GWAS Analysis Project– BTRY 4830

Emily Hur

**Introduction**

New advancements in genomic sequencing technology have transformed the world of biological research. Subsequently, many organizations and laboratories have spearheaded projects employing RNA sequencing on a large scale, given the newfound efficiency and throughput in which genetic information can be collected. Some of this data has been made publicly available, including findings from the Genetic European Variation in Health and Disease Consortium. This collection references sources from the 1000 Genomes Project, an effort to document human genetic variation by sampling hundreds of individuals. Specifically, researchers extracted lymphoblastoid cell lines from individuals from five different populations (Italy, British, Toscani, Utah, Finland) for RNA sequencing.

A subset of the aforementioned data is the focus of this study. Of the 462 samples, 344 were selected from 4 populations (Finland, Toscani, British, Utah). For each row/sample, 50,000 SNP (single nucleotide polymorphism) genotypes were selected (AjAk), encoded using 0, 1, or 2 based on the number of minor alleles. A SNP is some variant at a single base position of DNA, usually categorized as a mutation. Beyond this, there are 4 additional files used for analysis, the first being phenotype data providing expression levels for five genes, commonly known as MARCH7, FAHD1, PEX6, ERAP2, and GFM1. The gene info csv file contains additional information on these five genes, including start and end positions and the chromosome number. The covariates file has information on the 344 individuals themselves, specifically which of the four populations they belong to and their gender. Lastly, the SNP dataset identifies the chromosome number of the SNP, the physical location on that chromosome, and the rsID.

This project strives to use statistical techniques and quantitative methods to pinpoint locations of causal polymorphisms for the 5 genes. In other words, the ultimate goal is to identify regions where manipulating the DNA will lead to measurable effects. This will be done by conducting multiple Genome-Wide Association Studies, the steps of which will be outlined below, on each individual gene expression measurement. In summary, this analysis is repeated multiple times using different covariates, employing hypothesis testing under each model to assess whether the markers have causal effects on phenotypes. These results will provide insight as to which markers are statistically significant and how to proceed with future experiments that build off of these findings.

**Methods**

The first step involved loading the 5 csv files described above into dataframes and all required packages into R (ggplot2, factoextra, ggfortify, MASS). From there, histograms were generated for each phenotype (figure 1) to assess the distribution. It’s clear that the data is continuous, and all five phenotypes appear to be approximately normally distributed, centered at around 0, with no outliers. Since the data conforms to an expected distribution and there appear to be multiple states, linear regression analysis can be applied.

Next, the quality of the genotype and phenotype data was evaluated. Any individuals with >10% of their data missing across the genotypes as well as SNPs with >5% missing data across samples were to be excluded from the analysis. Each individual was also evaluated to detect missing expression levels for each of the five genes. Fortunately, none of the individuals, SNPs, or genes needed to be removed. From there, PCA was utilized to determine the presence of population structures. Although the plotting of dimensions revealed each component wasn’t responsible for a very large percentage of variance, it seemed the samples can be grouped into roughly three clusters (figure 2). This means there’s likely some sort of population structure, and that GWAS should be run with covariates.

The genotype data was converted into two matrices for additive and dominant codes. Xa encodings were determined using Xa(A1A1)=-1, Xa(A1A2)=0 and Xa(A2A2)=1, subtracting 1 from the original genotype data organized into 0s, 1s, and 2s. Meanwhile, for Xd encodings, each pair of alleles was mapped using Xd(A1A1)=-1, Xd(A1A2)=0 and Xd(A2A2)=-1, which can be calculated with the equation 1-2\*|Xa| for each of the Xa encodings.

GWAS was first performed using no covariates with the multiple regression equation Y=𝛽𝜇+Xa𝛽a+Xd𝛽d+𝜖, with each marker being considered one at a time. Subsequently, 50,000 hypothesis tests were performed and their corresponding p-values were recorded. This was repeated for each of the 5 phenotypes.

A 3-column matrix was generated for each SNP, including the Xa encodings, Xd encodings, and phenotype expression levels for all 344 samples. First, the maximum likelihood estimator identified the parameter values that maximize the likelihood of the input and returned a vector of 3 values for 𝛽𝜇, 𝛽a, and 𝛽d. The null hypothesis is that 𝛽a=0 and 𝛽d=0 (the marker is not associated with the outcome, Y), while the alternative hypothesis is that 𝛽a≠0 or 𝛽d≠0. In order to test these hypotheses, likelihood ratios tests were conducted for the 50,000 SNPs. First, the estimated phenotype for each sample was calculated using the output of the MLE. This was eventually used to calculate the sum of squares of the model, the sum of squares of the error, the mean squared model, and the mean square error. 3 degrees of freedom were used for the mean squared model, or the number of betas minus 1. 341 degrees of freedom wee used for the mean square error, or the number of individuals minus 3. Finally, the LRT was calculated to create an output that can be used to get a p-value: mean squared model/mean square error. The pf built-in function in R ultimately took in the three inputs, the two values for degrees of freedom and the LRT, to produce a p-value using the cumulative distribution.

Given the list of p-values, five QQ plots (figure 3) were made to represent each phenotype. Expected p-values were calculated using the qunif function and plotted on the x=axis against the observed p-values, on a -log10 scale. The shape of the points was compared against the line y=x. Five Manhattan Plots (figure 4) were also used to visualize the p-values, this time plotting the observed p-values against their corresponding indices (from 1 to 50,000).

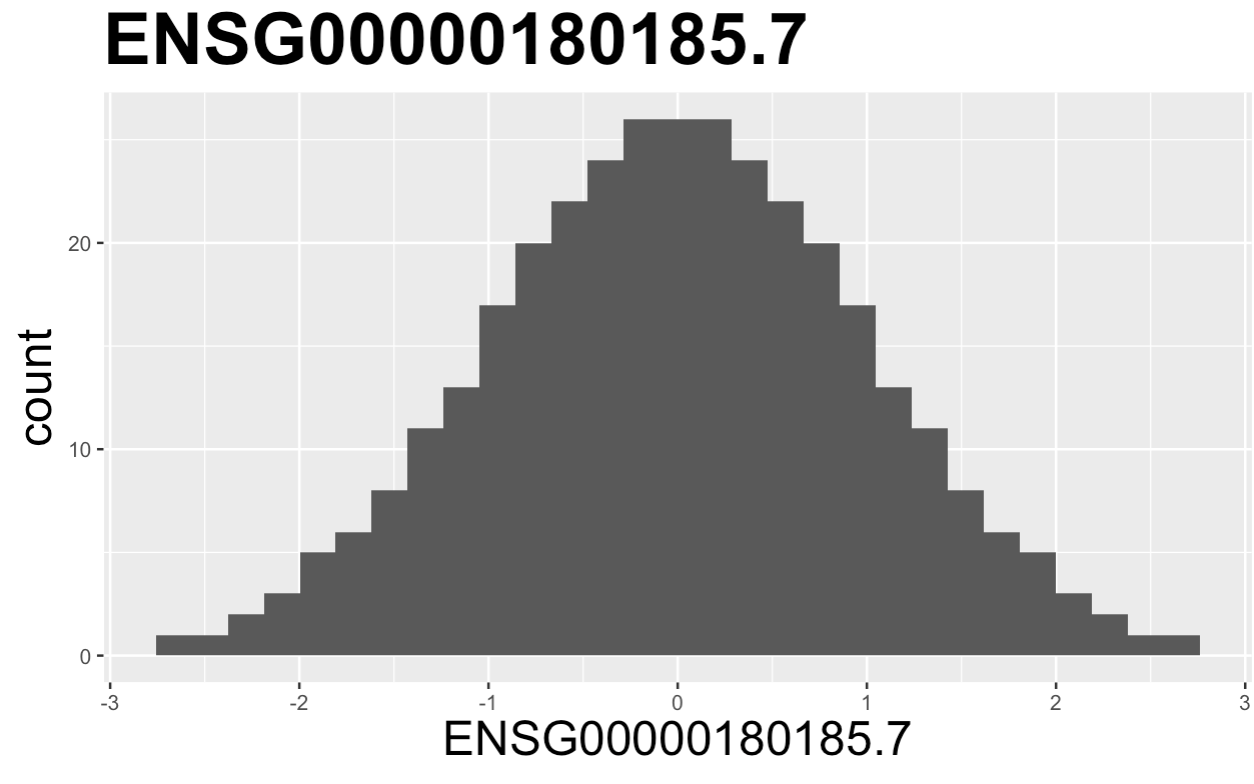
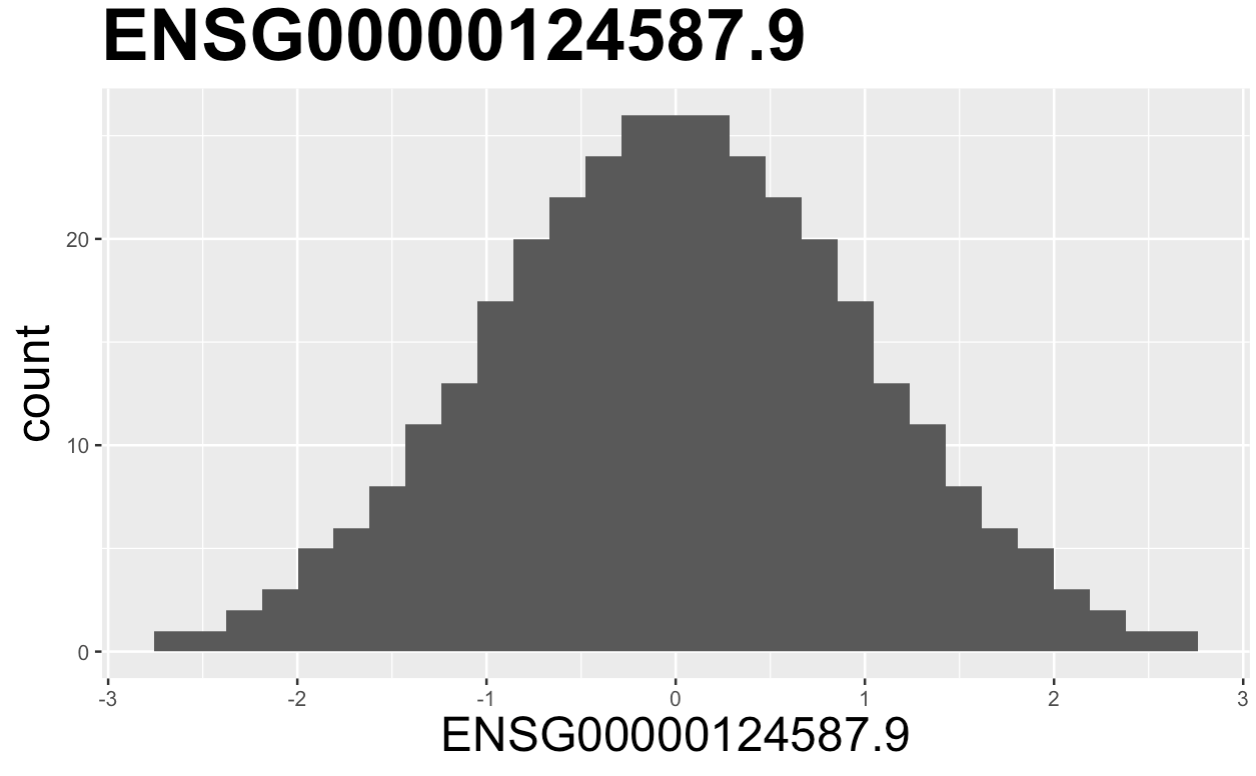
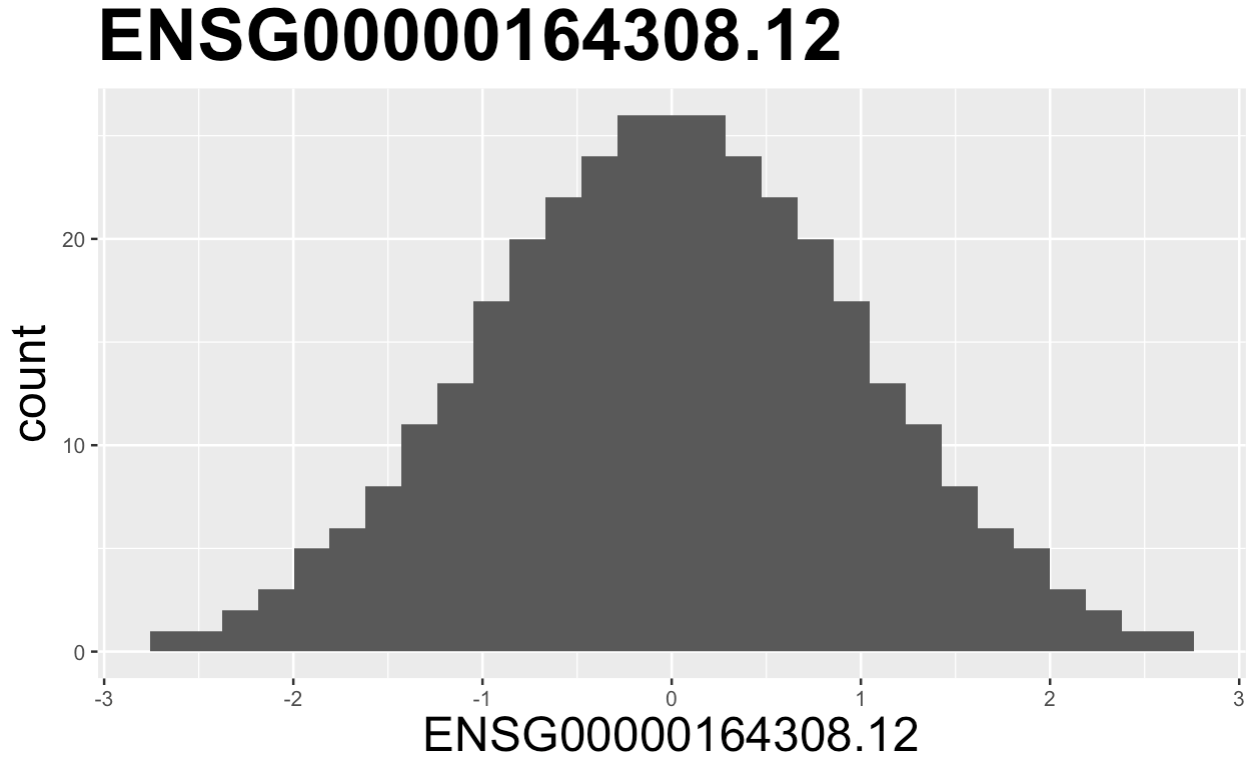
The significance of the p-values were evaluated using a threshold of .05, which yielded over 2,000 “hits” for each phenotype. After using a Bonferroni correction, this number was significantly reduced to 73 for ENSG00000164308.12, 29 for ENSG00000124587.9, 90 for ENSG00000180185.7, and 0 for the others. Local Manhattan plots (figure 5) just for the significant genotypes revealed that there was one notable outlier, but the markers were relatively contiguous otherwise.

The process outlined for GWAS above, with a few slight modifications, was repeated three more times: once using sex as a covariate, once using population as a covariate, and once with both covariates included. First, the covariate data was transformed into binary numbers. For gender, females were represented by 1 and males were represented by 0. For population, one-hot encoding was utilized to transform the one variable into four dummy variables, one for each of the four populations.

The regression equation was modified with the addition of covariates, incorporating each one as a factor: Y=𝛽𝜇+Xa𝛽a+Xd𝛽d+Xz,1𝛽z,1+Xz,2𝛽z,2+𝜖. Subsequently, the covariate model parameter is included in the LRT, meaning the MLE for the null hypothesis must now be calculated. Predictions are made under both the null and alternative hypotheses to be used in the sum of square errors equation for both. Finally, the F-statistic is calculated using the formula ((SSE(θ0)-SSE(θ1))/df(M))/(SSE(θ1)/df(E)), with the degrees of freedom varying depending on the number of betas. Again, along with the degrees of freedom, the F-statistic is used as an input into the pf function, generating corresponding p-values.

In summary, GWAS was conducted four separate times. Each individual analysis returned approximately the same amount of “hits.” For a sanity check, the rsIDs of the selected genotypes for ENSG00000124587.9 were compared across the four models and it was found that they only differed by one or two SNPs. Thus, the model with no covariates was used for the final interpretation of results as adding covariates didn’t improve the clarity or quality of work. Given the index of the significant markers, information about the chromosome number, id, and position on the chromosome was selected from the SNP file. From here, a data frame was created (table 1) using an inner join on the chromosome number with the gene info file. For each significant marker, this provided additional information on the transcript ID, gene of interest, and start and end positions of the gene of interest. Lastly, a column of boolean values appended to the dataframe, stating “TRUE” if the SNP is located in between the start and end positions of the gene of interest and FALSE otherwise. This procedure was repeated for the three genes that had significant markers remaining after the Bonferroni correction.

**Results**

****

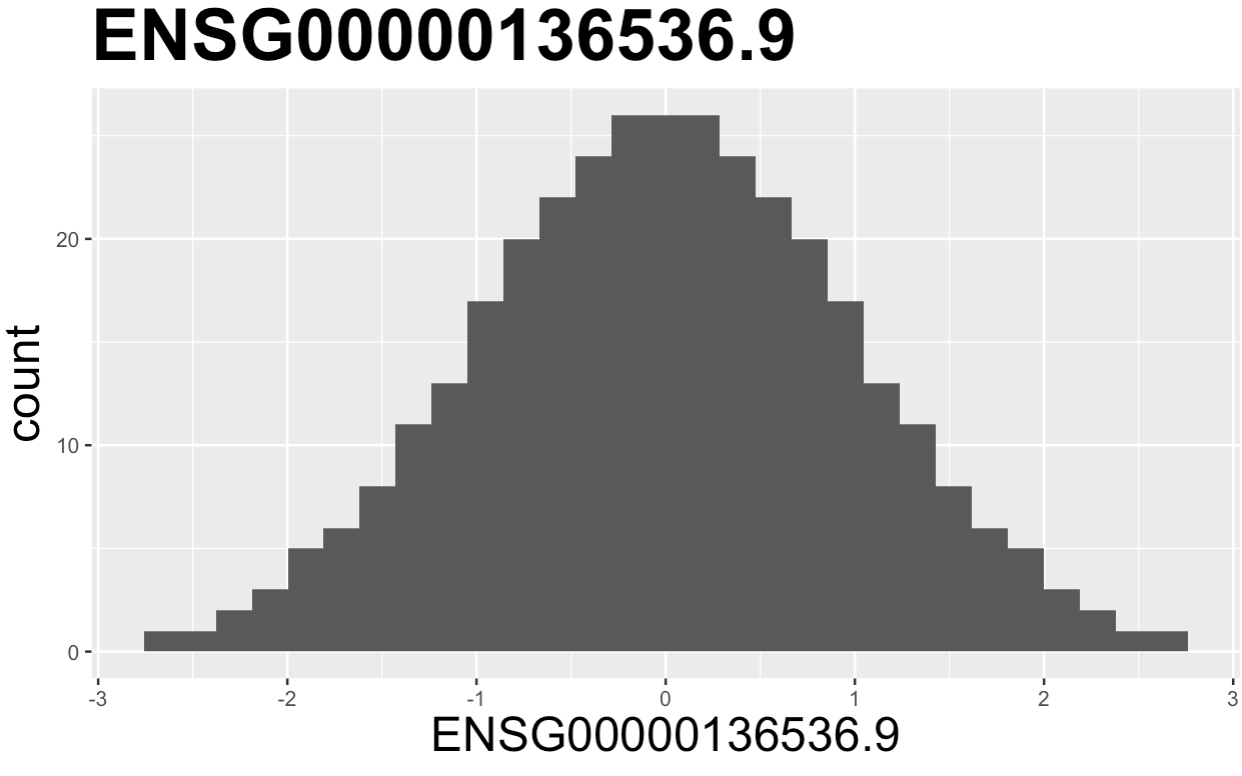
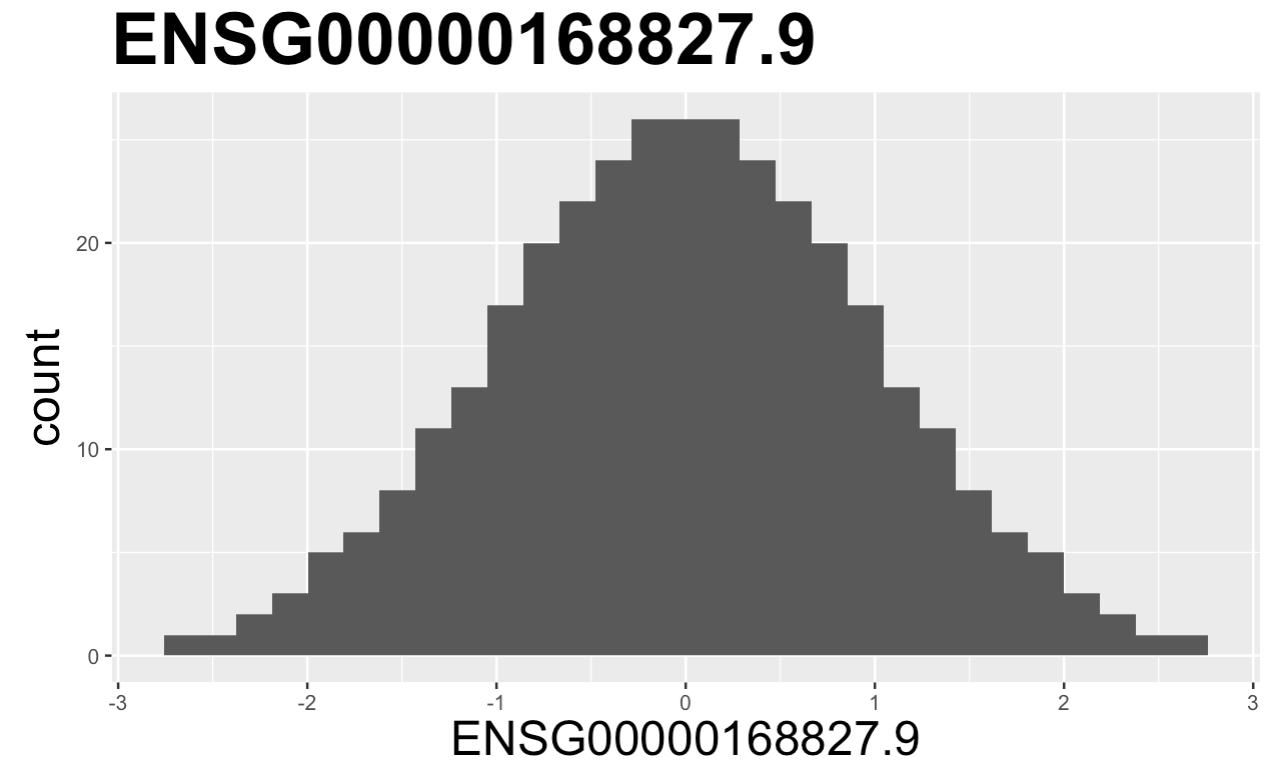
****

Figure 1: histogram of phenotype data for the five genes

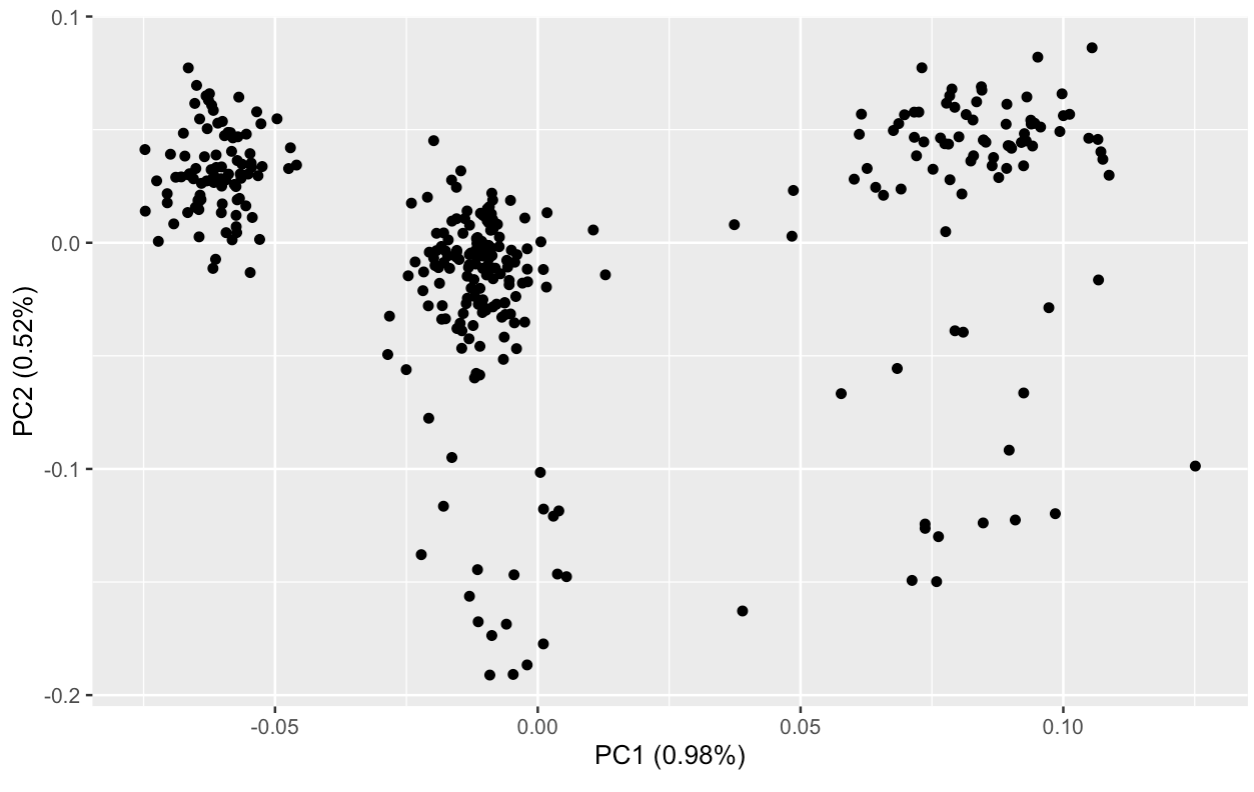
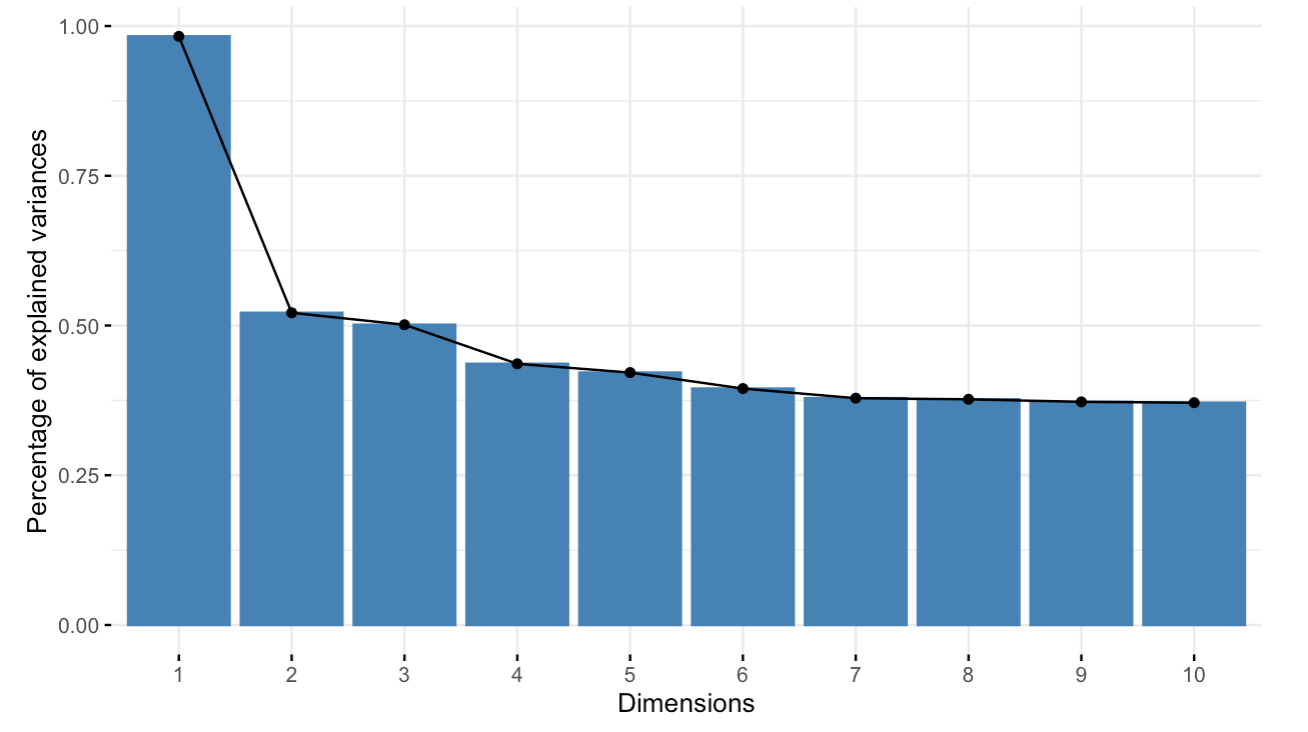
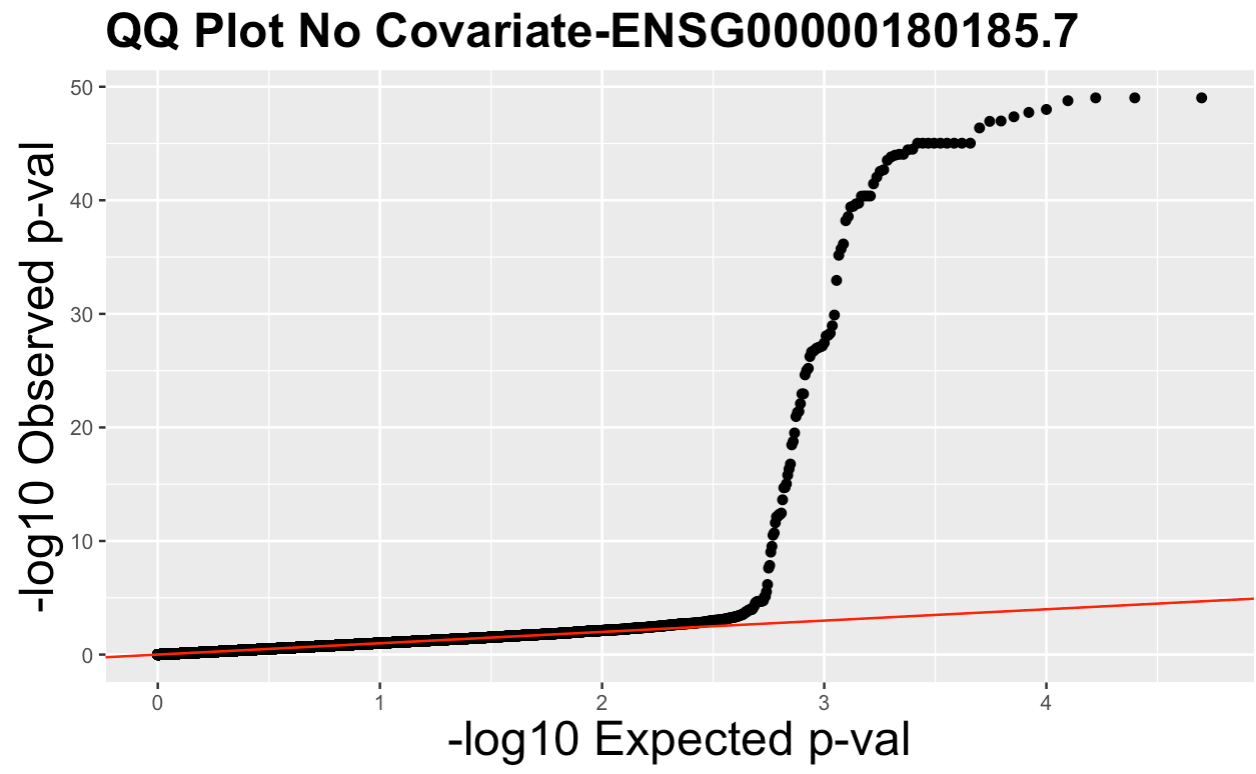
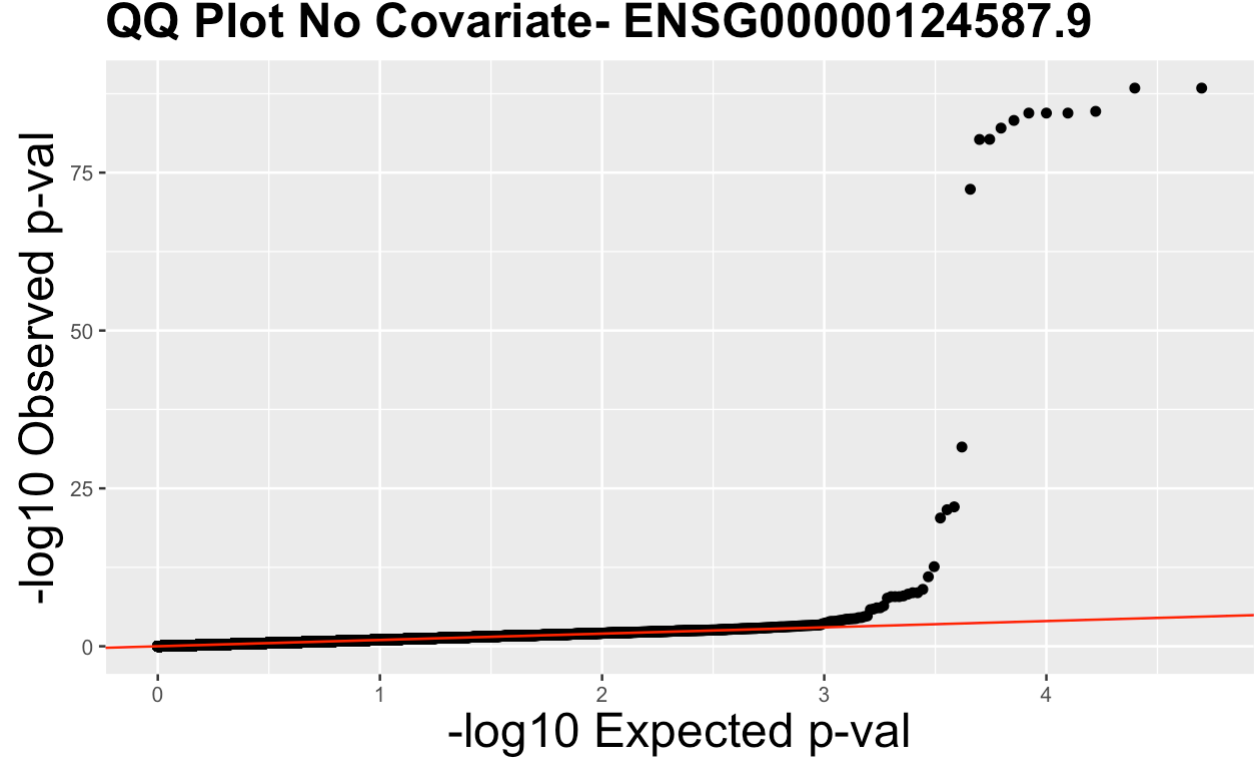
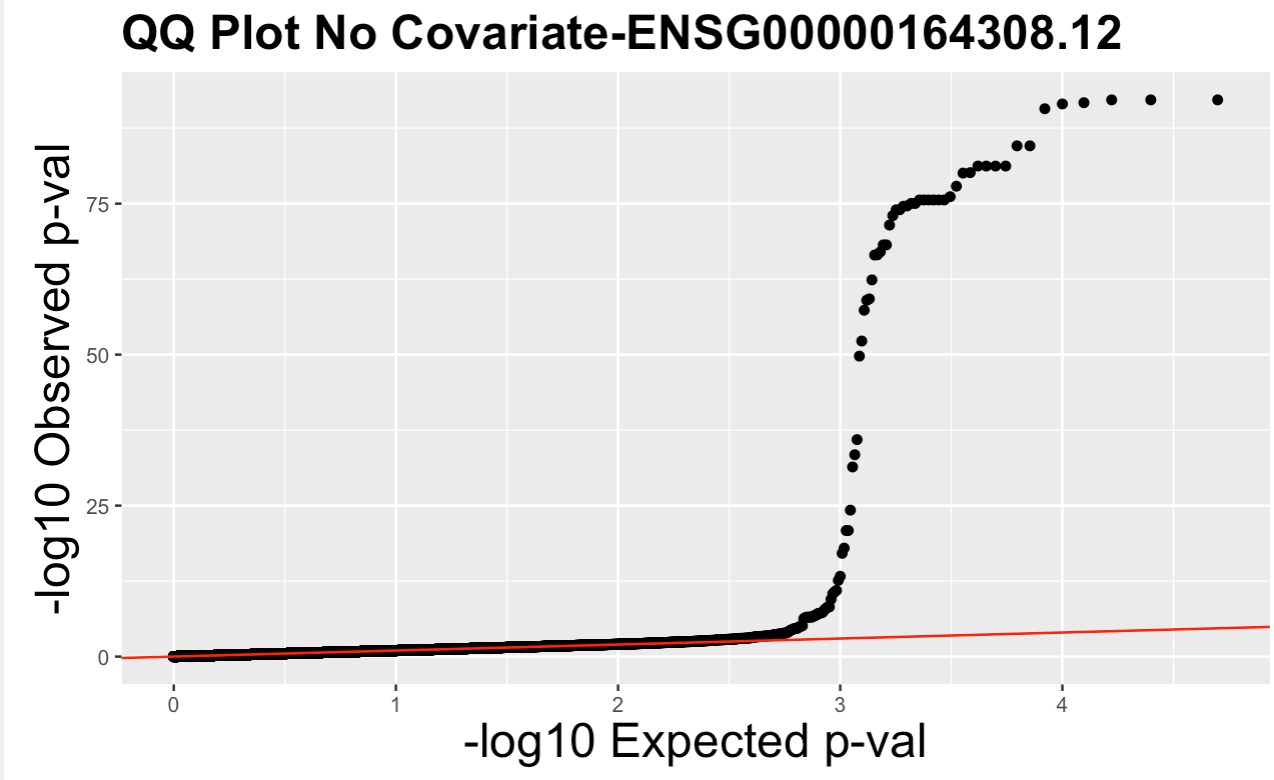
****

Figure 2: left) plot of percentage of explained variances vs. number of components right) plot of principal components 1 and 2



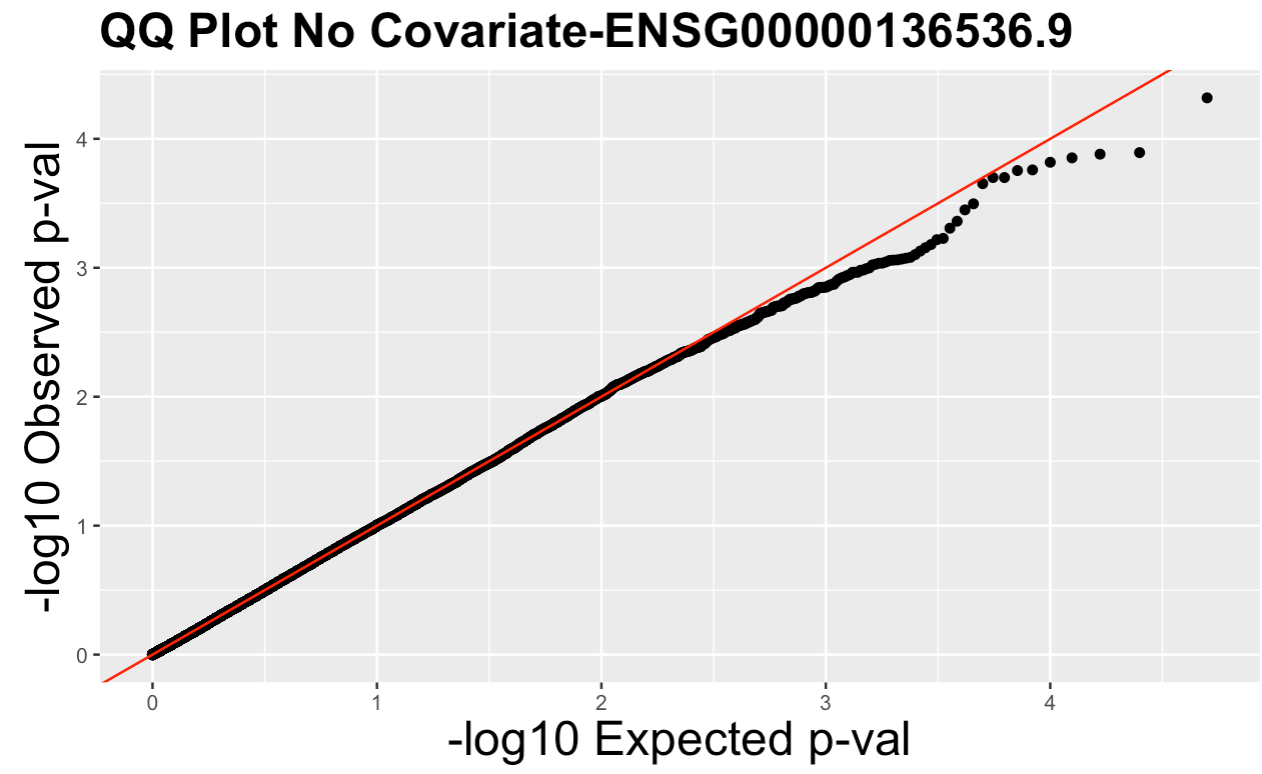
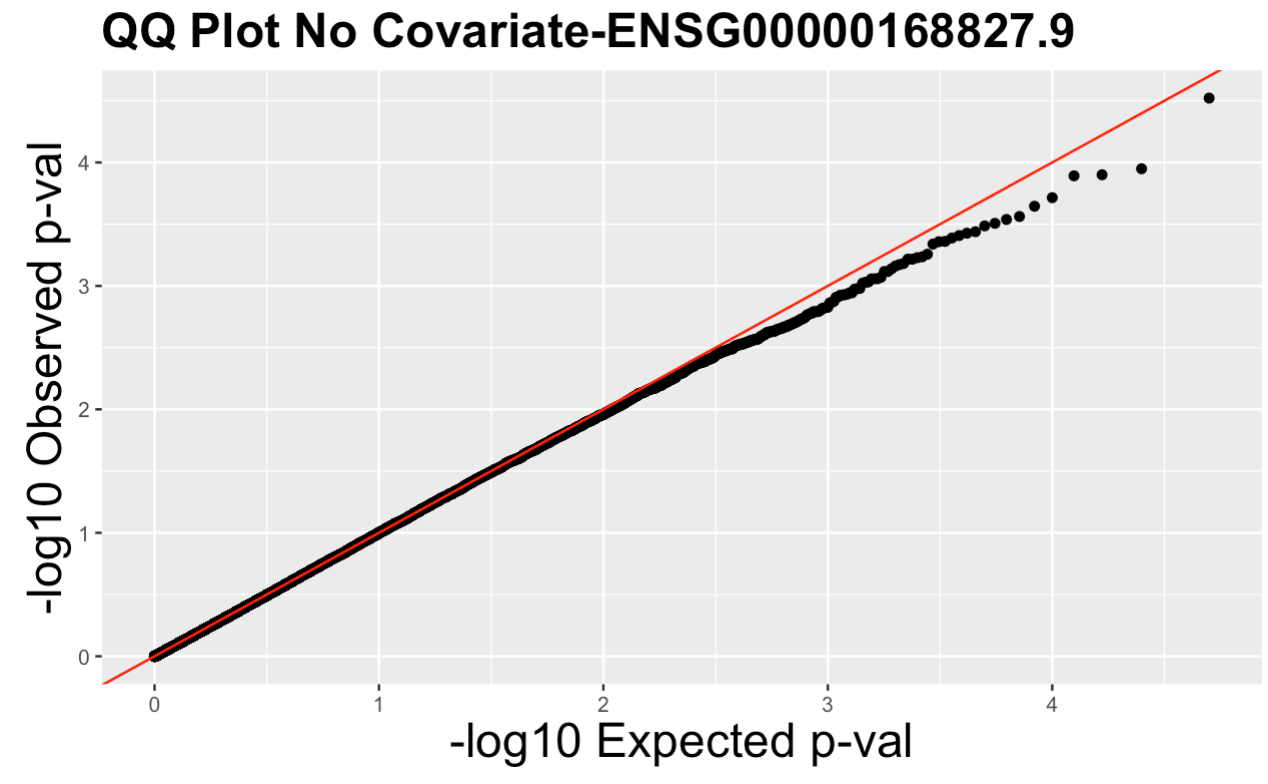
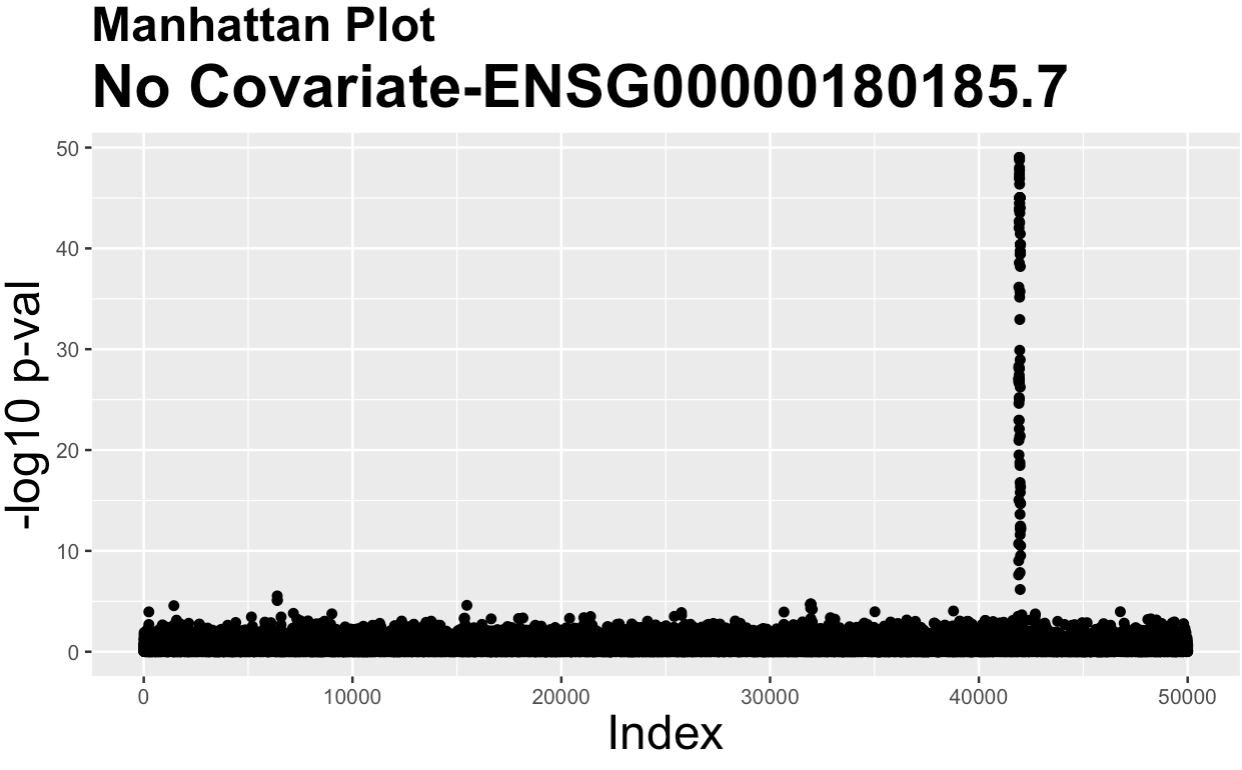
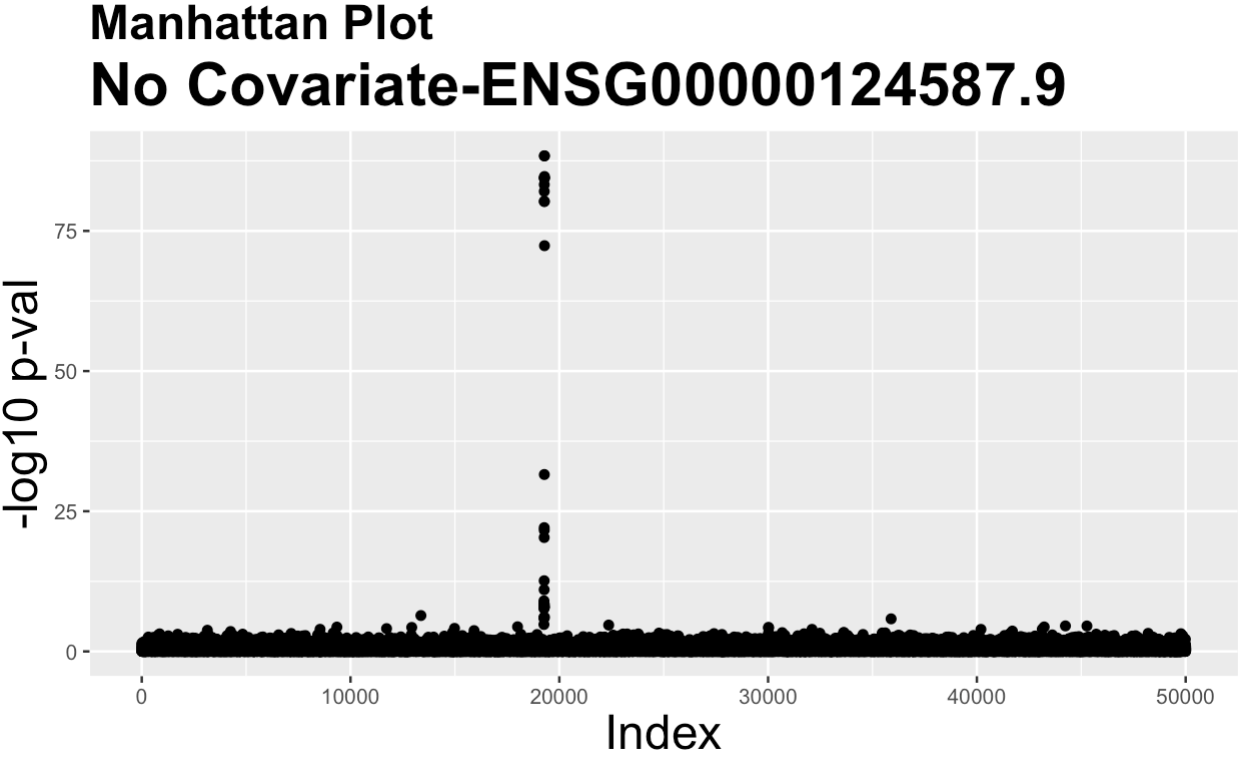
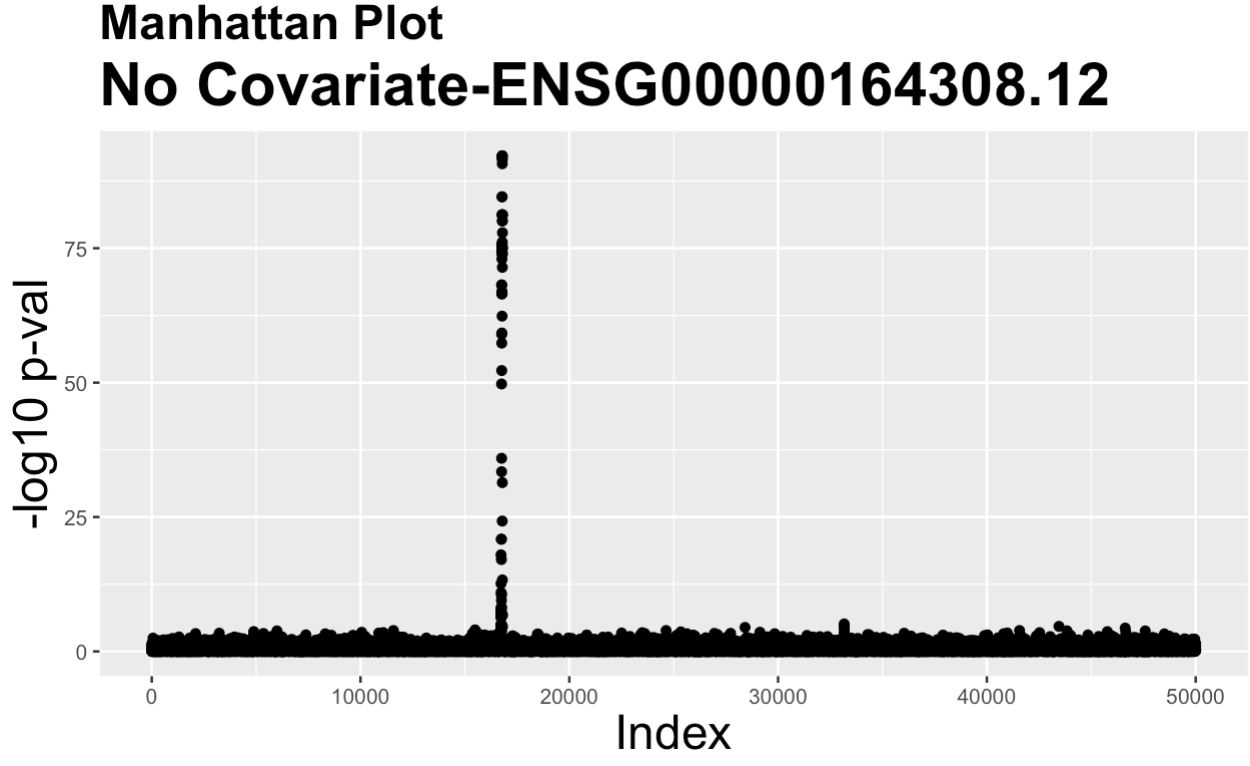


Figure 3: QQ plot of p-values for the five genes, no covariate. The x-axis is -log10(expected p-value) and the y-axis represents -log10(observed p-value).



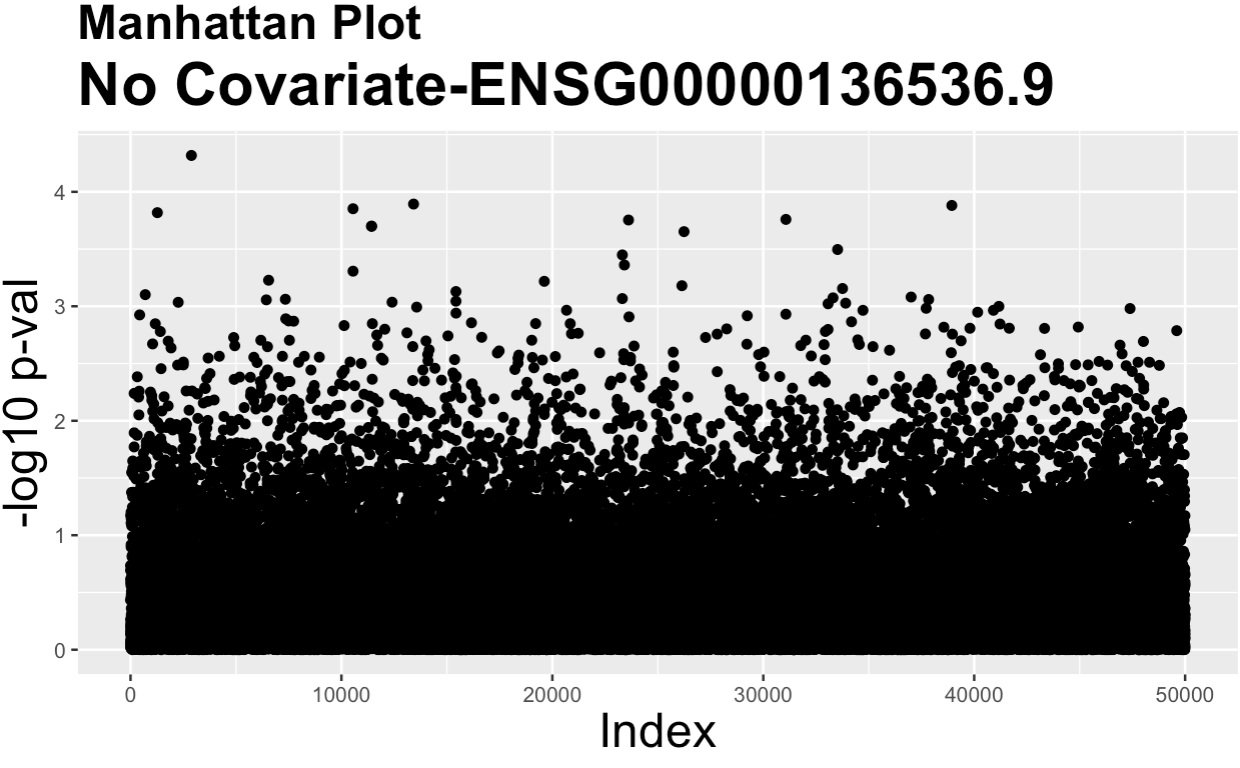
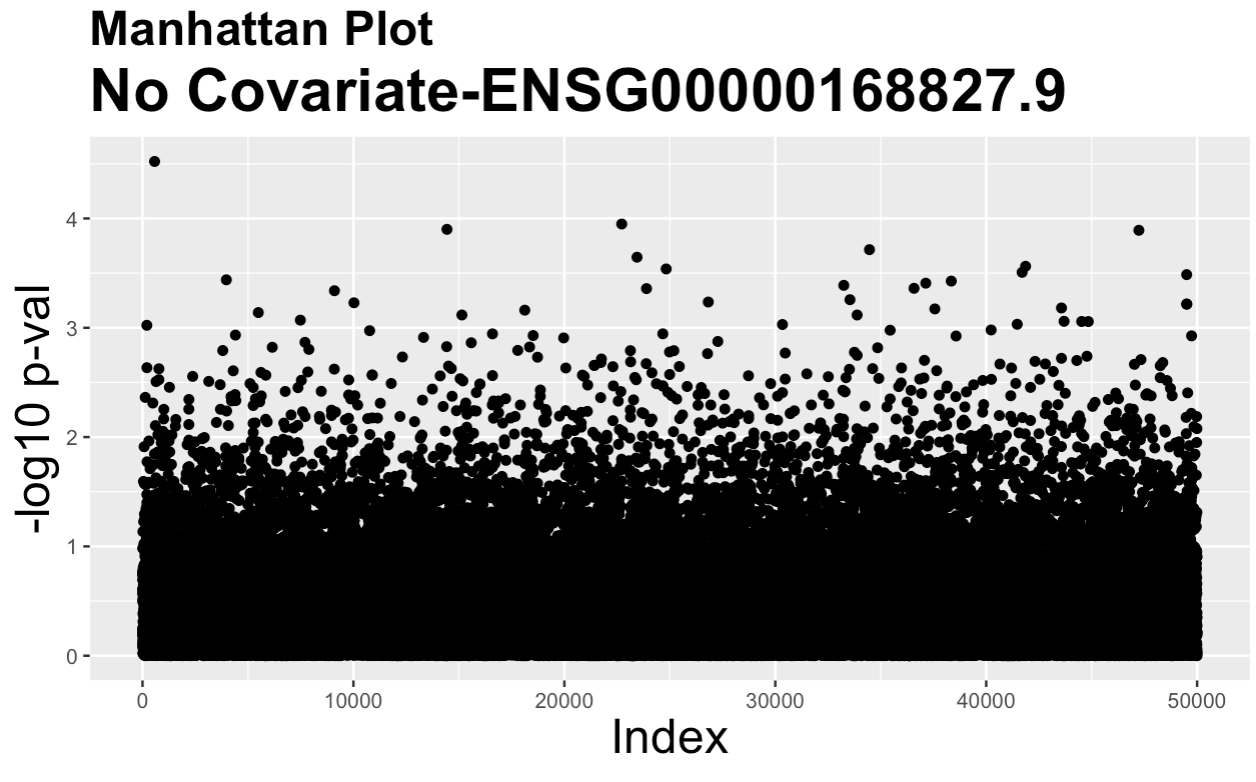


Figure 4: Manhattan plots for the five genes, no covariate. The x-axis contains the index and the y-axis is -log10(p-value).

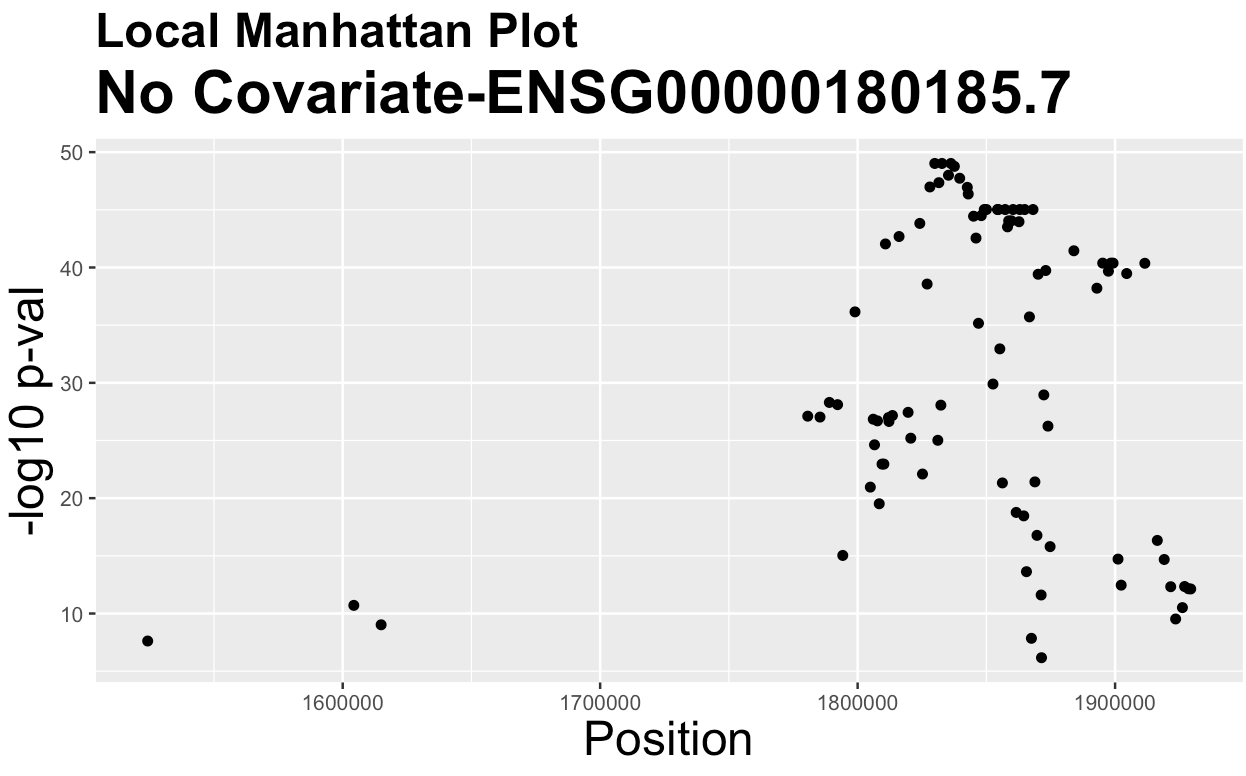
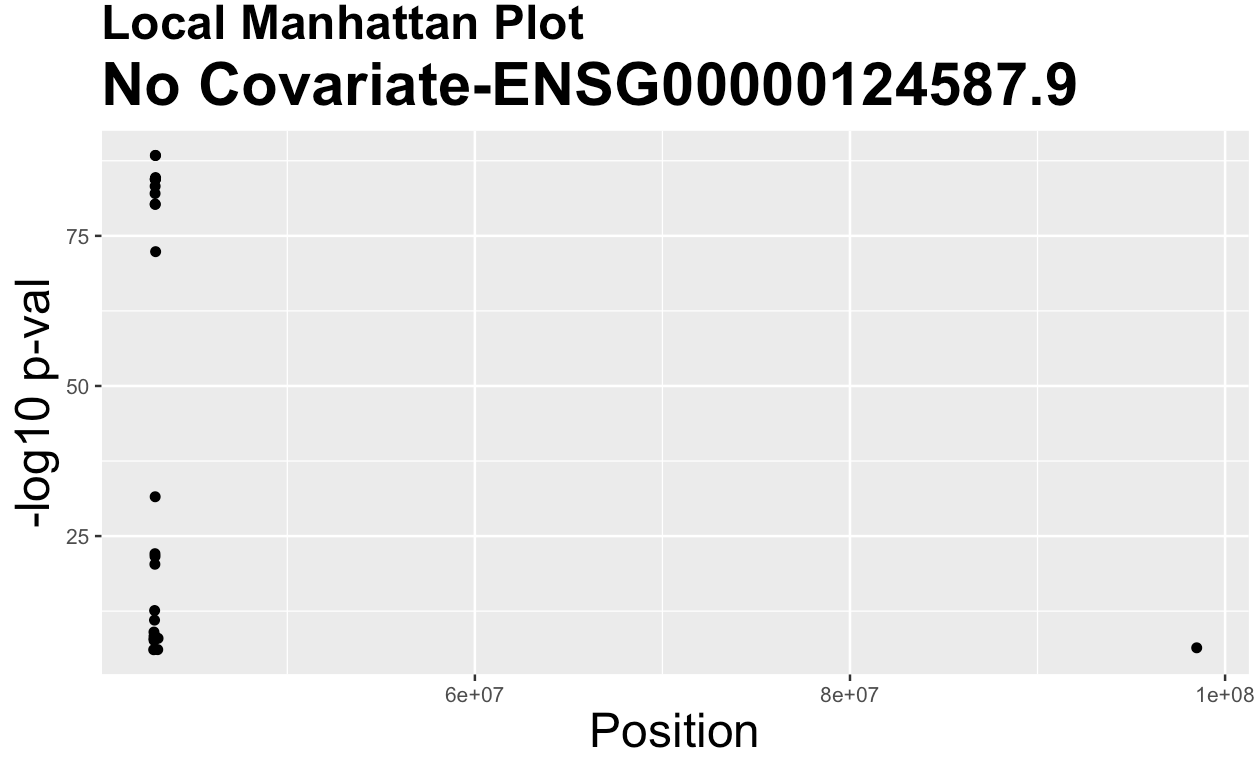
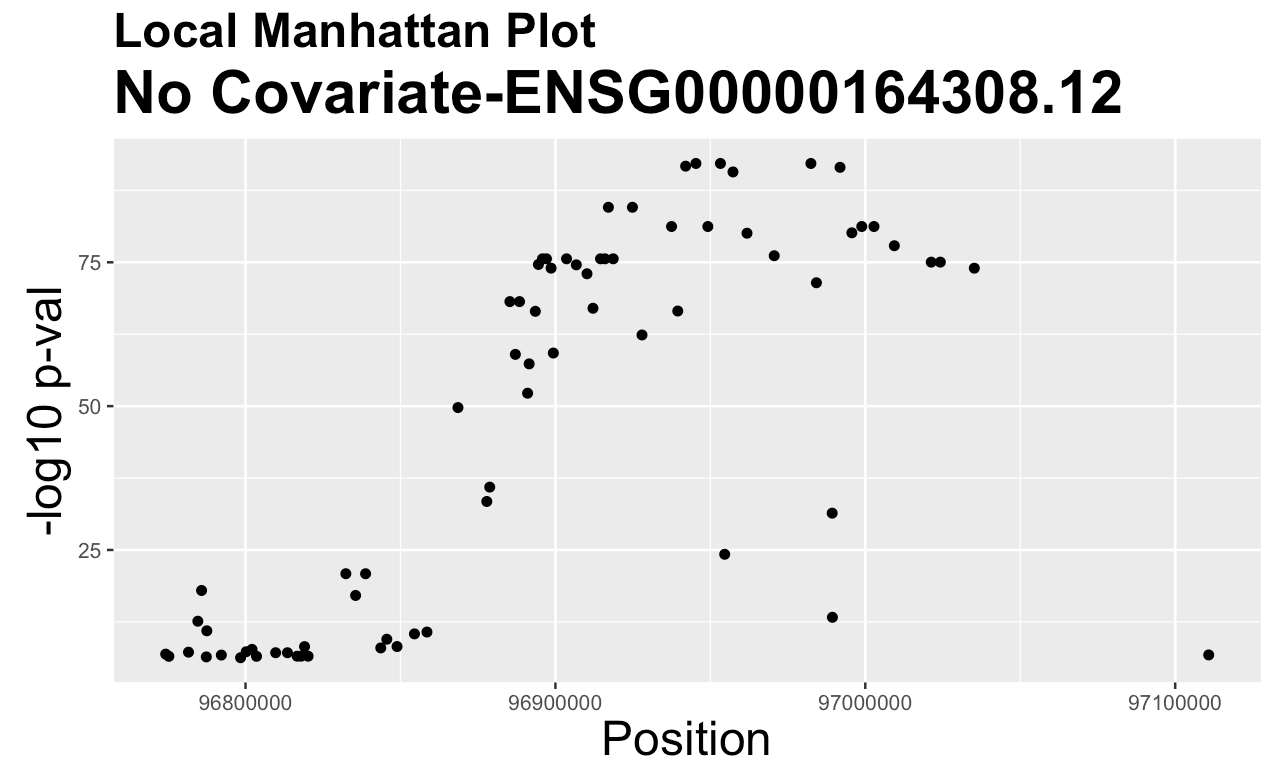
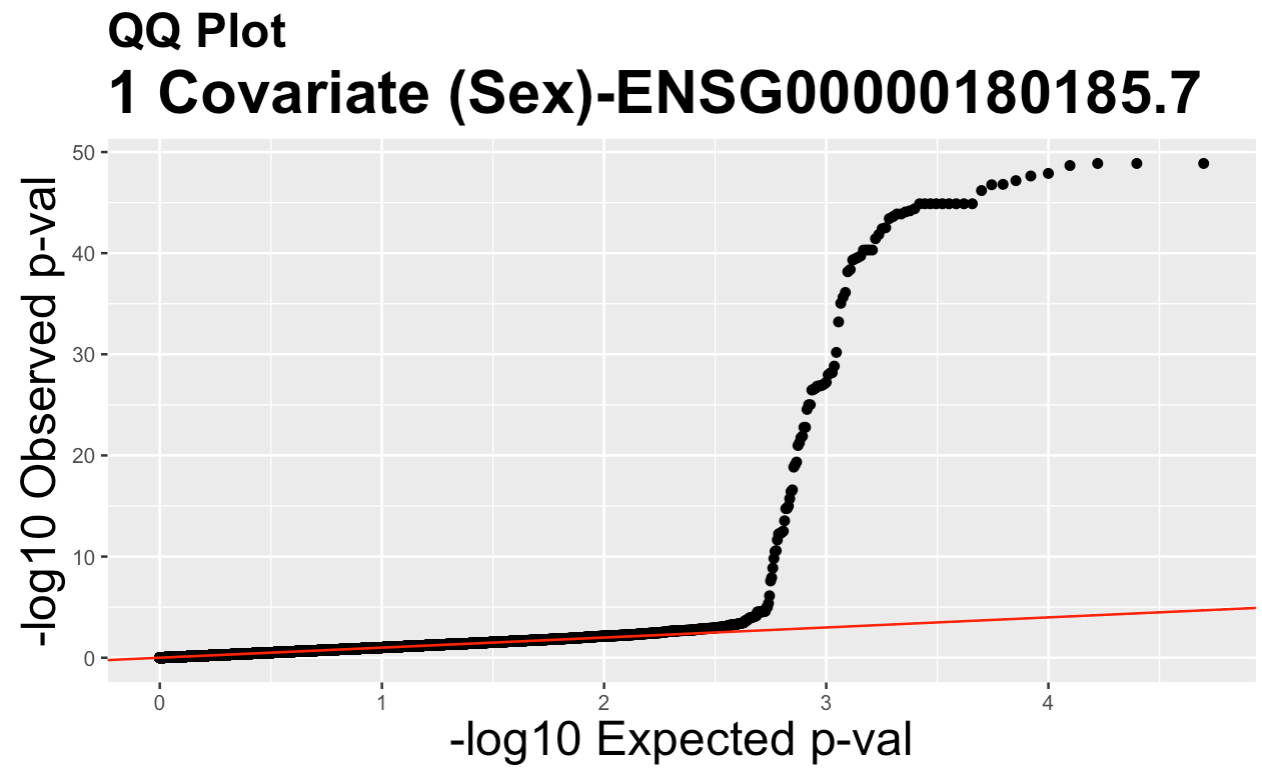
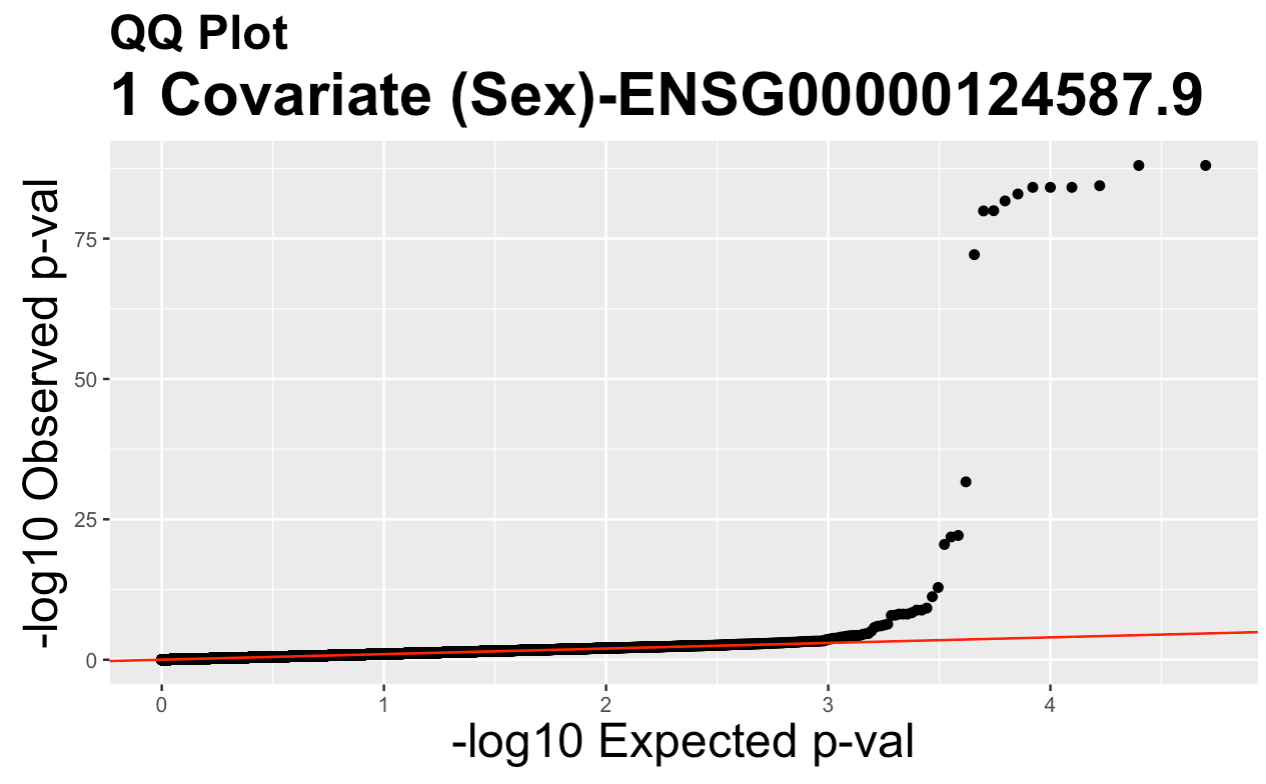
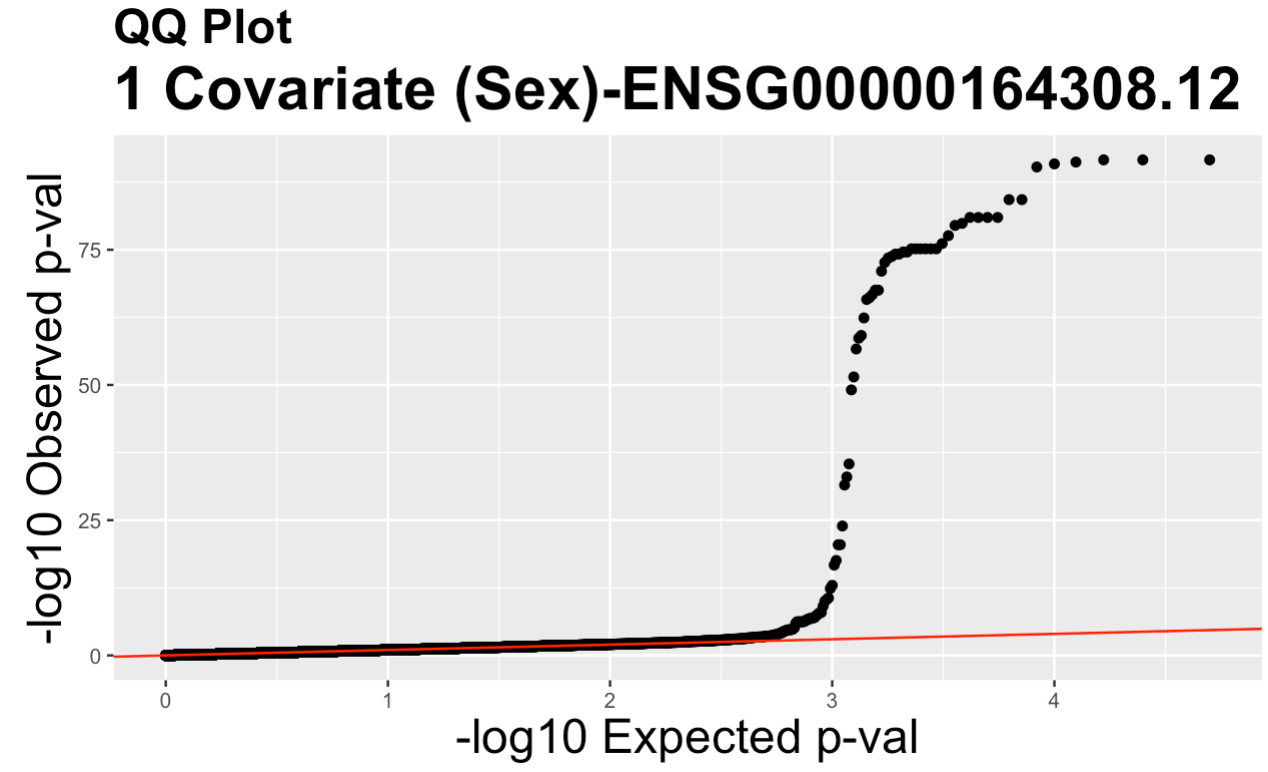


Figure 5: Local Manhattan plots for phenotypes with statistically significant markers, only displaying p-values corresponding to significant markers after Bonferroni correction. The y-axis is -log10(p-value) plotted against the position on the chromosome.



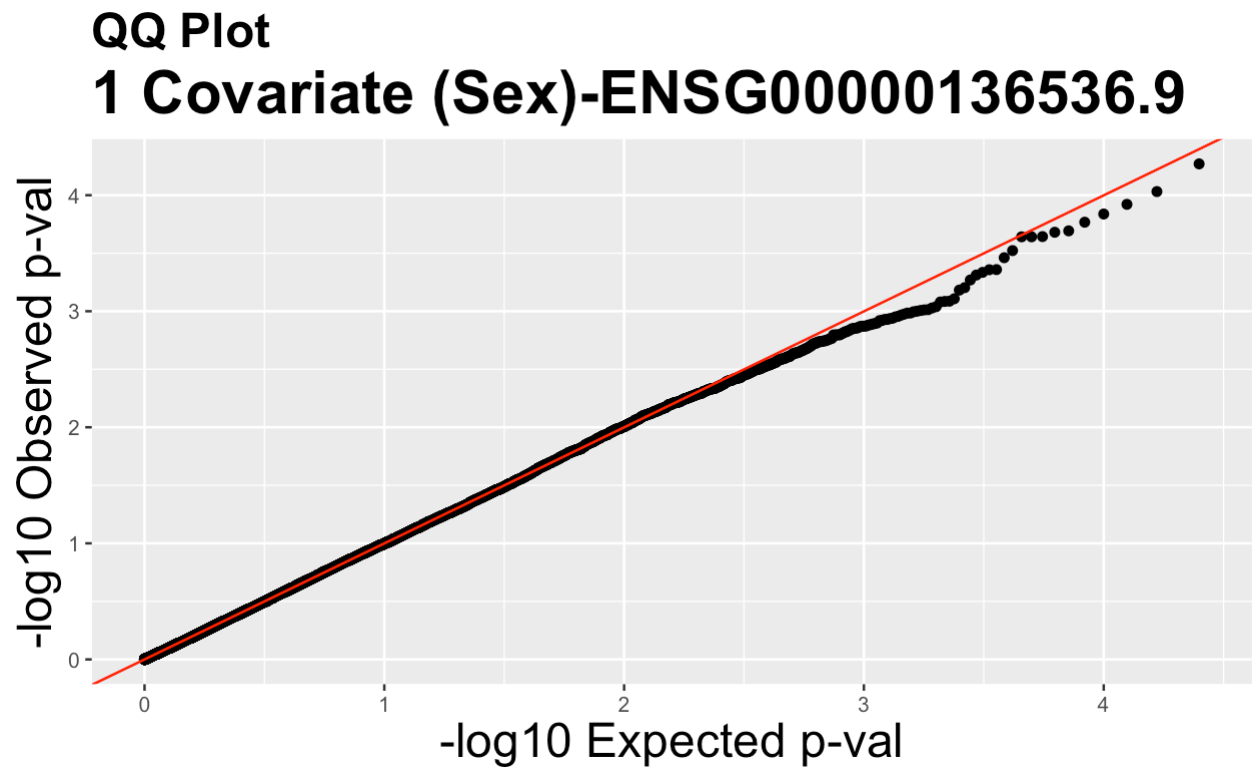
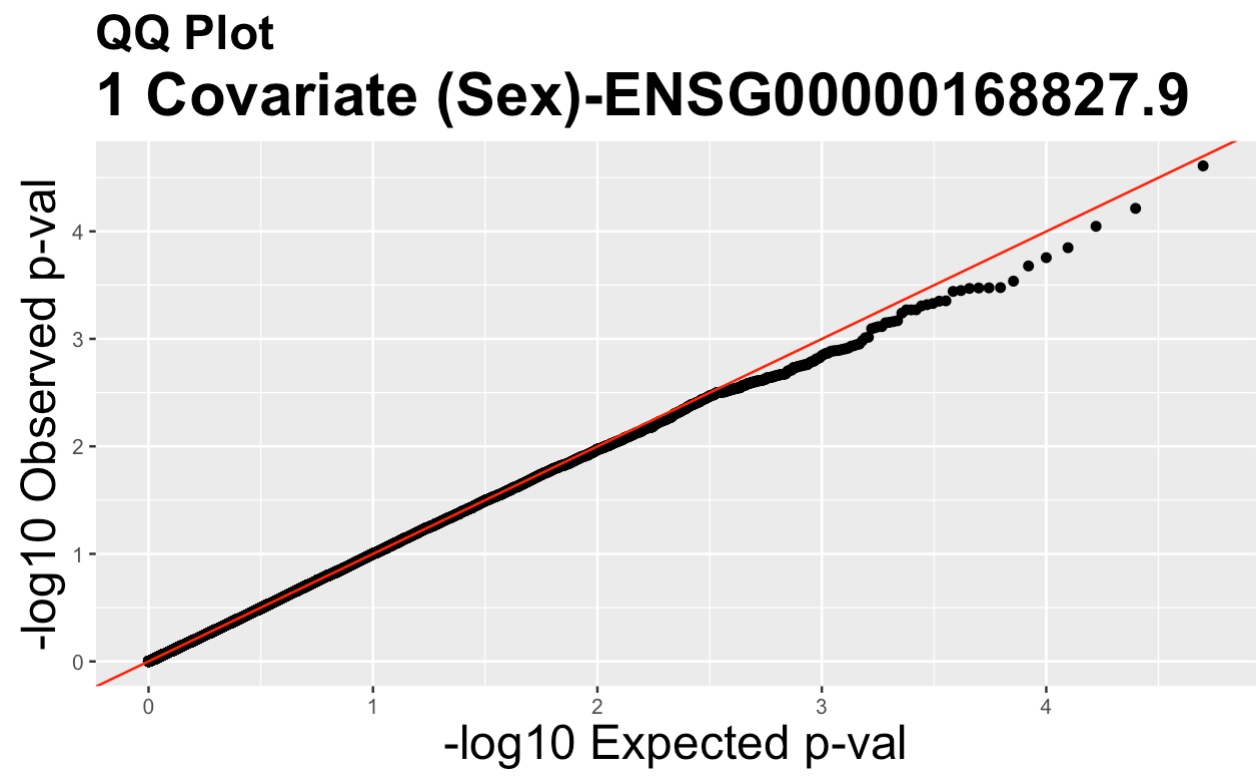
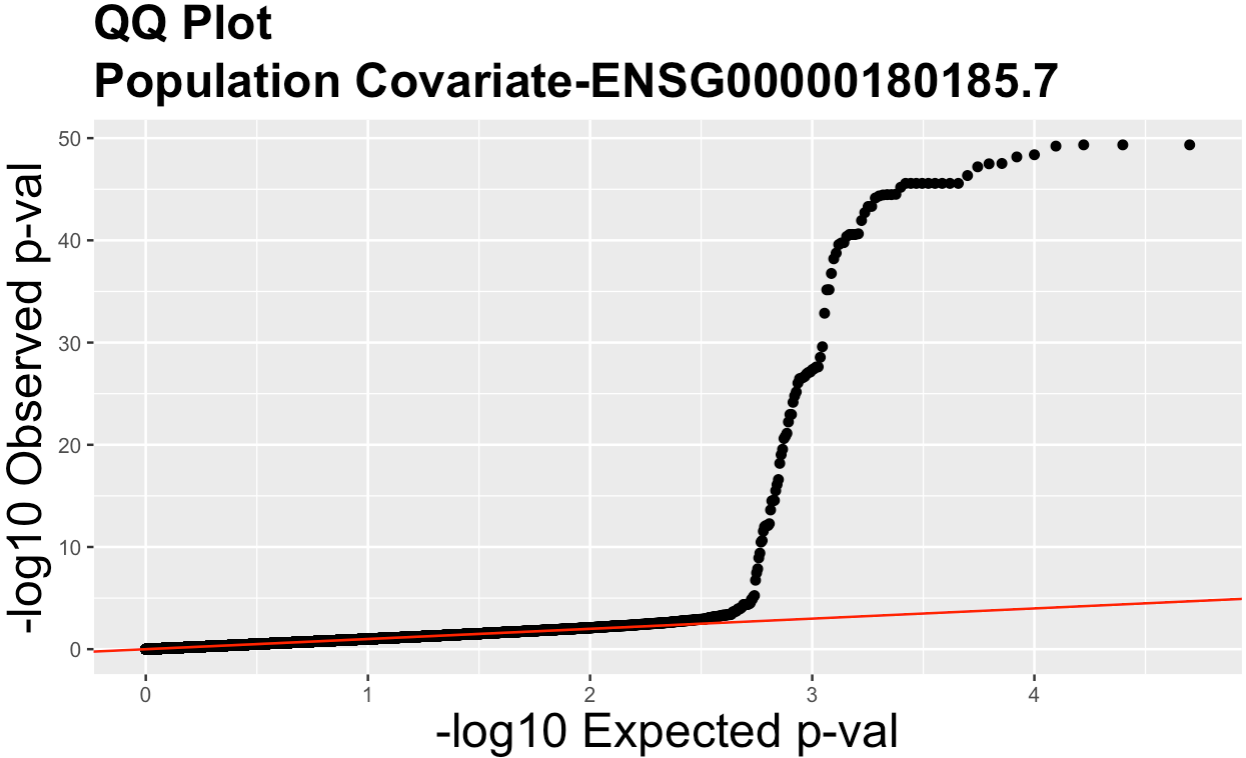
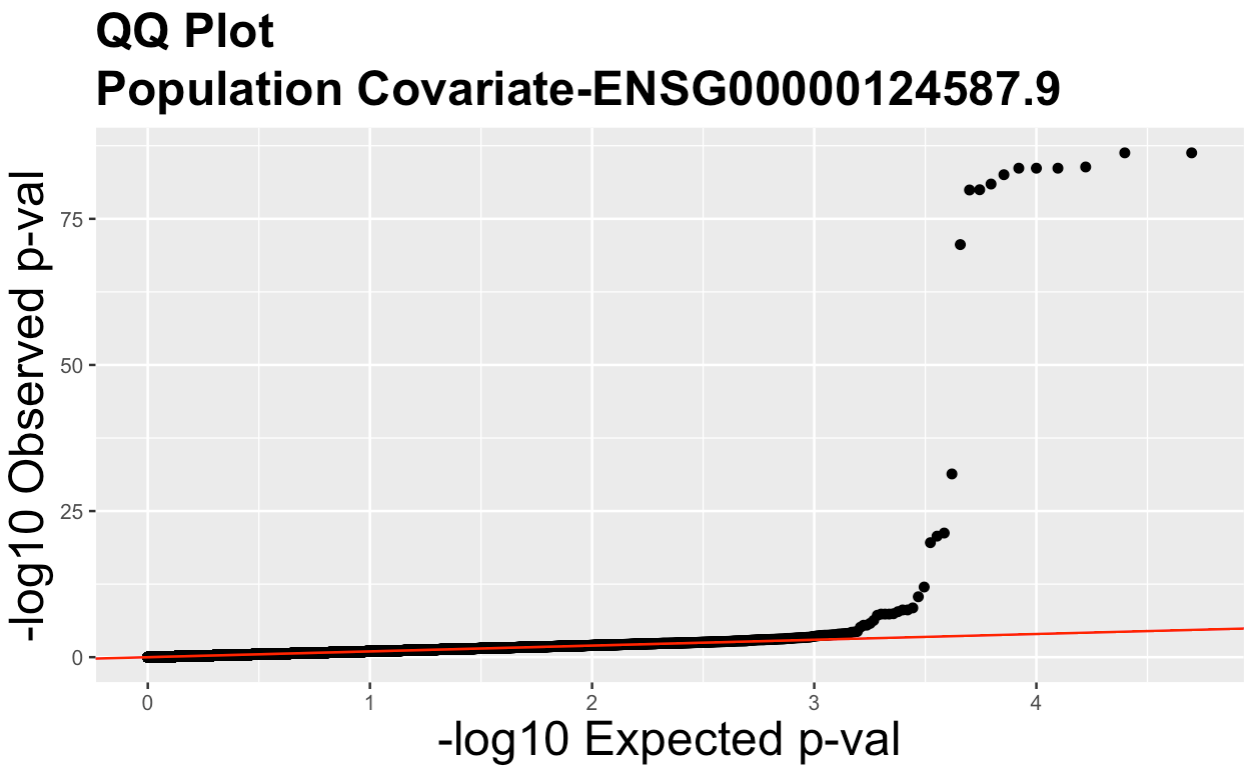
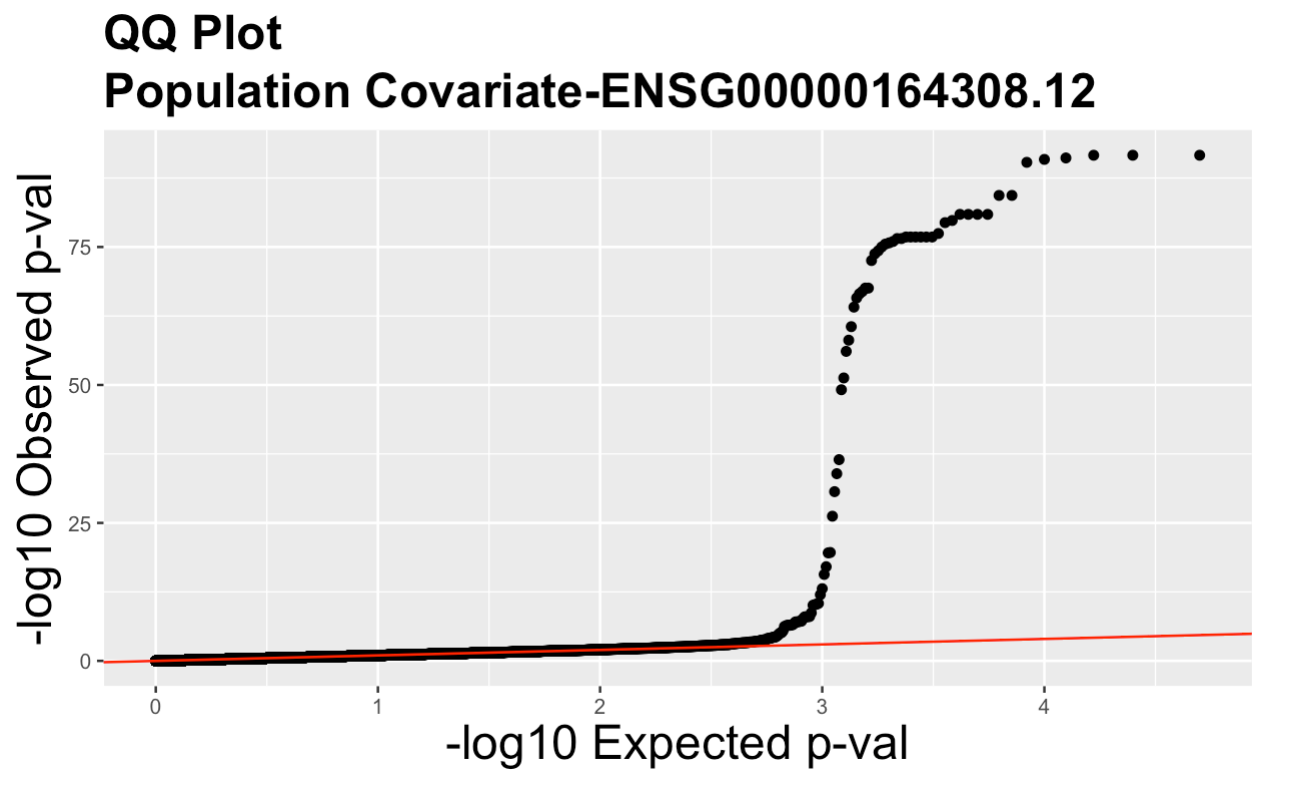


Figure 6: QQ plot of p-values for the five genes with sex as a covariate



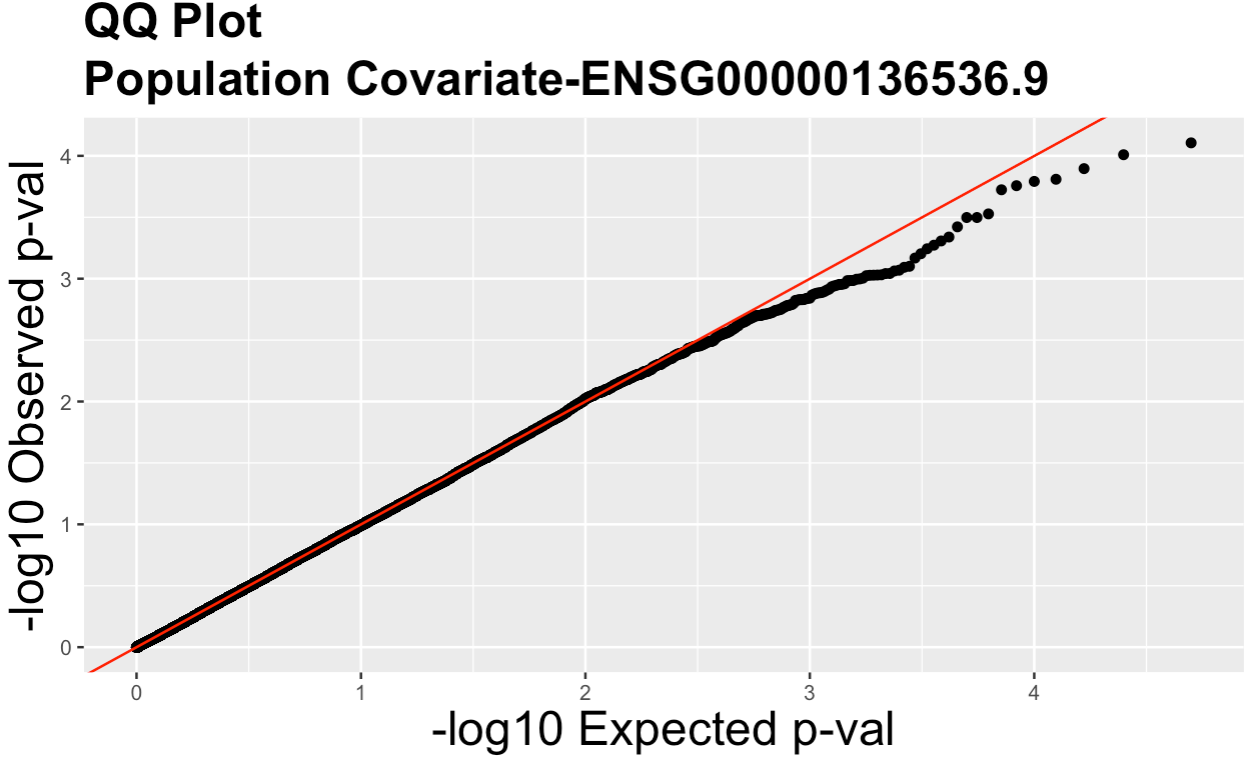
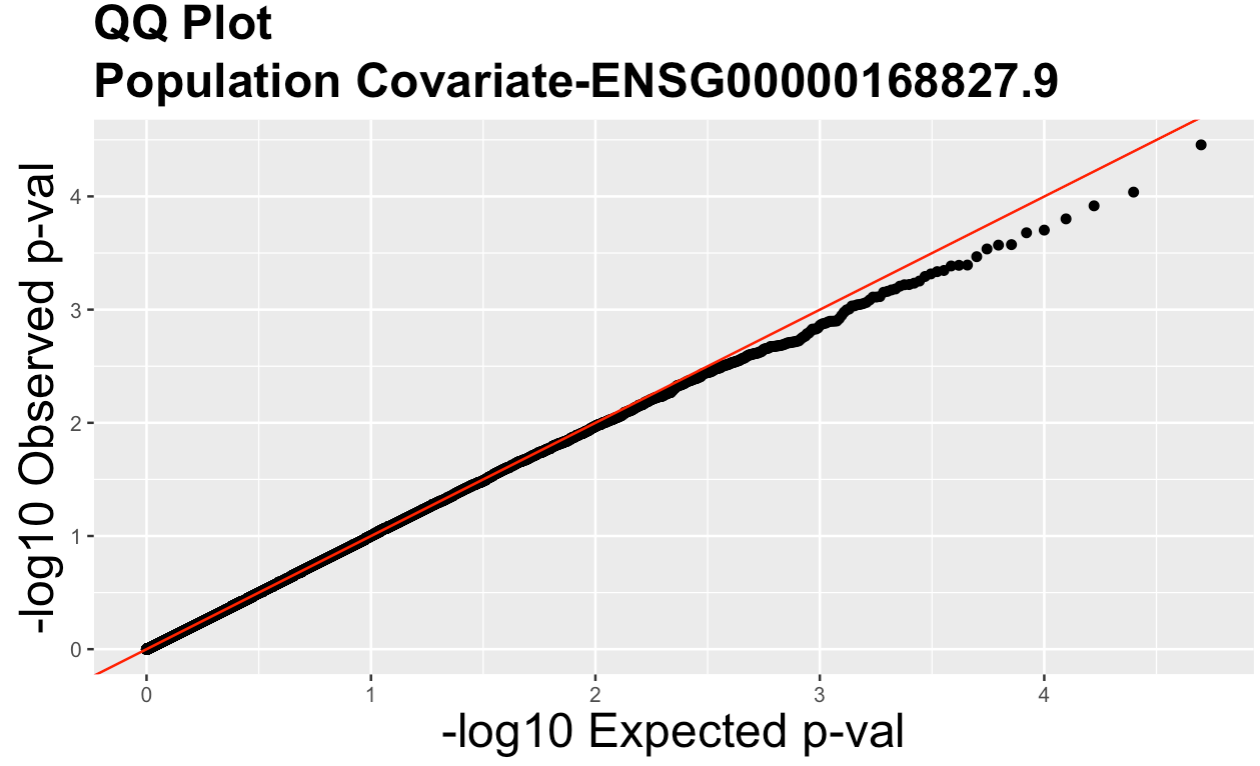
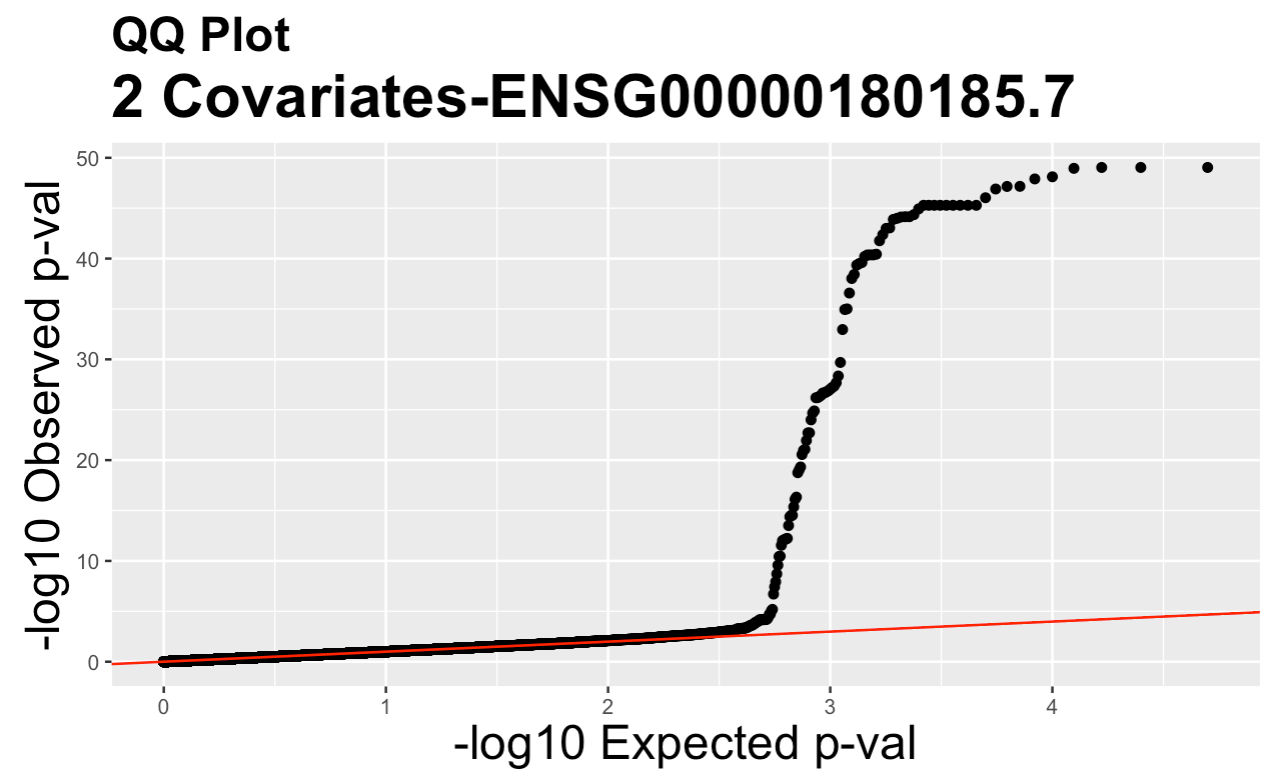
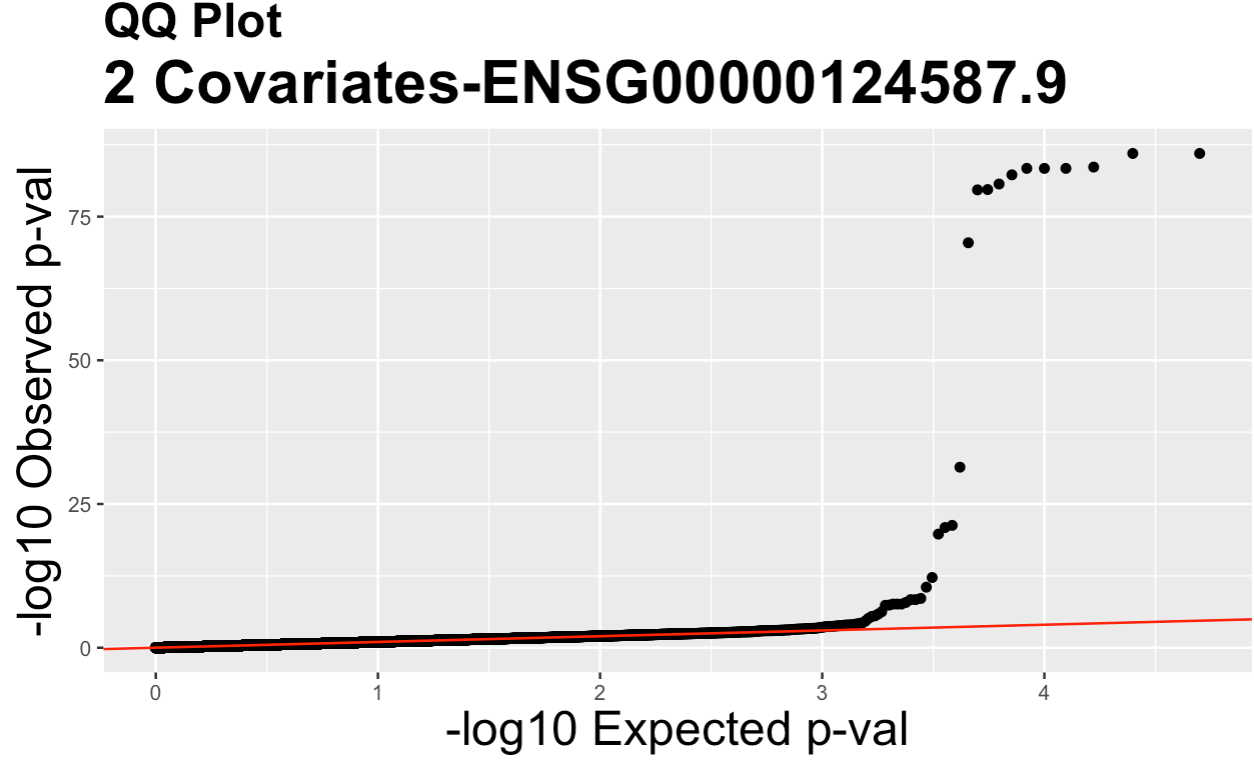
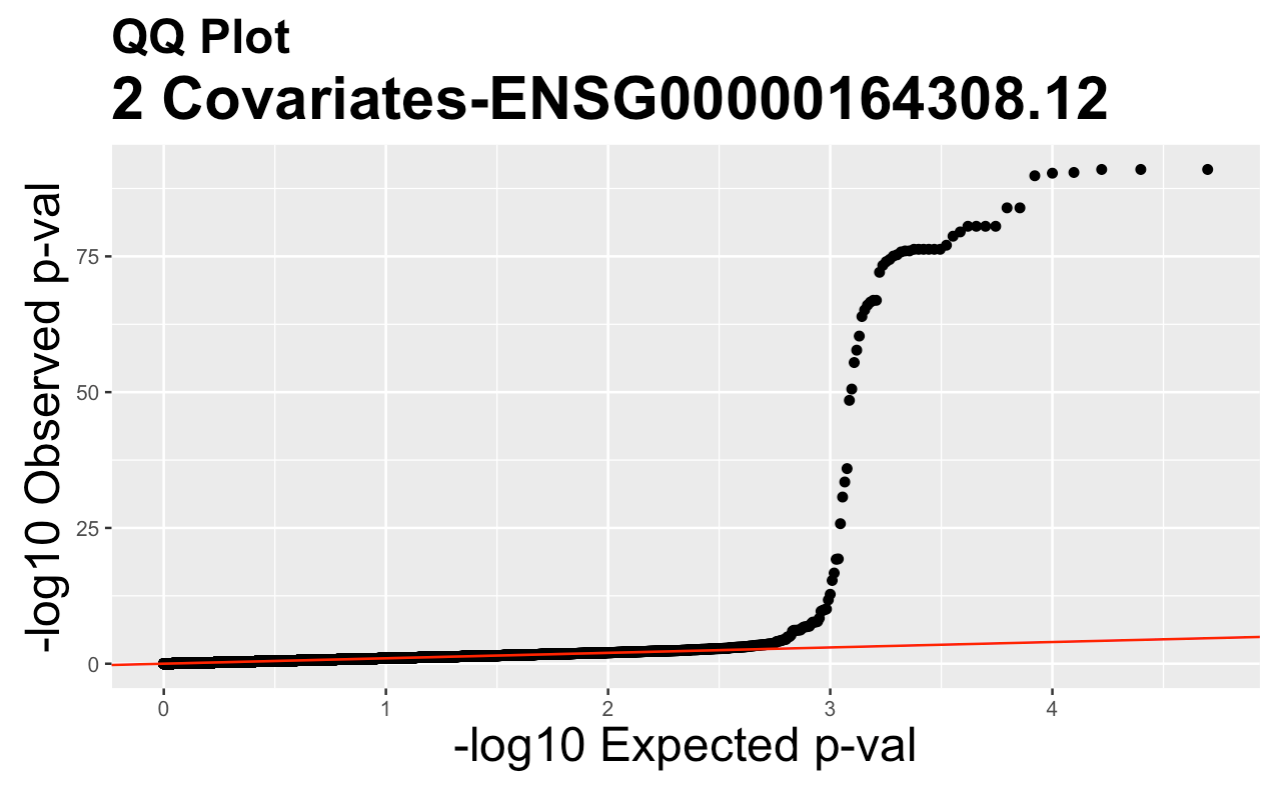


Figure 7: QQ plot of p-values for the five genes with population as a covariate.



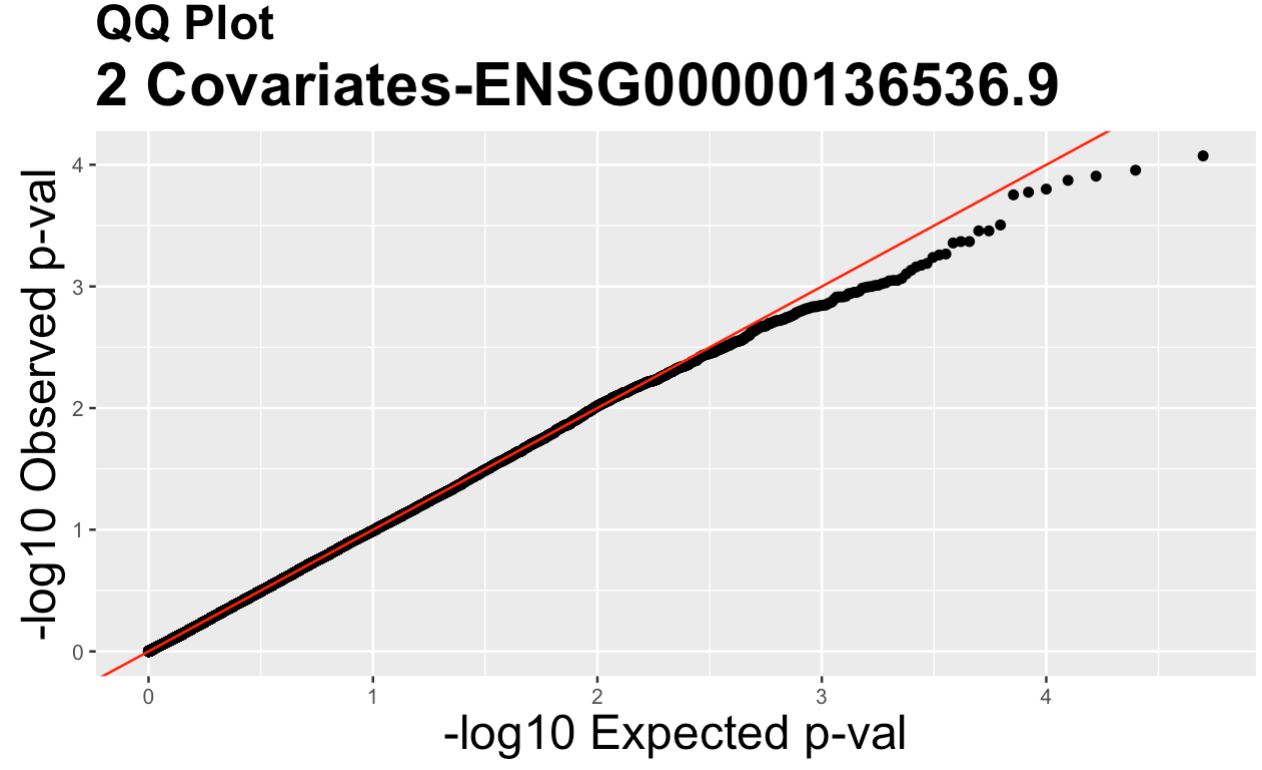
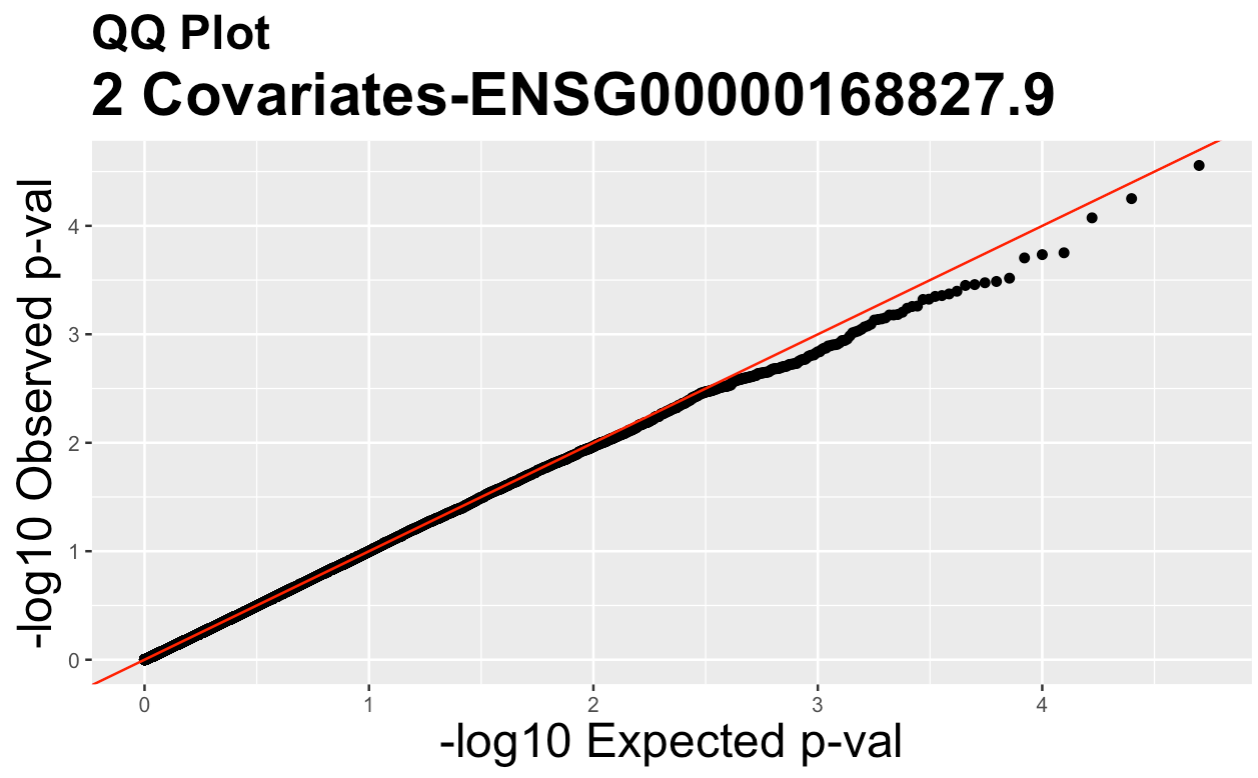
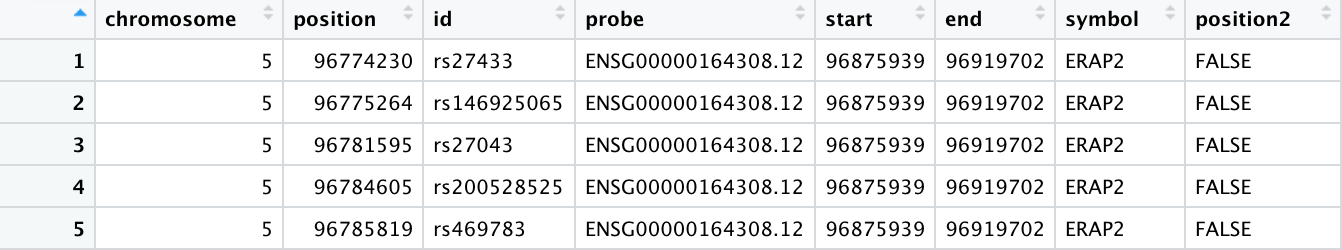


Figure 8: QQ plot of p-values for the five genes with both population and sex as covariates.

Table 1: First 10 significant markers for ENSG00000164308.12 identified using the model without covariates. For each SNP: contains the chromosome number, position on the chromosome, id, probe, start, end, common name for the gene of interest, and whether it’s located between the start and end of the gene of interest.



**Discussion**

In order to apply a statistical model to this dataset, two core assumptions were made: first, at least one causal mutation has affected each of the five genes of interest at some point within the evolutionary history of the human population. Second, there are at least two alleles at the locus being considered since humans are diploid organisms. From here, analysis was conducted as outlined above using multiple linear regression.

The p-values calculated at each marker represent the probability of obtaining a statistic (here, the F test, a form of the LRT statistic) or more extreme, conditional on the null hypothesis being true. The null hypothesis is an assumption of the value of the parameters, 𝛽a=0 and 𝛽d=0. On the other hand, the alternative hypothesis defines where the true parameters fall under the assumption that the null is incorrect. In this case, this is simply that 𝛽a≠0 or 𝛽d≠0, with 3 possibilities: there are only additive effects, only dominance effects, or both. Either way, it would suggest that the position corresponding to the marker likely contains a causal SNP.

The p-values were first visualized using QQ plots. These graphs fall into one of two categories. In the case where there are no causal polymorphisms, the QQ plot looks like a straight line at a 45 degree angle. This means the GWAS didn’t detect any hits and that the p-values follow the uniform distribution as expected. In contrast, when there are causal polymorphisms present, the QQ plot will be shaped like a straight line with a tail. This means while the majority of p-values do follow the uniform distribution, there are a few in linkage disequilibrium with the causal polymorphism that have significant p-values. For each of the four GWAS analyses, ERAP2 and GFM fell into the first category while MARCH7, PEX6, and FAHD1 fell into the latter. This is corroborated by either the lack of presence of significant markers after the Bonferroni correction.

The p-values were also visualized using Manhattan plots to identify regions spanned by a set of correlated SNPs. The genes MARCH7, FAHD1, and PEX6 clearly have sets of contiguous markers visible as “skyscrapers.” This likely maps the location of at least one single causal polymorphism. The local plot of just the significant p-values does reveal one outlier for PEX6 that is in equilibrium with the rest, located on chromosome 4 instead of 6. This was excluded from the final analysis.

The hits were first narrowed down using a critical value of 𝛂=.05. If the p-value<.05, there’s sufficient evidence that the null hypothesis is incorrect. Since the true parameter values remain unknown in this circumstance, it’s impossible to know with absolute certainty if the right decision was made. Furthermore, since the type I error is established to be 5%, the null hypothesis is erroneously rejected 5% of the time. This can have detrimental consequences on future experiments, as pursuing further research into these false positives is a waste of resources. Since there are 50,000 hypothesis tests being run simultaneously, the multiple testing problem occurs in which there will likely be chance findings just due to the quantity of inputs. This is illustrated by the presence of over 2000 p-values for each of the 5 genes without the correction, especially since the QQ plots for ERAP2 and GFM1 suggest that there shouldn’t be any significant positions out of the genotypes measured. The Bonferroni correction accounts for this by setting the threshold to .05/# of tests, essentially decreasing the significance level to weed out false positives. Here, it had a large impact on the number of significant markers, yielding 73 for MARCH7, 29 for PEX6, and 90 for FAHD1 with no covariates. This stayed consistent across different testing approaches, with the same markers passing the Bonferroni correction regardless of the multiple regression equation used (give or take a couple).

The power, or probability of correctly rejecting the null hypothesis when it’s false, can’t be controlled directly. It’s affected by factors such as sample size, magnitude of linkage disequilibrium, and size of effect of genotype on phenotype. While it’s difficult to control the identification of true positives, there are measures that can be taken to reduce the amount of false positives. Due to the vast nature of the genome, it’s impossible to perform a GWAS for every position on a chromosome. Thus, the majority of the time, the causal polymorphism itself isn’t even measured. Even if the statistically significant markers end up being non-causal, they’re correlated with the causal genotype due to linkage disequilibrium. This means they’re physically linked on the same chromosome and within the same region on that chromosome. The closer the polymorphisms are on a chromosome, the lower the probability of a recombination event between them, enhancing disequilibrium.

Another source of false positives and reduced power stems from unaccounted for covariates. Two factors were measured for each individual: sex and population. Including covariates in the multiple regression model means the covariate model parameter will be estimated and its correlation to expression measurements will be accounted for. However, the QQ plots between the no covariate model and all 3 covariate models look almost identical despite the PCA seemingly confirming the presence of population structures. Thus, the results from the original analysis with no covariates can be used.

**Conclusion**

More research needs to be done in order to identify the exact causal locus/loci for MARCH7, FADH1, and PEX6. However, the results of this experiment provide general regions for search and candidate loci. This experiment could be repeated looking at genotypes at other positions on chromosomes 5 and 3 because it’s possible that SNPs correlated with the causal polymorphism weren’t provided in this dataset. It’s also possible that there are covariates beyond the ones provided that are influencing the data. Performing GWAS with additional samples (such as the fifth population that was omitted) or new covariates could also lead to new findings.