNA Sequence Assembler, 2016).

Although it is possible that a specific SNP does not cause a disorder, SNPs act as a biological marker and can be associated with some diseases. This association helps scientists to discover an individual’s genetic predisposition towards developing a disease. When SNPs happen within or near a region gene, they can have 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.

SNPs can belong to a coding sequence of genes, a noncoding region of a gene or in the intergenic region between genes.

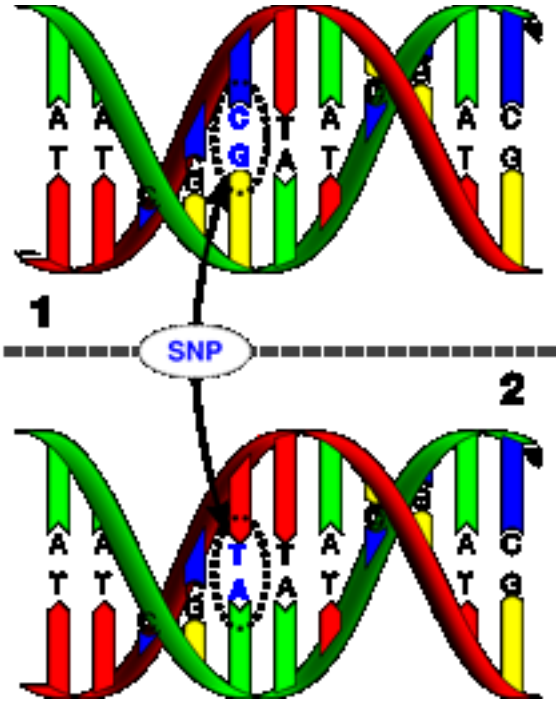


Figure 2-1Single nucleotide polymorphism (SNPs) [ISOGG Wiki]

2.2.2 Gene expression

A Gene is a small unit of genetic material written in code and carrying information from one generation to the next one. Genes are not used by an organism so they need to turn to a gene product. Gene expression is a process where information from a gene is used to direct protein synthesis and creates gene products. These gene products are mostly protein, but there are some non-

protein genes as well. Genes can be expressed as protein or RNA structure. Expressed genes consist of the genes that are transcribed into messenger RNA (mRNA) and then translate to protein. Some genes are involved in production of another format of RNA such as transfer RNA (tRNA) and ribosomal RNA (rRNA) (Nature, 2016).

Transcription is the process where a transcript mRNA of a gene is formed by the enzyme of a polymerase which creates an anti parallel RNA strand. In translation, the RNA sequence converts to a linear series of amino acids in a protein product. Cells are able to control which genes get transcribed and which transcript gets translated.

In genetics, gene expression is the important level where genotype instigates phenotype. The gene expression interprets the genetic code stored in DNA as a form of nucleotide sequence. So gene expression and genotype can be associated and help improve our understanding of genetic diseases.

2.2.3 RNA

Ribonucleic acid (RNA) is a polymeric molecule which is made of one or more nucleotide. These smaller nucleotides are called ribonucleotide bases: adenine (A), cytosine(C), guanine (G), uracil(U) , a ribose sugar, and a phosphate. The RNA structure looks like DNA nucleotide and it carries the same information as its DNA. So RNA is always compared with a reference or template (Nature, 2016).

During transcription process, RNA polymerase synthesizes RNA from DNA. Then the new RNA sequences are complementary to corresponding DNA template. After that RNA translates to protein by the ribosome. RNA plays an important role in the pathway from protein to DNA. Three types of RNA are involved in the translation process: mRNA, tRNA, and rRNA..

2.3 Quantitative trait loci method (QTL)

Quantitative trait loci (QTL) are determined by linking trait measurement and molecular markers; these markers are based on DNA polymorphisms. QTL mapping detects regions within the genome that contain genes linked to the specific quantitative trait (Collard et al, 2005). New technology provides more information, so the association between SNPs and gene expression will explain heritability in population and help study diseases. This association is referred as expression Quantitative trait loci (eQTL) (Petretto et al., 2006).

2.4 Expression quantitative trait loci method

Expression quantitative trait loci (eQTLs) are the genomic loci that influence genomic regions of a gene expression in the sample that was taken from a population of different individuals. The idea of genome-wide eQTL was proposed by Jansen et al. in 2001.

The expression of thousands of genes can simultaneously be measured with DNA microarrays, and latest sequencing-based profiling methods, which give a comprehensive portrait of cellular activity. The heritability of microarray gene expression traits in different species has been shown in a number of studies

(Yang et al., 2014). Two types of data are required from each individual to detect the variants that affect gene expression: high density genotype SNPs and microarray expression data (Albert et al., 2015). Fully sequencing the genome of individuals for known variants or using microarray for the unknown variants is the way to genotyping data. One of the important factors in eQTL is normalizing the microarray data. Normalizing helps to remove the biases from variation in microarray and enables us to compare the different levels of expression. This normalization can be done through Affy package (Gentleman et al., 2004). Also, the genotype data are subjected to quality control procedures to minimize false positive errors. Primarily, samples which have more untyped SNPs of a general cut-off value (e.g. 5%) are removed as their DNAs have low quality (Li et al., 2012). Also, it is necessary to determine the missing rate of SNPs. This rate is calculated based on the rest of the samples. Since missing rates cause uncertainty in eQTL, SNPs with a high proportion of missing rates should be filtered. Then in the final step, the SNPs with minor allele frequencies (MAF) less than a certain threshold (usually 0.05) are removed (Li et al., 2012).

eQTL mapping is similar to traditional QTL mapping since both determine genomic location (Kendziorski et al., 2006). So the gene is mapped to a related SNP region (Zhang et al., 2010). The eQTL studies are looking to test the association between a locus in genetic variation and expression variation of genes (Gilad et al., 2008). eQTL significance is assessed via a p_value of a null hypothesis test, and the log-odd score (LOD). It is essential to determine the prior significant threshold for values of test statistics (Abiola et al., 2003). Figure

2-2 describes the eQTL mapping as an association between SNPs and gene expression. eQTL mapping helps understand complex diseases mechanisms.

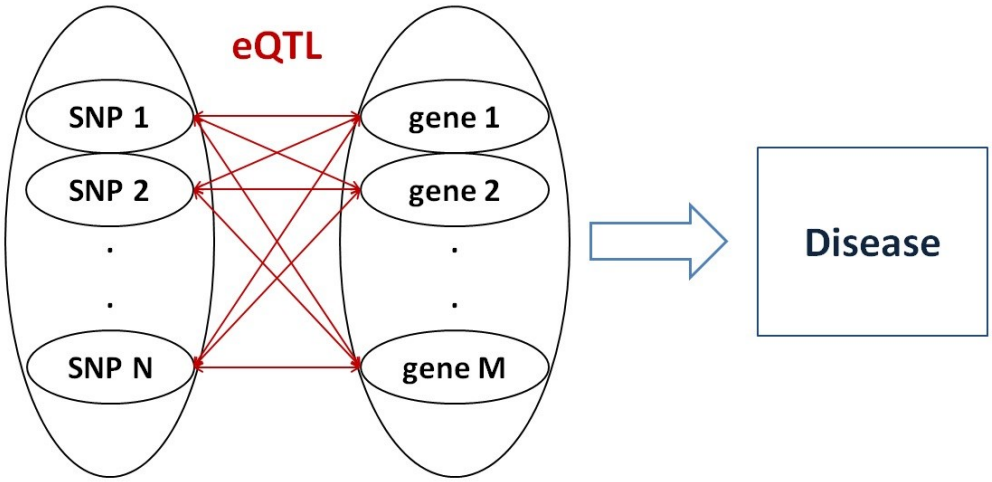


Figure 2-2 eQTL mapping.

2.5 Cis and trans eQTL

eQTLs are classified as cis-acting or trans-acting depending on the location of eQTL and distance to the gene (Li et al., 2012). The cis and trans terms are introduced by Haldane for the first time (Haldane, 1942).

Expression plays as a quantitative trait in eQTL mapping and when they belong to a genomic location in the gene, they are called eQTL. The DNA sequence polymorphism causes variation in expression level within or in the gene, the variation in gene expression causes variation in a gene or another gene. Then DNA sequence polymorphism affects gene expression called cis. But true cis_acting eQTL consider DNA in a specific location of the gene. So the DNA variation of a gene affects transcript level of the gene (Sieberts et al., 2007).

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 (Albert, 2015). 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 (Kliebenstein, 2009).

Distance based definition can have some problem with variation near gene count as cis or trans. So the distance should be defined clearly. Cis is classified when the distance of eQTL from a target gene is less than 1 Mb and for further than 100 kb is counted as trans (Sieberts et al., 2007: Li et al., 2012).

In reality, the variation of a gene can be regulated by a mixture of cis and trans eQTL.

Different eQTL studies need different sample size which their sample size differs from ten to hundred. Finding more trans eQTL depends on effect size, large sample size and allelic variation. Larger sample size gives a better estimate with existing statistical methods but high expenses should be considered too (Li et al., 2012: Gilad et al., 2008). Figure 2-3 shows how the eQTL classifies depend on the regulatory regions of a gene.

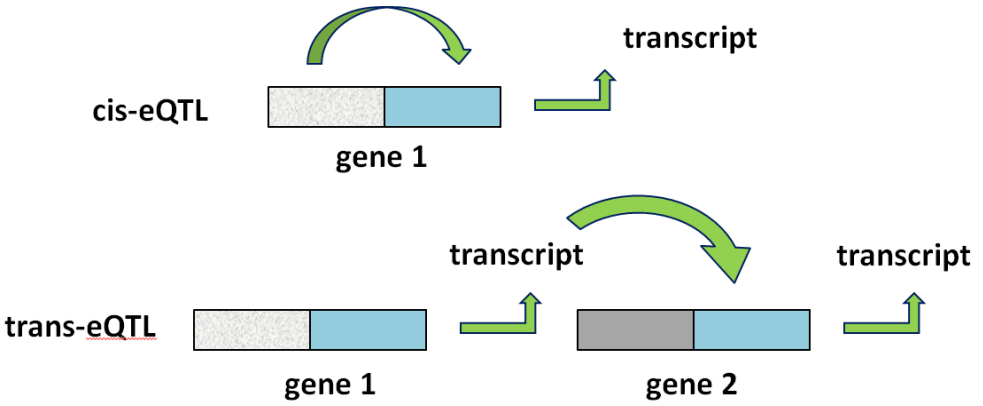


Figure 2-3cis_eQTL and trans_eQTL (Adopted from MacLellan et al. 2012).

2.6 eQTL mapping methods

There are different methods that can be used to find the association in eQTL mapping. R/eqtl and matrix eQTL methods apply linear regression and find an eQTL association by t-test. Another method is likelihood ratio test which identifies association by the model fitness. Also, a non parametric method such as Wilcoxon's rank test can be used to find eQTL.

2.7 Statistical challenges in analysis of eQTL

One of the main goals of eQTL is to find the association between SNPs and gene expression. Different populations were required in the first eQTL studies. The first eQTL studies overcame the low density of genotype by using different populations (Wright et al., 2012). But recent technology in microarray and sequencing can find the variety of genetic variation. However, when there are more than 1 million of SNPs and over tens of thousands of transcripts, billion statistical tests are needed (Shabalin, 2012). This huge data motivated eQTL

analysis, which came with various statistical challenges. One of these challenges is the large number of hypothesis tests for finding an association between markers with thousands of transcripts (Mackay et al., 2009). Most of the eQTL studies run a test statistic for each pair of transcript andSNP (Shabalin, 2012). This association can be tested through linear regression, ANOVA models or nonlinear techniques such as generalized linear and mixed model, Bayesian regression. Some studies criticize the use of nonlinear methods, because they cannot efficiently deal with the large sample size (Degnan et al., 2008). So it is essential to find a method that can work with large data set and run a billion tests. Also, multiple comparisons should be considered.

2.8 Computational challenges in analysis of eQTL

New microarray and sequencing technology provides huge data and help to identify the genetic variation and gene expression in the genome (Tian et al., 2014). There are a large number of gene expression and genetic variants in genome-wide eQTL. An important challenge is computational in nature. Most of the recent methods take a lot of time to complete the association between SNPs and gene expression.

Some studies considered a smaller sub-set of data and showed that nonlinear methods would be very slow even for these smaller data sets. These studies tested the association for transcript and SNP pairs, which usually take a couple of hours to finish the analysis. Since the dimension of data for recent eQTL studies has increased, there is a need for efficient eQTL analysis methods (Shabalin, 2012).

2.8.1 Comparing different tools for eQTL analysis

There are different software tools for eQTL analysis. We provide here a description and discuss limitations for each tool.

R/QTL is an R package for eQTL and QTL mapping. Initially, R/QTL was used for QTL but later it was extended to perform eQTL analysis. R/QTL uses hidden Markov model (HMM) technology for eQTL mapping (Broman et al., 2003). The HMM model can deal with missing genotype data. R/QTL under normal model can perform the one or two eQTL scan. Also, it can be used for imputation and genotyping error correction. The R package 'eqtl' can accommodate covariates and uses false discovery rate for correcting multiple comparisons in multiple eQTL.

Merlin (Multipoint Engine for Rapid Likelihood Inference) is a suitable tool for pedigree analysis. It uses sparse inheritance trees. Merlin can perform QTL analysis, linkage analysis, genotype error detection and haplotyping (Abecasis et al., 2002). eQTL analysis can be done in both population-based study and family-based analysis (Wright et al., 2012). Population stratification cannot be controlled for in Merlin. There are some options to avoid this problem in eQTL analysis, such as adjusting for population membership as a covariate or performing stratification before analysis (Tian et al., 2014). Also, Merlin shows slower performance when running in fast association mode (Tian et al., 2014).

eMap is an R package for QTL/eQTL analysis. The computational part is written in C and it requires installation of GSL library so it does not run in windows. It uses linear regression to find an association between gene expression and genetic

marker in eQTL (Sun, 2010). The model selection is based on backward selection, visualization and finding an eQTL hotspot (Tian et al., 2014). But for large data set, it shows average presentation.

PLINL is a tool for genome-wide association study (GWAS) and it is designed to handle large data set. PLINK performs different aspects such as data management, summary statistics, association analysis, population stratification and identify by decent estimation (Purcell et al., 2007). This package applies a compact binary file to show the SNP data. It also has the ability to merge two or more data sets and filter data. PLINK has the ability to give summary statistics like genotyping rates, allele and genotype frequencies, Hardy-Weinberg equilibrium tests (Purcell et al., 2007). It is essential to mention that PLINK is not ideal for multiple traits. The speed of PLINK decreases significantly when covariates are added to the model. It is ten times slower even when there is just a single covariate added to the model (Wright et al., 2012). PLINK performs multiple test corrections, such as Bonferroni correction and FDR (Purcell et al., 2007).

FastMap is a java based desktop software package which is used for eQTL analysis. It uses association mapping for population-based studies and it is designed for fast eQTL analysis. This fast calculation is made possible by the Subset Summation Tree (Gatti et al., 2008: Tian et al., 2014). Also, it has a graphical user interface. This package calculates the significant threshold and p_value for each gene. FastMap also calculates a FDR q-value via permutation

testing, to address the multiple comparison problems. FastMap does not accommodate covariates (Shabalin, 2012; Wright et al., 2012).

snpMatrix is R/Bioconductor package for eQTL/QTL analysis. It uses linear regression method for association analysis (Leung, 2007). It is computationally efficient and can handle covariates. snpMatrix supports different kinds of data formats such as HAP Map and PLINK (Shabalin, 2012; Wright et al., 2012).

In the next chapter we describe an existing computationally efficient eQTL method called matrix eQTL. We explain the advantages of this method over other methods. We chose this method over others to apply on analysis of a real dataset.

Chapter 3: Matrix eQTL

Matrix eQTL is a new tool for fast eQTL analysis. It provides efficient eQTL mapping. Matrix eQTL is 2-3 times faster than other eQTL/QTL tools. The matrix eQTL implements the matrix algorithm in user friendly software such as R.

Matrix eQTL 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. We describe below the matrix eQTL method, following Shabalin (2012).

3.1 Matrix eQTL Method

The primary step in eQTL analysis is finding the appropriate model to use. Matrix eQTL applies an algorithm and matrix operation and includes two different options. Simple regression is one of the options. It does not include covariates in the model and assumes uncorrelated homoskedastic errors. Another option is analysis of variance (ANOVA) model. It can handle covariates and heteroskedastic and correlated errors.

3.1.1 Simple Linear Regression

There are different approaches to eQTL analysis. One of the common ones is a 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 SNP *s*:

$$g = \alpha + \beta s + \varepsilon \quad where \quad \varepsilon \sim i.i.d. \quad N(0, \sigma^2). \qquad 3\text{-}1$$

SNPs are coded as 0, 1 and 2 according to the frequency of the minor allele.

Different variables are calculated in typical simple linear regression (Shabalin, 2012). These variables are sample mean. A test statistic is calculated. This test statistic can be a t_test, F_test or a likelihood ratio (LR) test. The statistical test determines the association between SNP and gene expression, where the null hypothesis is no association between genotype and phenotypes. Then p value is calculated for the test statistic. However, in matrix eQTL, p value is not calculated for each pair of SNP-transcript. A threshold is defined first, based on the test statistics. p_value is calculated only for those pairs whose test statistics exceed the threshold, leading to faster computational time.

One of the important factors is choosing a computationally efficient, yet powerful test statistic. We note that for simple linear regression in Equation **3-1**, the test statistics t, F, R^2 and LR are equivalent. All of these test statistics can be defined as functions of Pearson correlation $r = cor(g, s)$:

$$t = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}},$$

$$F = t^2 = (n-2)\frac{r^2}{1-r^2},.$$

$$R^2 = r^2,$$

$$LR = -n \quad \log(1-r^2) \qquad \text{3-2}$$

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 (n) and type I error (α) 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.

It is essential to mention that the standardization of genotype and gene expression variables does not affect the correlation value, but saves computation time.

The standardization is implemented once for each gene expression and genotype. So the sample correlation is calculated as the inner product between vectors *s* and *g*, after standardization:

$$r_{gs} = cor(s,g) = \frac{\sum (s_i - \bar{s})(g_i - \bar{g})}{\sqrt{\sum (s_i - \bar{s})^2 (g_i - \bar{g})^2}} = \sum s_i g_i = \langle s, g \rangle . \qquad \text{3-3}$$

3.1.2 ANOVA Model

eQTL analysis can also be done through ANOVA model. In this approach we consider 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:

$$g = \alpha + \beta_1 s_1 + \beta_2 s_2 + \varepsilon . \qquad 3\text{-}4$$

where s_1 and s_2 are dummy variables for SNPs. For testing joint significance of s_1 and s_2, the F statistic or LR can be used, as they are equivalent. Also for ANOVA model, the F statistic and LR are monotone functions of R^2, which is used by matrix eQTL with ANOVA option as test statistics. Then we can orthogonalize the regressors for calculating the test statistics. We describe here matrix eQTL algorithm:

I. Center variables g, s_1 and s_2 are assigned to eliminate α from the model.

II. Then variable s_2 is orthogonalized with respect to s_1 for every gene expression:

$$\tilde{s}_2 = s_1 - \langle s_1, s_2 \rangle s_1$$

III. Variable s_1 and s_2 are standardized.

IV. Test statistic of $R^2 = \langle g, s \rangle^2 + \langle g, \tilde{s}_2 \rangle^2$ are calculated using large matrix R functions.

V. 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 $k=2$, the number of regressors (s_1 and s_2).

This model can be generalized for testing joint significance of any subset of regressors.

3.1.3 Steps of Matrix eQTL Algorithm

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. Here we explained how Algorithm of Matrix eQTL in simple linear regression works. 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, then check if the absolute value of any correlation exceeds a predefined threshold or not. After the last step, matrix eQTL calculates and reports the test statistic, p-value, for Data matrices that are sliced in blocks of up to 10,000 variables meaningful correlations (Shabalin, 2012).

Figure 3-1 shows that the large matrices are sliced in blocks and then multiplied to each other into the correlation matrix.

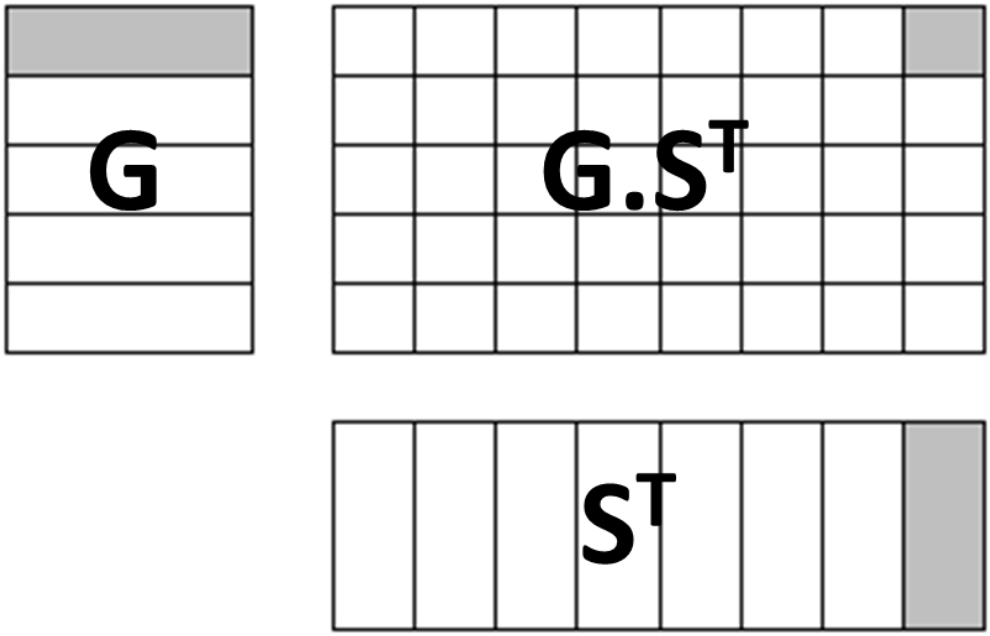


Figure 3-1Calculation of matrix of correlation in 10,000 × 10,000 blocks (from Shabalin, 2012).

3.2 Advantages of matrix eQTL over other existing methods

Compared to other techniques, Matrix eQTL is faster and can handle huge datasets. Shabalin (2012) 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 2 to 3 times faster than these methods. The computational time for Matrix eQTL will be unchanged when the covariates are added to the model. Moreover, Matrix eQTL implements FDR to account for multiple comparisons. FDR is separately estimated for both cis and trans eQTLs. Matrix eQTL is the only tool that implements ANOVA model to find associations between gene expression and SNPs.

Chapter 4: Statistical issues in the analysis of high-dimensional data

In this chapter, we review the multiple comparison problems in high dimensional data. Then we explain Bonferroni and False discovery rate methods to deal with adjustment in multiple comparisons.

4.1 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. This is called a type I error. When there is a set of hypotheses and we gather results from many simultaneous tests, we cannot test each hypothesis separately.

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).

FWER is the probability of at least one false rejection among multiple testing. Suppose we have m hypothesis to test in eQTL and type I error is α, then FWER for hypothesis testing is $1-(1-\alpha)^m$. For eQTL analysis, m is large, so this value becomes very high and close to one.

4.2 Bonferroni adjustment

One classic approach to correct for the multiple comparison problems is Bonferroni adjustment.

We assume that there are m different independent hypothesis to test and α is significant level to reject the null hypothesis. The probability of seeing no type I error will be $(1-\alpha)^m$ and the probability of seeing at least one false association would be:

$$1-(1-\alpha)^m. \qquad 4\text{-}1$$

which is approximately equal to $m\alpha$. It is necessary to decrease the false positive error for each test. Carlo Bonferroni suggested applying α/m instead of α for the significant threshold level. After applying the Bonferroni adjustment, we are able to control FWER at α (Glickman et al., 2014).

Although Bonferroni adjustment is used to control type I error and statistically reasonable, it comes with limitations. Firstly, when the null hypothesis is rejected in Bonferroni, at least one test rejects null hypothesis. Bonferroni adjustment is not able to find out which test is really significant. So the statistical power is decreased because of the possibility of incorrectly rejecting a null hypothesis in each test.

Secondly, all of the hypothesis tests are statistically independent in Bonferroni adjustment. When we deal with high dimensional data, this assumption is not true anymore. For example in GWAS data, each SNP is correlated with other SNPs that are close by. So the probability of seeing at least one type I error will be low and the power is reduced. As a result, the Bonferroni adjustment can be too conservative and cause high type I error (Pernger, 1998; Nakagawa, 2004).

4.3 False discovery rate (FDR) adjustment

Another approach for adjusting the multiple comparisons problem is FDR. FDR is defined as expected proportion of true null hypothesis among the entire rejected hypothesis. We assume that we want to test m null hypothesis while m_0 of them are true and R is the number of hypothesis rejected. The Table below shows the property and outcome of multiple testing:

Table 4-1Property of multiple testing.

	Non-significant test	Significant test	total
True null hypothesis	U	V	m_0
Non-true null hypothesis	T	S	$m-m_0$
	$m-R$	R	m

Assume that m hypotheses are known, and m_0 and $m-m_0$ are the numbers of the true and non-true hypotheses, respectively and they are unknown parameters. R is an observable random variable while U, S, T and, V are unobservable random variables (Benjamini et al., 1995). If we assume that each null hypothesis is tested individually at α level, then $R = R(\alpha)$ will increase when α is increasing. FDR by Benjamini and Hochberg (Benjamini et al., 1995) is defined as the expected proportion of false positives among all of the significant tests:

$$FDR = E(Q) = E(V/R). \qquad 4\text{-}2$$

Q=0 is defined where V+S=0, meaning that no error of false rejection can happen. Since U and V are unobservable variables, Q is unknown random variable (Benjamini et al., 1995).

Since the FDR calculates the number of Types I errors among the rejected hypothesis, when we use FDR, there is much smaller number of false positives (Benjamini, 2001).

4.3.1 False discovery rate procedure

Benjamini and Hochberg (Benjamini et al., 1995) proposed a multiple comparison procedure that controls FDR at q level.

If all of the null hypotheses are true then the FDR would be same as the probability of making no error. So for controlling the FDR, It is necessary to choose q at predictable level for α. Also the FDR would be smaller when some of the hypotheses are true and some are false. In controlling the FDR and when many of the hypotheses are rejected, we prefer to control proportion of errors instead of the probability of making one error (Benjamini et al., 2005)

Assume that there are $\mathrm{H}_1, \mathrm{H}_2, ... \mathrm{H}_\mathrm{m}$ hypothesis and we want to test them based on the corresponding p values $P_1, P_2, ..., P_m$. Let $P_{(1)} \le P_{(2)} \le ... \le P_{(m)}$ represent the ordered p values and assume they are independent. Define the Bonferroni type multiple testing procedure:

$$k = \max\left\{ i : P_{(i)} \le \frac{i}{m} q \right\}. \qquad 4\text{-}3$$

For $i = 1,2,...,k$ and reject $H_1^0, H_2^0, ..., H_k^0$.

So if there is no such *i*, then no hypothesis is rejected.

This procedure does not control FWER at q level when $m < m_0$ (Hommel, 1988).

To control FWER, same stepwise approach can be applied, but each $P_{(i)}$ is compared to $\frac{q}{m-i+1}$ (Benjamini et al., 2001; Benjamini et al., 1995).

Chapter 5: Data Description and Results

In this chapter, we described intestine data the RNA and SNPs that we used in this analysis. First, we explained the approach for processing data for analysis. In the result section, we report significant gene and SNPs association with their p-values.

5.1 Data Description

We apply our method on finding significant eQTL on an intestine data. The data consist of SNPs and RNA for 173 samples. Kabakchiev et al. (2013) have used a similar raw dataset for eQTL analysis. These samples are obtained from individuals who enrolled at Mount Sinai Hospital in Toronto, Ontario and they went under ileal pouch_anal anastomosis following colectomy. The cohort contains the samples with a diagnosis of ulcerative colitis or familial adenomatous polyposis. The individuals were enrolled at least 1 year after the end of their ileostomy. A board of clinical information and biospecimens is 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. At the end, the probe-level summarization files were

made and the RNA was background-adjusted, normalized and log transformed with the robust multiarray average algorithm in Affymetrix (Kabakchiev etal., 2013). In Table 5-1, we show an example of our RNA data which represents the RNA in rows and samples in the column. Each cell represents the measurement of RNA level.

Table 5-1An example of RNA data set.

Gene id	Sam_01	Sam_02	Sam_03	⋯	Sam_173
Gene 01	5.136644695	2.991858662	3.712783363	⋯	4.316683701
Gene 02	5.322985371	4.79872188	5.487924226	⋯	4.952076614
Gene 03	3.950133943	3.345143555	5.067203736	⋯	
⋮	⋮	⋮	⋮	⋱	⋮
Gene K	6.532202227	6.292107259	6.244795865	⋯	6.590339963

Also, the genomic DNA was extracted from whole blood samples from the same individuals. Separating and lysing white blood cells for purification of DNA from blood bio specimens is done by GentraPuregene Blood kit (Qiagen). Those extracted DNA were normalized at 50 ng/μL. Then samples are hybridized to HumanOmniExpress or HumanOmni2.5 Beadchips (Illumina). Arrays are scanned using iSCan system. Genome Studio (Illumina) called the genotypes. Data were genotyped with Illumina beadchips (Kabakchiev et al., 2013). An example of SNP data is given in Table 5-2.

Table 5-2An example of SNPs data set.

SNP id	Sam_01	Sam_02	Sam_03	⋯	Sam_173
SNP 01	BB	AB	AB	⋯	BB
SNP 02	AB	BB	BB	⋯	BB
⋮	⋮	⋮	⋮	⋱	⋮
SNP K	AA	AA	AB	⋯	BB

5.1.1 Minor allele frequency

SNPs are defined as single nucleotide variation in the DNA code. Minor allele frequency is the frequency of the second most frequent allele value occurring in a given population.

5.1.2 Hardy-Weinberg equilibrium

Hardy-Weinberg equilibrium principle states that the genetic variation in a population will remain constant from one generation to the next one when there is no disturbing factor. In a large population with no disruptive influences, when the mutation is random, then the model predicts that both genotype and allele will remain constant since they are in equilibrium.

5.2 Processing Data for matrix eQTL

To perform an eQTL analysis with Matrix eQTL, we need to transfer SNPs data to a matrix format. We processed SNPs data to a 0/1/2 matrix format. The 0/1/2 format cannot be simply assigned randomly to SNP by any software such as R. PLINK is used to recode the data. PLINK selects the minor and major allele for

each SNP data then recode into 0/1/2. So if A is a minor allele for a SNP then AA will recode to 2, AB to 1 and BB to 0. An example of matrix format is shown in Table 5-3.

Table 5-3An example of a 0/1/2 matrix format for SNPs.

SNP id	Sam_01	Sam_02	Sam_03	⋯	Sam_173
SNP 01	0	1	1	⋯	0
SNP 02	1	0	0	⋯	0
⋮	⋮	⋮	⋮	⋱	⋮
SNP K	2	2	1	⋯	0

5.2.1 Basic measure of data quality

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.

5.3 Results

In this section, we first report results from eQTL analysis using Matrix eQTL. Then we show the results for significant cis and trans eQTL.

5.3.1 Results for eQTL

After performing quality control and filtering, we examine how SNPs regulate RNA expression for this sample. First, we examine the eQTL analysis without considering the gene/SNP location. We use matrix eQTL with ANOVA option. We found 552,011 significant eQTL. Table 5-4 shows the eQTL identified from RNA data in all the subjects, using FDR thresholds of 0.01, 0.05, 0.10 and 0.25.

Table 5-4Number of eQTL mapping for various FDR thresholds.

	FDR<0.01	**FDR<0.05**	**FDR<0.1**	**FDR<0.25**
Number of eQTL	37310	77218	120864	295571

Figure 5-1 shows QQ plot for 19,732,641,238 p values. The QQ plot shows the eQTL association p-values for all SNP-gene pairs. The gray line is the identity line representing the null distribution under which there is no association between SNP genotype and gene expression level. Since the p-values deviate from the grey line, there is a significant association between SNPs and gene expressions.

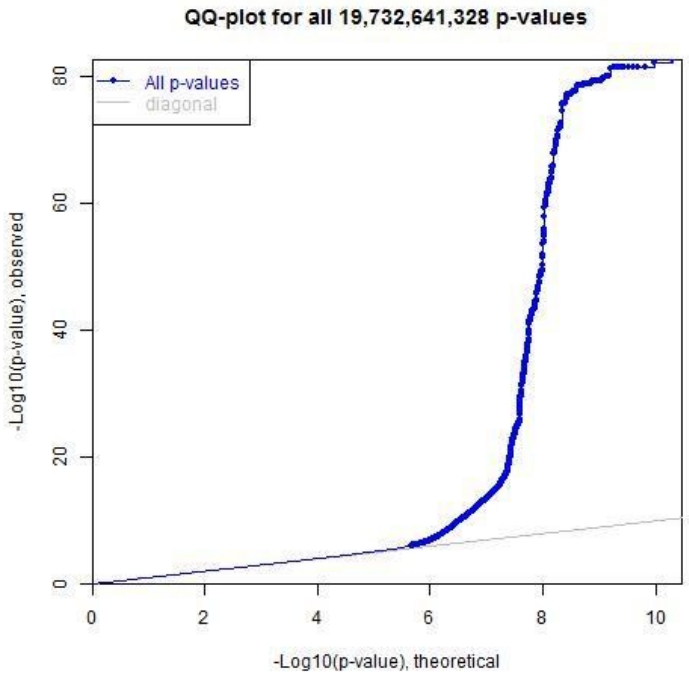


Figure 5-1QQ plot for 19,732,641,238 p values.

Figure 5.2 shows histograms of p values obtained from testing the effect of SNPs on gene expression. This plot shows the distribution of p-values in all the performed tests, even the tests that do not pass significance threshold.

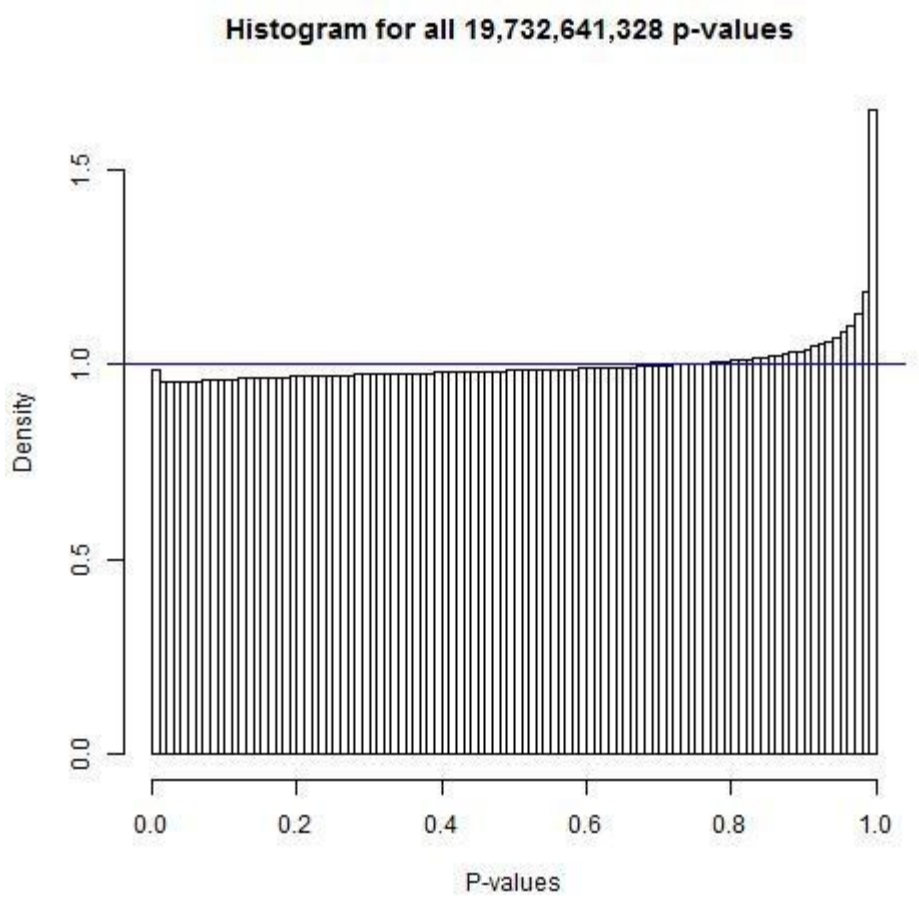


Figure 5-2 Histogram for 19,732,641, 238 p values.

5.3.2 Results for cis and trans eQTL

In this section, we separated the reporting of results by cis and trans effects. The ANOVA method is chosen to find the association between SNPs and gene expression. We chose the cis distance as 1 Mb and therefore cis are restricted to within 1Mb of transition starting site. The p value thresholds for cis and trans are 8×10^{-5} and1×10^{-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 5-5 presented the

number of cis and trans eQTL under different FDR values, after matching by chromosome.

Table 5-5Cis and trans eQTL for FDR thresholds of 0.01 and 0.05, after matching by chromosome.

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

Figure 5-3 shows QQ plot for 12,530,828cis and 19,720,110,500 trans p values. The QQ plot shows the eQTL association p-values for all SNP-gene pairs.

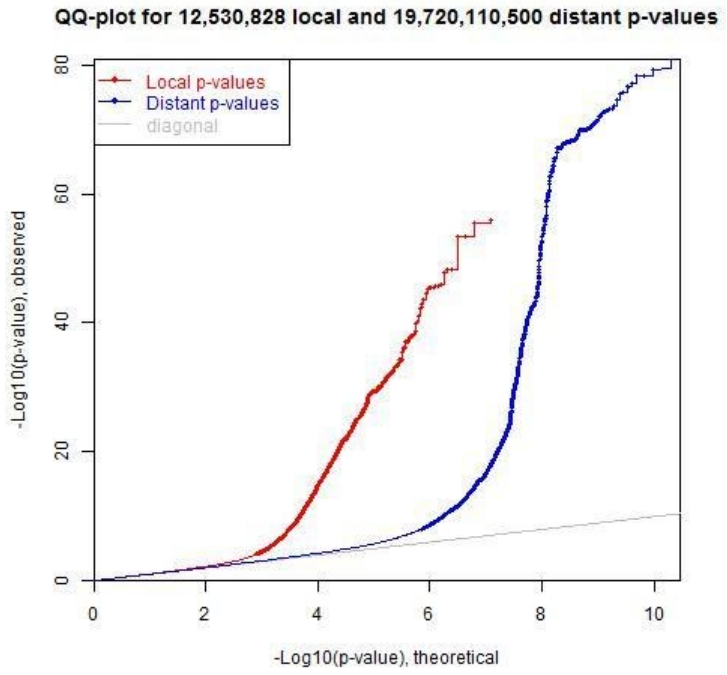
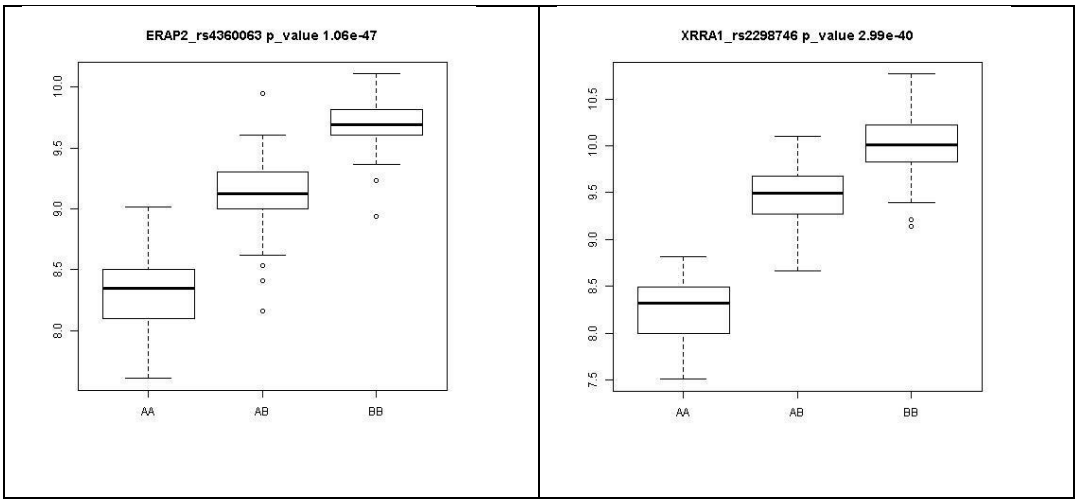
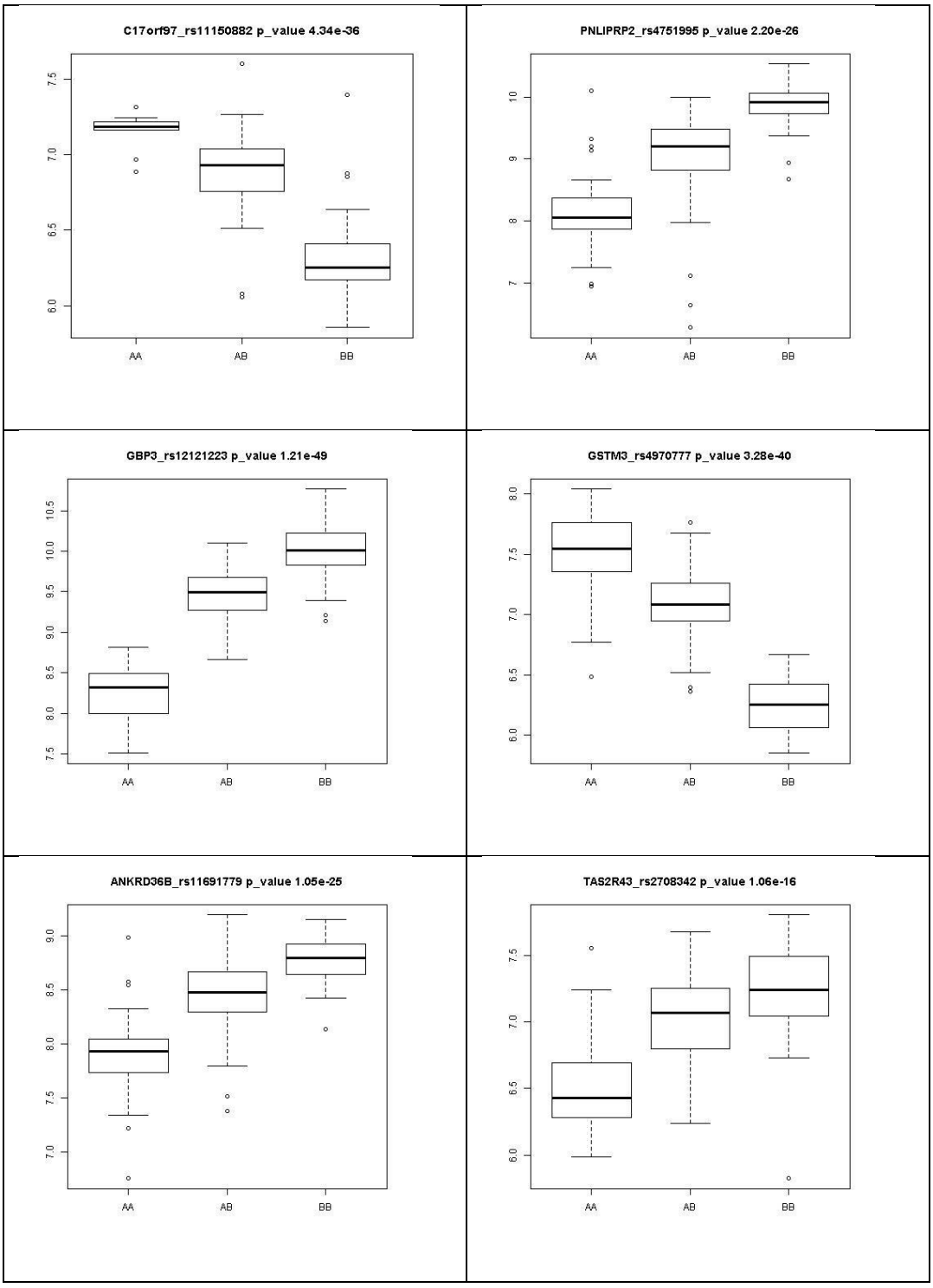


Figure 5-3 QQplot for cis and trans p values in all samples.

We focused our study on cis eQTL since they are more biologically interesting. The significant eQTL are matched by chromosome. Figure 5-4 shows the boxplot for 9 significant cis eQTL. The boxplots is a suitable way to present the different values of gene expression by a SNP.. The x-axes are related to the SNP and y-axes show the normalized and log transformed RNA. The horizontal line depicts the median RNA for each genotype. The points outside the whiskers are outliers. The boxplots enable us to identify the frequency of each SNP in the sample. Moreover, comparing the box plots in different studies help to see if a SNP is highly significantly correlated to a gene or not. . Kabakchiev et al. (2013) have presented the boxplots for 12 most significant cis eQTL for the same raw dataset.




C17orf97_rs11150882 p_value 4.34e-36
PNLIPRP2_rs4751995 p_value 2.20e-26
GBP3_rs12121223 p_value 1.21e-49
GSTM3_rs4970777 p_value 3.28e-40
ANKRD36B_rs11691779 p_value 1.05e-25
TAS2R43_rs2708342 p_value 1.06e-16
AA
AB
BB

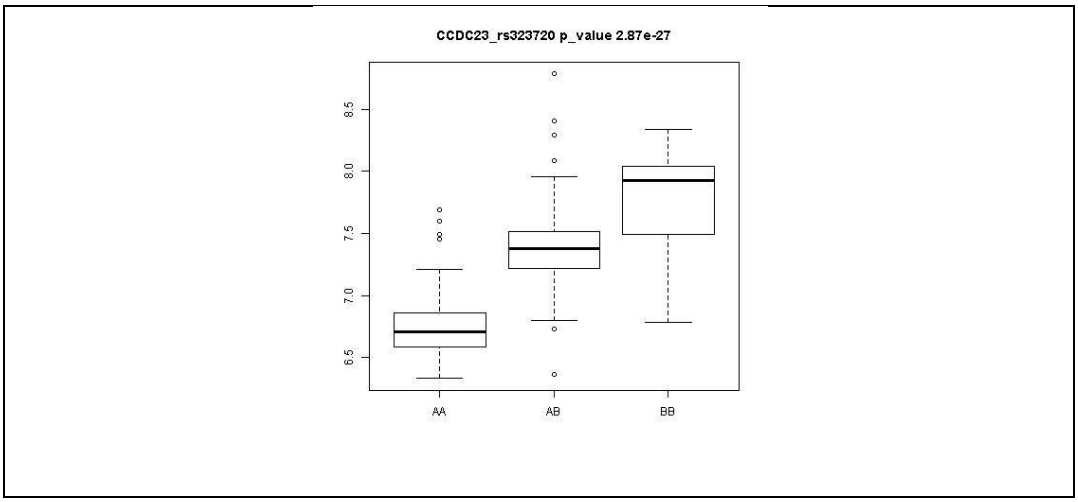


Figure 5-4Correlation between RNA and genotype for 9 significant cis eQTL.

5.4 Comparing results from eQTL and eQTLA methods

The raw data for this study is similar to the raw dataset used by Kabakchiev et al. (2013) implementing eQTLA tool. eQTLA is a custom standalone software application on C++. SNPs with GC score > 0.2 were considered in their studies. Also samples with call rate > 0.95, minor allele frequency <0.05 and Hardy-Weinberg equilibrium ($P > 10^{-6}$) were used in their dataset. 581,633 SNPs passed the quality control. RNA was removed with the miRNeasyMini Kit (Qiagen) in 2 sets; NanoDrop 1000 (Thermo Fisher Scientific, Waltham, MA) and Bioanalyzer 2100 (Agilent, Santa Clara, CA). After data quality control, 19,047 RNA data were studied. The p value thresholds for cis and trans were 1×10^{-3} and 2×10^{-7}, respectively. They found presence of 15,091 cis-eQTL which associated with 2,629 genes and 291 trans eQTL with cis distance 50_kb, and 14,535 significant cis eQTL associated with 1,811 genes when cis distance is 1MB.

Chapter 6: Conclusion

6.1 Conclusion

This thesis describes application of Matrix eQTL to investigate the association between genotype and gene expression. In this study, a large dataset is implemented which consists RNA data for intestinal tissue and SNPs from blood sample of same cohort individuals. There are different statistical method and tools to identify the eQTL. Most of the eQTL methods take days to complete the analysis. The modern datasets make computational challenges in eQTL analysis.

There are several benefits to using Matrix eQTL. Firstly, computational efficiency of matrix eQTL is superior compared to other tools. Matrix eQTL is a fast tool which can perform analysis in few minutes. Secondly, matrix eQTL is capable of handling and analyzing the association between SNPs and gene expression for large dataset.

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 are less than 0.05 and HWE ($P > 10^{-6}$). After processing data, the gene expression which is associated with SNPs can be identify with ANOVA model. 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. Also the significance of each factor (dummy variables for SNP) is tested with an F test or R^2 and instead of using regression

coefficients, pairwise comparisons are performed. We used ANOVA model to obtain the significant cis eQTL with 1 Mb cis distance. Several significant cis and trans eQTL under different FDR threshold have been identified in this research.

A lot of eQTLs are heritable so doing further gene ontology analysis is necessary to find the heritable genes. eQTLs delivers biological understanding of disease studies. GWAS of complex diseases brings great information about clinically and biologically association with different diseases. The gene expression provides valuable information which can be implemented to identify various types of diseases. If an SNP is associated with a specific disease, and at the same time it is associated with a gene, it provides information about that candidate gene. The eQTL data are capable to determine evidence about heritability by identifying genes.

6.2 Future Studies

The emergence of eQTL brings the insight in the field of genetics. Further studies are needed to find the significant genes which are associated with intestinal disease such as inflammatory bowel disease (IBD).

Also the eQTL loci can be compared with existing eQTL databases and compare it with other tissues. Matrix eQTL can include covariates and see the effect of other variables in eQTL analysis as well.

6.3 Software Packages

We used R software, version 3.3.1 for running the Matrix eQTL. Free Matrix eQTL code for performing eQTL analysis is available online. PLINK and SED are used to generate 0/1/2 matrix data using Single nucleotide polymorphisms (SNPs) data.

References

Abecasis, G. R., Cherny, S. S., Cookson, W. O., &Cardon, L. R. (2002). Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nature genetics, 30(1),* 97-101.

Abiola, O., Angel, J. M., Avner, P., Bachmanov, A. A., Belknap, J. K., Bennett, B. (2003). The nature and identification of quantitative trait loci: A community's view. *Nature Reviews Genetics, 4(11),* 911-916.

Closa, A., Cordero, D., Sanz-Pamplona, R., Solé, X., Crous-Bou, M., Paré-Brunet, L., ... & Biondo, S. (2014). Identification of candidate susceptibility genes for colorectal cancer through eQTL analysis. *Carcinogenesis*, bgu092.

Affymetrix, (2000). GeneChip Expression Analysis Technical Manual.

Albert, F. W., &Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nature Reviews. Genetics, 16*(4), 197-212.

Barrett, J. C., and Cardon, L. R. (2006). Evaluating coverage of genome-wide association studies. *Nature Genetics, 38*(6), 659-662.

Bartlett, C. W., Cheong, S. Y., Hou, L., Paquette, J., Lum, P. Y., Jäger, G., ... & Sakai, R. (2012). An eQTL biological data visualization challenge and approaches from the visualization community. *BMC bioinformatics, 13(8),* 1.

Benjamini,Y. & Hochberg,Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B Meth.,* 57, 289–300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 1165-1188.

Benjamini, Y., & Liu, W. (1999). A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *Journal of Statistical Planning and Inference, 82*(1), 163-170.

Benjamini, Y., & Yekutieli, D. (2005). Quantitative trait loci analysis using the false discovery rate. *Genetics, 171*(2), 783-790.

Bottolo, L., Petretto, E., Blankenberg, S., Cambien, F., Cook, S. A., Tiret, L., & Richardson, S. (2011). Bayesian detection of expression quantitative trait loci hot spots. *Genetics, 189(4),* 1449-1459.

Breitling, R., Li, Y., Tesson, B. M., Fu, J. J., Wu, C., Wiltshire, T., . . . Jansen, R. C. (2008). Genetical genomics: Spotlight on QTL hotspots. *PLoS Genetics*, *4(10),* e1000232.

Brem, R. B., Yvert, G., Clinton, R., & Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science, 296(5568)*, 752-755.

Broman, K. W. (2001). Review of statistical methods for QTL mapping in experimental crosses. *Lab Animal, 30(7),* 44.

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

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-Wide Association Studies. *PLoS Computational Biology, 8(12),* e1002822.

Buonaccorsi, J. P. (2010). Measurement error (1st ed.). *Hoboken: Chapman and Hall/CRC*.

Cardon, L. R., Cherny, S. S., Abecasis, G. R., & Cookson, W. O. (2002). Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics, 30*(1), 97-101.

Chatterjee, S., &Hadi, A. S. (2012). Regression analysis by example (Fifth edition. ed.). *US: John Wiley & Sons Inc.*

Chen, L., Tong, T., & Zhao, H. (2008). Considering dependence among genes and markers for false discovery control in eQTL mapping. *Bioinformatics, 24*(18), 2015-2022.

Chen, W., Brehm, J. M., Lin, J., Wang, T., Forno, E., Acosta-Pérez, E., . . .Celedón, J. C. (2015). Expression quantitative trait loci (eQTL) mapping in Puerto Rican children. *PloS One, 10(3),* e0122464.

Chun H, Keles S. (2009). Expression quantitative trait loci mapping with multivariate sparse partial least squares regression.*Genetics. 182(1):*79–909.

Church, J. M., Johnson, C., Mayo, J., & Parang, E. (2016). Reviews. *Serials Review, 42*(3), 272.

Collard, B., Jahufer, M., Brouwer, J., & Pang, E. (2005). An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica, 142*(1), 169-196.

Collier, Z. J., Humphries, L. S., Teven, C. M., Ginat, D., Reavey, P., & Reid, R. R. (2016). Abstract. *Plastic and Reconstructive Surgery - Global Open, 4*, 120-121.

Coriell Institute for Medical Research. What is genotyping and expression profiling? Available online:https://www.coriell.org/research-services/genotyping-microarray/what-is-genotyping-and-expression-profiling (accessed on 22 July 2016).

Cummins, M. (2013). Multiple comparisons problem. *Applied Economics Letters,* 20(7/9), 903-909.

Davis, J. R., Fresard, L., Knowles, D. A., Pala, M., Bustamante, C. D., Battle, A., & Montgomery, S. B. (2016). An efficient multiple-testing adjustment for eQTL studies that accounts for linkage disequilibrium between variants. *American Journal of Human Genetics, 98(1*), 216-224.

Degnan, J. H., Lasky-Su, J., Raby, B. A., Xu, M., Molony, C., Schadt, E. E., & Lange, C. (2008). Genomics and genome-wide association studies: An

integrative approach for expression QTL mapping. *Genomics, 92(3)*, 129–133.

Dennis S. Bernstein. (2009). Matrix mathematics (2nd ed..). *Princeton: Princeton University Press.*

Ding, J., Gudjonsson, J. E., Liang, L., Stuart, P. E., Li, Y., Chen, W., . . . Abecasis, G. R. (2010). Gene expression in skin and lymphoblastoid cells: Refined statistical method reveals extensive overlap in cis-eQTL signals. *The American Journal of Human Genetics*, 87(6), 779-789.

DNA Sequence Assembler, Single nucleotide polymorphism analysis and mutation detection. Available online: http://www.dnabaser.com/articles/SNP/SNP-single-nucleotide-polymorphism.html. (accessed on 15 July 2016).

Dudbridge, F., &Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. *Genetic Epidemiology, 32(3),* 227-234.

Flutre, T., Wen, X., Pritchard, J., & Stephens, M. (2013).A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genetics, 9(5),* e1003486.

Franke, B., Buitelaar, J. K., Cichon, S., Craddock, N., Daly, M., Faraone, S. V., . . . Sullivan, P. F. (2009). Genomewide association studies: History, rationale, and prospects for psychiatric disorders. *American Journal of Psychiatry, 166*(5), 540-556.

Gatti, D. M., Shabalin, A. A., Lam, T., Wright, F. A., Rusyn, I., & Nobel, A. B. (2009). FastMap: Fast eQTL mapping in homozygous populations. *Bioinformatics, 25*(4), 482-489.

Gelman, A., Hill, J., &Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness, 5*(2), 189.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., . . . Zhang, J. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology, 5*(10), R80.

George, A. W., Visscher, P. M., & Haley, C. S. (2000). Mapping quantitative trait loci in complex pedigrees: A two-step variance component approach. *Genetics, 156*(4), 2081-2092.

Gilad, Y., Rifkin, S. A., & Pritchard, J. K. (2008). Revealing the architecture of gene regulation: The promise of eQTLstudies. *Trends in Genetics, 24*(8), 408-415.

Glickman, M. E., Rao, S. R., & Schultz, M. R. (2014). False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. *Journal of Clinical Epidemiology, 67*(8), 850-857.

Goto, K., &Geijn, R. (2008). Anatomy of high-performance matrix multiplication. ACM *Transactions on Mathematical Software (TOM*S), 34(3), 1-25.

Haldane, JBS. (1942). New Paths in Genetics. *Harper& Brothers*.

Hamer, D. H. (1983). Gene expression.*New York: Liss*.

Hardy, J. & Singleton, A. (2009). Genomewide association studies and human disease. *N. Engl. J. Med.* 360, 1759–1768.

Hernandez, D. G., Nalls, M. A., Moore, M., Chong, S., Dillman, A., Trabzuni, D., … Cookson, M. R. (2012). Integration of GWAS SNPs and tissue specific expression profiling reveal discrete eQTLs for human traits in blood and brain. *Neurobiology of Disease*, 47(1), 20–28.

Hill, W. G., Wray, N. R., &Visscher, P. M. (2008). Heritability in the genomics era - concepts and misconceptions. *Nature Reviews Genetics, 9*(4), 255-266.

Hirschhorn, J. N. (2009). Genomewide association studies -- illuminating biologic pathways. *The New England Journal of Medicine, 360*(17), 1699-1701.

Hommel, G. (1988). A stage-wise rejective multiple test procedure based on a modified Bonferronitest. *Biometrika*75, 383–386.

Hunter, D. J., & Kraft, P. (2007). Drinking from the fire hose -- statistical issues in genomewide association studies. *The New England Journal of Medicine, 357*(5), 436-439.

Hwang, Y., Chu, S., &Ou, S. (2011). Evaluations of FDR-controlling procedures in multiple hypothesis testing. *Statistics and Computing, 21*(4), 569-583.

International Society of genetic geneology Wiki. Available online: http://isogg.org/wiki/Single-nucleotide_polymorphism (accessed on 30 August 2016).

Jansen, R.C. & Nap, J.P. (2001). Genetical genomics: the added value from segregation. *Trends Genet*. 17, 388–391.

John D. Storey. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B (Statistical Methodology), 64*(3), 479-498.

Kang, H. M., Ye, C., & Eskin, E. (2008). Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics, 180(4)*, 1909-1925.

Kao, C., Zeng, Z., & Teasdale, R. D. (1999). Multiple interval mapping for quantitative trait loci. Genetics, 152(3), 1203-1216.

Kendziorski, C. M., Chen, M., Yuan, M., Lan, H., & Attie, A. D. (2006). Statistical methods for expression quantitative trait loci (eQTL) mapping. *Biometrics, 62*(1), 19-27.

Kendziorski, C., & Wang, P. (2006).A review of statistical methods for expression quantitative trait loci mapping. *Mammalian Genome, 17*(6), 509-517.

Kliebenstein, D. (2009). Quantitative genomics: Analyzing intraspecific variation using global gene expression polymorphisms or eQTLs. *Annual Review of Plant Biology,60*(1), 93-114.

Korn, E. L., Troendle, J. F., McShane, L. M., & Simon, R. (2004).Controlling the number of false discoveries: Application to high-dimensional genomic data. *Journal of Statistical Planning and Inference, 124*(2), 379-398.

Kruglyak, L., & Rockman, M. V. (2006). Genetics of global gene expression. *Nature Reviews Genetics, 7*(11), 862-872.

Lan, H., Chen, M., Flowers, J. B., Yandell, B. S., Stapleton, D. S., Mata, C. M., . . .Attie, A. D. (2006). Combined expression trait correlations and expression quantitative trait locus mapping. *PLoS Genetics, 2*(1), e6.

Lathrop, M., Cookson, W., Moffatt, M., Liang, L., &Abecasis, G. (2009).Mapping complex disease traits with global gene expression. *Nature Reviews Genetics, 10*(3), 184-194. doi:10.1038/nrg2537.

L. Li, X. Zhang, H. Zhao. (2012). eQTL. *Methods Mol Biol*, 871: 265–279.

Leung,D. (2007). An R package for analysis of whole-genome association studies. *Hum.*

Hered., 64, 45–51.

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician, 54*(3), 217-224.

Mackay, T. F. C., Stone, E. A., & Ayroles, J. F. (2009). The genetics of quantitative traits: Challenges and prospects. *Nature Reviews Genetics, 10*(8), 565-577.

MacLellan, W. R., Wang, Y., & Lusis, A. J. (2012). Systems-based approaches to cardiovascular disease. *Nature Reviews. Cardiology, 9*(3), 172-184.

Majewski, J., &Pastinen, T. (2011). The study of eQTL variations by RNA-seq: From SNPs to phenotypes. *Trends in Genetics, 27*(2), 72-79.

Management Association, I. R. (2013). *Bioinformatics IGI Publishing.*

Manolio, T.A. (2010). Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.* 363, 166–176.

McCarthy, M. I., & Hirschhorn, J. N. (2008). Genome-wide association studies: Potential next steps on a genetic journey. *Human Molecular Genetics, 17*(R2), R165.

Michaelson, J. J., Loguercio, S., & Beyer, A. (2009). Detection and interpretation of expression quantitative trait loci (eQTL). *Methods*, 48(3), 265-276.

Morley, M., Spielman, R. S., Devlin, J. L., Molony, C. M., Weber, T. M., Ewens, K. G., & Cheung, V. G. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature, 430*(7001), 743-747.

Nakagawa, S. (2004). A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behavioral Ecology*, 15(6), 1044-1045.

Nam, D., & Kim, S. (2008). Gene-set approach for expression pattern analysis. *Briefings in Bioinformatics, 9*(3), 189-197.

Narum, S. (2006). Beyond Bonferroni: Less conservative analyses for conservation genetics. *Conservation Genetics, 7*(5), 783-787.

National Library of Medicine (NLM). What are single nucleotide polymorphisms (SNPs)? Available online: https://ghr.nlm.nih.gov/primer/genomicresearch/snp. (accessed on 11 July 2016).

Nature. Gene expression. Available online: http://www.nature.com/scitable/topicpage/gene-expression-14121669 (accessed on 22 July 2016).

Nature. SNP. Available online: http://www.nature.com/scitable/definition/single-nucleotide-polymorphism-snp-295 (accessed on 11 July 2016).

Nature. ribonucleicacid/RNA. Available online: http://www.nature.com/scitable/definition/ribonucleic-acid-45 (accessed on 20 July 2016).

Nica, A. C., & Dermitzakis, E. T. (2008). Using gene expression to investigate the genetic basis of complex disorders. *Human Molecular Genetics, 17*(R2), R134.

Nica, A. C., & Dermitzakis, E. T. (2013). Expression quantitative trait loci: Present and future. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 368*(1620), 20120362.

Nica, A. C., Montgomery, S. B., Dimas, A. S., Stranger, B. E., Beazley, C., Barroso, I., & Dermitzakis, E. T. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genetics, 6*(4), e1000895.

Noble, W. S. (2009). How does multiple testing correction work? *Nature Biotechnology, 27*(12), 1135-1137.

Peng, X., Li, S. S., Gilbert, P. B., Geraghty, D. E., & Katze, M. G. (2016). FCGR2C polymorphisms associated with HIV-1 vaccine protection are linked to altered gene expression of fc-γ receptors in human B cells. *PloS One, 11*(3), e0152425.

Pérez-Enciso, M., Quevedo, J. R., & Bahamonde, A. (2007). Genetical genomics: Use all data. *BMC Genomics, 8*(1), 69.

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *BMJ: British Medical Journal, 316*(7139), 1236-1238.

Petretto, E., Mangion, J., Dickens, N. J., Cook, S. A., Kumaran, M. K., Lu, H., . . . Aitman, T. J. (2006). Heritability and tissue specificity of expression quantitative trait loci. *PLoS Genetics, 2*(10), e172.

Psychiatric GWAS Consortium Steering Committee. (2009). A framework for interpreting genome-wide association studies of psychiatric disorders. *Mol. Psychiatry,* 14, 10–17.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ...& Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics,* 81(3), 559-575.

Qi, J., Asl, H. F., Björkegren, J., & Michoel, T. (2014). kruX: matrix-based non-parametric eQTL discovery. *BMC bioinformatics, 15(1),* 1.

Quitadamo, A., Tian, L., Hall, B., & Shi, X. (2015). An integrated network of microRNA and gene expression in ovarian cancer. *BMC Bioinformatics, 16.*

Sajuthi, S., Sharma, N., Chou, J., Palmer, N., McWilliams, D., Beal, J., . . . Das, S. (2016). Mapping adipose and muscle tissue expression quantitative trait loci in africanamericans to identify genes for type 2 diabetes and obesity. *Human Genetics, 135*(8), 869-880.

Scott A. Rifkin (ed.), (2012). Quantitative Trait Loci (QTL): Methods and Protocols, Methods in Molecular Biology, vol. 871. doi:10.1007/978-1-61779-785-9_14, 2012.

Schweizer, K. (2012). On correlated errors. *European Journal of Psychological Assessment, 28*(1), 1-2.

Servin, B., & Stephens, M. (2007). Imputation-based analysis of association studies: Candidate regions and quantitative traits.*PLoS Genetics, 3*(7), e114.

Shabalin, A.A. (2012). Matrix eQTL: ultra fasteQTL analysis via large matrix operations. *Bioinformatic,* 28, 1353–1358.

Sieberts, S. K., &Schadt, E. E. (2007). Moving toward a system genetics view of disease. *Mammalian Genome,* 18(6-7), 389-401.

Simon, L., Chen, E., Edelstein, L., Kong, X., Bhatlekar, S., Rigoutsos, I., . . . Shaw, C. (2016). Integrative multi-omic analysis of human platelet eQTLs reveals alternative start site in mitofusin 2. *The American Journal of Human Genetics, 98*(5), 883-897.

Stegle O, Parts L, Durbin R, Winn J. (2010). A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Computational Biology.*6(5):e1000770.

Stranger, B. E., Montgomery, S. B., Dimas, A. S., Parts, L., Stegle, O., Ingle, C. E., . . . Dermitzakis, E. T. (2012). Patterns of cis regulatory variation in diverse human populations. *PLoS Genetics, 8*(4), e1002639. doi:10.1371/journal.pgen.1002639.

Stranger, B. E., Tavaré, S., Koller, D., Deloukas, P., Dermitzakis, E. T., Montgomery, S., . . . Forrest, M. S. (2007). Population genomics of human gene expression. *Nature Genetics, 39*(10), 1217-1224.

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genome wide studies. *Proceedings of the National Academy of Sciences, 100(16),* 9440-9445.

Sul, J. H., Raj, T., de Jong, S., de Bakker, Paul I W, Raychaudhuri, S., Ophoff, R. A., . . . Han, B. (2015). Accurate and fast multiple-testing correction in eQTL studies. *American Journal of Human Genetics, 96*(6), 857-868.

Sun, W., Yu, T., & Li, K. (2007). Detection of eQTL modules mediated by activity levels of transcription factors. *Bioinformatics, 23*(17), 2290-2297.

Sun, W. (2009). eQTL Analysis by Linear Model.

Tian, L., Quitadamo, A., Lin, F & Shi, X. (2014). Methods for population-based eQTL analysis in human genetics. *Tsinghua Science and Technology, 19*(6), 624-634.

Troendle, J. F., & Westfall, P. H. (2011). Permutational multiple testing adjustments with multivariate multiple group data. *Journal of Statistical Planning and Inference, 141*(6), 2021-2029.

Van Nas, A., Ingram-Drake, L., Sinsheimer, J. S., Wang, S. S., Schadt, E. E., Drake, T., &Lusis, A. J. (2010). Expression quantitative trait loci:

Replication, tissue- and sex-specificity in mice. *Genetics, 185*(3), 1059-1068.

Warpole, K. N., &Kossoy, K. (2015). Articles. *Film Matters, 6*(3), 87-91.

William S Bush,& Jason H Moore. (2012). Chapter 11: Genome-wide association studies. *PLoS Computational Biology, 8*(12), e1002822.

Wright, F. A., Shabalin, A. A., &Rusyn, I. (2012). Computational tools for discovery and interpretation of expression quantitative trait loci. *Pharmacogenomics, 13*(3), 343-352.

Yang, S., Liu, Y., Jiang, N., Chen, J., Leach, L., Luo, Z., & Wang, M. (2014). Genome-wide eQTLs and heritability for gene expression traits in unrelated individuals. *BMC Genomics, 15*(1), 13.

Young, N. D. (1996). QTL mapping and quantitative disease resistance in plants. *Annual Review of Phytopathology, 34(1), 479-501.*

Zeng, Z. (1993). Theoretical basis for separation of multiple linked gene effects inmapping quantitative trait loci. *Proc. Natl Acad. Sci.,* 90, 10972.

Zhang, W., & Liu, J. S. (2010). From QTL Mapping to eQTL Analysis. *In Frontiers in Computational and Systems Biology* (pp. 301-329). Springer London.