UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Technology

Vladislav Tuzov

How sex influences the effect of genetic variants on gene expression?

Bachelor's Thesis (12 ECTS)

Curriculum Science and Technology

Supervisor(s):
Kaur Alasoo, PhD

Tartu 2020

How sex influences the effect of genetic variants on gene expression?

Abstract:

Complex human traits and disease prevalence and risks vary between women and men. It is important to understand and interpret the effect genetic variants have on cells to manage disease and differences in phenotypes. As most genetic variants involved in regulation, the effect of these variants (eQTLs) influences gene expression. Most previous research showed a small amount of sex-biased eQTLs. However, the majority of studies use whole blood samples, which are more likely to produce false positives than purified blood samples, because some blood cells are more abundant in one sex than the other. This work analysed 1187 eQTLs for the present sex-specific effect on gene expression using interaction tests. This approach shows reliability confirmed by the uniform distribution of P-values.

Keywords:

Bioinformatics, eQTLs analysis, interaction test

CERCS: B110 Bioinformatics, medical informatics, biomathematics, biometrics

Kuidas mõjutab sugu geenivariatsioonide efekti geeniekspressioonile?

Lühikokkuvõte:

Inimeste komplekssed tunnused, haiguste levimus ning nendega seotud riskid on meeste ja naiste vahel erinevad. Selleks et haiguseid ja nende erisusi fenotüüpides uurida, on oluline aru saada ning kaardistada erinevate geenide avaldumist rakkudes. Geeniekspressiooni mõjutavad varieeruvused geneetilistes markerites, nagu ekspressiooni kvantitatiivsete tunnuste lookused (eQTL). Varasemad teadustööd pole näidanud suurel hulgal inimese soost mõjutatud eQTLe, kuid enamik töödest on kasutanud täisvere proove. Need annavad suurema tõenäosusega valepositiivseid tulemusi kui rakutüübi kaupa eraldatud proovid, sest mõndasid vererakke esineb ühel sool rohkem kui teisel. Käesolev töö uuris 1187 eQTLi soost sõltuvate efektide esinemisest geeniekspressioonile kasutades interaktsiooniteste. Selle lähenemise usaldusväärsust kinnitab p-väärtuste ühtlane jaotus.

Võtmesõnad:

interaktsiooni test, eQTL-ide analüüs, interaktsioonitest

CERCS: B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

TABLE OF CONTENTS

1 LITERATURE REVIEW

1.1 Introduction

The human genome is the complete genetic material containing coding and non-coding genes, accounting for about 2% and 98% of the genome respectively, it consists of 3.1 billion nucleotides and two copies of each chromosome which are called homologous chromosomes. The coding genes are the DNA or RNA fragments that provide a protein blueprint. The non-coding portion of genetic material functions as regulation tools for transcription and translation, attachment regions, DNA replication origins, telomeres and centromeres. Genetically speaking, humans are identical and non-unique, as 99.9% of the genome is absolutely the same for every individual, however, the remaining 0.1% gives us our uniqueness and individuality. The difference between our genomes is described by a genetic variation which is the difference in nucleic acids among the population. It manifests mutations of various types: single-nucleotide polymorphisms, (SNPs), insertion-deletion mutation (indels) and large genetic recombination (Ginsburg 2013). The variability is relevant to the differential disease risk among individuals. It is fundamental to understand and interpret the effects genome variants have inside the cells to handle the biology of diseases and phenotypes of an organism. The variants likely to be involved in gene regulation are found in non-coding regions, to analyse such variants the studying of gene expressions in cells was implemented. These studies rely on Expression Quantitative Trait Loci (eQTLs) (Nica and Dermitzakis 2013).

1.2 Expression Quantitative Trait Locus

Expression Quantitative Trait Locus (eQTL) is a chromosomal region affecting the level of gene transcription, in other words, it can affect one or more gene expressions. These loci can be classified by their location (local or distant) and action mode (cis- or trans-). Local eQTLs reside near the genes that they impact, while distant ones are located further away. There is no exact physical or genetic distance that instructs an eQTL to be defined as distant; different studies determine it in different manners, ranging from 2Mb to regions on other chromosomes from the gene of interest. Local eQTLs can act in two modes. In *cis*, the allele is changed only on the copy of the same chromosome and its expression is affected, but the other copy stays intact on the homologous chromosome. The comparison of expression levels of the two alleles in heterozygous individuals can display the presence or absence of expression level balance, giving a clue about the effect of a cis-eQTL. The imbalance presence in expression levels between these two alleles hints that the gene is under the effect of *cis*-eQTL. However, the local eQTLs can also operate in *trans* mode, in which they alter the structure, function or expression of a diffusible intermediate resulting in altered gene expression levels. The presence of a diffusible factor means that both target gene alleles are affected, thus heterozygous individuals do not present allele-biased expressions, as evidenced in Figure 1. Due to the factor, *trans*-eQTLs are not restricted to be located close to the regulated gene and can be present anywhere in the genome. In the close proximity to the gene, they are described as local but not *cis*-acting.(Albert and Kruglyak 2015) A recent study of about 1000 individuals suggests that local eQTLs affect about 80% of expressed genes in whole blood(Battle et al. 2014).

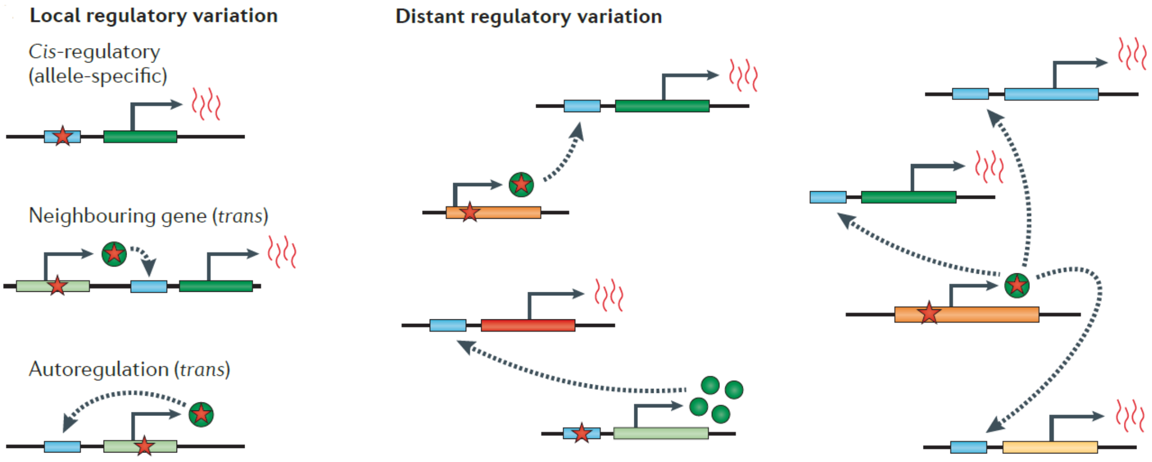

Local regulatory variation
Cis-regulatory (allele-specific)
Neighbouring gene (*trans*)
Autoregulation (*trans*)
Distant regulatory variation

Figure 1 eQTLs are categorised by their location (local and distant) and action mode (*cis*- and *trans*-). The figure is obtained from Albert and Kruglyak, 2015

1.3 Sex differences in humans at the molecular level

GWAS Data

The genetic effects at chromosomal loci can be identified by using data from genome-wide association studies (GWAS). Statistical interaction tests are implemented to investigate if sex correlated variables change the genotype effect on phenotype; the tests show if the genetic variant effect on phenotype has a difference between sexes, it also reveals if an effect of one variable relies on another variable value. In a gene-by-sex interaction test, there is a high chance for a false negative finding so that an incorrect null hypothesis fails to be rejected; these types of errors are called type II errors and they occur as large size samples are necessary to detect differences between two variables.

Sex difference in the regulation of the genome

There is a common characteristic shared among many species, which is that sex-biased gene expression is affected by the number of present copies and alleles as well as the quality of the expression. The sex-biased expression is present within and between tissues, cells and cell lines (Khramtsova, Davis, and Stranger 2019). There are several important subject matters that are shown in the work: genes with sex-biased gene expression located on both the sex chromosomes and autosomes, ChrX stands out for genes in which mRNA levels are different between two sexes; expression levels between males and females are likely to have a small fold-change for sex differentially expressed (sex-DE) genes; tissues, development stages and environment lead to various gene expression of the sex-biased genes. For example, a study of >5000 individuals looked at whole blood samples for sex-biased gene expression and identified that 51, 16 and 572 genes on chromosome X, Y and autosomes accordingly (Jansen et al. 2014). Male-biased genes were linked to renal cancer, while female ones - rheumatoid arthritis; that indicates a contribution of sex-biased gene expression leading to differences in diseases between males and females. Several recent reported studies, which assessed the distinct gene expression in various tissues and cell lines, showed that sex-biased gene expression is exhibited by 10-60% of autosomal genes depending on the tissues. (Khramtsova, Davis, and Stranger 2019)

Sex difference in regulation, the eQTLs effect

The variation in gene expression is partially heritable, this means gene expression variation can affect sex-differentiated traits. The indicator of genetic variation influenced on complex traits by way of regulation is that expression quantitative trait loci (eQTLs) are high among disease risks and complex traits loci. Some eQTLs are sex-biased and may affect transcriptional levels of genes in females, but not in males, for instance, or they could be present in both sexes but have a unique effect on each of them, therefore resulting in differences between sexes in mean expression levels or expression deviation. Figure 2 illustrates the way given genotype variants have various gene expression differences between females and males. The single-sex effect is present if there is a distinct mean expression difference in one sex but not in the other; the differential effect is shown when the genetic variants produce distinctive gene expression in each sex and there is no correlation between the two; the opposite effect exists when the same genotype variants induce reverse gene expression levels in females and males.

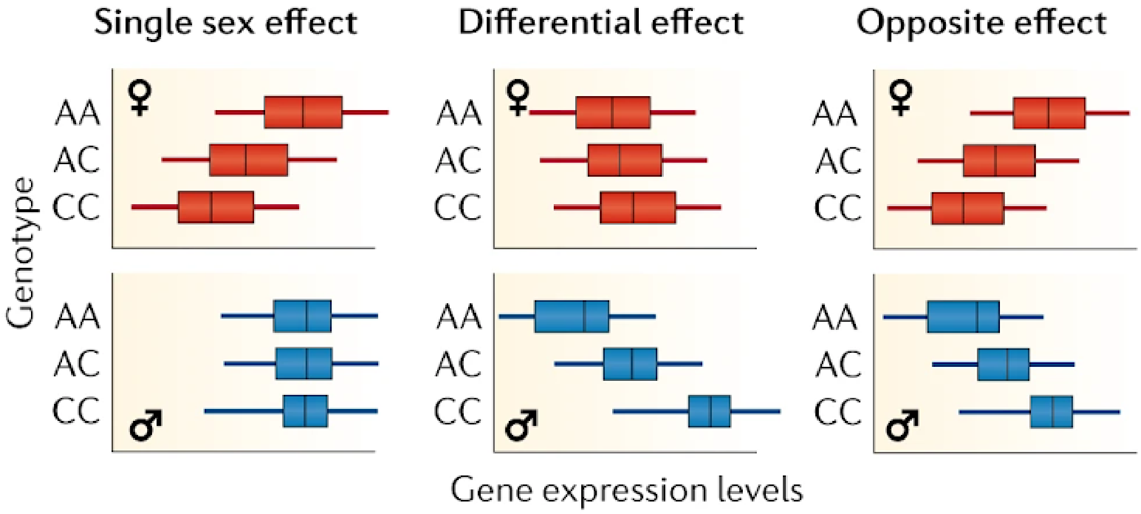


Figure 2 shows gene expression difference varies based on the genotype effect between males and females. The figure is obtained from Khramtsova, Davis, and Stranger 2019.

Cell type composition effect on eQTL analysis

A GWAS of sex-specific eQTLs in whole blood found 4 sex-specific eQTLs on autosomes and 2 eQTLs on chromosome X (Kukurba et al. 2016). Two of these sex-specific eQTLs were found in other individuals, there were eQTLs for *BSCL2* and *NOD2*. The latter *NOD2* sex-biased eQTL was identified in the whole blood analysis. However, purified blood cell analysis showed no sex-biased eQTL for *NOD2*. The reason for these findings is a sex-biased disbalance of some cell type numbers, for instance, neutrophils are present in larger numbers in females than in males, then resulting eQTLs will appear to be more female-specific, however, purified neutrophils samples show non-specific sex effect (Khramtsova, Davis, and Stranger 2019). The cell-type composition differences between males and females could lead to false positives. Thus, my thesis objective is to perform this analysis in purified cell types. As an example, I chose B-cell data from the CEDAR (Momozawa et al. 2018) from the eQTL Catalogue (Kerimov et al. 2020).

1.4 Sex-specific genetic effects on gene expression

A study investigated the influence of sex-biased eQTLs on the genetic variants of disease-associated traits (Yao et al. 2014a). As the way genetic variants influencing gene expression are unknown, the method to identify this type of interactions involving gene regulation is to examine the expression of genes being a quantitative trait and find genetic variants which correspond with the gene expression levels. The whole blood sample evaluation of 11 672 SNPs from 5254 individuals that may function as sex-biased eQTLs found 13 meaningful cis-eQTLs on autosomes and 1 on ChrX show genotype by sex interactions on gene expression, and there were no allele frequency differences between eQTLs. However, out of these 14 genes, only 7 demonstrated differences in the mean expression between females and males. The rest of the genes are regulated by eQTLs which either have an allele associated with higher expression in one of the sexes or have opposite-sex difference effect on the mean expression. The evidence demonstrates eQTLs interactions in Table 1, including an SNP (rs2605100 with P = 0.0009) at the *LYPLALI* locus to represent sex interaction towards obesity traits, and SNPs (rs4343 and rs4329 with P = 0.0015 for both) at the *ACE* locus to correspond with blood pressure, as well as two SNPs (rs10401969 and rs17750998 with P = 0.02 and P = 0.0021 respectively) to correlate with diversity of lipid traits between males and females.

eQTL	Chr:Position	Trait	Expected gene	P-value (interaction)
rs2605100	1 : 219 644 224	Adiposity	BPNT1	0.0009
rs4343	17 : 61 566 031	Angiotensin-converting enzyme inhibitors	PECAM1	0.0015
rs4329	17 : 61 563 458	Metabolism	PECAM1	0.0015
rs10401969	19 : 19 407 718	Cholesterol; triglycerides; cholesterol, LDL	ELL	0.02
rs17750998	19 : 19 388 446	Iron	MAU2	0.0021

Table1. Sex-interaction eQTLs in sex-dependent loci. The table is obtained from Yao et al. 2014b.

The study (Yao et al. 2014) indicates a failure to reproduce results of sex interaction of rs167769 and rs2872507 tested in the study, which looked at genotype-sex interactions in

1240 peripheral mononuclear cells and 379 lymphoblastoid cell lines, by Dimas et al. 2012. There are two main reasons explaining that: there is tissue specificity in eQTLs and a 10-fold increase in sample size is required to facilitate the interaction detection. However, as more than 5000 individuals are the sample size of the given study, the data shows the presence of sex interactions of eQTL genetic variants with gene expression.

Differences between the two sexes are also evident by the presence of a molecule, gene or characteristic which identify a particular physiological process or disease, these are called biomarkers; yet the degree of genetic influence to this is unknown. A study done by Flynn et al. 2019 aimed to investigate sex-biased heritability and genetic correlation over 33 quantitative biomarker traits in about 340 000 individuals, out of which approximately 180 000 were females and 160 000 were males. The method was to estimate the extent of genetic influence on both sexes and determine which genetic effects were common between women and men and which were shared to each. The UK Biobank data investigation of blood and urine biomarkers analysed sex-specific genetic effects of testosterone level and thus used the results to provide an explanation of involved biological mechanisms; selectively expressed genes in tissues and protein-altering variants, testosterone and other traits relationship and testosterone risk prediction models.

For the estimation of sex-biased contribution towards genetic variants, they build two models: one to analyse if variants influence the trait and another to determine genetic variants that have different effects between sexes. The latter resulted in the identification of genetic variants with shared (variants producing the same effect in males and females) and different (variants producing distinct effects in sexes) effects in males and females. 26 487 variants were found to have effects on the traits; the large portion of these variants had shared effects on sexes, only 463 and 146 genetic variants were identified to be with sex-biased effects in men and women accordingly, the majority of these variants influence sex-biased effects on testosterone levels. Figure 3 represents the effect sizes of the variants; both female-specific variants and males-specific variants have strong positive and negative effects on testosterone production. Further, the impact of sex-specific variants related to testosterone levels is investigated to regards to local gene expression of a specific tissue. In females, there was no significant enrichment in the tested genes, but in males, the enrichment was found with liver-specific genes, thus, it is suggested that liver diseases may have a different set of causes in males and females due to this enrichment effect.

The significance of these results is that testosterone is the only biomarker that has sex-biased effects, because the hormone is produced in different tissues in males and females, while other biomarkers show very little difference, leading that we should not find much for the sex-specific gene expression.

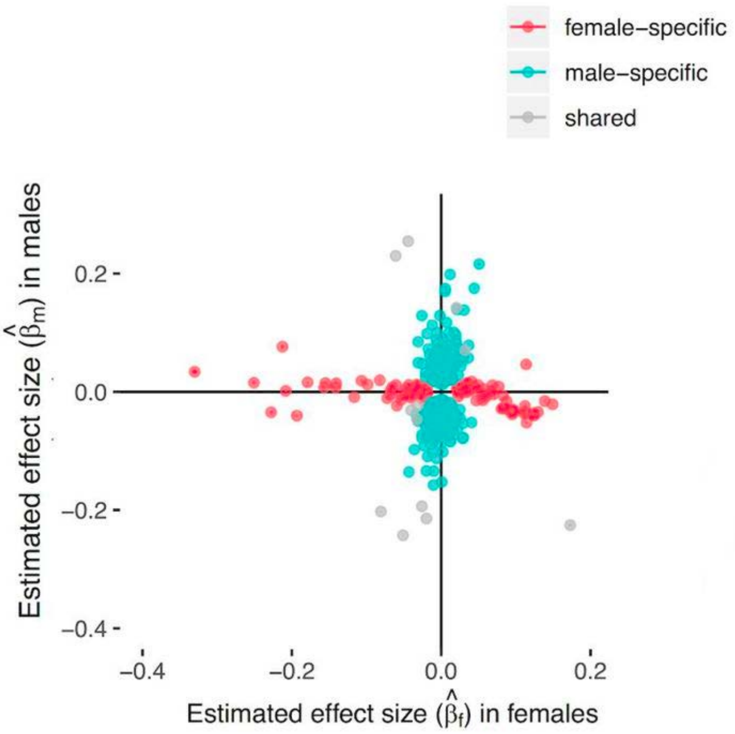


Figure 3. Sex-biased effects on genetic variants influencing testosterone levels. The figure shows an effect sizes estimation for variants on testosterone . The x-axis is an estimated effect size in females; the y-axis is an estimated effect size in males. Blue dots are variants with the male-specific effect; red are the female-specific effect; and gray dots are variants with shared effect. The figure is obtained from Flynn et al. 2019.

2 THE AIMS OF THE THESIS

- Analyse purified cell type samples for sex-biased genetic effect on gene expression

EXPERIMENTAL PART

2.1 METHODS

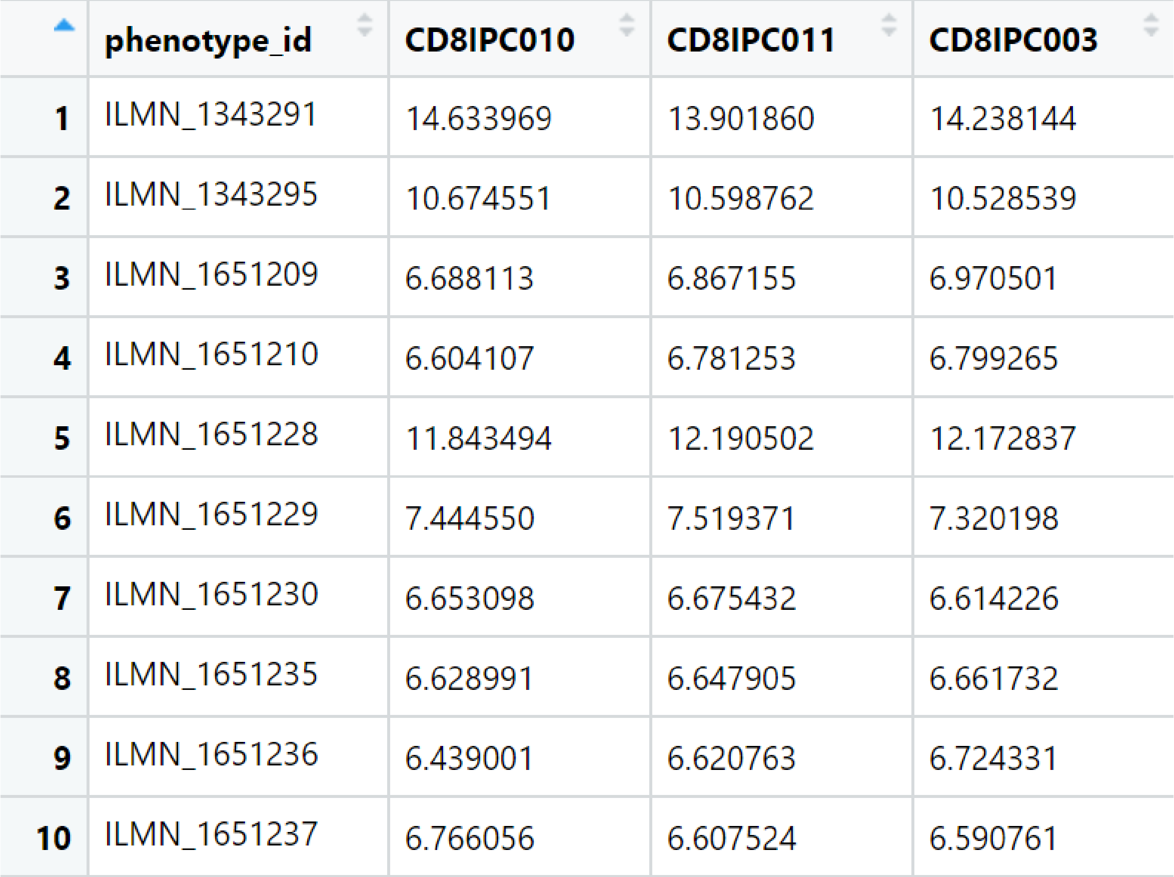
2.1.1 Dataset description

This work's dataset uses publicly available microarray data (CEDAR by Momozawa et al. 2018) from the eQTL Catalogue (Kerimov et al. 2020), a collaboration between University of Tartu and the European Bioinformatics Institute. The dataset consists of eQTL summary statistics and processed raw data. The raw data is passed through eQTL mapping and quality control. The dataset has 2388 samples from 322 donors. The data has various cell types or tissue, normalised gene expression data, genotype data and metadata. The available cell types or tissues are CD4 and CD8 T cells, monocytes, neutrophils, platelets, B-cells, ileum, rectum, transverse colon.

For this work, only 1187 eQTLs were used. The cell type data is obtained from eQTL lead variants (i.e. the genetic variant in the +/-1 Mb window around the promoter of each gene that had the smallest eQTL association p-value for that gene, when analysing males and females together) and their corresponding permutation P-values for B cells of CEDAR dataset taken from eQTL Catalogue (Kerimov et al. 2020). In order to identify eQTL, all SNPs in a 2 Mb window of the gene are scanned and the lead eQTL variants exhibit the strongest association with the corresponding the position of the gene, as characterised by P-values (Joehanes et al. 2017).

P-values are permuted that allowed to have all of the possible alternative assignments that could have been from a larger study with more samples. These are P-values obtained with the help of permutation tests (also called randomisation tests), which are non-parametric methods to determine statistical significance built on label rearrangement of a dataset. The significance of this 'reshuffling' of data is presented as its P-value. Thus, P-value represents the probability of acquiring a lead eQTL at least as extreme as the test statistic provided the null hypothesis is passed, which is that labels are interchangeable. Hence, the low P-values highlight labels which are not interchangeable, and permutation is relevant with respect to the original data(Knijnenburg et al. 2009).

Normalised gene expression data is a compressed text format file (tsv.gz). It provides a matched gene expression between phenotype id on the first column and sample id presented as column names starting from the second column in Table 2. The gene expression is normalised, increasing the quality of data; that is essential as low-quality samples can produce extreme outliers thus reducing the statistical power of the analysis. Low-quality samples occur due to contamination or minor human errors in the laboratories.



	phenotype_id	CD8IPC010	CD8IPC011	CD8IPC003
1	ILMN_1343291	14.633969	13.901860	14.238144
2	ILMN_1343295	10.674551	10.598762	10.528539
3	ILMN_1651209	6.688113	6.867155	6.970501
4	ILMN_1651210	6.604107	6.781253	6.799265
5	ILMN_1651228	11.843494	12.190502	12.172837
6	ILMN_1651229	7.444550	7.519371	7.320198
7	ILMN_1651230	6.653098	6.675432	6.614226
8	ILMN_1651235	6.628991	6.647905	6.661732
9	ILMN_1651236	6.439001	6.620763	6.724331
10	ILMN_1651237	6.766056	6.607524	6.590761

Table 2. Normalised gene expression data from CEDAR dataset.

Genotype data is stored as a compressed variant call format (VCF) file, which also has an index file allowing access to a specific genetic variant by its ID. That returns all genotype values with genotype IDs for this particular variant. The genotypes are imputed from a larger reference population. Raw genotype data has only three variants: 0, 1, 2. As the genotypes were subjected to imputation, they converted from discrete values to continuous ones from 0 to 2.

Metadata is a text file that has sample ID, genotype ID, sex, cell type, quality control test information and additional information about the participants. All these files were internally available from the eQTL Catalogue (Kerimov et al. 2020).

2.1.2 Languages and libraries

R packages (‘R: The R Project for Statistical Computing’) and Python language (python.org) were employed for the work on this thesis. R is a software environment for statistical analysis, computations and graphics. It was accessed with the help of Rstudio, an open-source tool for R, running on Windows 10 and later Ubuntu 18.04. R was used for the large portion computation for this thesis, with Python being an aid tool for accessing relevant genotype variant data.

In R the following libraries were imported ggplot2, dplyr, purrr, SNPRelate, GDSArray:

- ggplot2 creates graphs from the two identified gene-variant pairs (Wickham, Chang, et al. 2020);
- dplyr is highly exploited for generation and preparation of input data (Wickham, François, et al. 2020);
- purrr is used to combine two lists of data frames (Henry, Wickham, and RStudio 2020);
- SNPRelate (Zheng, 2013) and GDSArray are libraries from an open-source bioinformatics tool Bioconductor. They are initially used to import and extract genotype data (github.com/kauralasoo/MTAT.03.239_Bioinformatics/).
- To import genotypes, a script from (github.com/kauralasoo/MTAT.03.239_Bioinformatics/) provided by the supervisor, Kaur Alasso PhD.

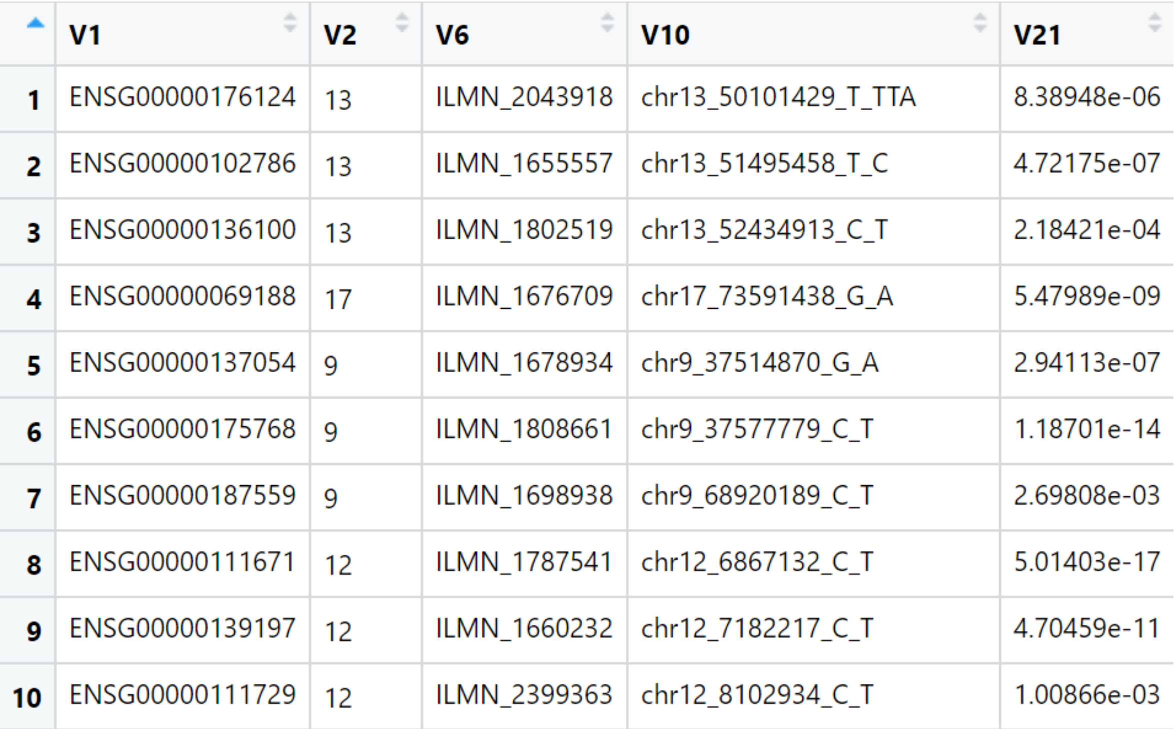
Python script used gzip, tabix and csv library:

- gzip is a module used to uncompress genotype data in text format;
- pytabix allows fast random access to an indexed file, therefore, extracts genotype values for a particular SNP ID (Li, 2011);
- csv is an exportation tool for the extracted genotype values to be written into a file and then imported to R for further data manipulation.

2.1.3 Generating input data

The dataset contains 4 files: cell type data, normalised gene expression data, genotype data and metadata. In order to proceed to eQTL analysis of the data, the required data from these files have to be coupled through specific parameters.

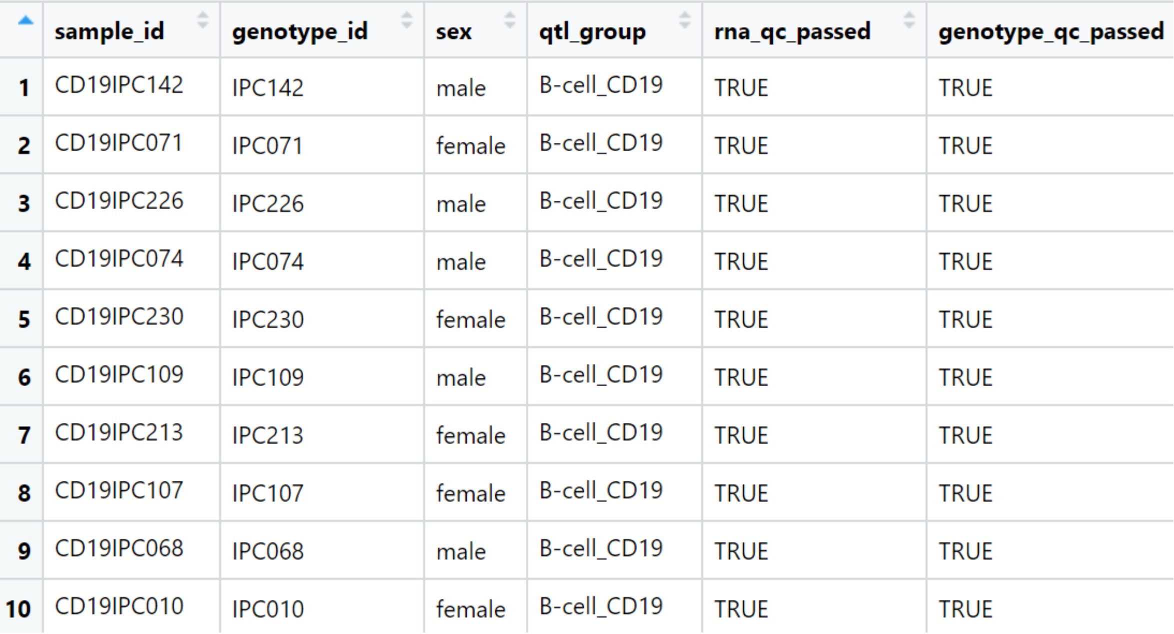
As cell type data has permuted P-values, the low P-values will signify that permutation data is relevant to the original one and high P-values provide false positives. To avoid false positives and improve statistical power of the analysis, P-values from cell type data are adjusted by the False Discovery Rate (FDR) method. All the FDR-corrected P-values that are larger than 0.05 are rejected that reduces the number of eQTL from almost 20 000 to about 1 187 and leaves only genes that are relevant with respect to original data in Table 3.



	V1	V2	V6	V10	V21
1	ENSG00000176124	13	ILMN_2043918	chr13_50101429_T_TTA	8.38948e-06
2	ENSG00000102786	13	ILMN_1655557	chr13_51495458_T_C	4.72175e-07
3	ENSG00000136100	13	ILMN_1802519	chr13_52434913_C_T	2.18421e-04
4	ENSG00000069188	17	ILMN_1676709	chr17_73591438_G_A	5.47989e-09
5	ENSG00000137054	9	ILMN_1678934	chr9_37514870_G_A	2.94113e-07
6	ENSG00000175768	9	ILMN_1808661	chr9_37577779_C_T	1.18701e-14
7	ENSG00000187559	9	ILMN_1698938	chr9_68920189_C_T	2.69808e-03
8	ENSG00000111671	12	ILMN_1787541	chr12_6867132_C_T	5.01403e-17
9	ENSG00000139197	12	ILMN_1660232	chr12_7182217_C_T	4.70459e-11
10	ENSG00000111729	12	ILMN_2399363	chr12_8102934_C_T	1.00866e-03

Table 3. B cell data with adjusted P-values; only genes with small P-value are left, signifying their relevance to the original data. B cell data for 4 genes provides gene, chromosome number, phenotype ID, SNP ID, the position of SNP on the chromosome and P-value.

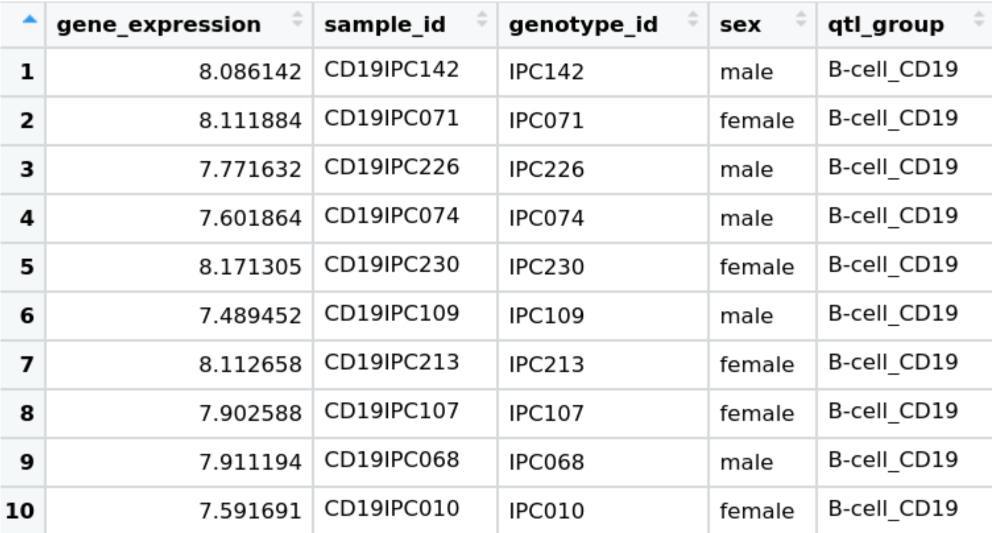
The metadata is used for quality control (QC) and select data for B cells only. QC is used to remove low-quality data and potential outliers, therefore samples, which do not pass both RNA QC and Genotype QC, are filtered out. This QC procedure leaves 2337 samples out of 2967. However, 2337 samples have data for all of the available cell types and filtering out all the cell types other than B cells leaves only 262 samples.



	sample_id	genotype_id	sex	qtl_group	rna_qc_passed	genotype_qc_passed
1	CD19IPC142	IPC142	male	B-cell_CD19	TRUE	TRUE
2	CD19IPC071	IPC071	female	B-cell_CD19	TRUE	TRUE
3	CD19IPC226	IPC226	male	B-cell_CD19	TRUE	TRUE
4	CD19IPC074	IPC074	male	B-cell_CD19	TRUE	TRUE
5	CD19IPC230	IPC230	female	B-cell_CD19	TRUE	TRUE
6	CD19IPC109	IPC109	male	B-cell_CD19	TRUE	TRUE
7	CD19IPC213	IPC213	female	B-cell_CD19	TRUE	TRUE
8	CD19IPC107	IPC107	female	B-cell_CD19	TRUE	TRUE
9	CD19IPC068	IPC068	male	B-cell_CD19	TRUE	TRUE
10	CD19IPC010	IPC010	female	B-cell_CD19	TRUE	TRUE

Table 4. Metadata for B cells and passed quality control, containing sample ID, genotype ID, sex, cell type (qtl_group) and two quality control results.

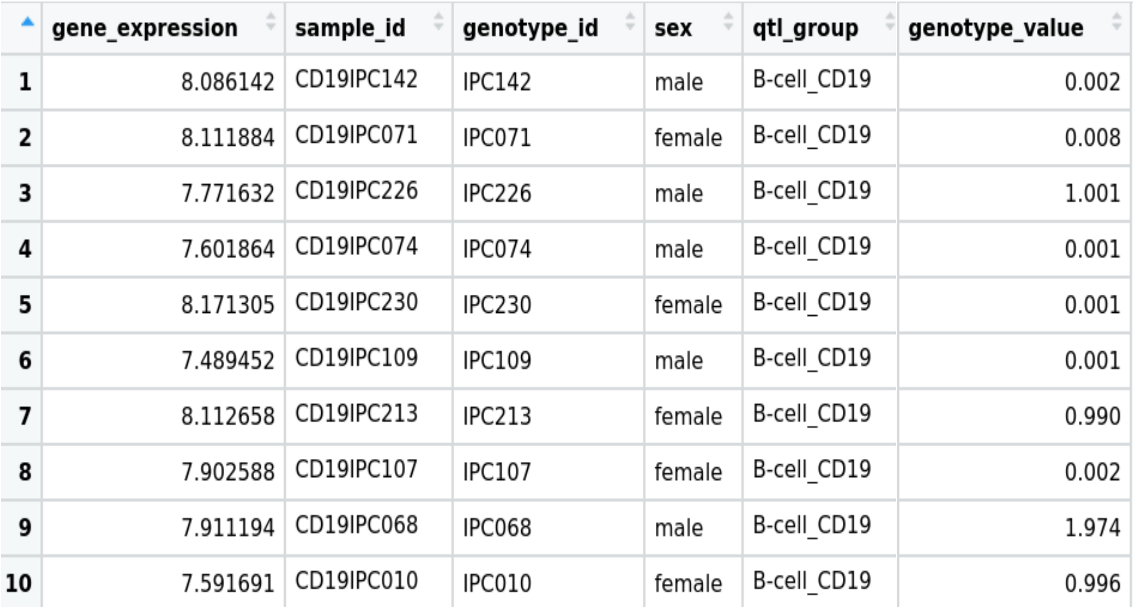
After importing normalised gene expression data into the R environment as a matrix with about 33 000 phenotype IDs by 2 337 samples, only preselected phenotypes with adjusted false discovery rate are selected creating a list of 1187 data frames (df). Each one df is joint with metadata containing gene expression information for 262 genotype ID with known sex in Table 5.



	gene_expression	sample_id	genotype_id	sex	qtl_group
1	8.086142	CD19IPC142	IPC142	male	B-cell_CD19
2	8.111884	CD19IPC071	IPC071	female	B-cell_CD19
3	7.771632	CD19IPC226	IPC226	male	B-cell_CD19
4	7.601864	CD19IPC074	IPC074	male	B-cell_CD19
5	8.171305	CD19IPC230	IPC230	female	B-cell_CD19
6	7.489452	CD19IPC109	IPC109	male	B-cell_CD19
7	8.112658	CD19IPC213	IPC213	female	B-cell_CD19
8	7.902588	CD19IPC107	IPC107	female	B-cell_CD19
9	7.911194	CD19IPC068	IPC068	male	B-cell_CD19
10	7.591691	CD19IPC010	IPC010	female	B-cell_CD19

Table 5. An example for one of the data frames of "ILMN_2043918" phenotype that has gene expression data joint with genotype ID.

To import with the genotype data into R on Windows 10, a compressed 6GB VCF file is converted into a binary GDS format from which variant names and coordinates (SNP ID) become easily accessible as well as genotypes for a specific variant or a matrix for all variants in a given region on the chromosome can be extracted. A list of 1187 data frames genotype and SNP ID for a specific variant is generated. As the two lists join, it generates a prepared input data, a list that contains 1187 gene-variant pairs that have gene expression, genotype value and sex in Table 6, which are required to perform an interaction test.



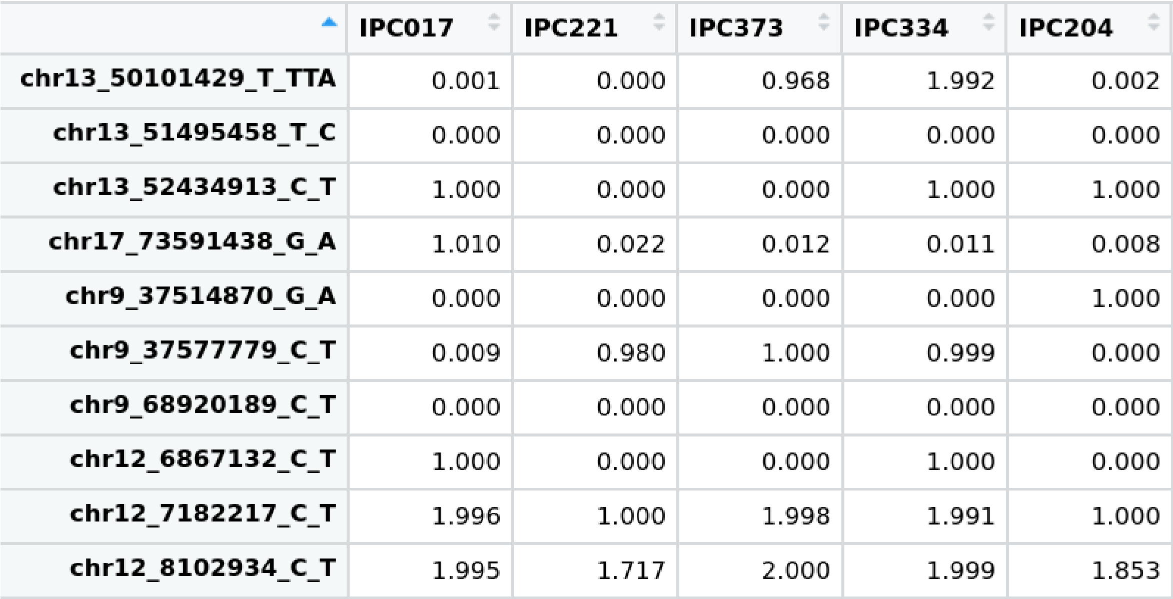
	gene_expression	sample_id	genotype_id	sex	qtl_group	genotype_value
1	8.086142	CD19IPC142	IPC142	male	B-cell_CD19	0.002
2	8.111884	CD19IPC071	IPC071	female	B-cell_CD19	0.008
3	7.771632	CD19IPC226	IPC226	male	B-cell_CD19	1.001
4	7.601864	CD19IPC074	IPC074	male	B-cell_CD19	0.001
5	8.171305	CD19IPC230	IPC230	female	B-cell_CD19	0.001
6	7.489452	CD19IPC109	IPC109	male	B-cell_CD19	0.001
7	8.112658	CD19IPC213	IPC213	female	B-cell_CD19	0.990
8	7.902588	CD19IPC107	IPC107	female	B-cell_CD19	0.002
9	7.911194	CD19IPC068	IPC068	male	B-cell_CD19	1.974
10	7.591691	CD19IPC010	IPC010	female	B-cell_CD19	0.996

Table 6. Prepared input data example for gene-variant pair "ENSG00000176124-chr13_50101429_T_TTA" that gene expression and genotype value for the interaction test.

However, I encountered challenges while extracting genotypes for a specific variant and a matrix for all variants. As initially all data manipulation in R were performed using Rstudio on Windows 10, the extraction of genotypes for a variant raised a problem as instead of finding the genotypes for the variant by index, it scanned from the whole file that took an enormous amount of time for about 20 to 30 minutes for each variant. Considering the 1187 variants, this approach was not viable. The extraction of a matrix for all variants per chromosome was memory extensive taking up to 6-7 GB of memory for each matrix. There are 22 chromosomes, thus 22 matrices were required to perform the interaction test, and taking into account that R does not remove unnecessary data from memory, this approach was not reasonable too.

To overcome this issue, the genotype VCF file was converted into a simple tab-separated genotype dosage file and indexed using tabix. I. To ease the memory load, a python script was created; it took an input text file with SNP IDs from B cells data frame after FDR correction and pulled all the genotypes into a dictionary using tabix module. The output of the python script in Table 7 was a text file (comma-separated values, CSV) that had all genotypes corresponding to SNP IDs from B cells data frame. This approach took a minimal amount of time of about 3 minutes.

From this genotype matrix, a list of 1187 data frames of genotype IDs and genotype values for each gene (or SNP ID).



	IPC017	IPC221	IPC373	IPC334	IPC204
chr13_50101429_T_TTA	0.001	0.000	0.968	1.992	0.002
chr13_51495458_T_C	0.000	0.000	0.000	0.000	0.000
chr13_52434913_C_T	1.000	0.000	0.000	1.000	1.000
chr17_73591438_G_A	1.010	0.022	0.012	0.011	0.008
chr9_37514870_G_A	0.000	0.000	0.000	0.000	1.000
chr9_37577779_C_T	0.009	0.980	1.000	0.999	0.000
chr9_68920189_C_T	0.000	0.000	0.000	0.000	0.000
chr12_6867132_C_T	1.000	0.000	0.000	1.000	0.000
chr12_7182217_C_T	1.996	1.000	1.998	1.991	1.000
chr12_8102934_C_T	1.995	1.717	2.000	1.999	1.853

Table 7. Genotype matrix as an output from the python script. It contains only relevant genotypes

2.1.4 Interaction test

The interaction test goal was to test a null hypothesis that gene expression of the gene was only effected by its genetic variant and by sex. An input data for the test was 1187 gene-variant pairs, where each gene-variant pair was tested for interaction between genetic variant (genotype value) and sex.

Applying linear regression, two models were created in Figure 4. Model 0 represents the null hypothesis. Model 1 is a modified model 0 version that besides genetic variant and sex, it is also affected by an interaction term between the genetic variant and sex. The interaction test is performed by analysis of variance (ANOVA) to test if model 1 with the interaction term fits the data significantly better than model 0, which only contains genotype and sex main effects.

```
     # Linear regression models
     model0 = lm(gene_expression ~ genotype_value + sex, gene_vari-
ant_df)
model1 = lm(gene_expression ~ genotype_value + sex + genotype_value*sex,
gene_variant_df)

     #interaction test
     test <- anova(model0, model1, test="LRT")
```

Figure 4 two linear models and interaction test written in R

2.2 RESULTS

2.2.1 Interaction test

The interaction test produces a P-value that demonstrates with null hypothesis passed or rejected. As the interaction test is applied to all the gene-variant pairs, P-values of those tests are collected in a vector. The vector of these P-values is illustrated on Figure 5 as a histogram, which has a uniform distribution meaning that for the majority of pairs the null hypothesis fails to be rejected. This type of P-values distributions is expected as previous studies (Yao et al. 2014; Jansen et al. 2014; Khramtsova, Davis, and Stranger 2019) there are few autosomal eQTLs that show sex-biased effects. However, to check if they are all null, FDR correction is applied and only values lower than 0.05 are passed.

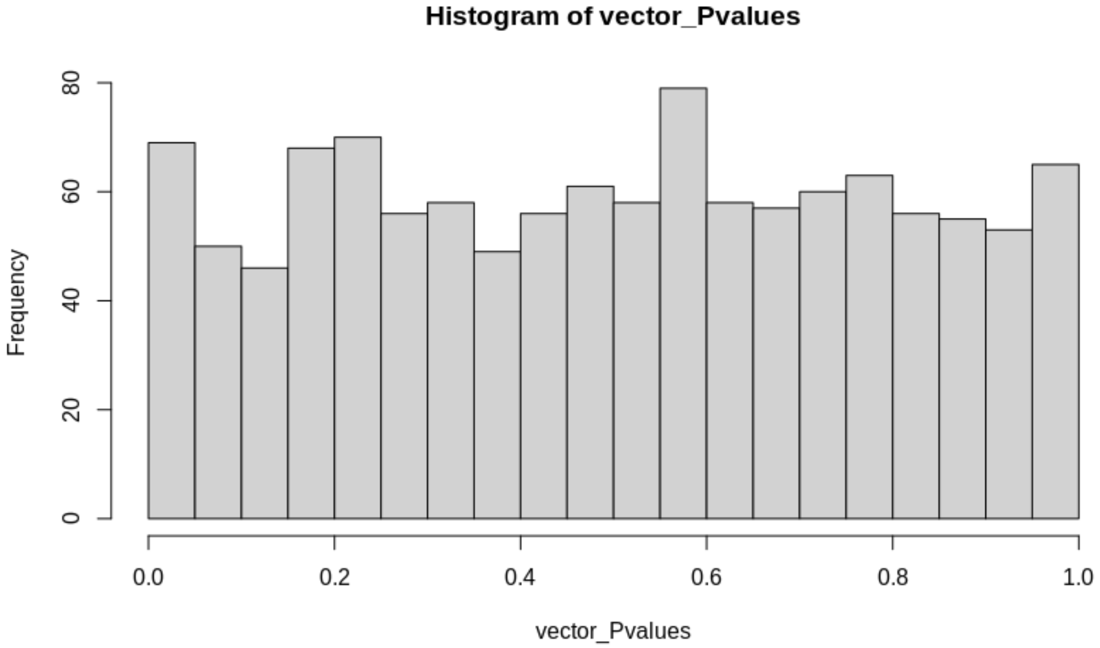


Figure 5 The Histogram of P-values for all interaction tests for every gene-variant pair.

The FDR correction reveals that two gene-variant pairs, which reject the null hypothesis meaning the alternative hypothesis (model 1) fits better for these two pairs: ENSG00000273802-chr6_26216356_C_T and ENSG00000188000-chr19_9163848_A_G

2.2.2 Identified gene-variant pairs: H2B clustered histone 8

The gene under ID, ENSG00000273802, is H2B clustered histone 8. Histones are nuclear proteins, their main role is to provide a nucleosome structure of the chromosomal fiber. The protein bears antifungal and antimicrobial activity. Figure 6 presents the first gene-variant pair (ENSG00000273802-chr6_26216356_C_T). The genotype distribution is almost discrete with values about 0 and about 1, but there is no genotype value 2.

There is also one outlier with genotype 1 in males, and thus it is likely to be a false positive. Due to this one outlier sample in males, the interaction test P-value could be inflated, that is the P-value is too small and thus it can be a false positive. However, there still seems to be a small difference in gradient between males and females.

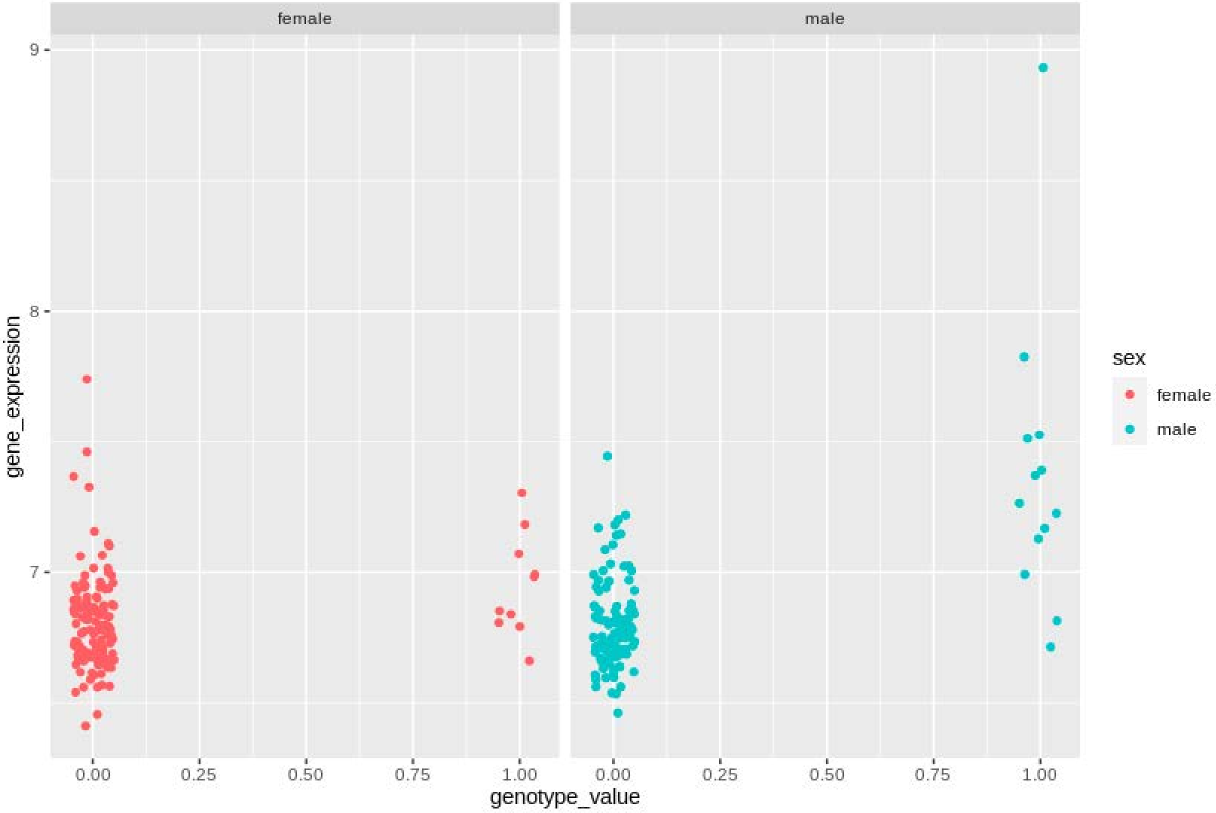


Figure 6. ENSG00000273802-chr6_26216356_C_T gene-variant pair. Gene expression on y axis and variant (genotype value) on x axis.

2.2.3 Identified gene-variant pairs: olfactory receptor

The gene under ID, ENSG00000188000, is olfactory receptor family 7 subfamily D member 2. Olfactory receptors initiate a neuronal response triggering the perception of a smell. The second gene-variant pair is for ENSG00000188000-chr19_9163848_A_G and shows more of the continuous nature in Figure 7, but still genotype value 2 is not presented only one outlier in females by x axis, that is likely to occur due to small sample size, and a larger dataset would contain all genotype values. The trend that for gradients in males is marginally larger for this pair is more clear, however there are still a few possible outliers in males.

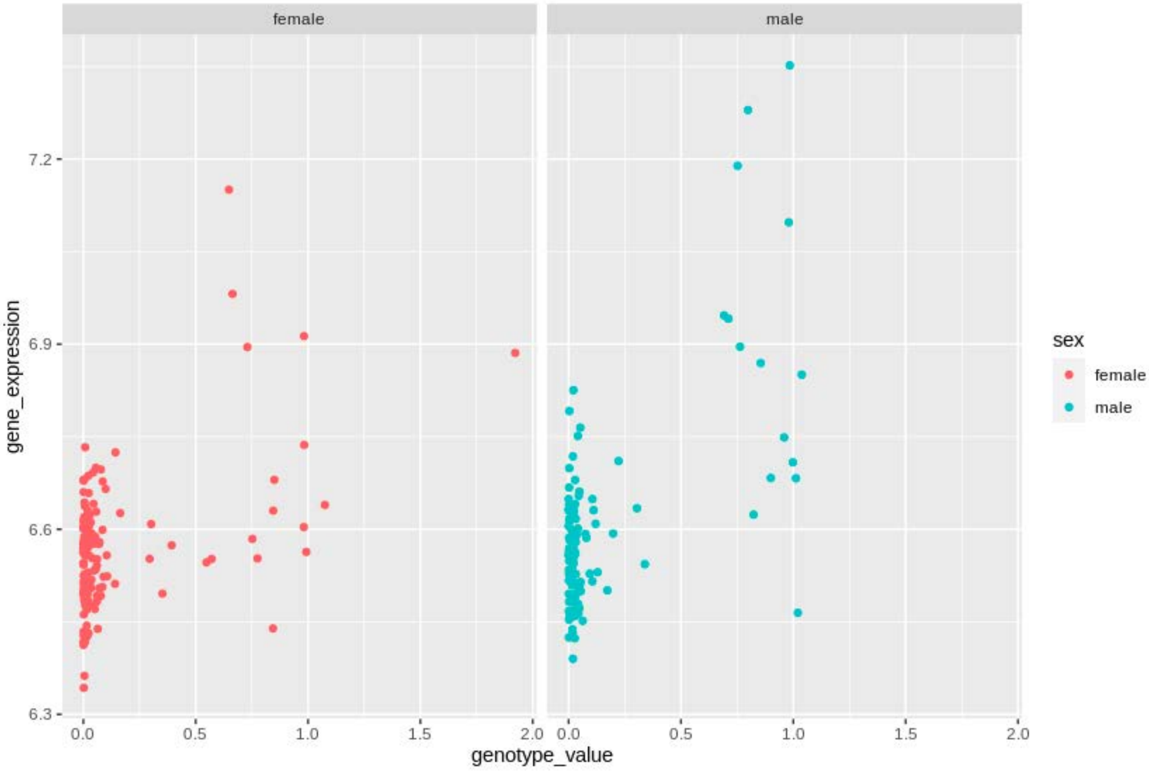


Figure 7. ENSG00000188000-chr19_9163848_A_G gene-variant pair. Gene expression on y axis and variant (genotype value) on x axis.

2.3 DISCUSSION

The method for discovery of sex-specific genentic effect on genotype should be considered as principally reliable due to the uniform distribution of P-values, since the null hypothesis fails to be rejected, meaning that for most pairs the gene expression in only influenced by the genetic variant and the sex but not a statistical interaction between the two, which is confirmed by the previous studies (Yao et al. 2014; Jansen et al. 2014; Khramtsova, Davis, and Stranger 2019). However, this method can find only single effect and differential genetic effects on gene expression (see Figure 1.2) since the analysis of variance is restrained to investigating gene-variant pairs that have a significant mean genetic effect between sexes.

The two gene-variant pairs that I identified should not be considered as reliable results and there are several reasons for that: mainly they may be considered to be false positives as the sample size is small (only 262 individuals). In addition, both gene-variant pairs have a few outliers as well as not fully represented range of genotype values. Moreover, taking into account the functions of one of the identified genes, H2B clustered histone (8ENSG00000273802), it is likely to be false positive as it has a basic structural function in eukaryotes.

The next steps of this work would be to increase the sample size that would increase the statistical power of the analysis by reducing the number of false positives and outliers. Further, eQTL analysis as well as other investigations involving quantitative traits in GWAS tend to have traits non normally distributed. Hence, by establishing the tests according to linear regression, they are conditioned to have reduced statistical power and inflated type I error, which is the presence of false positives in the results or rejecting a correct null hypothesis. The implementation of inverse normal transformation (INT) to traits with non-normal distribution leads to protection from the type I error and provides a better-powered association test for this type of data (McCaw et al., 2019.).

SUMMARY

The thesis aim was achived as 1187 gene-variant pairs were analysed using linear regration to model the null and alternative hypothesis and interaction tests to if alternative model fits significantly better than the null hypothesis. The null hypothesis is gene expression is only effected by genotype and sex. The alternative hypothesis is in addition to genotype and sex, it is affected by interaction term between sex and variant. The tests produced a uniform distribution of P-values that was expected and characterizes the method as reliable, as the null hypothesis, that sex-baised eQTLs are a rare, is supported by the recent studies (Yao et al. 2014; Jansen et al. 2014; Khramtsova, Davis, and Stranger 2019) as discussed in literature review.

False discovery rate indentified two sex-biased gene-variant pairs, however due to small sample size and few outliers on the data, it is likely that they are false positives, thus more investigations on larger datasets are neccassery to determine the correctness of these two hits.

REFERENCES

Albert, Frank W., and Leonid Kruglyak. ‘The Role of Regulatory Variation in Complex Traits and Disease’. *Nature Reviews Genetics*, vol. 16, no. 4, 4, Nature Publishing Group, Apr. 2015, pp. 197–212.

Battle, Alexis, et al. ‘Characterizing the Genetic Basis of Transcriptome Diversity through RNA-Sequencing of 922 Individuals’. *Genome Research*, vol. 24, no. 1, Jan. 2014, pp. 14–24.

Dimas, Antigone S., et al. ‘Sex-Biased Genetic Effects on Gene Regulation in Humans’. *Genome Research*, vol. 22, no. 12, Dec. 2012, pp. 2368–75.

Flynn, Emily, et al. ‘Sex-Specific Genetic Effects across Biomarkers’. *BioRxiv*, Cold Spring Harbor Laboratory, Nov. 2019, p. 837021.

Ginsburg, Geoffrey S. ‘Chapter 17 - Application of Human Genome Information to Clinical Practice’. *Genomic and Personalized Medicine (Second Edition)*, edited by Geoffrey S. Ginsburg and Huntington F. Willard, Academic Press, 2013, pp. 204–15.

Henry, Lionel, et al. *Purrr: Functional Programming Tools*. 0.3.4, 2020. *R-Packages*, https://CRAN.R-project.org/package=purrr.

Jansen, Rick, et al. ‘Sex Differences in the Human Peripheral Blood Transcriptome’. *BMC Genomics*, vol. 15, Jan. 2014, p. 33.

Joehanes, Roby, et al. ‘Integrated Genome-Wide Analysis of Expression Quantitative Trait Loci Aids Interpretation of Genomic Association Studies’. *Genome Biology*, vol. 18, no. 1, Jan. 2017, p. 16.

Kerimov, Nurlan, et al. ‘EQTL Catalogue: A Compendium of Uniformly Processed Human Gene Expression and Splicing QTLs’. *BioRxiv*, Cold Spring Harbor Laboratory, Jan. 2020, p. 2020.01.29.924266.

Khramtsova, Ekaterina A., et al. ‘The Role of Sex in the Genomics of Human Complex Traits’. *Nature Reviews Genetics*, vol. 20, no. 3, 3, Nature Publishing Group, Mar. 2019, pp. 173–90.

Knijnenburg, Theo A., et al. ‘Fewer Permutations, More Accurate P-Values’. *Bioinformatics*, vol. 25, no. 12, June 2009, pp. i161–68.

Kukurba, Kimberly R., et al. ‘Impact of the X Chromosome and Sex on Regulatory Variation’. *Genome Research*, vol. 26, no. 6, June 2016, pp. 768–77.

Li, Heng. ‘Tabix: Fast Retrieval of Sequence Features from Generic TAB-Delimited Files’. *Bioinformatics*, vol. 27, no. 5, Oxford Academic, Mar. 2011, pp. 718–19.

McCaw, Zachary R., et al. ‘Operating Characteristics of the Rank-Based Inverse Normal Transformation for Quantitative Trait Analysis in Genome-Wide Association Studies’. *Biometrics*, vol. n/a, no. n/a.

Momozawa, Yukihide, et al. ‘IBD Risk Loci Are Enriched in Multigenic Regulatory Modules Encompassing Putative Causative Genes’. *Nature Communications*, vol. 9, no. 1, 1, Nature Publishing Group, June 2018, p. 2427.

Murray, Charlotte. ‘Various Types of Variant: What Is Genomic Variation?’ *Genomics Education Programme*, 16 Sept. 2016. *www.genomicseducation.hee.nhs.uk*, https://www.genomicseducation.hee.nhs.uk/blog/various-types-of-variant-what-is-genomic-variation/.

Nica, Alexandra C., and Emmanouil T. Dermitzakis. ‘Expression Quantitative Trait Loci: Present and Future’. *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 368, no. 1620, June 2013.

Piasecka, Barbara, et al. ‘Distinctive Roles of Age, Sex, and Genetics in Shaping Transcriptional Variation of Human Immune Responses to Microbial Challenges’. *Proceedings of the National Academy of Sciences*, vol. 115, no. 3, National Academy of Sciences, Jan. 2018, pp. E488–97.

R: The R Project for Statistical Computing. https://www.r-project.org/.

Schmiedel, Benjamin J., et al. ‘Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression’. *Cell*, vol. 175, no. 6, Elsevier, Nov. 2018, pp. 1701-1715.e16.

Tukiainen, Taru, et al. ‘Landscape of X Chromosome Inactivation across Human Tissues’. *Nature*, vol. 550, no. 7675, 7675, Nature Publishing Group, Oct. 2017, pp. 244–48.

Wickham, Hadley, Romain François, et al. *Dplyr: A Grammar of Data Manipulation*. 0.8.5, 2020. *R-Packages*, https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Winston Chang, et al. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. 3.3.0, 2020. *R-Packages*, https://CRAN.R-project.org/package=ggplot2.

Yao, Chen, et al. ‘Sex- and Age-Interacting EQTLs in Human Complex Diseases’. *Human Molecular Genetics*, vol. 23, no. 7, Oxford Academic, Apr. 2014, pp. 1947–56.

Zheng, Xiuwen. *A Tutorial for the R Package SNPRelate*. p. 26.

NON-EXCLUSIVE LICENCE TO REPRODUCE THESIS AND MAKE THESIS PUBLIC

I, *Vladislav Tuzov,*

(*author's name*)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

How sex influences the effect of genetic variants on gene expression? ,

(*title of thesis*)

supervised by *PhD Kaur Alasoo*.

(*supervisor's name*)

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Vladislav Tuzov

dd/mm/yyyy