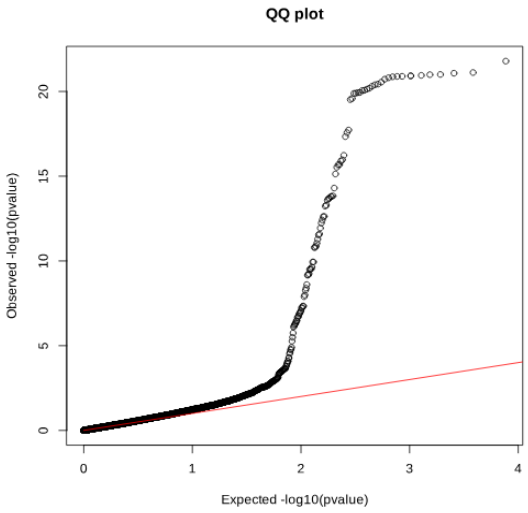
Peyton Smith

GWAS of Hypothetical Trait

In order to determine the single causal variant of a single gene expression hypothetical trait, I will use various methodologies and examine the genomic data of 671 individuals who all self-reported having European ancestry. First, I began my genome-wide association study (GWAS) by conducting data quality control, which I began by generating various files using PLINK: plink.imiss, plink.lmiss, plink.hwe, plink.log, and plink.frq. I then examined these files for any SNPs or individuals that may be outliers and heavily skewing our data, though I did not observe any values upon inspection. Upon inspection, the sample is composed of 430 males and 241 females. Next, I generated a list of SNPs with a missing rate greater than 5%, with a Hardy-Weinberg equilibrium (HWE) p-value less than 10E-6 (indicating a deviation from HWE), and with a minor allele frequency less than 2%. There were 124 SNPs, 7 SNPs, and 9 SNPs that met each of those characteristics respectively, and these are the SNPs I want to remove. I remove the 137 unique SNPs satisfying at least one of these characteristics because these characteristics are indicative of error (i.e. genotyping error) or make certain the SNPs more susceptible to error and removing them will aid in reduction of false positives.[1] Following this data quality control, I conducted linear regression association analysis with the hypothetical trait and each SNP. As a means of examining the results, I generated a QQ-plot, which is shown in Figure 1.

Figure 1: QQ-Plot from association analysis



The genomic inflation factor, or lambda, calculated from our association analysis was 1.212, indicating the existence of inflation and possible Type I error in our results.[2] Due to this inflation, I will examine the data for the existence of population stratification, which could be a source of confounding. Upon examining the MDS plot in Figure 2, there do not appear to be any obvious sub-clusters, but this does not imply that there are no sub-populations or that population stratification does not exist. To combat the inflation, I use the same methodology to conduct a linear association analysis, but I will account for two covariates in order to determine if adding such covariates reduces the inflation we see in the QQ-plot. After conducting such analysis, I obtained the QQ-plot seen in Figure 3, which has a lambda value of 0.970, which is below the threshold of 1.1

[1] Pongpanich M, Sullivan PF, Tzeng JY. A quality control algorithm for filtering SNPs in genome-wide association studies. Bioinformatics. 2010 Jul 15;26(14):1731-7. doi: 10.1093/bioinformatics/btq272. Epub 2010 May 25. PMID: 20501555; PMCID: PMC2894516.

[2] Williams, C.J., Li, Z., Harvey, N. *et al.* Genome wide association study of response to interval and continuous exercise training: the Predict-HIIT study. *J Biomed Sci* 28, 37 (2021). https://doi.org/10.1186/s12929-021-00733-7.

and is indicative of no inflation in our results. There also does not appear to be any systematic bias in the QQ-plot, as there are no early departures from the red diagonal line.

Figure 2: MDS Plot from association analysis

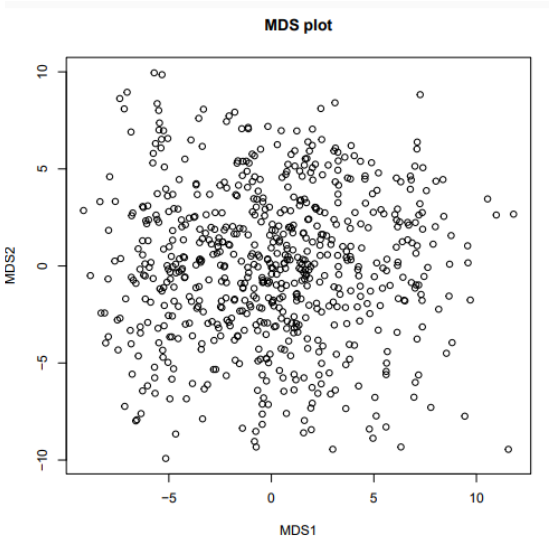


Figure 3: QQ-Plot from association analysis with added covariates

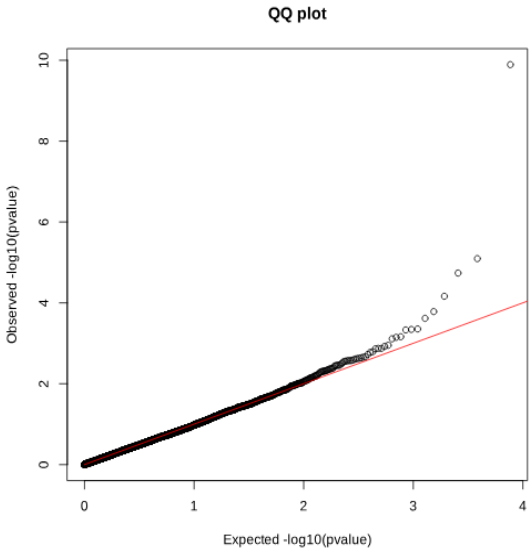
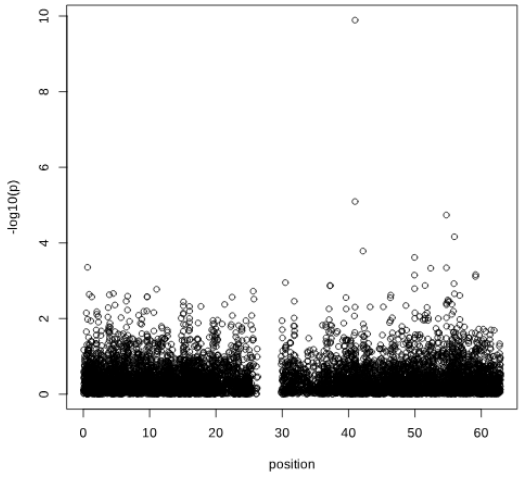


Figure 4: Manhattan Plot from association analysis with added covariates



In examining the Manhattan plot, it appears there is one SNP that is above the conventional p-value threshold of 5E-8, as indicated by the point at the top of the plot in Figure 4.[3] This is likely to be the single causal variant, though further analysis is needed to determine whether or not this is the case. Next, I examined the most significant SNPs, which are SNPs from the association analysis with low p-values, as this indicates the association between the hypothetical trait and this significant SNP is highly likely not due to chance.[4] The two most significant SNPs, in order, were rs6102795 and with p-values 1.285E-10 and 8.060E-6 respectively. It appears that the significant causal variant is located at rs6102795,

[3]Ehret GB. Genome-wide association studies: contribution of genomics to understanding blood pressure and essential hypertension. Curr Hypertens Rep. 2010 Feb;12(1):17-25. doi: 10.1007/s11906-009-0086-6. PMID: 20425154; PMCID: PMC2865585.

[4]Lazzeroni LC, Lu Y, Belitskaya-Lévy I. P-values in genomics: apparent precision masks high uncertainty. Mol Psychiatry. 2014 Dec;19(12):1336-40. doi: 10.1038/mp.2013.184. Epub 2014 Jan 14. PMID: 24419042; PMCID: PMC4255087.

In order to gain more details about the regions with significant SNPs, I carry out *in-silico* fine mapping. I chose a 1Mb window, specifically from 40 Mb to 41 Mb, for pre-phasing and genotype imputation because the most significant SNP is located at base pair 40982342 and the second most significant SNP is located at base pair 40982105. Before pre-phasing, I examined the genome region around the most significant SNP in the UCSC Genome Browser on hg19, which is seen in Figure 5.

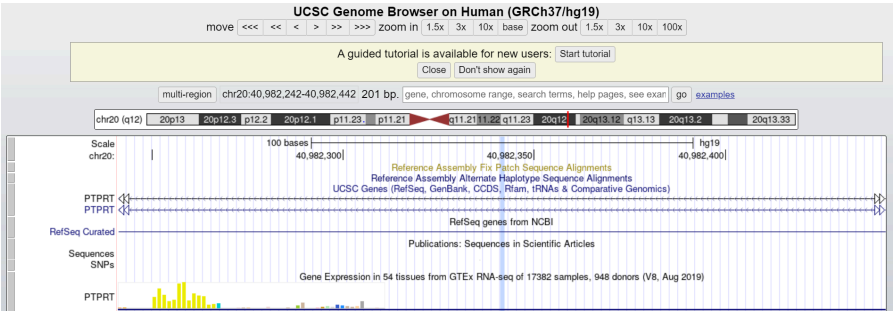


Figure 5: UCSC Genome Browser view for rs6102795

From the genome browser, it appears that the SNP rs6102795 is related to the expression of the PTPRT. Following this examination, I pre-phased using a reference panel, then used Minimac 3 to conduct genotype imputation on the pre-phased dataset. Following the imputation, I tested for association using the SNPs I imputed and including the two covariates previously mentioned using the program SNPTEST. Following this genotype imputation, I obtain the summary statistics about the quality score of the imputation shown in Figure 6.

```
    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00344 0.65947 0.88136 0.79260 0.97155 1.00000
```

Figure 6: R-Squared Imputation Quality Score

A high R-squared value signifies there is a high correlation between the imputed genotypes and the actual genotypes.[5] In this scenario, the mean R-squared value is 0.7926, indicating that the average imputed genotype in this region is highly accurate. The imputed SNPs are shown in blue and the genotyped SNPs are shown in green in Figure 7. It appears that the imputed SNPs are superior in showing the SNP that corresponds to the causal variant for the hypothetical trait, as signified by the tall, extended line of green dots around position 41.0.

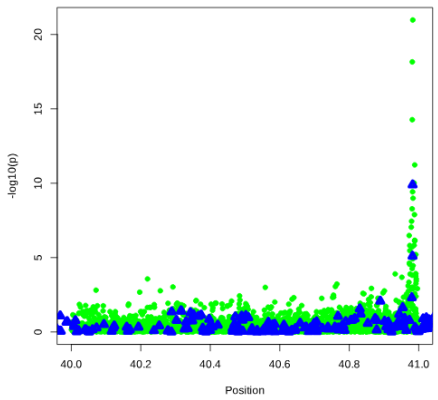


Figure 7: Plot of Genotyped and Imputed SNPs

[5] Sargolzaei, M., Chesnais, J.P. & Schenkel, F.S. A new approach for efficient genotype imputation using information from relatives. *BMC Genomics* 15, 478 (2014). https://doi.org/10.1186/1471-2164-15-478