Janis Corona

BIOL 59000: Data Science Project for Life Sciences

Week 4 Progress

June 29, 2019

Meta-Analysis of the Ubiquitous Genes Associated with Human Uterine Leiomyoma Development in Healthy Human Tissue – Week Three Progress

During this fourth week of research on uterine leiomyoma (UL) genes ubiquitous to the current research studies published on UL gene risk factors in healthy females, more analysis and visualizations were developed for the materials and methods as well as for the results section of this research. R was used to analyze the gene expression data from the five studies obtained from the Gene Expression Omnibus (GEO) and to also map out the 130 genes in common between the five studies of 121 samples consisting of 70 UL and 51 non-UL samples. Many PNG plots and additions to the R script of commands were made as the analysis and synthesis of the gene expression data unraveled. All of these additions can be obtained at <https://github.com/JanJanJan2018/Better-Cleaned-Version-UL-Research> for the plots, data tables and script. The script has the notes that go with the tables developed for analyzing the gene expression data.

Some of the plots made were better annotations of the genes that are expressed more (up regulated) and expressed less (down regulated) in UL tissue compared to non-UL tissue samples. Tables were made that could be added to the methods or results section showing up and down regulated genes per chromosome, as well as if those genes are up or down regulated as a majority of other genes on that chromosome that were up or down regulated. The chromosomal locations were the filtering method to shrink down the very large data table of genes from gigabytes to megabytes of data when merging the five studies together by genes in common. These five studies were selected from six studies because they all had the six genes ubiquitous to UL research of TNRC6B, BET1L, CYTH4, HMGA2, CCDC57, and FASN. An R package, lattice, was used to show a pairwise gene relationship of scatter plots on non-UL samples, but the image couldn’t show any definite relationships that could be linear. When more than five genes were compared to each other at a time, the plot looked like a giant mess of tiny squares in an array having splotches in each square. These images of non-UL pairwise gene comparisons were done on 10 samples each out of some of the 51 non-UL samples and uploaded to github as Pairwise\_10\_Most\_DE\_in\_nonUL.png and Pairwise\_10\_Least\_DE\_in\_UL.png. The data was further filtered so that the six genes ubiquitous to the UL research studies and the top 10 genes having the highest magnitude of expression in UL compared to non-UL samples in up or down regulation is the final data set. This could also be good as a table for the methods. Some bootstrap simulations of a population mean from random sampling of the first gene in the 16 top genes table was done on FSCN2 for non-UL and UL samples to get a better approximate mean, a standard deviation, and a two-tail 95% confidence interval that is in the process of being created for all of the 16 genes. Using this data, the goal is to have the data fit a model such as linear regression to the 16 genes original output for each of the 121 UL and non-UL GEO samples in predicting 30 per cent of the samples as either UL or non-UL.

The packages used in R thus far are Gviz, lattice, ggplot2, UsingR, and dplyr. More work must be done on building the table of simulated means for each gene in the UL and non-UL samples from the simulation of 10,000 sampling observations with replacement using each gene as a vector of values for each sample. The genes from the non-UL group have 51 values or columns and 10,000 simulated observations, and the genes from the UL group have 70 values and 10,000 simulated observations.

When grouping the genes per chromosome that were up or down regulated, filtering was done within those groups to show genes as part of the majority of genes on that chromosome up or down regulated or part of the minority using a Boolean value for a created field called ‘majority’. It was curious to find that three of the six genes are not part of the majority on chromosome 11 and 17, while those on chromosome 12 and 22 are up or down regulated as part of the majority. The chromosome plots were able to be annotated with the up and down regulated genes and having the gene symbol. These plots are in the github web link as PNG images having majority/minority, Down/Up, chr, and 11, 12, 17, or 22 in the file names, such as: majorityDownChr11.png or minorityUpChr17.png.