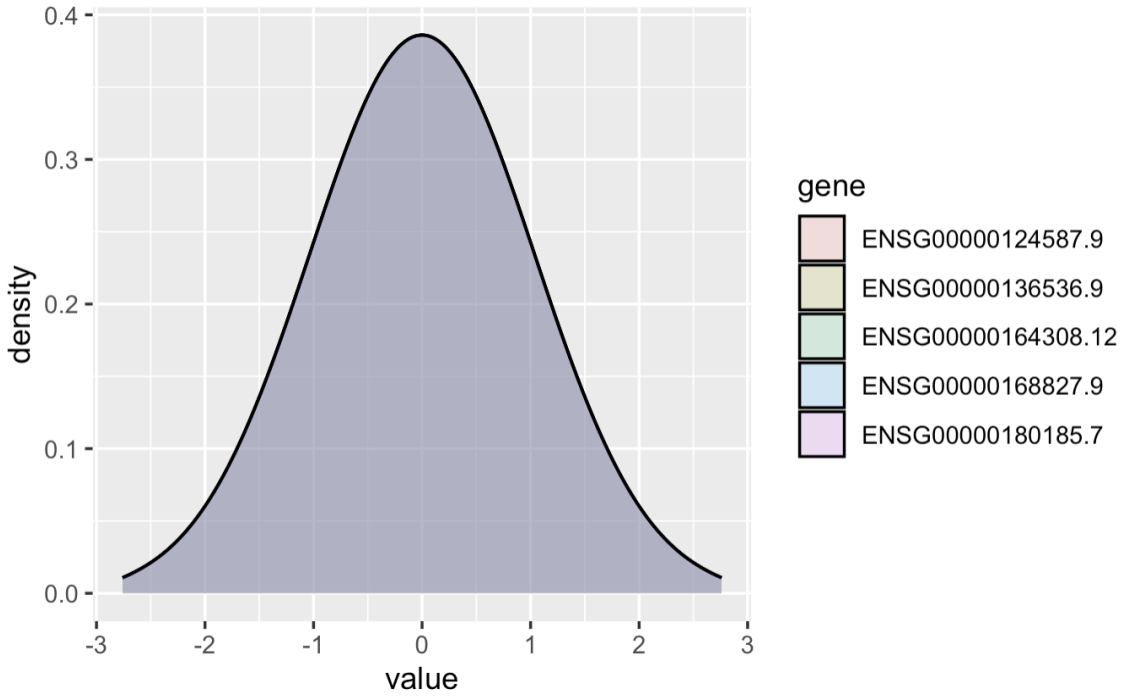
**Quantitative Genetics and Genomics**

Final Project

Noela Wheeler

In this project I looked at the mRNA expression of five genes (MARCH7, FAHD1, PEX6, ERAP2, and GFM1) across 344 different samples of European descent. I also had 50,000 SNPs for these 344 individuals. My aim was to find out if there were any significant loci associated with the expression of these 5 genes. The steps I took to answer this question are outlined below.

First I loaded the phenotype and genotype data along with the supplementary datasets - gene_info, which has the position of each gene for which we are measuring gene expression, snp_info, which has the position and id of each SNP we are measuring, and the covariate dataset, which gave the sex and country of origin for each individual. I looked at the first entries of each dataset to check that everything was loaded correctly. I also checked that the dimensions of the datasets were correct given that we have 344 samples and 50,000 SNPs. I plotted the distribution of each gene expression data to get a better understanding of the data. I was surprised to see that they all had the same, normal, distribution. We can see this in the plot below where the colored section is a combination of all the gene label colors. Although this doesn't make very much sense for biological data, I assumed that this data was simulated on the same normal distribution.



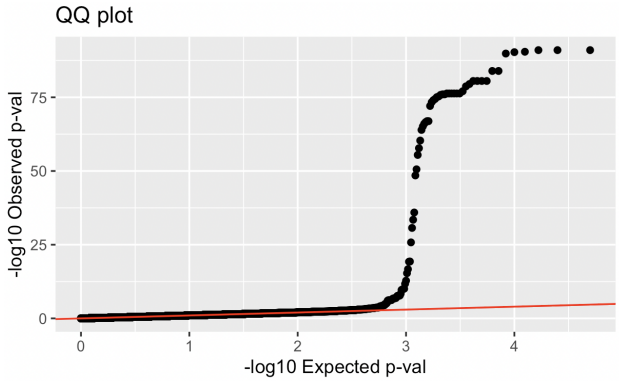
Next I converted the genotype matrix to Xa and Xd dummy variable matrices for linear regression. These variables followed the pattern shown in Professor Meazey's slides.

$$X_a(A_1A_1) = -1, X_a(A_1A_2) = 0, X_a(A_2A_2) = 1$$

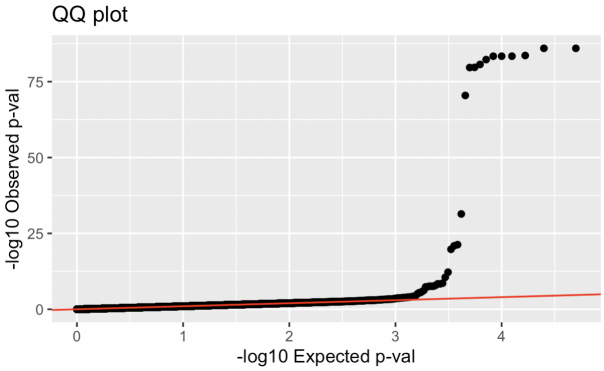
$$X_d(A_1A_1) = -1, X_d(A_1A_2) = 1, X_d(A_2A_2) = -1$$

After I did this I constructed a likelihood ratio test for each SNP. I calculated an F-statistic using Fstat = MSM/MSE. MSE was calculated using MSE = SSM/2 and MSM = SSE/n-3. I compared each Fstat to the F distribution to find p-values for each SNP. I visualized these p-values using a QQ-plot to ascertain if we are getting the expected distribution.

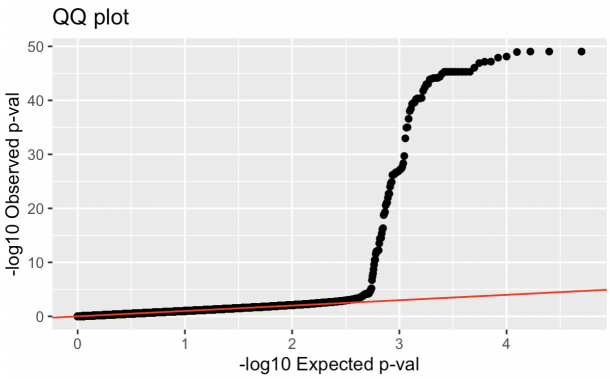
ERAP2:



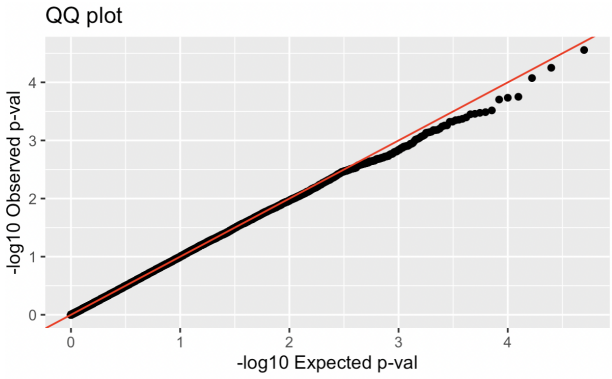

PEX6:



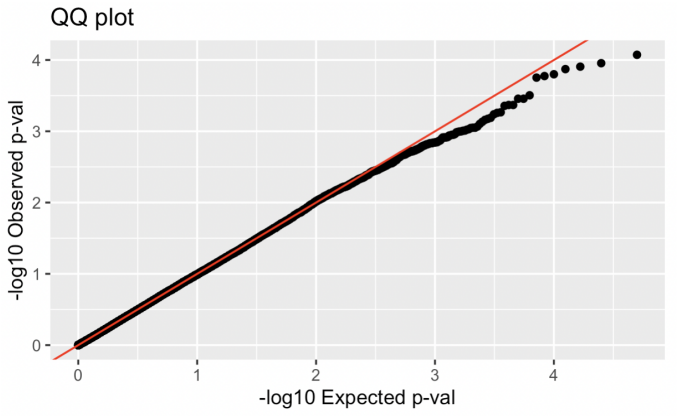

FAHD1:




GFM1:

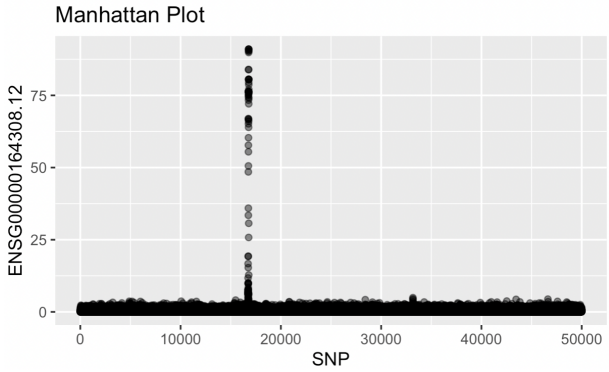


MARCH7:

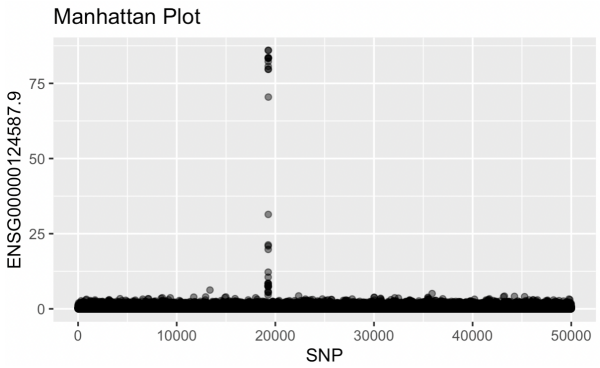



These results show that the first 3 genes show highly significant loci. These results are a bit surprising because we see more SNPs that are very significant than we would expect. This made me think that maybe these high significance locations are due to the covariates that I did not account for yet. The next two genes show almost no significant locations. This is to be expected as not all gene expression will be related to the SNPs we measured. Our Manhattan plots show results that correlate with the QQ plots where the first 3 genes have very significant regions and the last two have none.

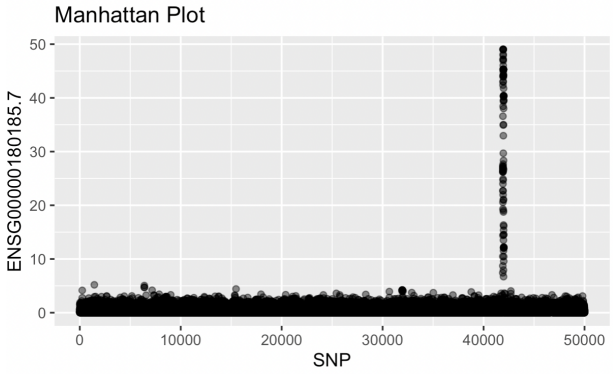
ERAP2:



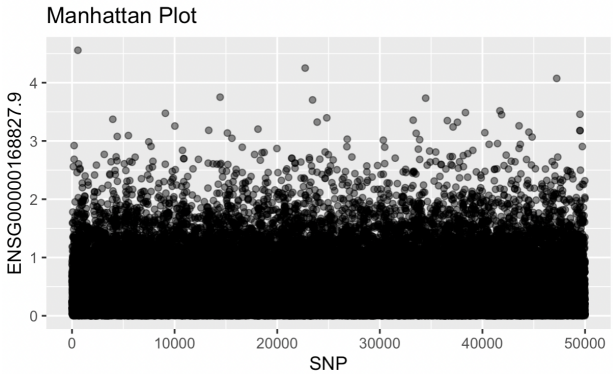

PEX6:



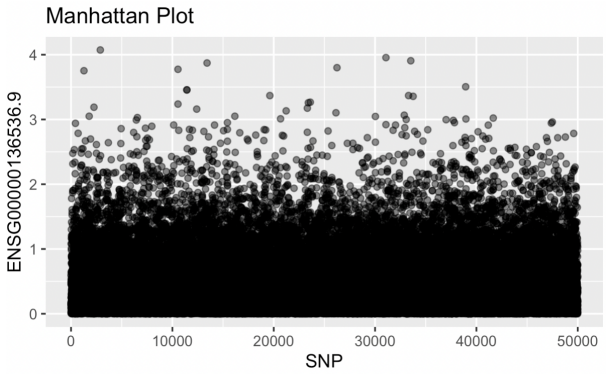
FAHD1:




GFM1:




MARCH7:




Since the QQ plots showed somewhat surprising results, I next decided to add in the covariate data that we were given. The covariate data frame has sex and origin. I encoded sex as a binary variable where female=0 and male=1. I one-hot encoded the place of origin variable. I converted each location to a separate column and gave the sample 1, if they were from there, or 0, if they were not. My covariate data frame had 5 columns to represent sex and the 4 possible countries of origin. I calculated a new F statistic for each SNP using the formula below.

$$F_{[2,n-\#(\hat{\theta}_1)]}(\mathbf{y}, \mathbf{x_a}, \mathbf{x_d}) = \frac{\frac{SSE(\hat{\theta}_0)-SSE(\hat{\theta}_1)}{2}}{\frac{SSE(\hat{\theta}_1)}{n-\#(\hat{\theta}_1)}}$$

The null hypothesis, as before, is that there is no correlation between a SNP and the phenotype. The alternative hypothesis is that there is correlation. I compared each F statistic to the chi-squared distribution to calculate p-values for each SNP. After I did this I redid my QQ plot and Manhattan plots to compare. Surprisingly, my plots were exactly the same as before. I will not include these plots because they are redundant. This result shows that the covariates were

actually not correlated with our phenotype at all. They did not cause any false positive significant SNPs.

Finally I did Bonferroni corrections to account for multiple hypothesis testing. After this, my number of significant SNPs went down. For reference, ENSG00000136536.9 = MARCH7, ENSG00000180185.7 = FAHD1, ENSG00000124587.9 = PEX6, ENSG00000164308.12 = ERAP2, ENSG00000168827.9 = GFM1.

```
"There are 2482 significant SNPs for gene ENSG00000164308.12"
"After bonferroni correction there are 73 significant SNPs for gene ENSG00000164308.12"
"There are 2530 significant SNPs for gene ENSG00000124587.9"
"After bonferroni correction there are 27 significant SNPs for gene ENSG00000124587.9"
"There are 2540 significant SNPs for gene ENSG00000180185.7"
"After bonferroni correction there are 90 significant SNPs for gene ENSG00000180185.7"
"There are 2491 significant SNPs for gene ENSG00000168827.9"
"After bonferroni correction there are 0 significant SNPs for gene ENSG00000168827.9"
"There are 2409 significant SNPs for gene ENSG00000136536.9"
"After bonferroni correction there are 0 significant SNPs for gene ENSG00000136536.9"
```

I wanted to see where these SNPs were in the genome so I checked the reference file. I found that the position of the significant locus for ERAP2 was on chromosome 5 around position 96774230. The significant locus for PEX6 was on chromosome 6 around position 42889467. The significant locus for FAHD1 is on chromosome 16 around position 1604317 and position 1806559-1929366. I did not include the last 2 genes since after corrections they have no significant SNPs.

I did a little research to try to understand why gene expression for only certain genes is highly correlated with these SNPs. I found that MARCH7 is a ubiquitin ligase and GFM1 is a mitochondrial translation elongation factor. Since these genes have to do with the transient state of the cell, it makes sense that these genes would be more determined by cell activity and not genotype. ERAP2 is an aminopeptidase involved in antigen presentation. PEX6 is a predominantly cytoplasmic protein, which plays a direct role in peroxisomal protein import. FAHD1 is in cytosol, mitochondrion, and nucleoplasm. Some of its related pathways are Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins. and TCA cycle. I'm not sure why these proteins would be more associated with the SNPs we have chosen than the others. It is interesting that ERAP2 is associated with the immune system as this could be more genetically linked than some of the other proteins.

In conclusion, I found that the gene expression of ERAP2, PEX6, and FAHD1 were each significantly correlated with one specific loci in the SNPs measured. MARCH7 and GFM1 seemed to have no significant associations with the SNPs. In addition, accounting covariates had no impact on the significant associations. This shows that sex and place of origin were not linked to gene expression of these particular genes and so did not produce any false positives. More research could be done by measuring more SNPs or looking deeper into the biology of why some

loci had such strong associations with certain gene’s expression.