GWAS Final Report

Jake Sauter

5/14/2021

Background

The data associated with the following analyses represents the mRNA expression levels of five different transcripts of interest (ERAP2, PEX6, FAHD1, GFM1 and MARCHF7), measured from four European or European-related populations. mRNA expression profiles were quantified through Next-Generation Sequencing (NGS) of reverse-transcribed cDNA of RNA samples dervied from Lymphoblastic Cell Lines (LCLs) generated from each individual in the study, originating from the 1000 Genomes Project.

The phenotype data for this study was generated by the Genetic European Variation in Health and Disease (gEUVADIS). A published Nature Article (Transcriptome and genome sequencing uncovers functional variation in humans) resulted from these data, and the data can be accessed via an International Genome Sample Resource Data Portal. The individuals included in the following analyses are apart of four seperate populations, **CEU** (Utah (U.S.A) residents with European ancestry), **FIN** (Finns), **GBR** (British) and **TSI** (Toscani).

The goal of the following analysis is to reproducilby locate positions of causal single nucleotide polymorphisms (SNPs) for the five LCL-mRNA related phenotypes aquired in the reference study. In order to accomplish this goal, a standard Genome Wide Association Study (GWAS) will be performed, in which independent hypothesis tests for association will be performed for each genotype-phenotype pair. Our aim is to identify genomic locations in which we calculate a significant association between a difference in genotype and observed phenotype, indicated by a significant ($\alpha < 0.05$) multiple testing corrected p-value for these locations. Due to the nature of the complex biological processes occuring with the information stored in genomic DNA, a significant p-value (for other than statitical reasons) may not indicate a true causual polymorphism is present at the exact location, more specifically due to linkage disequilibrium, or less formally, observed correlation between closely co-located genomic positions.

To begin these analyses, an inspection of the reference study data will be performed, aiming to pinpoint missing, outlier, or additional trends in the data that may break necessary assumptions in the analyses to follow. After data inpsection and validation, we will inspect the distribution of phenotype data. The distributions of all measured RT-cDNA phenotypes appeared to be continuous and sufficiently normally distributed, allowing for analysis using linear regression. Lastly, we present the results of the five GWAS analyses by means of QQ plots (determining if the resultant p-value distribution is fairly uniform as expected) and both global and localized Manhattan Plots. The localalized Manhattan plots are then overlayed to determine possible concordance or disconcordance of significant genomic locations between the five genotypes analyzed in this study.

Model Setup

In order to analyze the given genetic data, we will make use of the qunantiative genetic model for a muliple regression. Both an additive ($X_a(A_1A_1) = -1, X_a(A_1A_2) = 0, X_a(A_2A_2) = 1$ and dominant ($X_d(A_1A_1) = -1, X_d(A_1A_2) = 1, X_d(A_2A_2) = -1$ genetic coding will be utitlized in our multiple regession model equation ($Y = \beta_\mu + X_a\beta_a + X_d\beta_d + \epsilon$) where $\epsilon\ N(0, \sigma^2)$. This set of model and encodings allow us to capture the relationship between the genotype information ($X_a,\ X_d$) and the continous phenotype (Y), while preserving our genetic understanding of additive and dominant allele effects.

In the following analyses, we will use this model to perform a hypothesis test with a null hypothesis $H_0 : \beta_a = 0 \cap \beta_d = 0$ (indicating that additive or dominant genetic effects were not found to be assocaited with the phenotypic difference) vs our alternative hypothesis $H_a : \beta_a \neq 0 \cup \beta_d \neq 0$ (indicating that either one or both of addiditive genetic effects are associated with the observed phenotypic difference).

Understanding the Data

Phenotype Data

The phenotype data for this study has been provided in a relatively common and expected format, with one row per individual meansured, and one column for each of the measured phenotypes. A sample selection of the

phenotype data for six individuals is shown below.

	ERAP2	PEX6	FAHD1	GFM1	MARCH7
HG00096	-1.166337	-0.6859358	-0.3062421	0.9275844	0.3600088
HG00097	-1.045803	-0.1789576	-1.0333305	-0.6676644	-1.7782530
HG00099	1.045803	-0.8945747	-2.5242603	-2.7590424	-2.5242603
HG00100	0.223446	-0.0982433	1.1521112	0.2533471	0.0545189
HG00101	-0.223446	0.8625121	1.3782830	1.0842344	0.2608569
HG00102	-1.457685	-0.3291676	-0.1863454	-1.1663369	-2.3783289

Genotype Data

Below we see that we have **50,000** genotype markers for each of our **344** individuals. Each individual is represented in a **row** in our `genotype` data frame, with each **column** corresponding the the number of minor alleles the individual has present in their DNA.

	rs10399749	rs62641299	rs115523412	rs75932129	rs10900604
HG00096	0	2	0	1	1
HG00097	0	2	0	2	0
HG00099	1	1	0	0	1

We can now convert this encoding to our standard X_a and X_d encodings (additive and dominant genetic codings) in the following way.

```
Xa <- as.data.frame(genotypes) - 1
Xd <- 1 - 2*abs(Xa)
```

Covariate Information

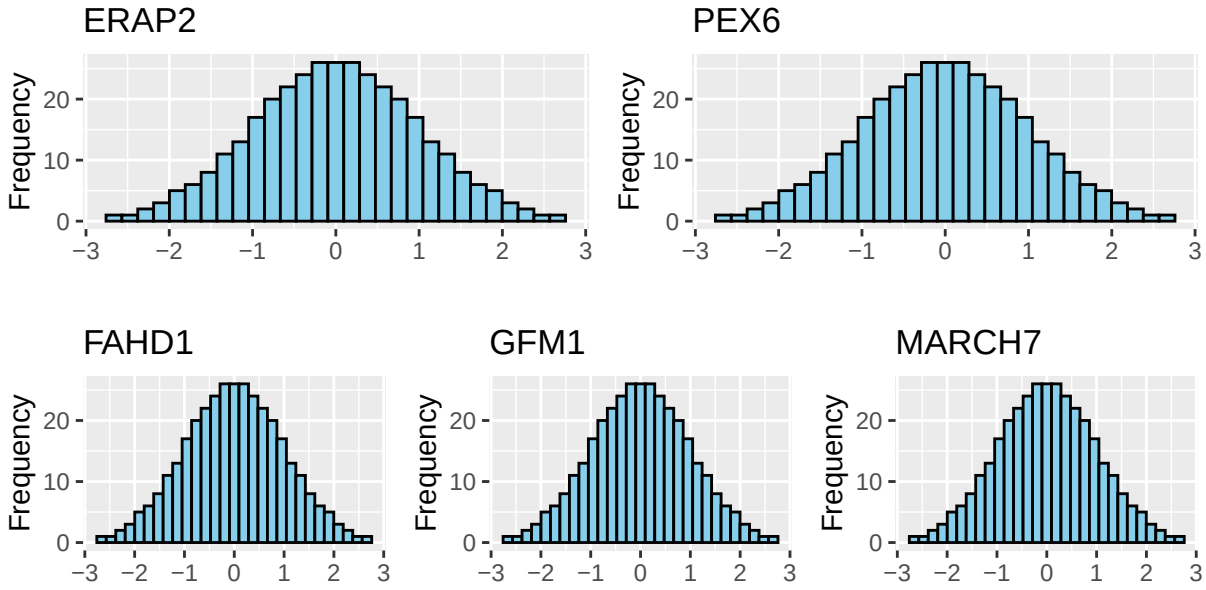
Covariates of **Population** and **Sex** were measured in this study. In this report we will model the genotype-phenotype relationship both with and without including these covariates.

An important aspect of including covariates while modelling is ensuring a stable and informative model. Practially, this required a sufficient sample size per covariate-defined population. Below we see that each covariate controlled population has over **30** samples, indicating that we should have enough data/information here in order to control for both covariates with introducing non-robust model characteristics.

Population	Sex	Num. Observed
CEU	FEMALE	37
CEU	MALE	41
FIN	FEMALE	55
FIN	MALE	34
GBR	FEMALE	46
GBR	MALE	39
TSI	FEMALE	43
TSI	MALE	49

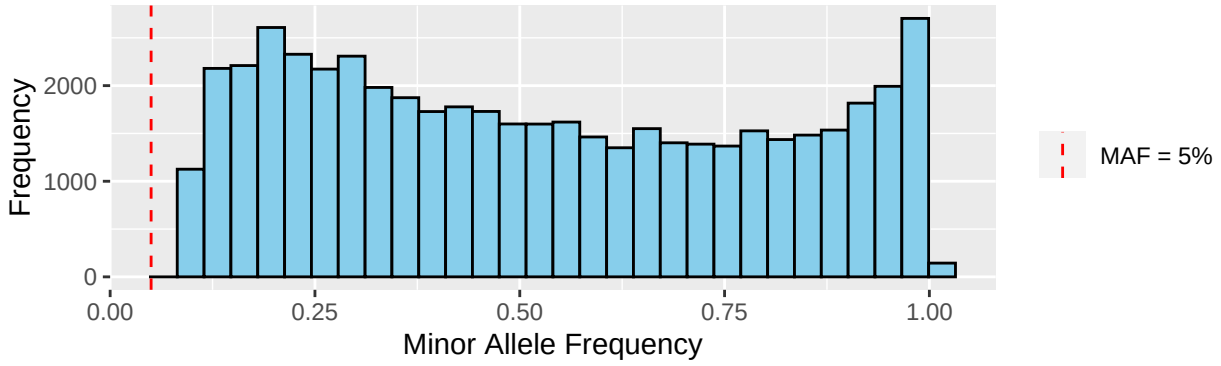
Phenotype Data Quality Control

From the information above, it is clear that we will be working with **5** phenotypes from **344** individuals, each phenotype describing the continuous variable of mRNA expression of the genes of interest. No missing values were identified in the phenotype data, and all phenotypes appear to be be normally distributed between -3 and 3. This fits our expection of Normality in GWAS.



Genotype Data Quality Control

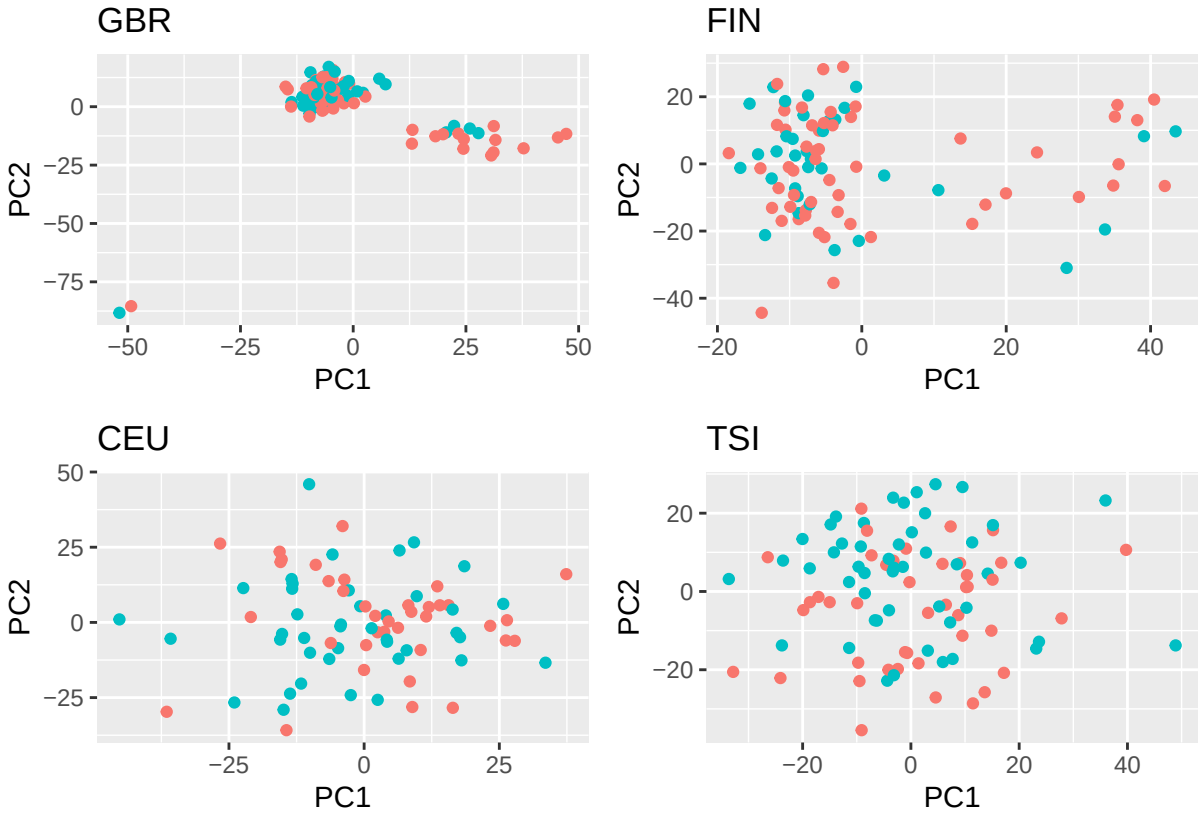
Common techniques used to filter GWAS study data usually include removing **individuals** with more than 10% missing data across all genotypes, removing **genotypes** with more than 5% missing data across all individuals, and removing **genotypes** with an MAF less than 5%. No missing values were identified in the gentoype data, thus the first two points are satisfied. From the histogram below, we can see that no genotypes with an MAF of less than 5% are present in our data.



Principle Components Analysis

In order to understand if any other population structure exists in the data other than the captured population identifiers in the covariates, we now perform Principle Components Analysis (PCA) on the genotype data of each population seperately, and plot the population-seperated data by the two most discriptivive principle components.

Below we see that **CEU** and **TSI** populations are fairly uniform, and slight seperation is seen in both **FIN** and **GBR**. Both of these observed sub-populations do not seem to be extreme or due to the sex of the individual (indicated by the color of the points).

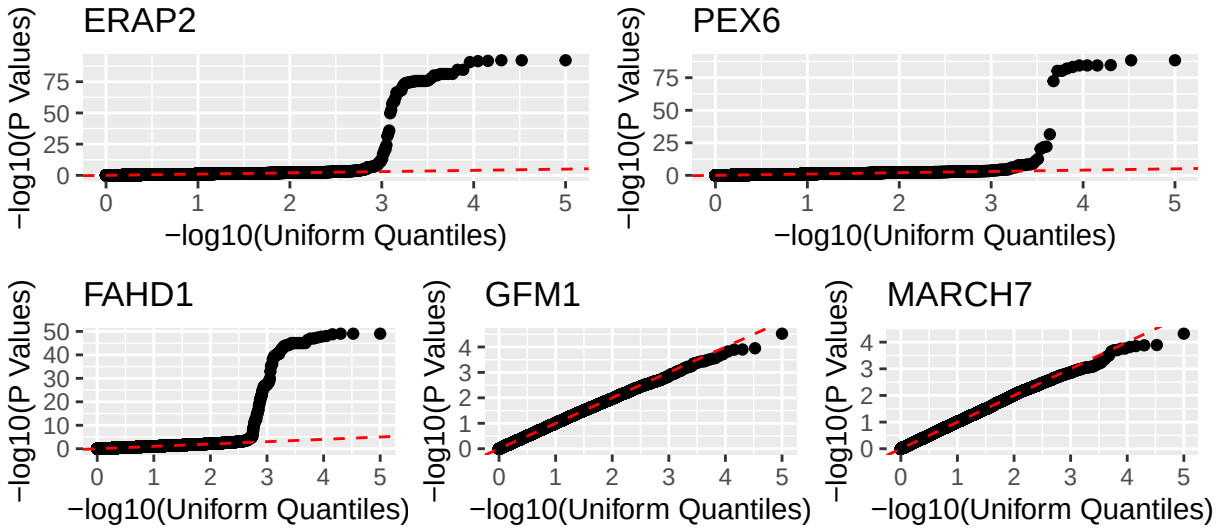


Performing GWAS Without Covariates

We will first perfrom the five GWAS analyses **without** any included covariates, thus the linear models we will be fitting are of the form $Y = \beta_\mu + X_a\beta_a + X_d\beta_d + \epsilon$ where $\epsilon\ N(0, \sigma^2)$. The models were fit using the `lm()` function in the R programming language, followed by calculation of the p-value through the models F-statistic on a distribution with the appropriate degrees of freedom.

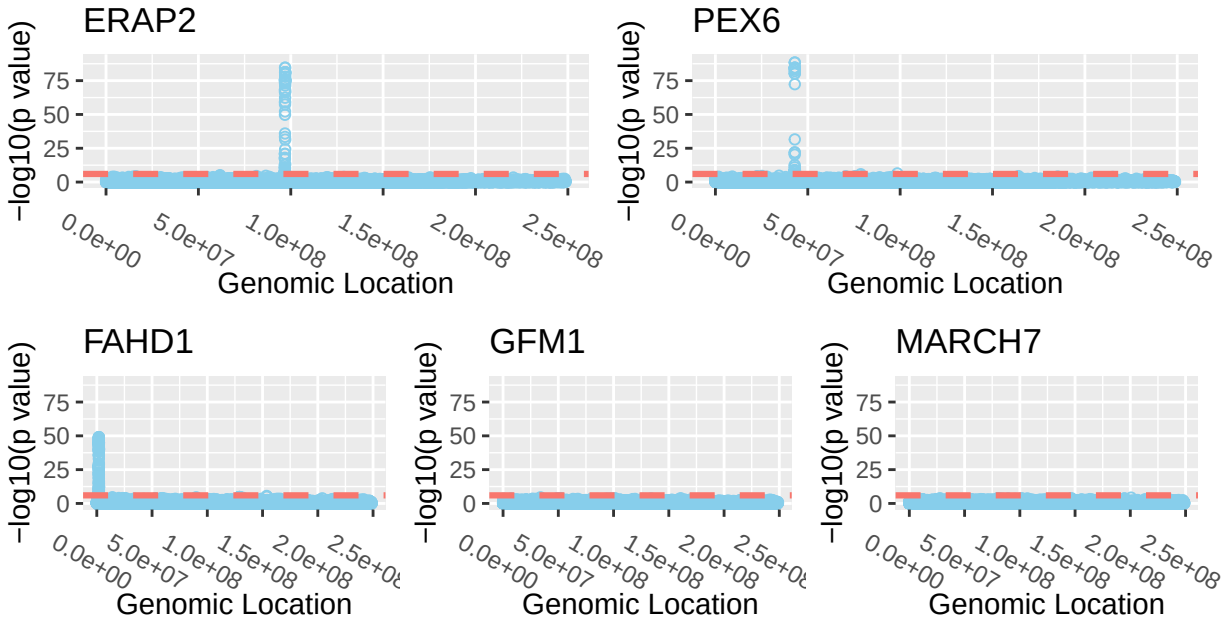
QQ Plots

In order to observe if the resultant ditribution of p-values for each GWAS is approximately normal, as we would expect under H_0, we produce a Quantile-Quantile plot comparing the quantiles of the p-value distribution with that of a uniform distribution.

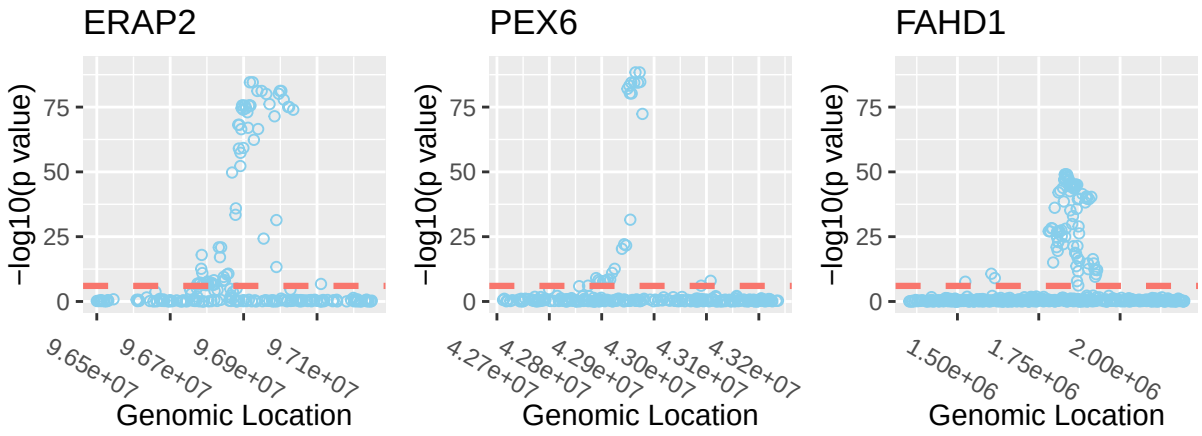


Manhattan Plots

To illistate the location and significance of the detected hits in the analyses perform, we plot the $-log_{10}(pvalue)$ of each genotype at the position where the genotype is found in the human genome.



Of the three phenotypes that we have observed significant hits for, we will take a closer look at the local regions where the hits were found to be.

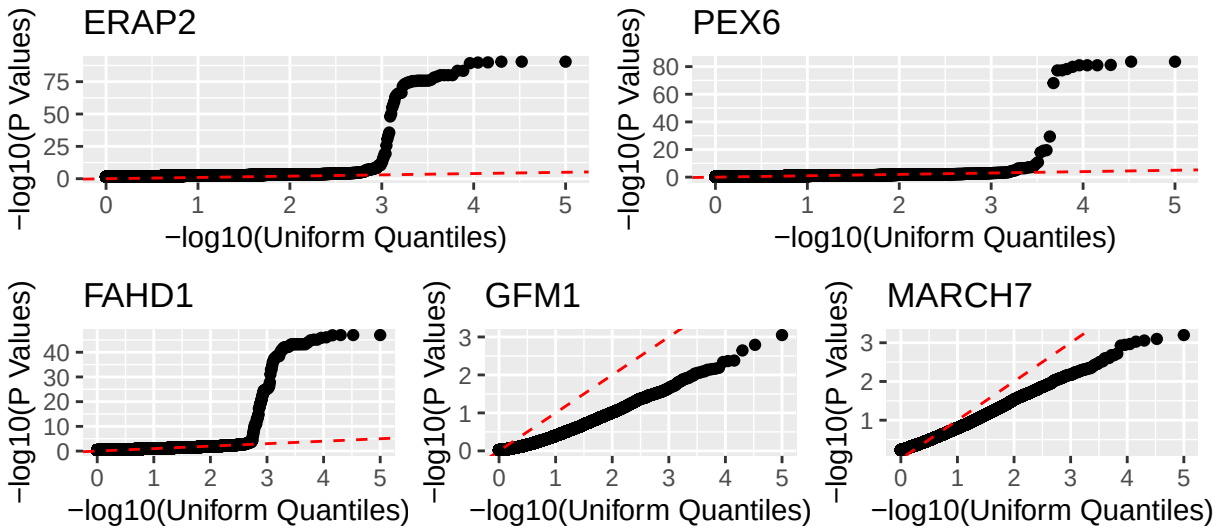


Performing GWAS With Covariates

We now perfrom the same five GWAS analyses **with** the inclusion of the collected **Population** and **Sex** covariates, thus the linear models we will be fitting are of the form $Y = \beta_\mu + X_a\beta_a + X_d\beta_d + X_{Sex}\beta_{Sex} + X_{Population}\beta_{Population} + \epsilon$ where $\epsilon\ N(0, \sigma^2)$. The models were fit similarly using the `lm()` function in the R programming language, followed by calculation of the p-value through the models F-statistic on a distribution with the appropriate degrees of freedom.

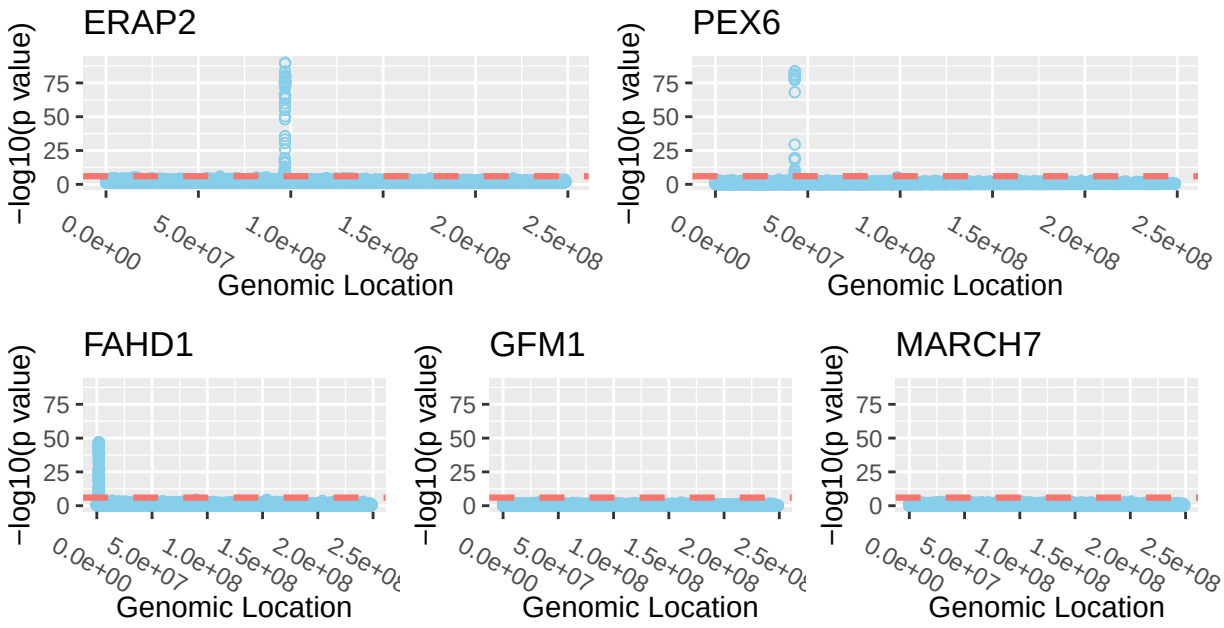
QQ Plots

In order to observe if the resultant ditribution of p-values for each GWAS is approximately normal, as we would expect under H_0, we produce a Quantile-Quantile plot comparing the quantiles of the p-value distribution with that of a uniform distribution.



Manhattain Plots

To illistate the location and significance of the detected hits in the analyses perform, we plot the $-log_{10}(pvalue)$ of each genotype at the position where the genotype is found in the human genome.



Of the three phenotypes that we have observed significant hits for, we will take a closer look at the local regions where the hits were found to be. We see below that we have found the same hits with and without covariates included in our model.

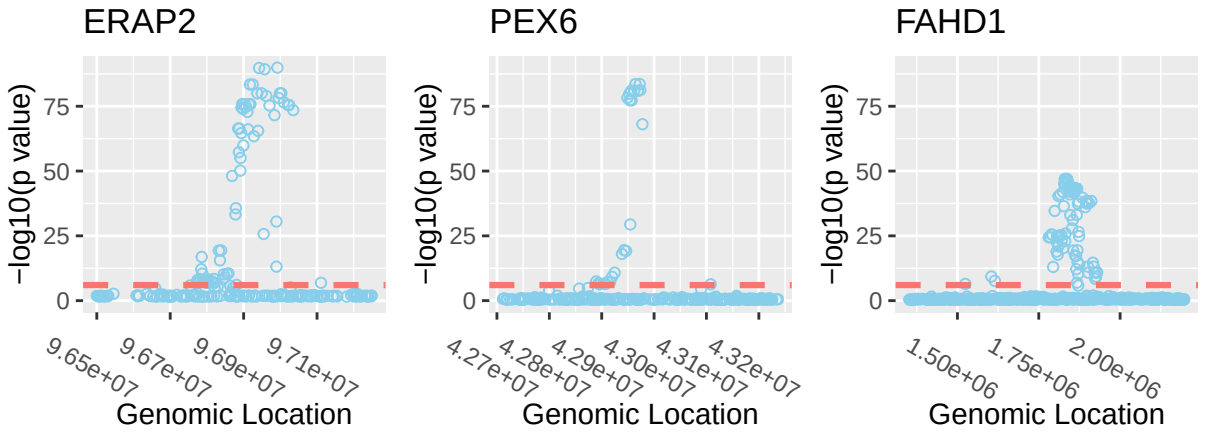


Table of Most Significant Hits

	Phenotype	Num.Sig..without.covs	Num.Sig..with.covs	Top.3.Sig..without.covs	Top.3.Sig..with
ERAP2	ERAP2	73	75	rs7726445, rs7731592, rs2161548	rs7726445, rs77
PEX6	PEX6	29	26	rs1129187, rs10948061, rs9471985	rs1129187, rs10
FAHD1	FAHD1	90	89	rs11644748, rs140254902, rs9652776	rs11644748, rs1

Biological Support of Found Hits

TODO: Maybe just look up Genecards for the genes that corespond to the most significant hits?

Using the Manhattan plots above, we can search through published research online for SNPs with known biologicalfunctions that are located where our GWAS identified a locus for a causal polymorphism. Since we have onlyreceived a sample of the SNPs in the genome, we may have found the location of a specific genotype that isnot

included in our small sample. Therefore, we search for known causal genotypes determined by eQTLexperiments that are located in the regions where there are many hits in the ERAP2, PEX6, and FAHD1genes.7

For ERAP2, the eSNP rs2927608 has been previously discovered as a causal genotype [**1**]. This SNP is locatedat position 96916728, which is in the region where many of our significant genotypes lie for ERAP2, and isassociated with inflammatory bowel disease through an increased gene expression of ERAP2. The genotypeof this locus alone accounted for 74.5% of the total variability of ERAP2 expression in blood in this study.

For PEX6, the eSNP rs10948059 is located at position 42960723 in the genome. This corresponds to thelocus where our significant genotypes were found using our GWAS analysis. This SNP has been associatedwith prostate cancer from previous association studies [**2**] . The results of the literature imply that variantsassociated with prostate cancer can be identified through expressional change in the PEX6 gene.

For FAHD1, the eSNP rs1065656 is located at position 1788835 in the genome. We have found significantgenotypes at this location in our GWAS analysis. Literature has demonstrated that rs1065656 is associatedwith circulating IGF-I and IGFBP-3 concentration, which have been implicated in risk of human diseasesincluding cardiovascular diseases, diabetes, and cancer [**3**].

References

Di Narzo, A., Peters, L., Argmann, C., Stojmirovic, A., Perrigoue, J., Li, K., . . . Hao, K. (2016, June23). Blood and Intestine Eqtls from AN Anti-TNF-Resistant Crohn's DISEASE Cohort inform IBDgenetic association loci. Retrieved May 04, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4931595/#sup12.

[**2**] Lee, J., Ryu, J., & Lee, C. (2016, December). Strong cis-acting expression quantitative trait loci for thegenes encoding snhg5 and pex6. Retrieved May 04, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5207599/

[3] Teumer, A., Qi, Q., Nethander, M., Aschard, H., Bandinelli, S., Beekman, M., . . . Kaplan, R. (2016,October). Genomewide meta-analysis IDENTIFIES loci associated WITH IGF-I and IGFBP-3 levelswith impact on age-related traits. Retrieved May 04, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5013013