Valeria Duran

MATH 6388

April 10, 2020

Final Project Proposal

The data set I plan on analyzing comes from the BioProject Tuberculosis Meningitis in Pediatric Patients (PRJNA577283). This project uses RNA-Seq to compare whole blood transcriptional profiles in children with or without tuberculosis meningitis (TBM). In the project, a total of 43 whole blood samples were sequenced. The samples used consist of paired-end reads. These 43 samples consist of 15 tuberculosis meningitis cases, 4 non-tuberculosis meningitis cases, and 24 non-infection healthy controls. Because there are 3 types of cases, I plan on comparing only tuberculosis meningitis cases with healthy controls. I will select either 6 or 8 samples; depending on the sample size, it will consist of either 3 or 4 TMB samples and 3 or 4 healthy samples. Since there are 39 samples to choose from, I selected the samples based on how many reads were sequenced. Therefore, each group of TBM and healthy cases will consist of the samples containing the smallest reads sequenced, which are samples: SRR6809855, SRR6809856, SRR6809862, SRR6809864, SRR6809883, SRR6809886, SRR6809890, and SRR6809894. I will compare gene expression between the two groups in order to determine any notable differences or similarities.

I will use the SRA toolkit in order to download the fastq files into my directory in order to conduct a kallisto run on my samples and afterwards do a downstream analysis on the data using DESeq2. I will use MA plots as well as tidyverse functions in order to analyze log-fold change values and p-values. Apart from this, I will use heat maps and PCA plots in order to see sample similarity. I will have to transform the data for clustering and visualization using the logarithm transformation.