GWAS Final Report

Jake Sauter

4/29/2021

Background

Analysis	ERAP2 # of Significant Genotypes	ERAP2 3 Most Significant Genotypes	PEX6 # of Significant Genotypes	PEX6 3 Most Significant Genotypes	FAHD1 # of Significant Genotypes	FAHD1 3 Most Significant Genotypes
Without Covariates	73	rs7726445, rs7731592, rs2161548	29	rs1129187, rs10948061, rs9471985	90	rs11644748, rs140254902, rs9652776
—	—	—	—	—	—	—
With One Hot Encoded Covariates	73	rs7726445, rs7731592, rs2161548	27	rs1129187, rs10948061, rs9471985	90	rs11644748, rs140254902, rs9652776
—	—	—	—	—	—	—
With Linearly Encoded Covariates	74	rs7726445, rs7731592, rs2161548	28	rs1129187, rs10948061, rs9471985	90	rs11644748, rs140254902, rs9652776

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