Triana San Miguel
Introduction To Computational Biology

Work Documentation Two

Week of February 20-February 22

This week I began to drive into the world of understanding gene expression. This concept seemed easy to me at first because I had already learned about the concept of gene expression in my Biology based courses. However at a second glance, using tidyverse and other coding functions seemed to make understanding gene expression much more confusing to me. This week I focused on a certain gene expression coding worksheet where we had to focus on using normalization techniques to calculate (TPM)s otherwise known as Transcripts Per Million. I imported two different data sets into R one of them was a gene length and another which was the read count.I learned how to analyze transcriptomics data sequenced from the mid brain of an african frog species. After importing each data set and loading them as libraries the nice code I ran was actually calculating the TPMs. I had to divide the read counts by gene length and then sum the reads. After that I had to Loop the function. This was a very interesting worksheet and coding aid and I feel like I learned a lot from analyzing it. Another fact that I learned was that instead of using the loop function I could have used the “apply” function to perform this function over arrays instead!

Week of February 27-February 29

This week I found myself to be very confused about the concept of gene expression and how it works. Even though I had learned about it in Biology and had focused on the RNAseqIntro in class I still felt confused about the science behind it all. I Rewatched the youtube video that was provided to us titled ;Difference between Global and Local Sequence Alignment, and learned more about the entire process such as pairwise and multiple sequence alignment and why we align sequences at all. However I still decided to do a much deeper dive so I could prepare myself for the future assignments I would have to do in relation to gene expression. I found two very helpful youtube videos that I watched and took notes on. The first one was titled “Gene Expression and Protein Synthesis The Genetic Code tutorial” this video showed and explained the DNA and RNA bases in depth; it also clarified on amino acid sequencing and anticodons. The next youtube video explained concepts such as silencing gene expression and the specialization of cells. The video was titled Regulation of Gene Expression: Operons, Epigenetics, and Transcription Factors.

Week of March 5- March 7

This week we got introduced to the concepts that were centered around RNAseq clustering. I was very interested when it came to learning about these topics because I was able to understand the summarization of data when it came to very large data sets. To focus more on these concepts out of class I decided to re-read the worksheets and try to perform the coding tasks on my own. For

example, a topic that confused me was log transforming the data. I was trying to figure out how to properly log the data so it would not be as skewed towards the genes that had high variance. I was tasked with loading the same data sets from earlier and completing the TPM calculations and the first round of filtering. I got confused when I had to create a new object that transforms my data. I kept on putting spaces where I did not need to and get on getting different errors. However after a few fixes I then named my data trinis_log_tpm. I then decided to challenge myself by subsetting the data frames to only show me the genes that were high variance. I struggled with this code at first and eventually had to ask a mentor to aid me in subsetting the data frame. Outside of class I also began to analyze the heat lamp clusters and learned about co-clutersting and dendrograms.

Week of March 19- March 21

I was mainly focused on two different things this week PCA and also trying to read more about the DESeq2. My main focus of work outside of this class was the PCA worksheet and trying to perform the worksheet on my own by really understanding the worksheet that Dr. Young had provided us with that is called 'PCA brain region'. This code was simple yet intricate at the very same time. I made many careless mistakes along the way when trying to perform this code. For example, I had completely forgotten to load both the tidyverse library and the readr libraries into R studio. I also needed a bit of a refresher about the entire ggplot sequence. However after doing this worksheet I learned a lot about PCA and how it works

Week of March 26- March 28

This week was full of preparing for my presentation on Thursday. Me and my groupmates all planned and prepared our slides for the upcoming presentations and tried to break down our paper in a way where all of us could understand, represent, and present our ideas to the class. My task was the methods which were extremely lengthy in our paper. There were many methods that the authors of the paper used to represent and locate their data such as bulk sequencing, dna sequencing, snp analysis, and then represented that data using t-tests and statistical graphs. This paper taught me many things about clonal cancer cells and how each sequencing task can have different effects on the data. For example bulk sequencing was falling behind on the representation of data compared with bulk sequencing. Me and my group also focused on trying to represent our data in a way that everyone in the class would understand. I also spent the whole day on Monday practicing what I was going to say. Due to the fact that English is not my first language I tend to stutter and trip over some words or misinterpret them so I tried my best to practice beforehand.