WARSAW UNIVERSITY OF TECHNOLOGY

FACULTY OF MATHEMATICS AND
INFORMATION SCIENCE

FIRST SEMESTER 2021/2022

Computational Genomics

PROJECT 2

BY:

Andra Umoru (324334)

SUBMITTED TO:

Tojek Zofia (DOKT)

1. Table of Contents

2. List of Figures

3. Project Description

Project 2 consists in performing SV-eQTL analysis on RNA-Seq data and a list of population structural variants. The output from the analysis should be the list of discovered SV-eQTL along with the p-value of the given regression (SV / expression relationship).

Requirements:

- Working algorithm for eQTL analysis.
- Unit tests (e.g., unit tests can be on RNA-Seq / SV data slice; they can also be on other organisms; however, must work for SV data from the 1000 Genomes Phase 3 project, and for RNA-Seq data from gEUVADIS).
- Comparison with other results (published / state-of-the-art).
- Final report
- Presentation of the program operation and results during the last/second to last classes.

4. Brief Overview on genomics

4.1 What is an eQTL?

A gene expression phenotype is explained by an eQTL, which is a locus that explains a percentage of the genetic variance. A direct association test between markers of genetic variation and gene expression levels commonly assessed in tens or hundreds of people is used in standard eQTL study [2], [4]. This association study might be done close to the gene or far away from it. One of the key benefits of employing the GWAS approach for eQTL mapping is that it allows for the discovery of new functional loci without the need for prior knowledge of specific cis or trans regulatory areas [2].

To put it another way, an expression quantitative trait is the amount of mRNA transcript or protein present. These are usually the result of a single gene that is located on a certain chromosome. This distinguishes expression quantitative features from the majority of complex traits, which are not the result of a single gene's expression. eQTLs are chromosomal locations that explain variation in expression characteristics. Local eQTLs, also known as cis-eQTLs, are eQTLs that are found near the gene of origin (the gene that creates the transcript or protein) [2].

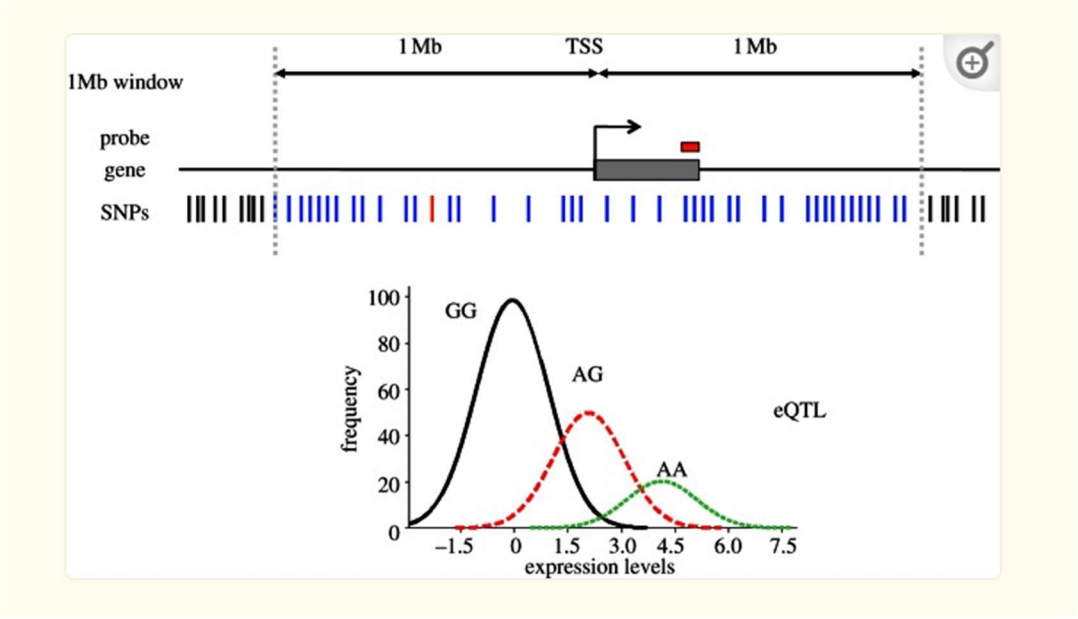

Figure 1: A typical eQTL; Source **[4]**

4.2 SV-eQTL Analysis

By considering gene expression quantifications as quantitative features, eQTL analysis intends to investigate the impact of genetic variation on gene expression. Genotypes of genetic variations and gene expressions are both analyzed for the same set of samples in a population in a typical eQTL setup. eQTL analysis, in particular, allows for integrative analysis by discovering significant connections between genetic variations and gene expressions [1].

5. Algorithm for SV-eQTL Analysis

The algorithm for performing the SV-eQTL Analysis on **RNA-Seq data** and **list of population structural variants.** This algorithm is structured to read the data from the above-mentioned files written in bold. The expected output of this algorithm is a list of discovered structural variants and **p-value** of the given expressions (SV/expression relationship). The algorithm for this project is a pseudocode and a flowchart.

5.1 Pseudocode

Keys:

- **Input:** RNA-Seq data file (**RNA-Seq**), list of population structural variants (**LPSV**)

i. Start

ii. Load the input file **(RNA-Seq)** and (**LPSV**).

iii. Read the data from the two files loaded in step (ii).

iv. Map the Transcript Start Site (TSS) and get the Gene Position from the files (WGS VCF file and RNA-Seq file).

v. From the Genes locations, get the variation close to the genes, for each, calculate linear regression and get the P-Values. Return the regression with P-Values > 0.05.

vi. Select, sort and list the structural variants.

vii. Print the list of discovered SV-eQTL and the P-Values of the expression.

viii. End

5.2 Flowchart

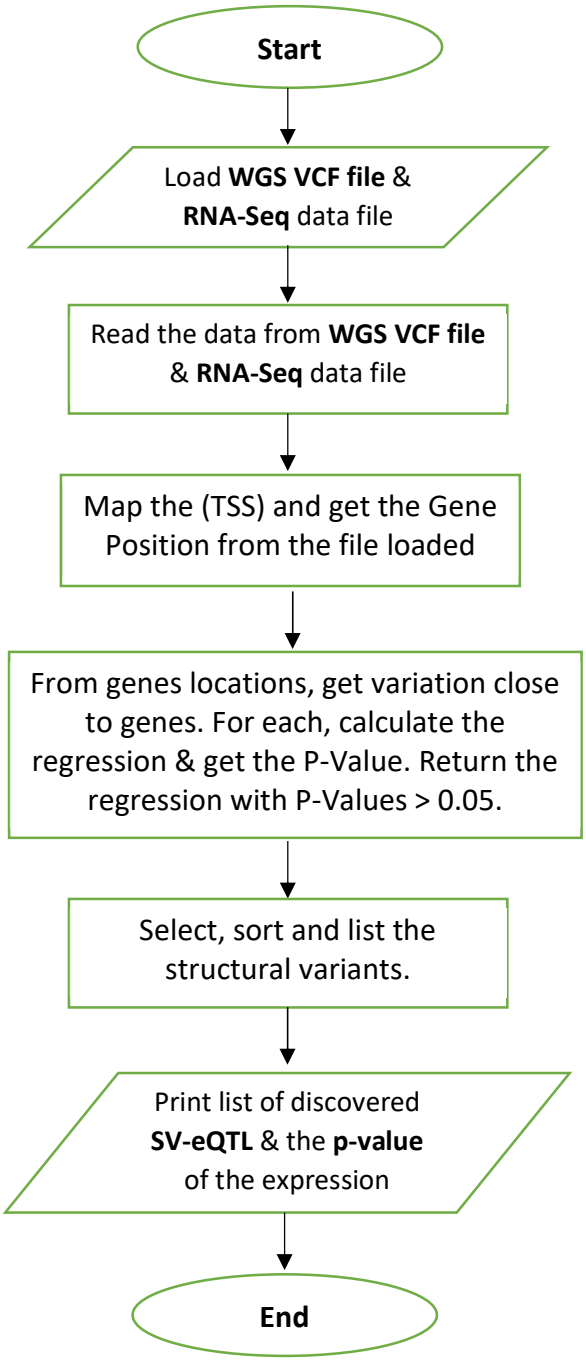

Figure 2: Algorithm (Flowchart)

6. Implementation

This program was implemented using R programming language. In the course of implementation, the following R libraries are used: library(ggplot2), library (SNPlocs.Hsapiens.dbSNP.20101109), library (devtools), library (Biobase) and library (VariantAnnotation).

7. Tests

The test was conducted on the SV data from the 1000 Genomes and for RNA-Seq data. The test produced some considerable results which I am not completely satisfied with the outcome of the program. The results of the tests are hereby presented below:

8. Results

The following are the results obtained:

8.1 Genotype Dosage

The figure below shows the genotype dosage as fetched from the SV data from the 1000 Genomes given.

```
> DS[1:10,]
           HG00096 HG00097 HG00099 HG00100 HG00101
rs7410291        0    0.00       1       0    0.00
rs147922003      0    0.00       0       0    0.00
rs114143073      0    0.00       0       0    0.00
rs141778433      0    0.00       0       0    0.00
rs182170314      0    0.05       0       0    0.00
rs115145310      0    0.00       0       0    0.00
rs186769856      0    0.00       0       0    0.05
rs77627744       0    0.00       0       0    0.00
rs193230365      0    0.00       0       0    0.00
rs9627788        0    0.00       1       0    0.00
```

Figure 3: Genotype Dosage

8.2 DNA String Set

The figure below shows sequence of ten (10) DNA strings set.

```
> ref(my_vcf)[1:10]
DNAStringSet object of length 10:
     width seq
 [1]     1 A
 [2]     1 C
 [3]     1 G
 [4]     1 C
 [5]     1 C
 [6]     1 G
 [7]     1 T
 [8]     1 G
 [9]     1 G
[10]     1 T
```

Figure 4:DNA String Set

8.3 SNPs Quality Table

The figure below also shows the quality of the SNPs. As it can be seen below, most have a quality value of 100 with the exception of some few SNPs which have more different quality value.

```
> qual(my_vcf)[1:466]
  [1] 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100
 [20] 100 100 100 100 100 100 100 100 100 100 100 100 100 100 425 100 298 100 100
 [39] 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100
 [58] 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100
 [77] 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100
 [96] 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100
[115] 100 100 100 100 100 100 100 100 100 100 100 100 497 100 100 100 100 100 100
[134] 100 100 100 100 100 100 100 100 213 100 100 100 100 100 100 100 100 100 100
[153] 100 100 100 100 100 100 100 100 100 100 100 100 242 100 100 100 100 100 100
[172] 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100
[191] 100 100 100 100 100 100 100 100 100 100 100 100 100 305 100 384 100 100 100
[210] 183 100 100 100 100 100 249 100 100 100 100 100 100 100 100 100 100 100 100
```

Figure 5: The SNPs Quality

8.4 GG Plot (Input Quality)

The plot below shows the input quality plotted against the density:

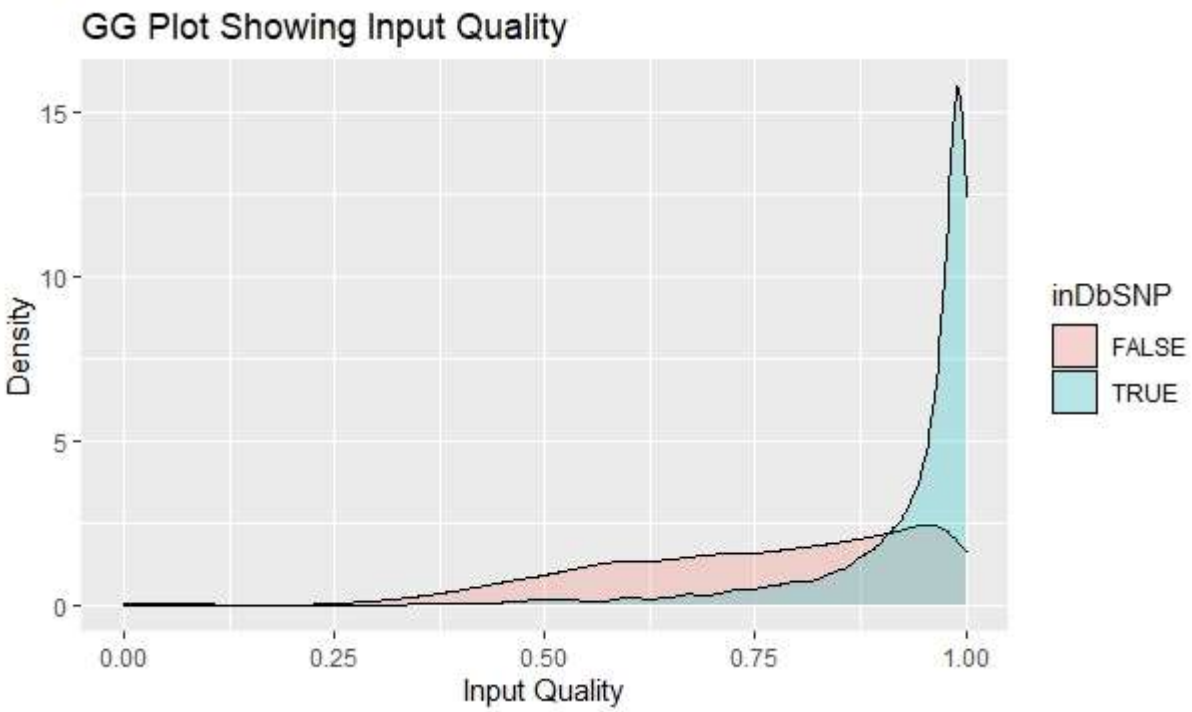

Figure 6: GG Plot Showing Input Quality

8.5 Histogram for the P-Values

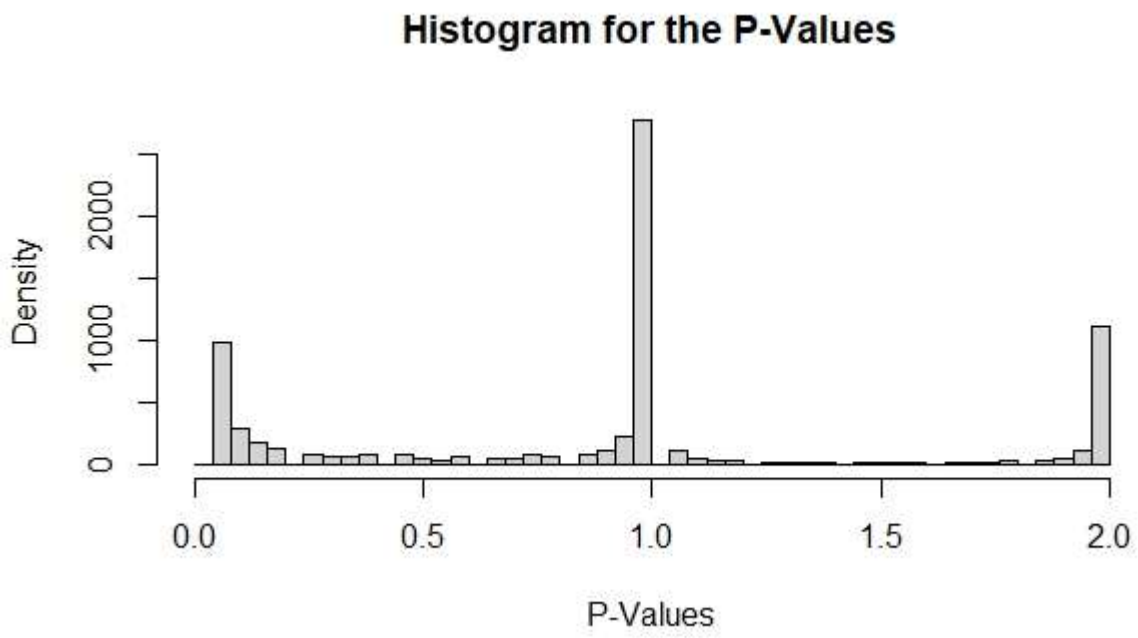

Figure 7: Histogram for the P-Values

8.6 Histogram for Genotype Dosage (DS)

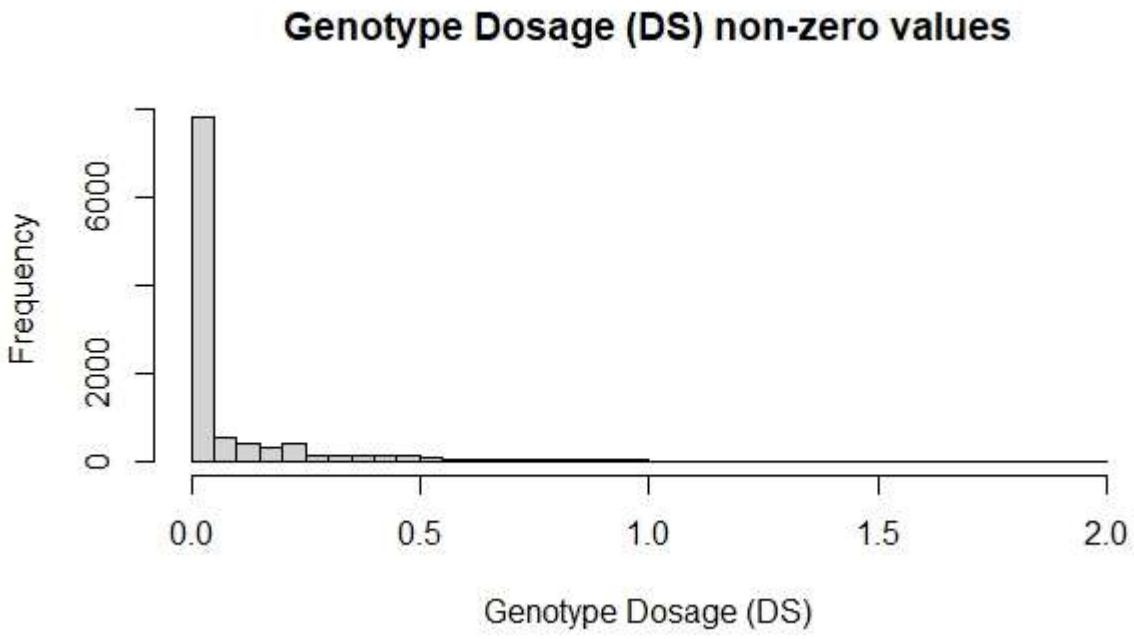

Figure 8: Histogram showing Genotype Dosage (DS)

8.7 QQ Plot for Genotype against the Expression

The QQ Plot below shows the number of genotypes against the gene expression.

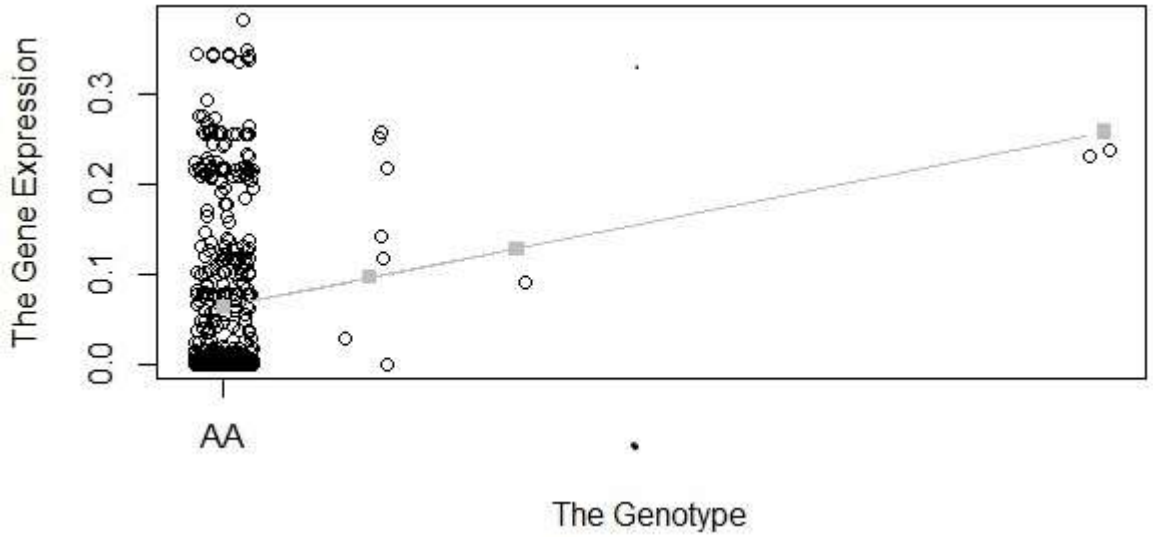

Figure 9: A QQ Plot for The Genotype against the Expression

8.8 P-Value, Intercept and Error

The table below shows the p-value, intercept, standard error, the highest and the lowest confidence interval obtained after running the linear model. Also, it can be observed that the p-value is **P < 0.05;** this means that we accept the alternate hypothesis and indicates that there is a relationship between the genotype and the gene expression.

```
> tidy(lm1, conf.int = TRUE, conf.level = 0.95)
# A tibble: 2 x 7
  term        estimate std.error statistic  p.value conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
1 (Intercept)   0.0636   0.00432      14.7  1.44e-40   0.0551    0.0721
2 My_SNPS       0.648    0.205         3.16 1.66e- 3   0.246     1.05
```

Figure 10: A Table showing the P-Value and Intercept

8.9 Histogram for Gene TSS > 0.05

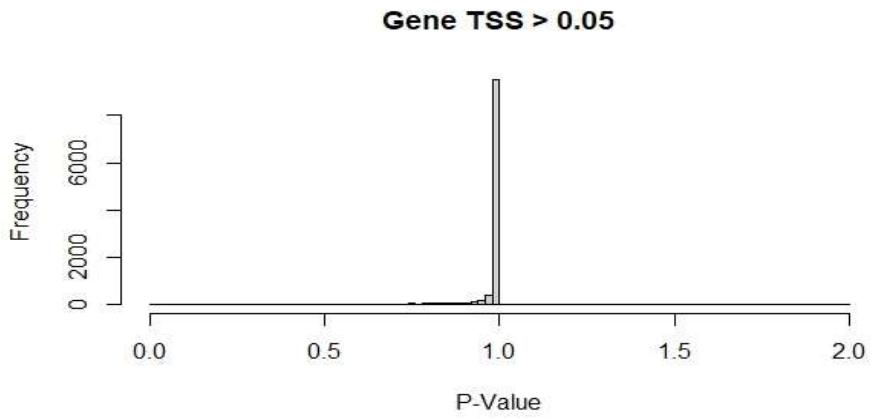

Figure 11: Histogram for Gene TSS > 0.05

8.10 Results Analysis status

```
75.00% done, 613,588 eQTLs
77.77% done, 640,682 eQTLs
80.55% done, 667,191 eQTLs
83.33% done, 667,901 eQTLs
86.11% done, 668,592 eQTLs
88.88% done, 669,290 eQTLs
91.66% done, 669,992 eQTLs
94.44% done, 670,665 eQTLs
97.22% done, 671,348 eQTLs
100.00% done, 671,367 eQTLs
Task finished in 34.56 seconds
```

Figure 12: Results Analysis Status

8.11 Histogram showing all P-Values

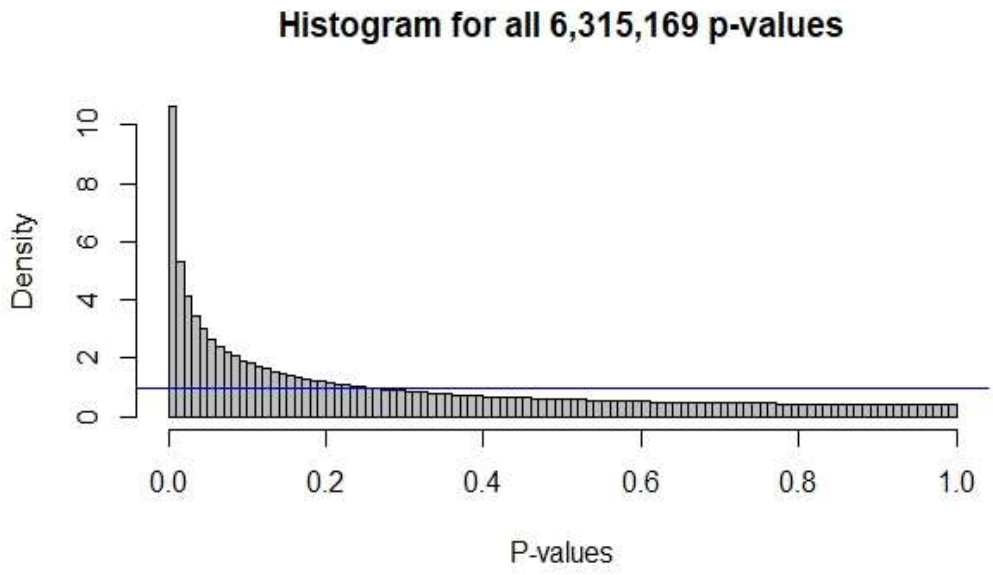

Figure 13: Histogram for all P-Values

8.12 Q-Q Plot for all P-Values

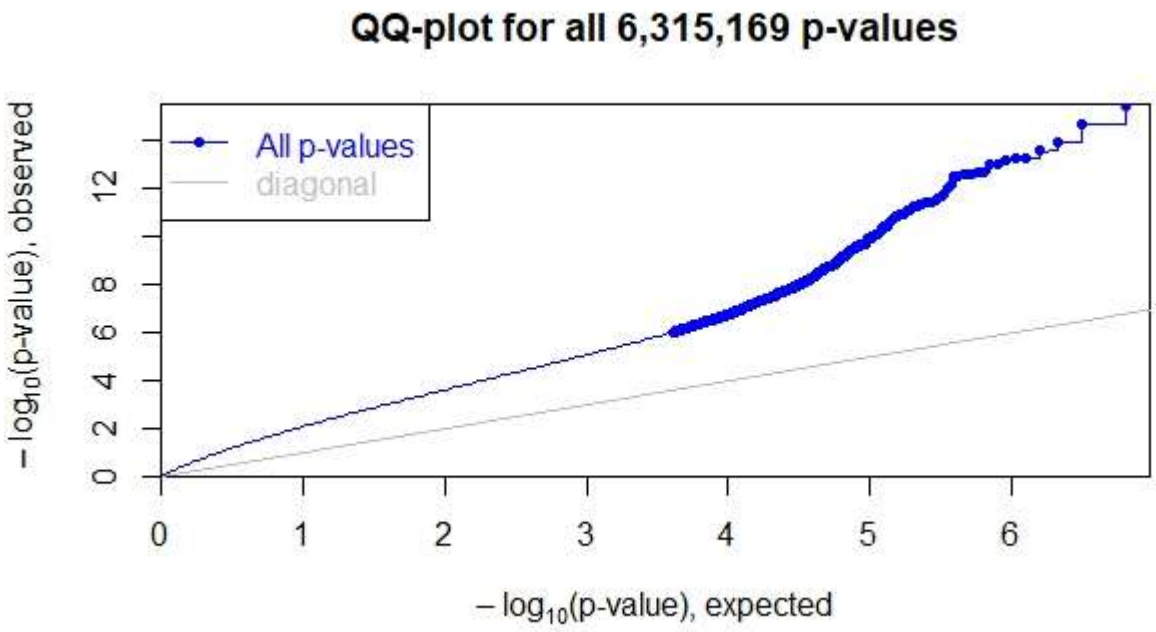

Figure 14: Q-Q Plot for all P-Values

9. Existing solution

Even though my implementation is not completed and built to the required standard, some of the results obtained shows some decent similarities with the results obtained from some of the estate of the earth e-QTL analysis programs like the Matrix e-QTL.

10. Conclusion

In conclusion, I was able to develop my own algorithm (Pseudocode and Flowchart) for the e-QTL analysis and tried implementing it using R programming language. The results of the implementation are presented above. There is need for more work to be done in order to fine-tune the program so that it can stand shoulder to shoulder with other state of the art programs for performing e-QTL analysis.

11. References

1. Andrew Q., James J., Xinghua S. (2017). Bayesian Hyperparameter Optimization for Machine Learning Based eQTL Analysis. Genomic Variation and Disease. Pp. 1-9.
2. https://en.wikipedia.org/wiki/Expression_quantitative_trait_loci

3. A Step-by-Step Explanation of Principal Component Analysis (PCA) https://builtin.com/data-science/step-step-explanation-principal-component-analysis
4. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3682727/#:~:text=An%20eQTL%20is%20a%20locus,tens%20or%20hundreds%20of%20individuals.
5. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1728-x
6. http://jtleek.com/genstats/inst/doc/04_10_eQTL.html