Juan Ibarra

Briana Banks

ECEV-32000

3/25/20

Coding Final Project Report

The purpose of this code was to organize data from RNA sequencing. RNA sequencing is a frequently used technique to determine the presence and quantity of RNA in a given sample. Using next generation sequencing, RNA-seq presents a number representative of relative expression of a given gene in a sample. For biologists, results usually come in different data file formats (in this case excel file), however an analysis can be composed of 20000+ genes with varying expression scores. This can make parsing through data very challenging due to the unorganization and range of scores. Our code is very useful for our research because it allows us to organize the RNA-seq data, set a threshold, determine which genes of interest are above or below the set threshold (upregulated or downregulated), and graph our results. Another issue that biologists often face is to determine whether a set of genes involved in specific pathways or family of genes are enriched in their RNA-seq data since it's difficult to look up every single gene one by one in their RNA-seq data. To overcome this we generated a code to match a gene set (gene families/pathways) with the genes present in the RNAseq, and determine their significance based on their expression score.

The RNAseq used throughout the code was obtained from BaF3 cells comparing cells carrying the protein calreticulin (CALR) as WT vs. CALR mutant. Calreticulin is an endoplasmic reticulum (ER) protein which functions as chaperone. However, it has been shown that when mutated it loses its chaperon function and instead activates the unfolded protein response (UPR) pathway. The UPR pathway is responsible for restoration of function of the cell by: halting protein translation, degrading misfolded proteins, and activating the signalling pathways that lead to increasing the production of chaperones. When this pathway is abnormally upregulated it is associated with endoplasmic reticulum stress that is associated with neurodegenerative diseases and hematopoietic cancers. Thus we matched the RNAseq against UPR-associated genes obtained from the Reactome database. In addition, we also matched the RNAseq against glycolysis-associated genes as an alternative validation of the code developed.

Using Jupyter Notebook on Python, we analyzed a RNA seq (RNAseqdata.xls) set which contained the different genes and the given expression score. First we uploaded the excel file of 16101 genes with scores ranging from -44 to 31. Using the xlrd program, we were able to upload the excel sheet and display the top genes with the highest score and the bottom genes with the lowest score. Based upon the literature or arbitrarily we determined that upregulated genes had a score of > 20. Statistical threshold was set and named separately therefore users can decide the name of the file and the threshold to make this part of the code a function. This threshold means that, any score greater than or equal to 20 is considered to be upregulated in this sample. Once we were able to parse this through our data set, we were left with 79 genes. We labeled these genes as dataframe upregulated. Of the upregulated genes, we wanted to determine which of these genes were associated with the UPR pathway. Using the Reactome gene database, we found a set of UPR-associated genes. We incorporated this list of 100+ genes onto our RNAseqdata.xls file. Then we matched the Upregulated RNAseq genes and the UPR-associated genes based on their name. From there we were able to determine that there were 9 UPR-associated upregulated genes. Using matplot, we plotted the upregulated genes and performed the same analysis for downregulated genes. Based upon the literature, it was determined that a score < -20 was associated with downregulation. We parsed our dataframe to include only genes that had this score. Of this dataframe, we performed the same analysis by matching via gene name. We saw that there was only one UPR-associated downregulated gene. We also plotted these genes using mat.plot. Finally, we followed the same approach to test genes associated with Glycolysis that were obtained similarly from the Reactome gene database. We found that there were 3 Glycolysis-associated upregulated genes and 0 glycolysis-associated downregulated genes.

Overall, our code has proven and validated. Biologists can adjust the thresholds for a given gene based upon the literature, and have the ability to decipher which genes of interest are associated with upregulation and downregulation. We were able to parse through the dataframe to allow for convenient and organized extraction of data as well as visualization to look into overall trends.

We are grateful for this final coding project because it is a technique that we have and will continue to use for our thesis projects. It has also allowed us to put into practice some of the concepts and techniques learned throughout the course, which overall have encouraged us to become more confident in our ability to code and manage data.