Janis Corona

BIOL 59000: Data Science Project for Life Sciences

Week 2 Assignment: Research progress

June 15, 2019

Progress Made During Week 2

This week more work was completed using R coding to combine the Gene Expression Omnibus (GEO) data series into one data set of uterine leiomyomas (UL) and non-UL samples with other genes sharing the same chromosome location as the top genes’ chromosomes. Originally there were six series to combine, but this is now five series of 121 microarray gene expression samples instead of 133. These samples consist of 70 UL samples and 51 non-UL samples. All have the six ubiquitous genes in the microarray gene expression data from GEO of TNRC6B, BET1L, CYTH4, CCDC57, FASN, and HMAG2. The excluded data series is GSE764 and it only had FASN and HMGA2. This series is still attached as a reference as a limitation for the discussion section. There are 183 genes from data that originally had between 22,000 and 55,000 genes depending on which GEO series the data was combined from. More coding was going to be done to lay down tracks of the genes like what ENSEMBL does with genes, but the operating system this research was being developed on crashed and a replacement computer was purchased the following few days.

Other progress towards this research was on making it clear that the focus of this research is not on the genotypes or single nucleotide polymorphisms (SNP)s of the ubiquitous genes found in current population studies to pose a risk of UL in females. This research is on the genes expressed in the microarray data from the five GEO data series to compare the top six ubiquitous genes with other top genes and determine how the gene expression compares across a set of UL samples to a set of non-UL samples.

Corrections from feedback of the Week 1 submittal were made in progressing through this research. The GEO direct link to the accession ID for the platform and series data obtained for this research work was placed into each GEO reference. The PMID from GEO series referenced items in the report were omitted. A run-on sentence on the first page was corrected in the proposal.

Further expansion and explaining of other items were made to the research. The relationship between estrogen and UL growth and treatment involving an estrogen inhibitor to cease UL growth was explained better than previously reported. An explanation of how the training and testing sets would be selected as partitions in the R caret package was made. The ubiquitous genes’ SNPs reported as significant in studies were based on Minor Allele Frequency count and it is not known if these genes have a low or high expression in UL or non-UL data from the studies currently reviewed for this research. Although, one exception is the ubiquitous gene CYTH4 found to be expressed low in thyroids of African American females who have UL. The identifiers in the data only have sequence, chromosome, and cytoband information in one of the five data sets. The other identifiers are the alternate IDs to combine data or merge the data together by, then filter by chromosome location. This data is all human samples as only one series had rat samples, but those samples were omitted from this data collection by sample name. Possible models to build are based algorithms to train using the training data set such as linear, logistic, Bayesian, Random Forest, and K-Nearest-Neighbor. These models would each be tested on the testing data partition for accuracy in prediction

Some more exploring with R and Bioconductor of this completed data set on UL and non-UL samples is needed to further progress towards completing research on the ‘Meta-Analysis of the Ubiquitous Genes Associated with Human Uterine Leiomyoma Development in Healthy Human Tissue.’ Once the data has been modeled accurately to predict a UL or non-UL sample based on gene expression, the genes will be layed out in a chromosome plot to show the top differentially expressed genes between UL and non-UL samples in unidentified race demographic samples.