Quantative Genomics Project Report

Jennie Li (Weill-xil4009)

2023-05-09

Overview

Introduction

This study analyzed a subset of data from the large scale human genomics resources Genetic European Variation in Health and Disease (gEUVADIS) (Altshuler et al., 2015). The 344 samples were collected from subjects from 4 different European populations. Each of these individuals were sequenced and analyzed to identify SNP genotypes. For expression profiling, lympoblastoid cell lines (LCL) were generated from each sample and mRNA levels were quantified through RNA sequencing (Lappalainen et al., 2013). Each of these gene expression measurements is considered as a phenotype, which is called an "expression Quantitative Trait Locus" or "eQTL" analysis.

The data `genotypes.csv` contains 50,000 of the SNP genotypes for 344 samples from the CEU (Utah residents with European ancestry), FIN (Finns), GBR (British) and, TSI (Toscani) population. The expression levels of five genes for each individual is also provided in the `phenotypes.csv`. Information on the population and gender of each of these individuals is in `covars.csv`, and information regarding the position of each gene and SNP in the genome are in `gene_info.csv` and `SNP_info.csv`.

All the codes used in generation of the plots and results in this report can be found in file `qg_project_codes.Rmd`.

Results

Information about the phenotypes of interest in this study are listed below:

probe	chromosome	start	end	symbol
ENSG00000136536.9	2	159712456	159768582	MARCH7
ENSG00000180185.7	16	1827223	1840206	FAHD1
ENSG00000124587.9	6	42963872	42979242	PEX6
ENSG00000164308.12	5	96875939	96919702	ERAP2
ENSG00000168827.9	3	158644496	158692571	GFM1

Figure 1-5 shows the Manhattan plot and QQ plot of Genome-Wide Association Study (GWAS) for each of the five phenotypes. The Manhattan Plot displays the -log10 transformed p-values for each genetic variant plotted against its indexed position. Significance thresholds are indicated by the red horizontal line. Each dot represents a single nucleotide polymorphism (SNP). The SNPs are colored by chromosome in alternating colors. Significant loci were identified based on a threshold of bonferroni-corrected p-value. The QQ Plot compares the observed p-values to the expected p-values under the null hypothesis of no association. Both plots provide insights into the SNP associations with a genetic phenotype. The expression levels of PEX6, FAHD1, and ERAP2 were found to be correlated with some loci in the genome after accounting for the gender and demographics. In contrast, GFM1 and MARCH7 expression levels did not show significant correlation with any genetic loci that were tested in this study. The higher resolution Manhattan plots displaying only significant loci are shown in Figure 6.

A causal polymorphism is when variations in a particular DNA sequence would produce an effect on the phenotype under specific conditions.The peaks in a Manhattan plot may indicate the genomic position of a causal polymorphism because a causal variant that associates with a phenotype tends to be in linkage disequilibrium with nearby variants, and these variants can be used as proxies for the causal variant in GWAS. But the peak may not necessarily represent the actual causal polymorphism because non-causal sites linked to causal sites will carry a similar association with the phenotype the causal variant impacts.

PEX6 (peroxisomal biogenesis factor 6) gene is located on chromosome 6 from position 42963872 to 42979242. This gene encodes a type of ATPase known as a member of the AAA family (National Library of Medicine [NIH], 2023c). The protein is present in the cytoplasm and has a direct function in importing proteins into peroxisomes. Mutations in this gene can cause disorders related to the formation of peroxisomes. The GWAS results of PEX6 is shown in figure 1. A wide peak at chromosome 6 spans position 42889467 to 43108015. Since we consider a set of markers with highly significant p-values to span the region where a causal polymorphism is located, the causal SNPs are likely to be within the region. Therefore, the true causal SNPs which correlate with PEX6 expression level are likely to fall within the gene body or regulatory regions of PEX6.

Interestingly, one SNP on chromosome 4 at position 98486048 which is also correlated with expression of PEX6. This might be a artifact or a false positive. If this site is a true positive, then the two regions might be in linkage disequilibrium (LD) with one another. LD refers to nonrandom association of alleles at different loci, i.e., two sites are likely to be inherited together (Slatkin, 2008). Even though polymorphisms on different chromosomes tend to be in equilibrium and they are not physically linked, if they are both required to show a given phenotype or under selection, they might be in LD (Koch et al., 2013). However, without further information, it is hard to draw a solid conclusion.

The FAHD1 (fumarylacetoacetate hydrolase domain-containing protein 1) gene spans from position 1827223 to 1840206 on chromosome 16. The protein coded by FAHD1 gene is related to acetylpyruvate hydrolase activity, fumarylpyruvate hydrolase activity, and oxaloacetate decarboxylase activity (National Library of Medicine [NIH], 2023b). It is primarily involved in the citric acid cycle (TCA cycle). The GWAS analysis revealed that a peak on chromosome 16 correlates with FAHD1 expression (Figure 2). This peak spans from position 1524250 to 1929366, covering the gene body and potential regulatory regions of FAHD1. Thus, it is reasonable to propose that some causal polymorphisms are within or near this region.

ERAP2 (endoplasmic reticulum aminopeptidase 2) gene locates on chromosome 5 from position 96875939 to 96919702. This particular gene produces a type of enzyme that belongs to the M1 protease family. It is located in the endoplasmic reticulum and involves in trimming the N-terminus of antigenic epitopes, allowing these epitopes to be presented by MHC class I molecules (National Library of Medicine [NIH], 2023a). Some mutations in this gene have been linked with ankylosing spondylitis and pre-eclampsia. The GWAS result showed that a set of loci on chromosome 5 strongly correlates with ERAP2 expression, pointing to possible locations of causal SPNs that affect ERAP2 gene expression level (Figure 3). This peak spans from position 96774230 to 97110808. Thus, the causal SNPs are likely to be in the gene body or regulatory regions of this gene.

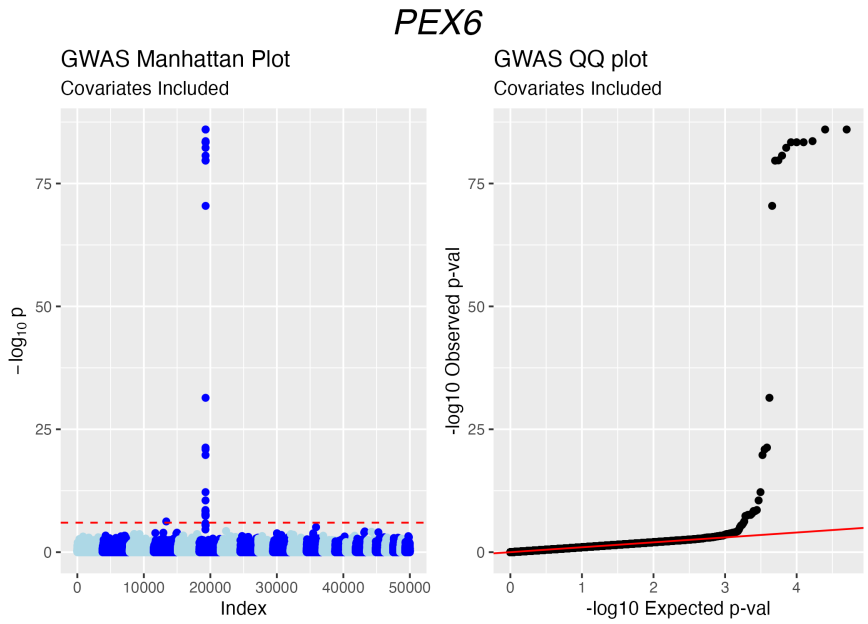

Figure 1: he Manhattan Plot and QQ Plot of the PEX6 GWAS with sex and population covariates included. The peaks are found on chromosome 4 and 6.

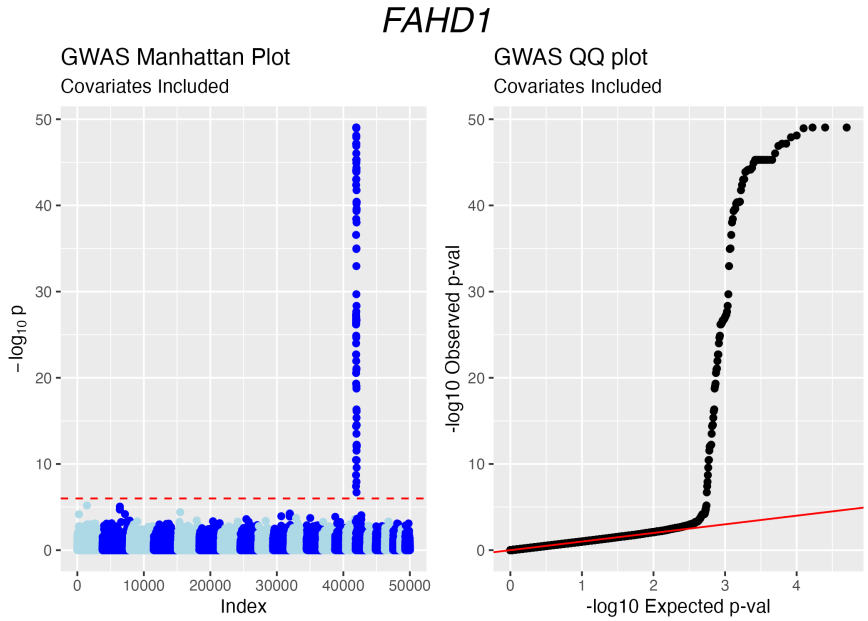

Figure 2: The Manhattan Plot and QQ Plot of the FAHD1 GWAS with sex and population covariates included. The peak is located on chromosome 16.

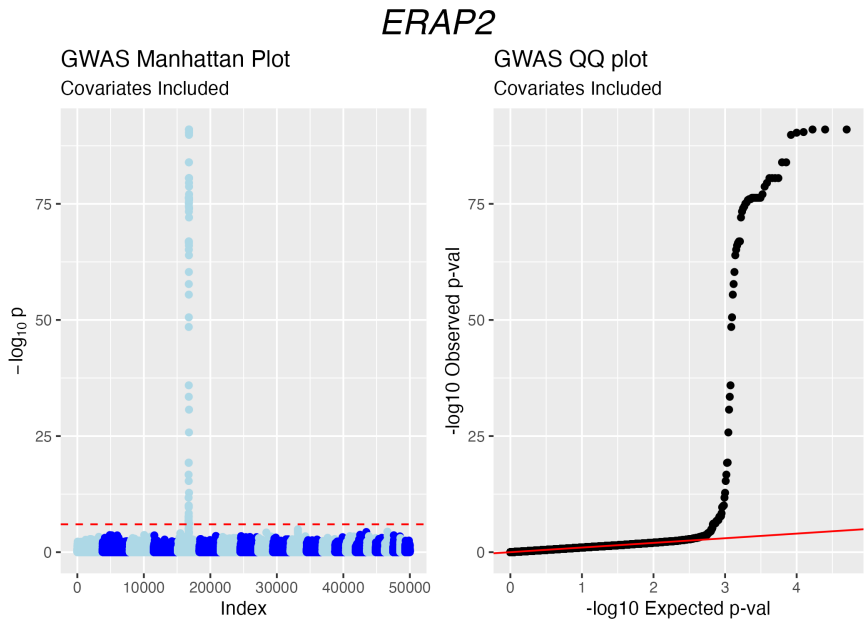

Figure 3: The Manhattan Plot and QQ Plot of the ERAP2 GWAS with sex and population covariates included. The peak is located on chromosome 5.

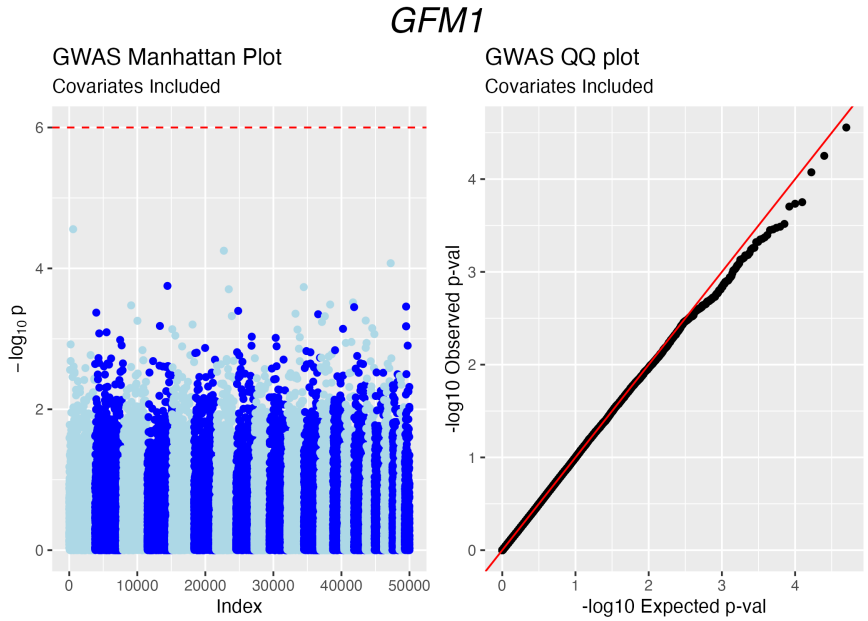

Figure 4: The Manhattan Plot and QQ Plot of the GFM1 GWAS with sex and population covariates included. No significant SNP is identified.

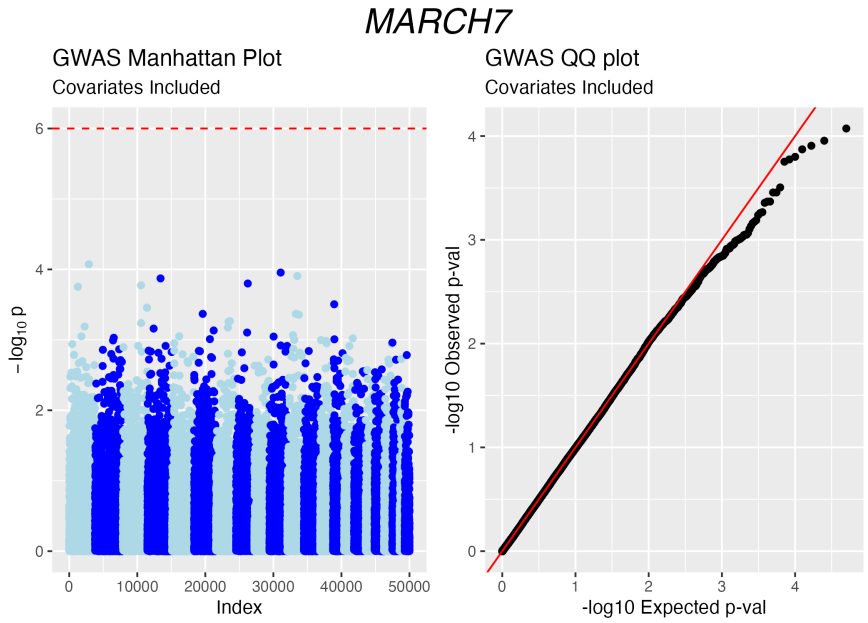

Figure 5: The Manhattan Plot and QQ Plot of the MARCH7 GWAS with sex and population covariates included. No significant SNP is identified.

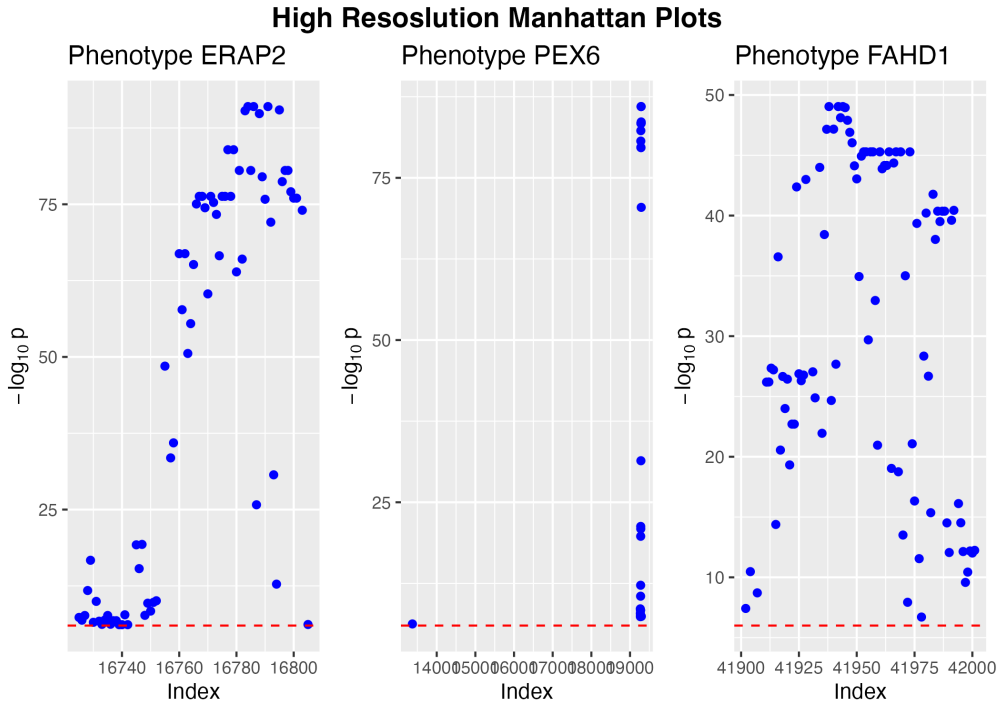

Figure 6: High resoslution Manhattan plots display GWAS results for expression level of ERAP2, PEX6, and FAHD1 genes. A close-up view highlighting the loci that reached genome-wide significance.

Discussion

This study reveals some genetic loci that are correlated with the expression of five genes of interest. Specifically, PEX6, FAHD1, and ERAP2 were found to have a correlation with some loci in the genome, while GFM1 and MARCH7 did not show any significant correlation. The Manhattan plots and QQ plots were used to illustrate these results. The results suggest that the true causal SNPs which correlate with PEX6, FAHD1, and ERAP2 expression levels are likely to fall within the gene body or regulatory regions of these genes. However, while the peaks in a Manhattan plot may indicate the genomic position of a causal polymorphism, it may not necessarily represent the actual causal polymorphism.

In conclusion, this study provides insights into the relationship between genetic loci and gene expression levels for PEX6, FAHD1, and ERAP2 genes. These findings could lead to further research into the underlying mechanisms of these genes and their potential implications for disease. The limitations of this study should also be acknowledged, including the small sample size and limited number of genes analyzed. Future research with larger sample sizes and more comprehensive gene expression analysis could provide a more complete understanding of the genetic basis of these phenotypes.

Methods

Basic reformatting was performed after loading the data tables. For clear interpretation, gene IDs were replaced by gene names. The categorical covariates were coded by one-hot encoding. The Xa and Xd matrices are generated accroding to the quantitative genetic model.

The quantitative genetic model is a multiple regression model with the independent variables:

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

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

and the multiple regression model:

$$y = \beta_\mu + X_a\beta_a + X_d\beta_d + \epsilon = X\beta + \epsilon$$

The parameter that we want to estimate is β, where $\beta = [\beta_\mu, \beta_a, \beta_d]$. We define a statistic called MLE that takes the sample (y, x_a, x_d) and outputs an estimate $(\hat{\beta})$. In linear regression case, we construct MLE where $MLE(\hat{\beta}) = (x^Tx)^{-1}x^Ty = [\hat{\beta}_\mu, \hat{\beta}_a, \hat{\beta}_d]$.

Use sample information and the MLE equation to get the estimates of Beta parameters:

```
y <- pheno_input
x <- cbind(rep(1,length(xa_input)), xa_input, xd_input)
# The three columns of `X` are $\beta_\mu$, $X_a$, and $X_d$.

MLE_beta <- ginv(t(x) %*% x) %*% t(x) %*% y
```

The hypothesis is set up as $T(y, x_a, x_d|H_0 : \beta_a = 0 \cap \beta_d = 0)$.

To perform hypothesis testing, we need to use a likelihood ratio test (LRT). In the case of linear regression, we use F-statistic as the LRT to obtain a p-value.

1. Use MLE to obtain a prediction of the phenotype: $\hat{y} = xMLE(\hat{\beta})$.
2. Calculate "model sum of squares" and "sum of squares due to error": $SSM = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$, $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.
3. Calculate df(M) and df(E):

- df(M) = number of betas under H1 - number of betas under H0
- df(E) = sample size - number of betas under H1

4. Calculate the "mean square of means" with degrees of freedom $df(M)$ and "mean square of error" with $df(E)$: $MSM = \frac{SSM}{df(M)}$, $MSE = \frac{SSE}{df(E)}$
5. Calculate F-statistic with degrees of freedom [2, n-3]: $F_{[2,n-3]}(y, x_a, x_d) = \frac{MSM}{MSE}$
6. Calculate p-value of the genetic model.

The six steps in codes:

```
y_hat <- x %*% MLE_beta

SSM <- sum((y_hat - mean(y))^2)
SSE <- sum((y - y_hat)^2)

df_M <- n_betas_h1 - n_betas_h0
df_E <- n_samples - n_betas_h1

MSM <- SSM / df_M
MSE <- SSE / df_E

Fstatistic <- MSM / MSE

pval <- pf(Fstatistic, df_M, df_E,lower.tail = FALSE)
```

In a GWAS, we perform the above process for different loci across the genome, and obtain a p-value for each locus. The Manhattan plot depicts the corrected p-value for each locus.

To include the covariates, we add additional columns to the beta matrix and use different degrees of freedom given the number of parameters: $Y = \beta_\mu + X_a\beta_a + X_d\beta_d + \Sigma_{i=i}^{n} X_{z,i}\beta_{z,i} + \epsilon$.

Reference

Altshuler, D., Albers, C., Abecasis, G., & Et al. (2015). A global reference for human genetic variation. Nature (London), 526(7571), 68-74.

Koch, E., Ristroph, M., & Kirkpatrick, M. (2013). Long range linkage disequilibrium across the human genome. PloS one, 8(12), e80754. https://doi.org/10.1371/journal.pone.0080754

Lappalainen, T., Sammeth, M., Friedländer, M., 't Hoen, P., Monlong, J., Rivas, M., . . . Dermitzakis, E. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature (London), 501(7468), 506-511.

National Library of Medicine (US). (2023a). ERAP2 endoplasmic reticulum aminopeptidase 2 [Homo sapiens (human)]. https://www.ncbi.nlm.nih.gov/gene/64167

National Library of Medicine (US). (2023b). FAHD1 fumarylacetoacetate hydrolase domain containing 1 [Homo sapiens (human)]. https://www.ncbi.nlm.nih.gov/gene/81889#summary

National Library of Medicine (US). (2023c). PEX6 peroxisomal biogenesis factor 6 [Homo sapiens (human)]. https://www.ncbi.nlm.nih.gov/gene/5190

Slatkin M. (2008). Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. Nature reviews. Genetics, 9(6), 477–485. https://doi.org/10.1038/nrg2361