Quantitative Genomics and Genetics - Spring 2020
BTRY 4830/6830 Project

Kevin Van Vorst - kpv23

May 12, 2020

1 Introduction

Since the completion of the Human Genome Project in 2003, more advanced research methods have been developed to study the genome for genetic variations [2]. Genome-wide association study (GWAS) is one method that analyzes a genome for locations of polymorphisms which can be potentially linked to diseases. A polymorphism is a position in the genome where variation of the DNA at this location produces an effect on the phenotype expression. This project focuses on using a kind of polymorphism called single-nucleotide polymorphisms (SNPs) to identify these locations of genetic variation. One major limitation of GWAS is that it solely provides information on the location of these SNPs. Therefore, further experimentation and research must be done to properly synthesize conclusive evidence for linking significant SNPs to risk factors. Expression Quantitative Trait Locus (eQTL) analysis is a type of GWAS where the phenotype studied is gene expression. This project uses an eQTL analysis to analyze a subset of data provided by the Genetic European Variation in Health and Disease (gEUVADIS) collection. Statistical modeling methods are used to identify locations of significant SNPs and to justify that these SNPs have an effect on phenotype expression.

2 Datasets

The genotype dataset is a subset of the gEUVADIS collection which contains 344 samples of 50,000 genotypes from 4 European ancestral populations: CEU (Utah residents with European ancestry), FIN (Finns), GBR (British), and TSI (Toscani). The phenotype dataset contains 344 samples of the expression levels of the 5 genes of interest: ERAP2 (ENSG00000164308.12), PEX6 (ENSG00000124587.9), FAHD1 (ENSG00000180185.7), GFM1 (ENSG00000168827.9), MARCH7 (ENSG00000136536.9). The covariate dataset includes the sex and European descent each sample is associated with. Additionally, gene info and snp info datasets are provided for further analysis of significant SNPs.

Before starting the GWAS, the genotype and phenotype datasets are put through a filtering process. First, any samples with missing data or empty cells are identified and removed. Second, genotypes with a low minor allele frequency (MAF) are removed as well. MAFs are associated with the power of the statistical test which is the probability of rejecting the null hypothesis when it is in fact false. As the MAF decreases, so does the power of the test, therefore it is critical to filter out low MAFs to minimize the false positive rate. MAFs strictly less than 5% are deemed as rare variants in the genome and filtered out. After filtering both datasets, it is found that no samples or genotypes need to be removed.

Potential outliers in the phenotype data are identified when each phenotype is visualized by a histogram. Each histogram should follow a normal distribution and have no unusual spikes by the tails. As seen below,

the phenotype data for the five genes follow a normal distribution and have no extreme outliers that need to be removed.

3 eQTL Method

The goal of this eQTL analysis is to take the large and complicated genotype and phenotype datasets and model them such that there is a simplified relationship. Linear regression models are a great way to find relationships between many independent variables and a single dependent variable. A model used for the project is as such:

$$Y = \beta_\mu + X_a\beta_a + X_d\beta_d + X_z\beta_z + \epsilon$$

To input our data into this model, we know that phenotype expression is the dependent variable so the phenotype dataset is the y matrix. The Xa matrix only takes certain inputs (-1, 0, 1) that our genotype set does not reach yet (0, 1, 2). To convert this into the Xa matrix, the genotype dataset will simply be subtracted by 1. The Xd matrix, which maps homozygotes and heterozygotes, is dependent on Xa and is calculated as follows:

$$X_d = 2\cdot \mid X_a \mid -1$$

The Xz matrix contains all the covariates that are to be accounted for in the regression. The two identified covariates are the sex and ancestry group of each sample taken. For the sex covariate, males are coded as -1 and females are coded as 1. For the ancestry groups, each population is assigned a unique number: -1 for CEU, 0 for FIN, 1 for GBR, and 2 for TSI. The two newly coded covariate matrices are then appended together to create a single Xz matrix.

Since the betas are unknown they must be estimated with the information already given in this model. Therefore, a maximum likelihood estimator will be used since it will produce the most accurate beta values.

$$MLE(\hat{\beta}) = (x^T x)^{-1} x^T y$$

This model defines the null hypothesis stating that there is no significant genetic variation in the gene that leads to the variation of phenotype expression. Therefore, the alternative hypothesis states that there is some kind of genetic variation that can be located and those significant SNPs do cause variation of phenotype expression.

$$H_0 : \beta_a = 0 \cap \beta_d = 0 \quad H_A : \beta_a \neq 0 \cup \beta_d \neq 0$$

Statistical tests are performed to inform justify whether to accept or reject the null hypothesis. A popular statistical test used in population genetics is called the F-test, which is in this project to determine whether to accept or reject the null hypothesis. An F-statistic is then used to produce p-values for each corresponding genotype. The p-values are critical to identifying significant SNPs. If a p-value is lower than a set alpha value, then it is significant evidence against the null hypothesis. The set alpha value of 0.05 which is normally used to represent 95% confidence (1-α). Since 50,000 tests for each phenotype are performed, a bonferroni correction on the alpha value must be done.

$$Bonferroni\ Corrected\ p\ value\ = \frac{\alpha}{N} = \frac{0.05}{50000}$$

The new cut off will be $1.0 \cdot 10^{-6}$ which means any p value strictly greater than -$\log_{10}(1.0 \cdot 10^{-6})$ is significant. The p values are then plotted on a bonferroni corrected manhattan plot to visualize any significant outliers. Also, observed p values of each gene are plotted against p values generated by a uniform distribution in a Quantile-Quantile Plot (QQ-Plot) to further show if there are any significant SNPs that exist.

4 Results

After performing an eQTL analysis on the five genes of interest, three of them show potential of clear location to significant SNPs specifically the genes ERAP2, PEX6, and FAHD1. All QQPlots for the genes' corresponding p values did not show an exact 45 degree line, however further investigation of the manhattan plots for genes GFM1 and MARCH7 showed there are no significant SNPs hit above the bonferroni corrected cutoff. ERAP2, PEX6, and FAHD1 have more than one significant SNP, but the most significant or the largest $-log_{10}(pvalue)$ is the most important to analyze. Further SNP information was searched and extracted from the snpinfo dataset provided.

According to the ERAP2 Manhattan plot, the most significant SNP occurs at index 16,784 of the genotype dataset which has the header genotype of "rs7726445." The ERAP2 QQ-Plot shows it trends to an S-shape tail meaning the associated genotypes are in linkage disequilibrium with the SNP. rs7726445 is located on chromosome 5 at position 96945338. Mutations on the ERAP2 gene are associated with the inflammatory arthritis syndrome ankylosing spondylitis and pre-eclampsia[3].

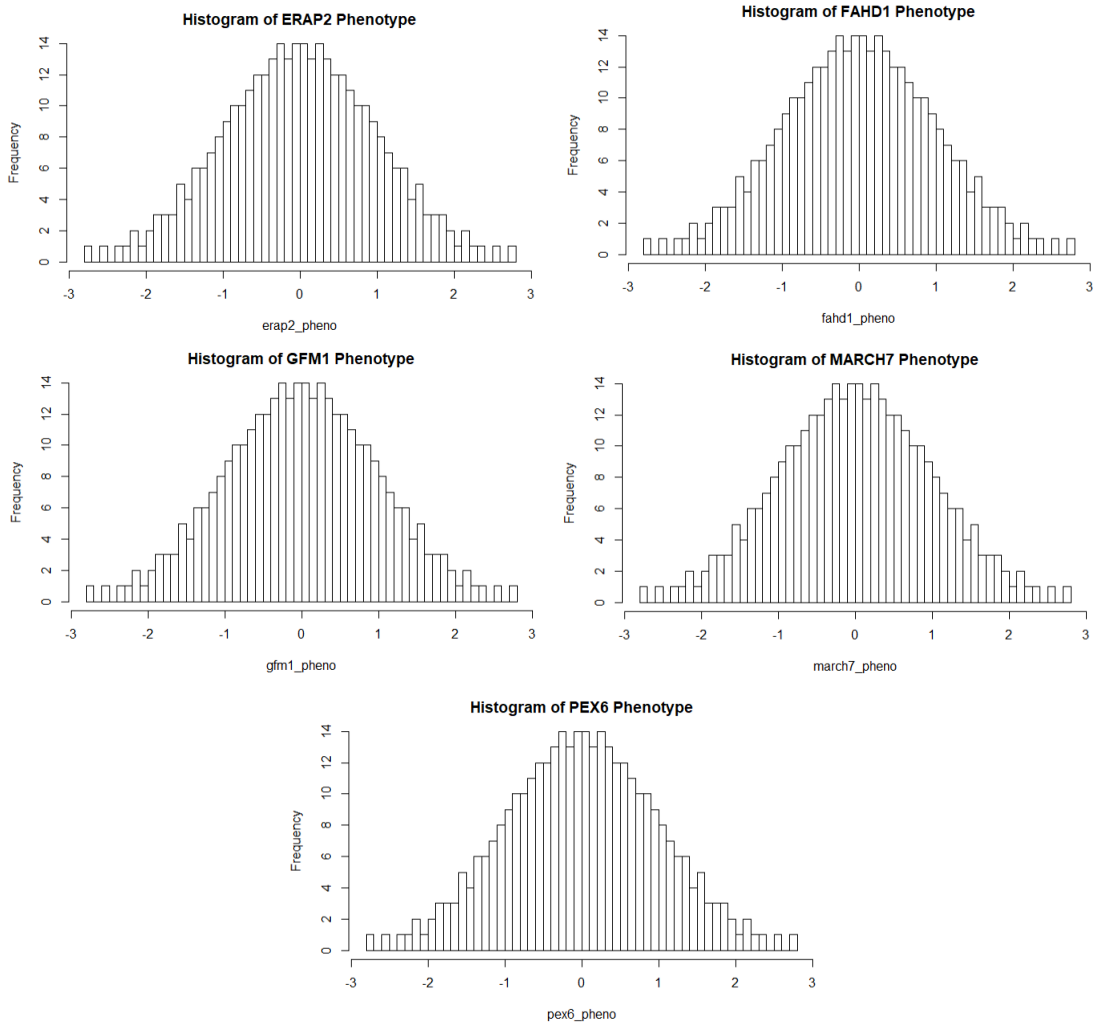
According to the PEX6 Manhattan plot, the most significant SNP occurs at index 19,286 of the genotype dataset which has the header genotype of "rs1129187." The PEX6 QQ-Plot shows it trends to an S-shape tail meaning the associated genotypes are in linkage disequilibrium with the SNP. rs1129187 is located on chromosome 6 at position 42964461. Mutations on the PEX6 gene cause peroxisome biogenesis disorders such as Zwelleger syndrome[1][5].

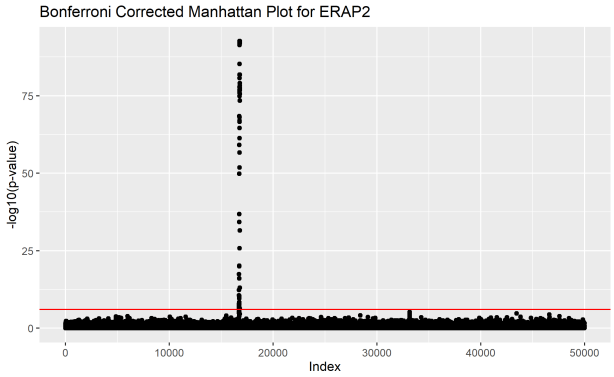
For the FAHD1 gene, the most significant SNP occurs at index 41938 of the genotype dataset which has the header genotype of "rs11644748". The FAHD1 QQ-Plot shows it creates an S-shape tail meaning the associated genotypes are in linkage disequilibrium with the SNP. rs11644748 is located on chromosome 16 at position 1829958. Diseases associated with FAHD1 include Ecthyma and Cerebral Lymphoma[4].

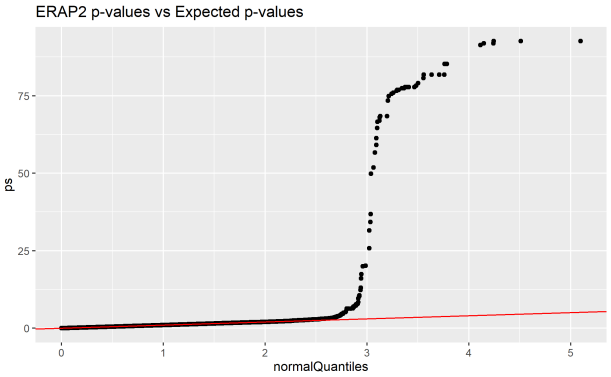
5 Conclusion

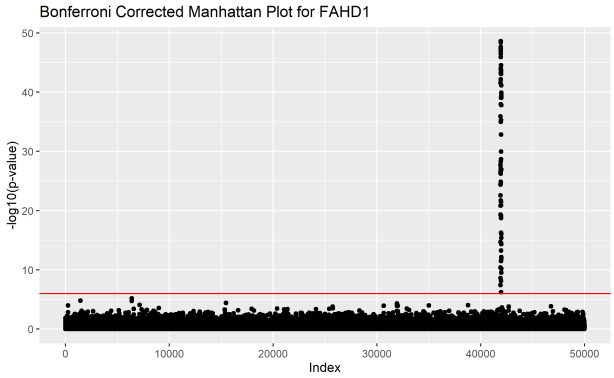
An eQTL analysis on part of the gEUVADIS dataset yielded 3 genes with numerous significant SNPs. A linear regression model was used to parse relevant datasets, analyze, and locate significant SNPs on each gene. This study only provides insight on probable locations of genetic variation greatly affecting phenotype expression and in no way explains how these variations are done or what their effects are, but only slightly explores possibilities considering their associated genes and already existing research on them. Results recommend that the genes ERAP2, PEX6, and FAHD1 should go under further experimentation and research.

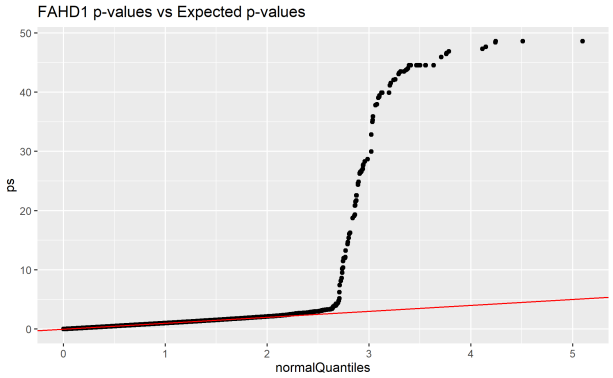
6 Figures

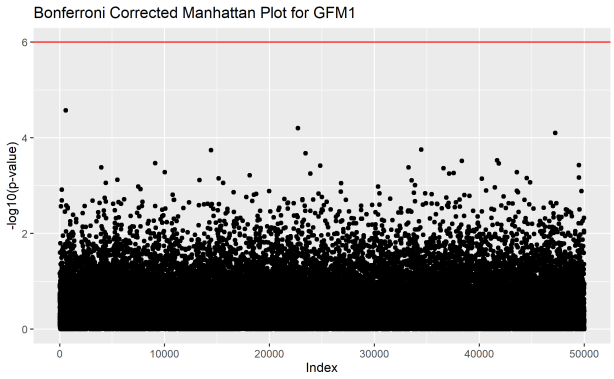

Histogram of ERAP2 Phenotype
Frequency
erap2_pheno
Histogram of FAHD1 Phenotype
Frequency
fahd1_pheno
Histogram of GFM1 Phenotype
Frequency
gfm1_pheno
Histogram of MARCH7 Phenotype
Frequency
march7_pheno
Histogram of PEX6 Phenotype
Frequency
pex6_pheno

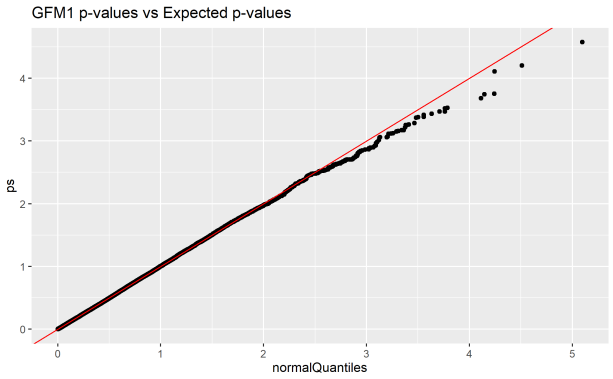

Bonferroni Corrected Manhattan Plot for ERAP2
-log10(p-value)
Index

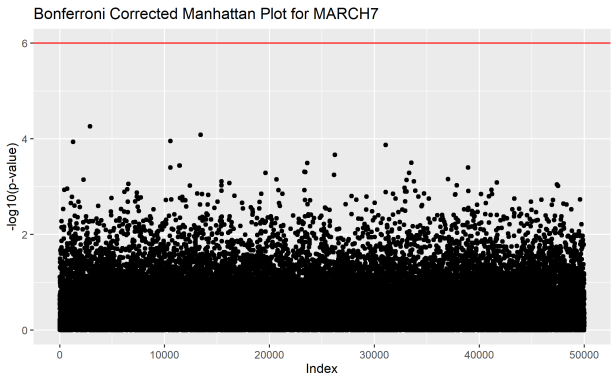

ERAP2 p-values vs Expected p-values
ps
normalQuantiles

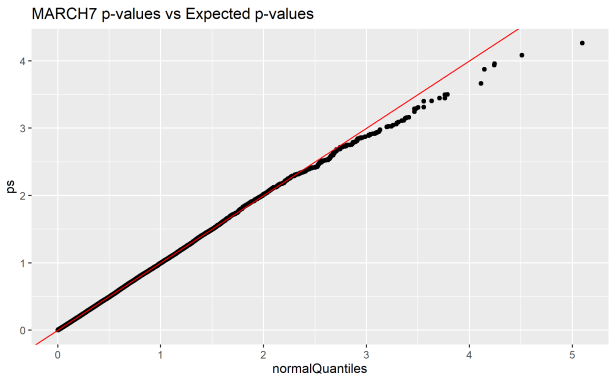

Bonferroni Corrected Manhattan Plot for FAHD1
-log10(p-value)
Index

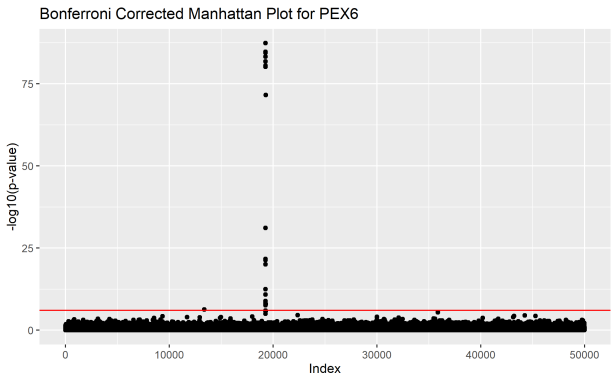

FAHD1 p-values vs Expected p-values
ps
normalQuantiles

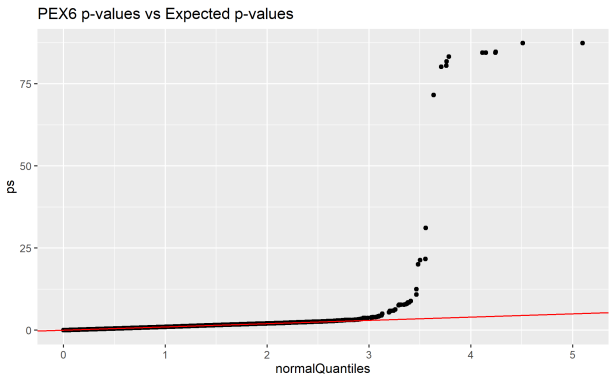

Bonferroni Corrected Manhattan Plot for GFM1
-log10(p-value)
Index


GFM1 p-values vs Expected p-values
ps
normalQuantiles


Bonferroni Corrected Manhattan Plot for MARCH7
-log10(p-value)
Index


MARCH7 p-values vs Expected p-values
ps
normalQuantiles


Bonferroni Corrected Manhattan Plot for PEX6
-log10(p-value)
Index


PEX6 p-values vs Expected p-values
ps
normalQuantiles

Whole bibliography

[1] *Orphanet: Peroxisome biogenesis disorder*. Dec. 2012. URL: https://www.orpha.net/consor/cgi-bin/OC_Exp.php?lng=en&Expert=79189.

[2] *Genome-Wide Association Studies Fact Sheet*. Aug. 2015. URL: https://www.genome.gov/about-genomics/fact-sheets/Genome-Wide-Association-Studies-Fact-Sheet.

[3] *ERAP2 endoplasmic reticulum aminopeptidase 2 [Homo sapiens (human)] - Gene - NCBI*. May 2020. URL: https://www.ncbi.nlm.nih.gov/gene/64167.

[4] *FAHD1 fumarylacetoacetate hydrolase domain containing 1 [Homo sapiens (human)] - Gene - NCBI*. May 2020. URL: https://www.ncbi.nlm.nih.gov/gene/81889.

[5] *PEX6 peroxisomal biogenesis factor 6 [Homo sapiens (human)] - Gene - NCBI*. Mar. 2020. URL: https://www.ncbi.nlm.nih.gov/gene/5190.