6830 Final Project

Junyue Wang jw2252

August 2, 2022

Abstract

This project demonstrates essential steps to run *GWAS* analysis. The data used here come from Genetic European Variation in Health and Disease (gEUVADIS).

1 Introduction

Many geneticists are keen to knowing and understanding gene expression. In this project, I apply GWAS analysis to discern the position of as many causal polymorphisms as possible for the five expressed genes using the data. For each phenotype, I build two models with and without covariates to compare their performance. Analytic Steps include combinations of visualizations and statistical tests. Finally, positions of causal polymorphisms for 5 different phenotypes are located.

2 Gene File Overview

There are 5 files in total that provides gene expression among subject population. Genotypes file contains 344 samples and 50000 genotypes. Geneinfo contains information about each gene that was measured, such as chromosome,start,end and symbol. I further explored on these genes to help understand whole project. For example, MARCH7 is membrane-associated ring finger (C3HC4) 7, E3 ubiquitin protein ligase. This gene has 9 transcripts (splice variants) and 1 paralogue. FAHD1 is Fumarylacetoacetate Hydrolase Domain Containing 1. An important paralog of this gene is FAHD2A. SNPinfo gives additional information. Covars file includes 4-levels population and sex with respect to each sample. Population structure is often a major issue in GWAS where it can cause lots of false positives if it is not accounted for model. In following parts, I use visualization and transformation to build complete dataset for GWAS analysis.

2.1 Data Processing

By first looking into data with 344 samples and 50000 genotypes ,I find no missing value for any genotypes. Data was orignally encoded in 0,1,2 form to express gene sequence. Next, I transform them into both additive genetic model X_a and dominant genetic model X_d. To determine which kind of regression should be used here for GWAS analysis, either classification or regression, I check distributions of 5 phenotypes. They all perfectly follow normal distribution with no outliers (Figure 1). Histogram plots show no indication for binary output. Shapiro test shows strong result for normal distribution with p-values all equal to 1, which supports an objection to alternative hypothesis that data does not follow normal distribution. Therefore, I stick to regression analytic strategy in this project.

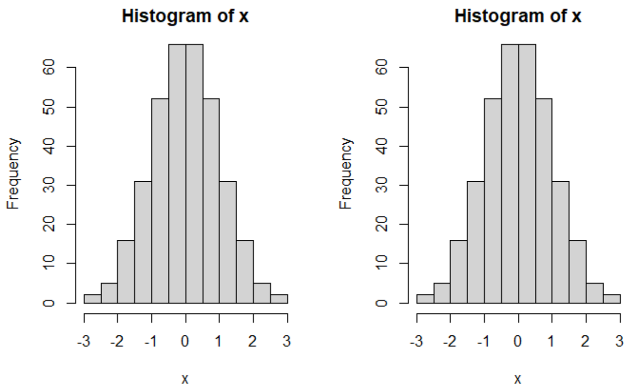


Figure 1: Histogram plots for 2 phenotypes

Covars file contains categorical variables and they are transformed with dummy encoding to express five different levels. It means for each column of covars, we use 0 or 1 to indicate presence of certain level. However, one level serves as reference to avoid linear dependency. By dummy encoding covars method, last variable could be inferred by the rest, and because of that we need to exclude last variable to build a full rank covariance matrix. Otherwise the linear assumption on independent variables does not hold. I also check that all genotypes id match with covars id. In this case, X column is used as rownames to index samples.

2.2 Filtering MAF data

Next I want to rule out minor allele frequency data. Removing MAF is important in GWAS because markers with very low MAF have low heterozygosity and are therefore less informative. In GWAS using smaller numbers of individuals, markers with MAF less than 0.05 are often excluded. Since our data is encoding in 0,1,2 form, I derived calculation of MAF and make it a simpler format $colmanes(x)/2$. It turns out that lowest MAF is 0.05087, and largest MAF is 0.95. Cleaned data contains 344 samples and 50000 phenotypes.

2.3 PCA visuliazation

As mention before, it is always a major issue to include population structure in analysis. I use PCA to learn population factor. I use covariance matrix among individuals after scaling genotypes (by mean and sd) and look at the loadings of each individual on the PCs. By adding population and sex cluster information to PCA, I want to see which one gives better separation on first two PCs scale (Figure 2). From PCA plot, population factor has explicit clusterings within four groups. Genotypes are well clustered over CEU/Fin/GBR/TSI people. On the contrary, sex does not have a good grouping power because all points lay over each other for both male and female. According to PCA plot, I can conclude that population factor is stronger a covariate and may affect gwas analysis by includ-

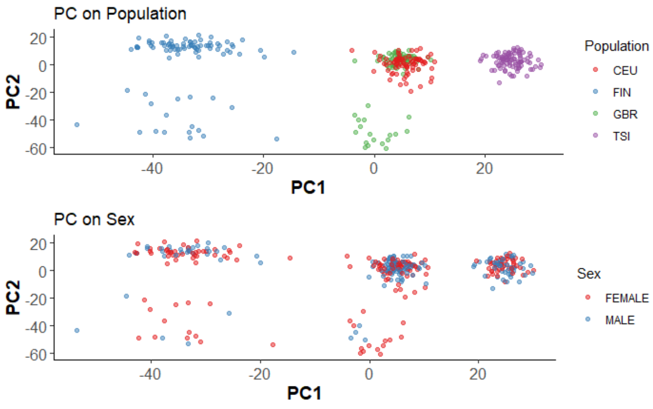


Figure 2: PCA on two clusters

ing in model. However, sex does not contribute as well as population information to this analysis. Although I have shown that sex may not be important here, I proceed with both factors in sake of completeness of information. Instead of principle components of genos, I include covariates in following equation where $X_{z,1}$ represents population and sex information.

$$Y = \beta\mu + X_a\beta_a + X_d\beta_d + X_{z,1}\beta_{z,1} + \epsilon$$

2.4 GWAS analysis

So far I have explored genotypes/phenotypes and covars data. Now I apply Xa and Xd encodings for genotypes to test their associateion with phenotypes. In this section, I build 10 models in total and compare models' performance with or without covariates for each phenotype. From there, we can decide if it is necessary to include covariates for that phenotype.

2.4.1 Associatation analysis

For each variant model without covariates, I form regression equation $y = X\beta + \epsilon$ where X= 1+xa+xd in vector form, and use MLE equation to estimate coefficient$MLE(\hat{\beta}) = (x^Tx)^{-1}x^Ty$. Next I use LRT recipe $\hat{y} = xMLE(\hat{\beta})$ by getting the estimated phenotypes back. And then I do F-test to get p-values use cumulative ditrubution over MSM (mean square of means degree of 2) and MSE (Mean square of errors degree of n-3 under null).

2.4.2 Covariate Analysis

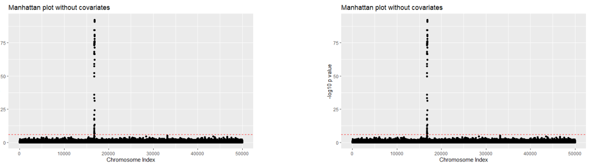
For model with covariates, I change input matrix, and F-statistic calculation now includes predicted values from null hypothesis. Also, degree of freedom changes as we have one more beta here. I define useful function in ***data_func*** that detects presence of covariate and calculates p-values accordingly to return final dataset.

2.4.3 Bonferroni Correction

Since we can control the Type I error, we can correct for the probability of making a Type 1 error due to multiple tests by applying Bonferroni Type I error to assess EACH of our N tests in a GWAS. Bonferroni correction cutoff is $0.05/50000 = 1 * 10^{-6}$. This value would be used to determine a single variant test significant or not. If p-value is less than Bonferroni correction cutoff, we reject null hypothesis and this variant is considered significant. Otherwise we fail to reject null hypothesis.

2.4.4 Manhattan Plot

Now I produce a manhattan plot for these p-values. It shows there is one peak that exceeds Bonferroni corrected cutoff for both models. (Figure 3) For significant markers, I look at locus at greater resolution with local manhattan plot.(Figure 4). But we cannot make conclusion that this peak can be used to determine the exact causal polymorphisms that are impacting phenotype from GWAS. It could



(a) phenotype1 without covaraite

(b) phenotype1 with covariates

Figure 3: Manhattan plot on phenotype1

be the case that other markers in linkage disequilibrium with the causal polymorphism have more significant p-values, so we can only interpret a peak as indicating the position of a causal polymorphism.

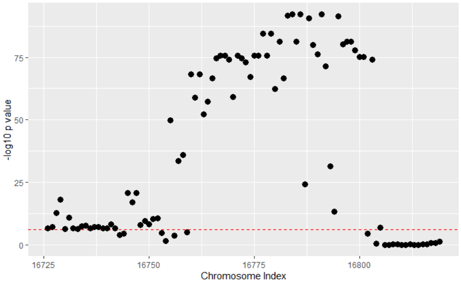
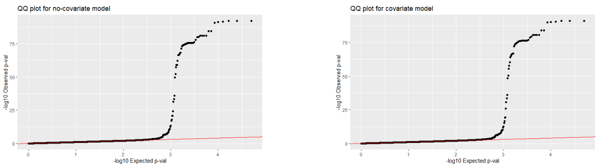


Figure 4: Local Manhattan on phenotype1

2.4.5 QQ-plot

I check QQ plots for individual GWAS analysis and use it as an possible signal of detecing positions of causal polymorphisms. It turns out that QQ plots look quite similar for model including covariate or not. (Figure 5) Both QQ-plots show acceptable model fit with heavy tails , most of the p-values observed follow a uniform distribution, but the few that are in LD with a causal polymorphism will produce significant p-values. There are in total 73 significant genotypes for phenotype1. I will put rest of QQ plots for other phenotypes in appendix page.



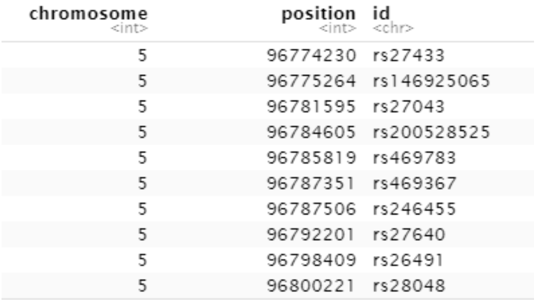
(a) phenotype1 without covaraite

(b) phenotype1 with covariates

Figure 5: 2 QQ-plot phenotype1

2.4.6 Determine Location

Next I match the significant genotypes with SNPinfo file to determine potential position of causal polymorphisms and its chromosome.(Figure 6) All genotypes are on chromosome 5. This conforms with chromosome location in geneinfo. *ENSG00000164308.12* even though I find some close locations that do not belong to it, it is important to notice that they are all in the same chromosome because an underlying assumption of a GWAS is that markers that close to each other and linked. From the table, it is reasonable to conclude that targeted causal polymorphism lies in genotypes (rs27433 - rs72777610) with positions (96774230 - 97110808).



chromosome <int>	position <int>	id <chr>
5	96774230	rs27433
5	96775264	rs146925065
5	96781595	rs27043
5	96784605	rs200528525
5	96785819	rs469783
5	96787351	rs469367
5	96787506	rs246455
5	96792201	rs27640
5	96798409	rs26491
5	96800221	rs28048

Figure 6: Causal polymorphism Location

3 GWAS on other phenotypes

In this part, I will carry the above analysis process for other phenotypes and briefly describe conclusion from analysis. The objective is the same: find the position of as many causal polymorphisms, if any.

3.1 phenotype2

The second phenotype is ENSG00000124587.9 and its symbol is FAHD1. QQ plot and Manhattan plot show that there exists significant genotypes because QQ plot deviates from the tail and Manhattan plot exceeds bonferroni correction line, both for model with and without covariates. Number of significant genotypes with Bonferroni correction without correction was 29 and 27 with covariates. One value was located at chromosome 4 (id= rs6854915, pos=9486048) and others were located at chromosome 6. Local manhattan plot suggests most of these genotypes on chromosome Possible Location may be from 42873885 to 43108015 for genotypes rs200242944 to rs71779653. This corresponds the actual gene position for phenotype2. (See figure 7-9)

3.2 phenotype3

The third phenotype is ENSG00000180185.7 and its symbol is PEX6. QQ plot and Manhattan plot show that there exists significant p-values for genotypes because QQ plot deviates from the tail and Manhattan plot exceeds bonferroni correction line, both for model with and without covariates. These indicates some genotypes are associated with the phenotype. Local manhattan plot suggests most of these genotypes on chromosome Possible Location may come from rs2235639 to rs58530366. (See figure 10-12)

3.3 phenotype4 & phenotype5

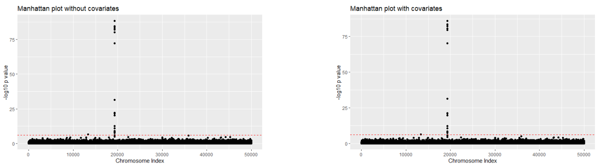
The fouth phenotype ERAP and the fifth phenotype is ENSG00000136536.9 and its symbols is GFM1. Through GWAS analysis, even after I included covariates, I didn't find significant causal polymorphisms for these two phenotypes, as suggested by QQ plot and Manhattan Plot. Most of the observed values correspond to the expected values in QQ plot (a line close to diagnoal) with no deviation at tail part. Manhattan plot does not show an obvious peak out of the boundary. This does not mean causal polymorphisms about ERAP/GFM1 does not exist. Maybe our GWAS is under power to detect the existence or we did not include powerful covariates variable to differentiate significant genotypes regarding this GFM1. (See figure 13-16)

4 Conclusion

In this GWAS analysis project, I tried to find possible location of causal polymorphism for different phenotypes. I also tried to group in different covariates into my analysis. I tried population, population*sex, sex, however, these covariates don't seem to effect result much. Therefore, a simple model might be robust enough to detect relative position of causal polymorphisms. Since GWAS only aims to find possible locations of causal polymorphisms, we could not make concrete conclusion on exact position. However, this analysis shows that all detected significant genotypes lie on expected chromosome together with phenotypes. This model may not be powerful enough to be applied to phenotype 4 or 5 since no significant p-values were found. Later we may need to consider more covaraites to include in the analysis steps that hopefully can help us differentiate causal genotypes from others.

Appendices

A Graph on GWAS phenotype 2



(a) phenotype2 without covaraite

(b) phenotype2 with covariates

Figure 7: Manhattan plot on phenotype2

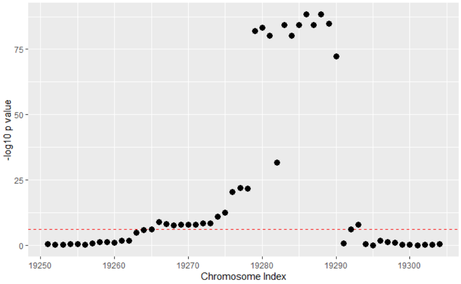
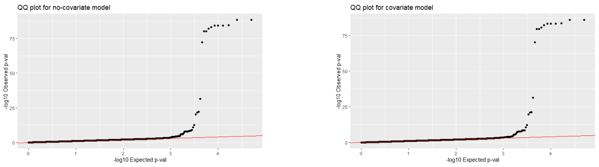


Figure 8: Local Manhattan on phenotype2

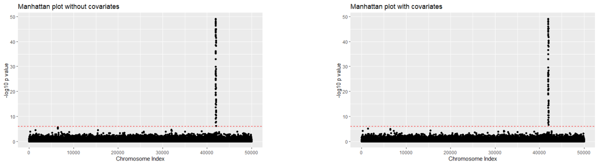


(a) phenotype2 without covaraite

(b) phenotype2 with covariates

Figure 9: 2 QQ-plot phenotype2

B Graph on GWAS phenotype 3



(a) phenotype3 without covaraite

(b) phenotype3 with covariates

Figure 10: Manhattan plot on phenotype3

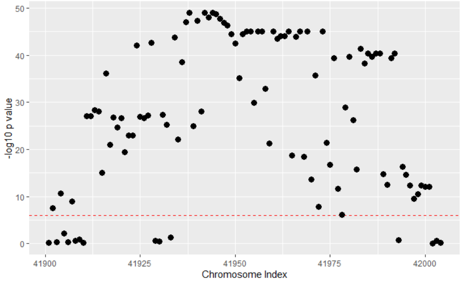
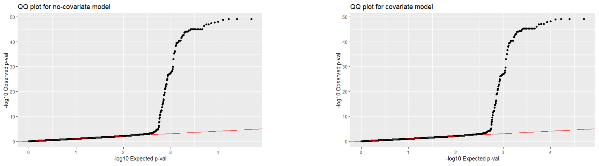


Figure 11: Local Manhattan on phenotype3

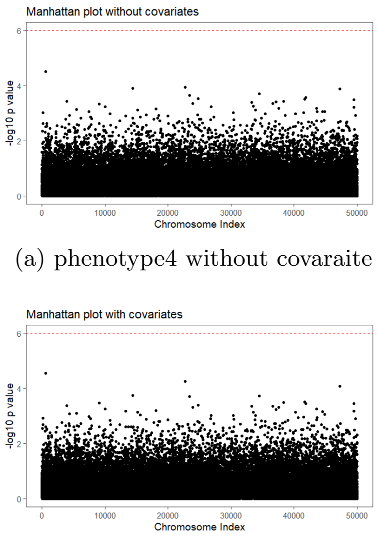


(a) phenotype3 without covaraite

(b) phenotype3 with covariates

Figure 12: 2 QQ-plot phenotype3

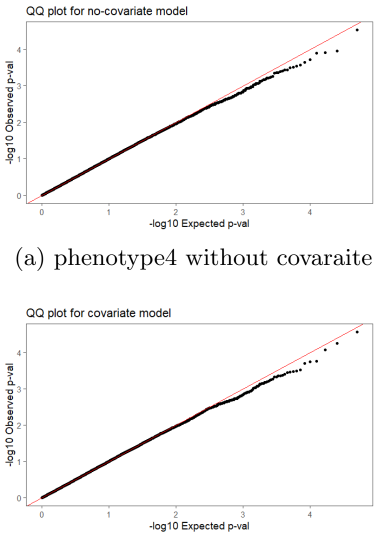
C Graph on GWAS phenotype 4



(a) phenotype4 without covaraite

(b) phenotype4 with covariates

Figure 13: Manhattan plot on phenotype4

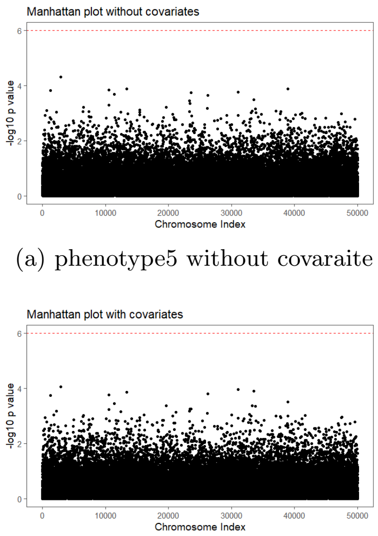


(a) phenotype4 without covaraite

(b) phenotype4 with covariates

Figure 14: 2 QQ-plot phenotype4

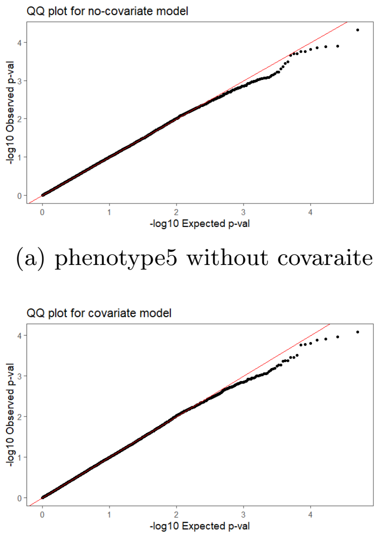
D Graph on GWAS phenotype 5



(a) phenotype5 without covaraite

(b) phenotype5 with covariates

Figure 15: Manhattan plot on phenotype5



(a) phenotype5 without covaraite

(b) phenotype5 with covariates

Figure 16: 2 QQ-plot phenotype5