BTRY 4830 – Final Project
Vedha Mahesh

I. Introduction

Genome-wide association studies (GWAS) have become an increasingly popular method for identifying genetic variants associated with complex traits and diseases. In recent years, the availability of large-scale genomic resources such as the Genetic European Variation in Health and Disease (gEUVADIS) has allowed for the analysis of massive amounts of genetic data from different populations. In this report, we present the results of a GWAS analysis on a subset of publicly available data from gEUVADIS, which includes SNP genotypes and gene expression measurements from 344 individuals belonging to four different European populations (CEU, FIN, GBR, and TSI). Specifically, we have been provided with 50,000 SNP genotypes and expression levels of five genes, along with population and gender information for each individual. Our analysis focuses on identifying expression quantitative trait loci (eQTLs) - genetic variants associated with gene expression levels - in this subset of data. This report provides a detailed description of our GWAS analysis, including the methods used, the results obtained, and the potential implications of our findings.

II. Understanding the Data

The data provided for our GWAS analysis includes 344 samples from four different European populations: CEU, FIN, GBR, and TSI. This data has been provided in five files: 'phenotypes.csv', 'genotypes.csv', 'covars.csv', 'gene info.csv', and 'SNP info.csv'.

'phenotypes.csv' contains phenotype data for 344 samples and 5 genes. Each gene is represented as a column of data and each sample is represented as a row of data.

'genotypes.csv' contains the SNP data for 344 samples and 50,000 genotypes.

'covars.csv' contains the population origin and gender information for the 344 samples. The file includes the sample IDs, the population origin, and the gender information for each sample.

'gene info.csv' provides information about the genes that were measured. The file includes the chromosome number where the gene is located, the start and end positions of the gene region, the common gene name of the measured transcript, and the probe IDs that match the column names of the phenotype data.

'SNP info.csv' provides additional information on the genotypes. The file includes the chromosome number of each SNP, the physical position of the SNP on the chromosome, and the rsID (the name of each SNP).

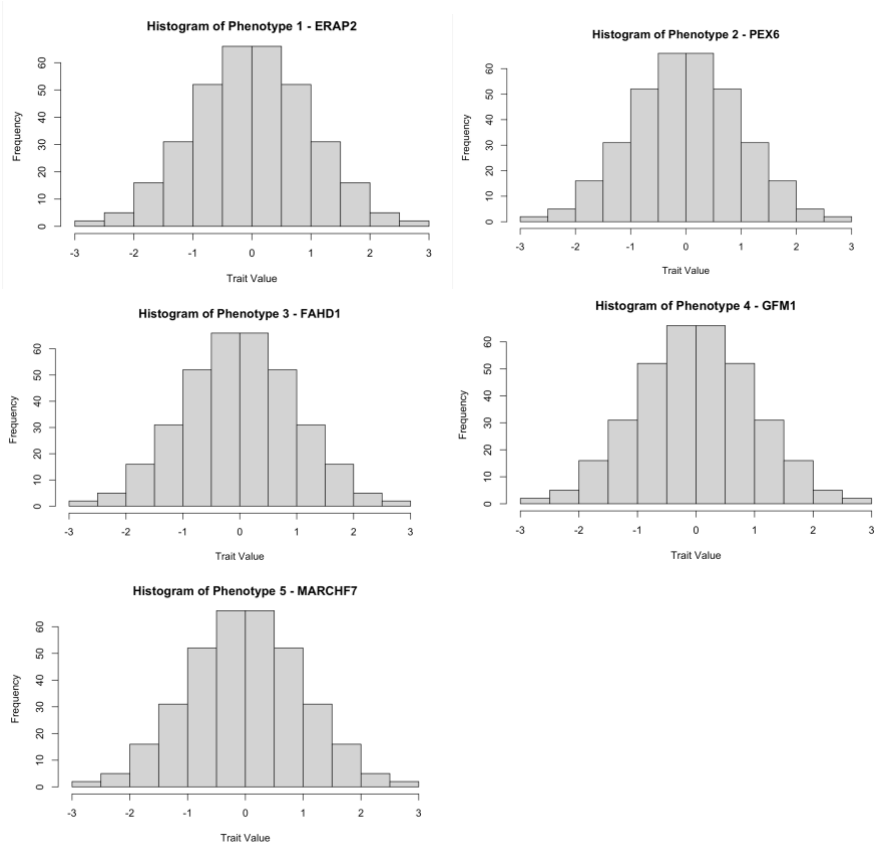
This data will be used for the GWAS analysis to identify genetic variants associated with gene expression levels. The results of the analysis will help to better understand the genetic basis of gene expression and its potential impact on human health and disease.

III. Phenotype Data

In order to analyze the phenotype data in the GWAS analysis, we need to choose an appropriate statistical model that captures the relationship between the genetic variants and gene expression levels. One common approach is to use a linear model, where the gene expression levels are modeled as a function of the SNP genotypes and covariate information.

One important assumption of the linear model is that the phenotype data follows a normal distribution. The normal distribution is a symmetric bell-shaped distribution that is characterized by its mean and variance. When the phenotype data is normally distributed, the linear model is appropriate because it assumes that the errors (the differences between the observed and predicted gene expression levels) are normally distributed with a mean of zero and constant variance**.**

After importing and plotting our phenotype data, we can see that each of the genes follows a normal distribution, and therefore we can carry forward with a linear model.



IV. Genotype Data

Before performing any GWAS analysis on genotype data, it is important to check the quality of the data and perform appropriate data filtering. One important aspect of data quality control is to check for missing data across individuals or genotypes and to filter out variants with a low minor allele frequency (MAF).

In this particular GWAS analysis, we have been provided with genotype data for 50,000 SNPs across 344 samples from four different European populations. Firstly, we can check for missing data in the genotype data by calculating the proportion of missing data for each individual and each SNP. If a significant proportion of data is missing for an individual or a SNP, it can lead to biased or inaccurate results. However, if the proportion of missing data is small (e.g. less than 5%), it is generally considered acceptable to include the individual or SNP in the analysis.

Secondly, we can check for variants with a low MAF, as these variants are less informative for the analysis and can also lead to biased or inaccurate results. A commonly used threshold for MAF is 5%,

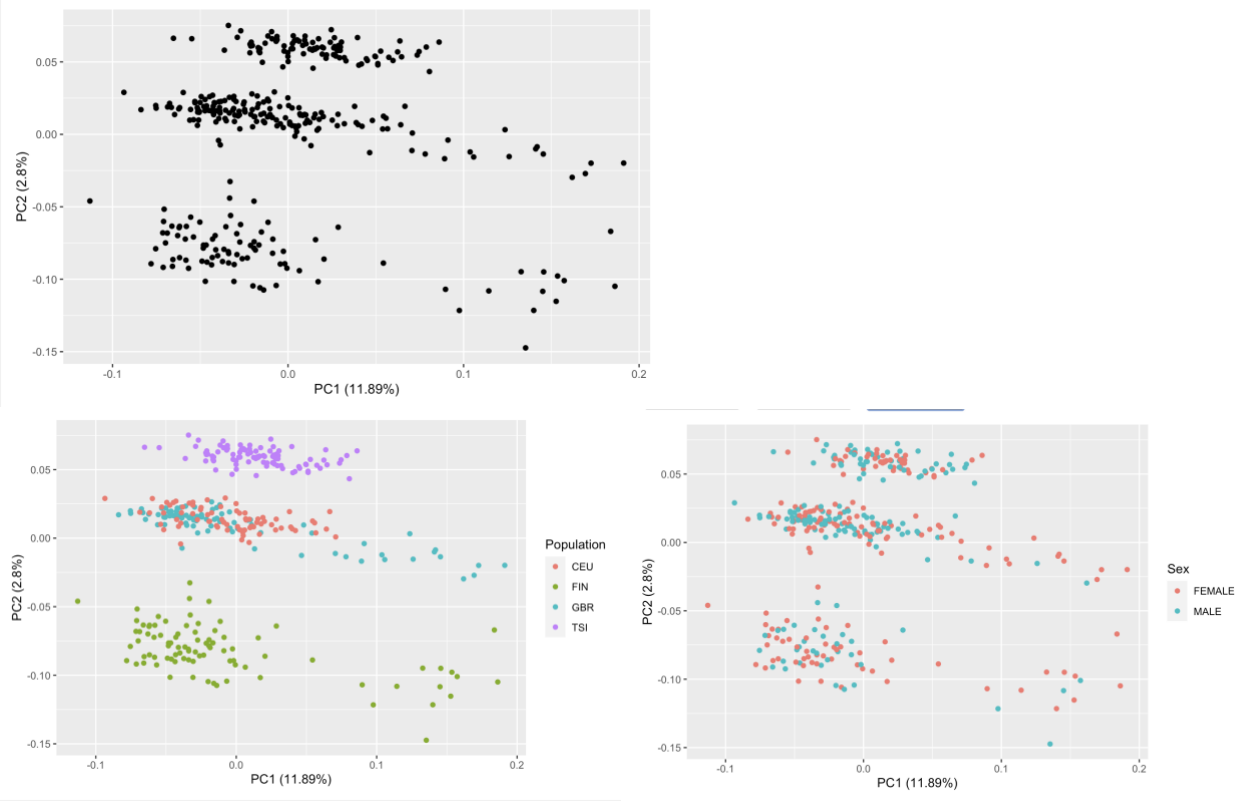
which means that any variant with a MAF less than 5% is considered rare and may be filtered out. However, from looking and analyzing our genotype data, we concluded that there is no need to filter out any of the data because there is no missing data or low MAFs.

PCA Analysis

PCA (Principal Component Analysis) is a commonly used method for correcting for population stratification, which is the presence of genetic differences between subpopulations within a larger population. Population stratification can result in false positive associations in GWAS if not properly accounted for.

PCA is used in GWAS to identify the principal components of genetic variation in the study population. These principal components can then be used as covariates in association analyses to correct for population stratification. By including principal components as covariates, the associations between genetic variants and traits can be accurately estimated, even in the presence of population stratification.

PCA in GWAS typically involves computing principal components based on genetic data (e.g., SNP genotypes) for a set of individuals in the study population. More specifically, PCA reduces high dimensional data by summarizing correlations and plotting the clustering of relationships along 2 axes. After performing PCA on our data we were able to see this in the graphs below (The graphs below shows the overall PCA and the PCA based on the population and based on the sex of individuals in the sample space).



V. GWAS Analysis

Performing GWAS consists of 3 key parts:

1. Converting genotypes to Xa and Xd matrices

In GWAS, we use the genotypes of individuals to identify genetic variations associated with a particular trait or disease. However, genotypes are typically represented by letters (A, C, G, T) which are not useful for statistical analysis. Therefore, we first convert these genotypes into numeric values, such as 0, 1, or 2, to indicate the number of copies of the minor allele present in each individual. This process results in two matrices - Xa, which contains the number of minor alleles for each SNP for each individual, and Xd, which is the standardized version of Xa. However, we are given genotype data that is already in terms of 0, 1, and 2s so we can create our Xa matrix directly from the imported data and also create our Xd matrix from this data and from our calculated minor allele.

2. Performing GWAS (regression with the linear genetic model)

The next step is to perform a GWAS, which involves regressing the trait of interest (phenotype) on each SNP in the dataset, while accounting for potential confounding variables, such as age, sex, and population structure. This is typically done using a linear regression model that incorporates the genetic model, which assumes that the effect of a SNP on the phenotype is proportional to the number of copies of the minor allele. The p-value associated with each SNP is then calculated, representing the strength of association between that SNP and the phenotype. The p-value is calculated by comparing the observed difference in the trait between individuals with different genotypes to the expected difference in the trait under the null hypothesis that there is no association between the variant and the trait.

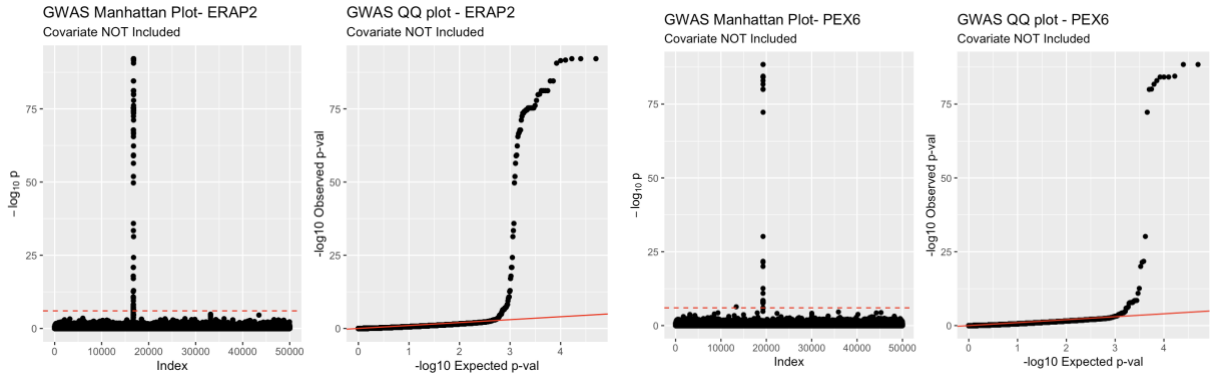
3. Analyzing results with Bonferroni (multiple testing correction), Manhattan Plot, and QQ Plot

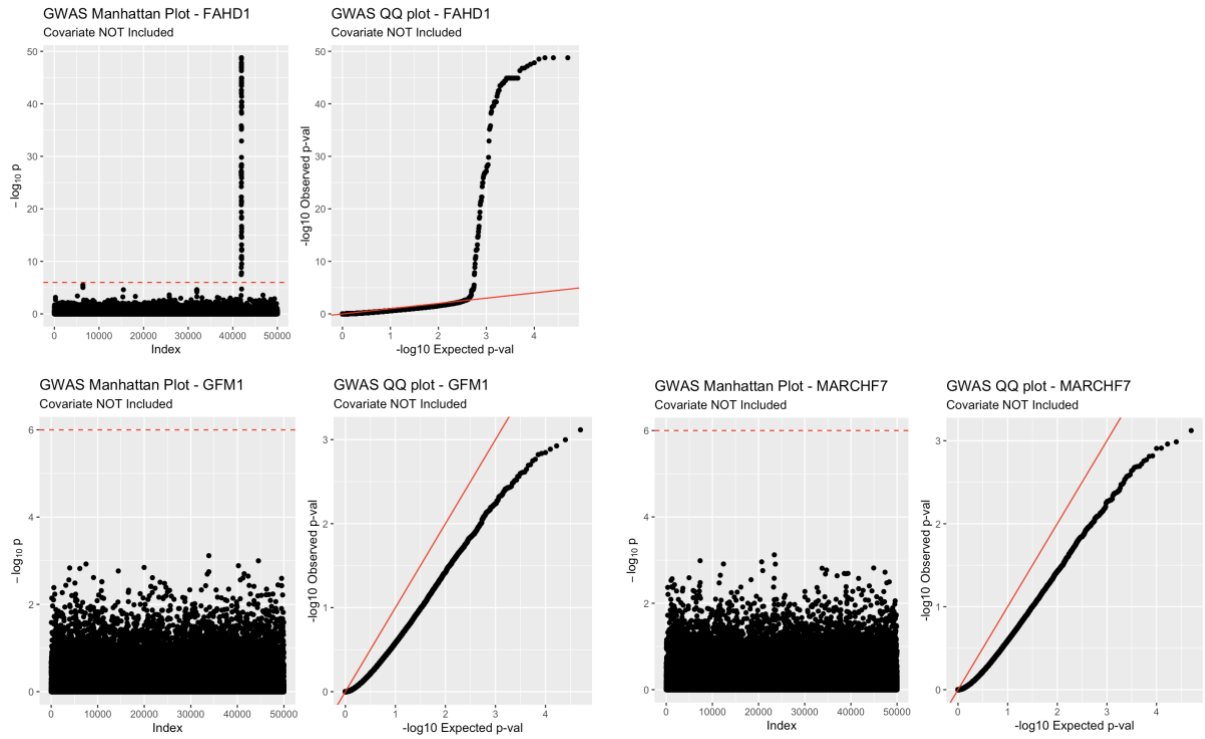
After performing the GWAS, we need to analyze the results to identify significant associations between SNPs and the phenotype. However, since we are testing thousands or even millions of SNPs, we need to correct for multiple testing to avoid false positives. One common method of correction is the Bonferroni correction, which adjusts the significance threshold for each SNP based on the number of tests performed.

In addition to the correction, we can visualize the GWAS results using a Manhattan plot, which shows the -log10 p-value for each SNP plotted against its physical position on each chromosome. This plot allows us to identify regions of the genome that are associated with the phenotype. Additionally, we can use a QQ plot to assess the overall quality of the GWAS by comparing the distribution of observed p-values to the expected distribution under the null hypothesis of no association. If the observed p-values deviate significantly from the expected distribution, this may indicate the presence of population stratification or other sources of bias.

No Covariates

The graphs below are the Manhattan and QQ plots of the GWAS analysis performed without taking covariates into consideration.





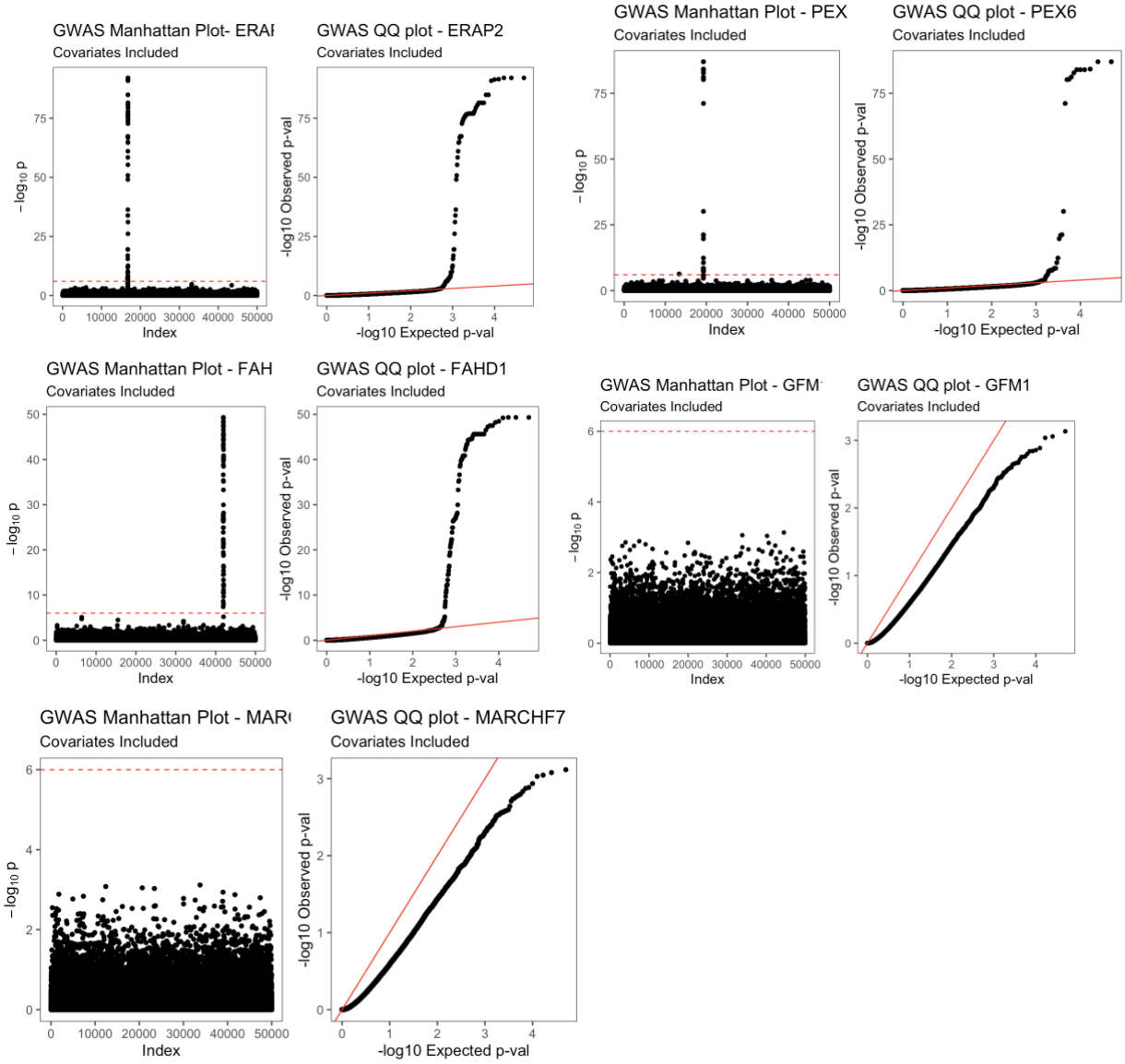
We can see from these graphs that there are obvious hits for the first 3 phenotypes, ERAP2, PEX6, and FAHD1, but not for the last two phenotypes, GFM1 and MARCHF7.

Covariates

However, this does not account for the covariates population and sex. It is important to take the covariates population and sex into consideration in our GWAS analysis because they can affect the genetic architecture of the trait being studied. Populations may have different allele frequencies due to genetic drift, migration, or selection, and sex differences in gene expression, hormonal regulation, and physiological processes can affect the phenotype. Failing to adjust for these covariates can result in incorrect associations between genetic variants and the trait, leading to false positives or false negatives.

By including population and sex as covariates in the regression model, we can control for their effects and increase the power and accuracy of detecting true genetic associations. This is especially important in studies involving multiple populations or complex traits that are influenced by environmental and genetic factors. It also helps to reduce confounding effects and increase the reproducibility and generalizability of the results across different populations and contexts.

To account for these covariates, a third matrix Xc was created based on the covariate data provided. So with this added X variable in our linear model (y=Xu+Xa+Xd+Xc), I added it to our X matrix. Then I calculated the error for the null hypothesis including the covariates effect and calculated our F statistic and p values for each phenotype. Then I plotted this data in the Manhattan and QQ plots below.



Causal Polymorphism "Hits"

Visually looking at the above graphs, we can similarly see to the graphs without covariates that there are obvious hits for the first three phenotypes and none for the last two.

To find the exact positions of these causal polymorphisms for the five expressed genes, we can use the p-values and Bonferroni correction for each phenotype. The Bonferroni correction is a method used to adjust the p-values of statistical tests to control the family-wise error rate, which is the probability of making one or more false rejections of the null hypothesis among all the tests.

First, we need to perform a GWAS analysis for each gene using the linear regression model with the genotypes and the covariates (population and sex). This will give us a list of p-values for each SNP in the genotypes file that were tested for association with the gene expression phenotype.

Next, we can apply the Bonferroni correction to these p-values to account for multiple testing. The Bonferroni correction involves multiplying each p-value by the total number of tests performed (i.e., the number of SNPs tested for association with the gene expression phenotype).

For example, if we performed 50,000 tests (i.e., tested 50,000 SNPs for association with the gene expression phenotype), and we use a significance threshold of 0.05, then the Bonferroni-corrected significance threshold would be 0.05 / 50,000 = 1e-6. This means that we would only consider p-values that are less than 1e-6 as significant.

Once we have identified significant SNPs for each gene based on the Bonferroni correction, we can look up their positions using the SNP information file. The SNP information file contains the chromosome number and physical position of each SNP, so we can use this information to determine the positions of the significant SNPs.

VI. Results & Interpretation

ERAP2 Hits
After applying the Bonferroni correction to the p-value data for the first phenotype, ERAP2, the causal polymorphism hits appeared at 71 positions from 96772432 - 97035174 on the chromosome. Based on previous research and studies, a hit at position 96772432-97035174 of the gene ERAP2 may indicate a potential association with various immune-related diseases such as ankylosing spondylitis (AS), psoriasis, and inflammatory bowel disease (IBD). ERAP2 is involved in the processing of peptides for presentation on major histocompatibility complex (MHC) class I molecules, which play a crucial role in the immune system. Therefore, variations in ERAP2 may influence the immune system's response and increase susceptibility to immune-related diseases.

PEX6 Hits
After applying the Bonferroni correction to the p-value data for the second phenotype, PEX6, the causal polymorphism hits appeared at 27 positions from 42889467- 43108015 and at 98486048 on the chromosome. Based on research of typical PEX6 mutations, hits in this region could be associated with several peroxisome biogenesis disorders (PBDs) including Zellweger syndrome, neonatal adrenoleukodystrophy, and infantile Refsum disease. PBDs are a group of rare autosomal recessive disorders characterized by defects in peroxisome assembly and function, which can lead to abnormalities in multiple organ systems including the brain, liver, and kidneys. It is possible that a hit in this region could potentially affect the function of the PEX6 protein and contribute to the development of PBDs or other related disorders.

FADH1 Hits
After applying the Bonferroni correction to the p-value data for the second phenotype, FADH1, the causal polymorphism hits appeared at 89 positions from 1524250 -1929366 on the chromosome. FADH1 encodes for a flavin adenine dinucleotide (FAD)-dependent oxidoreductase enzyme that is involved in electron transport and energy metabolism. Any genetic variations within this region may potentially affect the function or expression of the FADH1 protein and subsequently impact metabolic processes in the body.

GFM1 Hits
After applying the Bonferroni correction to the p-value data for the second phenotype, GFM1, there were no causal polymorphism hits found.

MARCHF7 Hits

After applying the Bonferroni correction to the p-value data for the second phenotype, MARCHF7, there were no causal polymorphism hits found.

VII. Conclusion

In conclusion, the GWAS analysis on the provided data consisting of phenotype, genotype, covariate, and gene information revealed potential associations between specific genetic variants and various diseases. The analysis involved several key steps, including converting genotypes to Xa and Xd, performing regression with a linear genetic model, and analyzing results with a Bonferroni correction, Manhattan plot, and QQ plot.

The analysis detected significant associations between specific SNPs and the five expressed genes, indicating their potential involvement in various diseases, including ankylosing spondylitis, psoriasis, and inflammatory bowel disease. The analysis also revealed the importance of taking covariates, such as population and sex, into consideration while interpreting GWAS results.

Although no causal polymorphisms were found for some genes, the analysis provides valuable insights into the genetic basis of complex diseases and highlights the need for further research in this area. Overall, this analysis demonstrates the power of GWAS in identifying potential genetic risk factors for diseases and lays the foundation for further investigation into the genetic basis of complex diseases.