Bahaa Noori DSP 539 Dr. Rachel Schwartz

**Differential Expression of Genes in Presence and Absence of FANCD2**

An experiment was done on Fanconi Anemia ear fibroblast cell line. This cell line (+G) is missing FANCD2 gene which encode for FANCD2 protein that play a key role in the DNA repair when inter-strand breaks occur. We made a complement cell line to the (G) and called it (D2). This experiment is RNAseq experiment in which we treated those cell lines with Aphidicolin (induces double strand breaks in DNA) for 24h. In this experiment, we had 4 replicates of each cell line. The data was analyzed and graphed.

The sequencing data are very complex and huge compared to the data that we human can work with and handle. The sequencing data are impossible to analyze by humans. We need computational method to deal with this size of data. Furthermore, this data consists of multivariable dataset that our traditional method of two-dimension analysis of data would not work with multivariable dataset. So, in order to analyze a multivariable dataset, we ought to use computational method called principal component analysis (PCA). PCA is a statical method that allows us to summarize the large data by means of a smaller set of “summary indices” that can be easily analyzed and visualized. This will allow us to observe trends, jumps, clusters, and outliers. This might uncover the relationships between observations and variables, and among the variables.

In this project, I did a principal component analysis (PCA) on a .csv file containing all the differentially expressed genes. This table include all the replicates, p-values, log2fold change and gene names. The data was analyzed using DESeq2, which is a count-based statistical method. This method requires input data obtained from RNA-seq or other high-throughput sequencing experiment in the form of a matrix of un-normalized counts. After analysis with DESeq2, a PCA can be performed then visualize the sample-to-sample distance. In this method, the data points are projected on a 2D plane where they spread out into two directions that explain most of the differences. The x-axis is the direction that separate the data points the most (written as PC1), and the y-axis is the direction that separate the data the second most (written as PC2).

I started by performing a PCA analysis on the table and got many PCs that account for most of the variance between the samples. These PCs do not all account for the variance between the cell lines. So, in order to see which PC account for most of the variance I analyzed the PCs by plotting the variance on the y-axis and the PC on the x-axis (Figure 1). This showed that more than 90% of the data are explained by PC1 and 2. Therefore, I proceeded by selecting those two PCs. After PC selection, I looked at the structure of my principal component and selected the PC 1 and 2 and attached those two PCs to the original data as new columns. Then, I set a p-value to be less than 0.05 to omit any data that is not significant in this analysis. Then, I plotted them using volcano plot to have PC1 on the x-axis and PC2 on the y-axis. This graph showed most of the variance and differential expression of the genes of interest. I further customized the graph by specifying p-cut off and f-cut off because the basic volcano plot will attempt to label genes that pass the threshold for statistical significance. Also, it will only label as many as of these that can reasonably fit on the plot space. Then, I adjusted the cut-off lines by adding extra threshold lines. After that, I was trying to label the red dots with the gene name that correspond to these dots, but I wasn’t able to get the code to work for me.