Adalee Koshiol

BF591

4/29/2022

Final Project Write Up: Project 1 Programmer Role

**Introduction**

The programmer role for this project was meant to normalize microarray data for further use of the other roles. This is an important step in analyzing microarray data, as it ensures differences in expression are significant because they are actually differentially expressed. Visualization of the data in this role included exploring histograms of Relative Log Expression and Normalized Unscaled Standard Error scores. These tools are integral in visualizing unwanted variation in the data.

**Methods**

All analysis methods were implemented in R Statistical Analysis Software (version 4.1.2.) with the following packages: affy, affyPLM, sva, AnnotationDbi, hgu133plus2.db, ggplot2, and tidyverse. CEL files were read into R and normalized using a Robust Multiarray Averaging algorithm. These files were then fit to Affymetrix genechip data, and the Relative Log Expression (RLE) and Normalized Unscaled Standard Error (NUSE) scores of the microarray samples were calculated. There were 134 samples in the Affymetrix. Median RLE and NUSE scores were then plotted in respective histograms for visualization. From there, the data were corrected for batch effects using a given annotation. Correcting of batch effects included use of Center and RNA extraction methods and features of interest. Lastly, Principal Component Analysis was performed on normalized, centered, and scaled data. A plot of the first two principal components was produced to visualize the percent of variance within these two principal components.

**Results**

Histograms of RLE and NUSE scores show scores to be clustered around zero, suggesting heterogeneity (Fig. 1A & 1B). Though the scores are clustered around zero, the variation in the median RLE and NUSE scores do still show a little variation.

Principal Component Analysis of the samples shows the variance explained by the first principal component is 14.5%, and the variance explained by the second principal component is 9.54%. When plotting the first PC against the second, the samples are spread semi-evenly but do somewhat cluster toward the left side of the graph (Fig. 2). However, a significant cluster is not visible (Fig. 2). Therefore, no obvious outliers are visible.

**Discussion**

The purpose of this segment was to normalize the data and visualize the results of that normalization. Visualization of the normalized data helps in doing a sanity check, but also provides further insight for roles further down the pipeline. It can help determine if filtering is necessary and the validity of results discovered.

RLE and NUSE scores suggested that there is little variation in the data, which means the normalization methods used were successful. The PCA visualization helped to identify that the data is not clustered, meaning there were no obvious outliers. However, it would still be useful to perform filtering methods on the data to remove unwanted outliers. Thus, the next role in the pipeline should filter the data.

**Figures and Deliverables**

**Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated**

Figure 1. Histogram of RLE and NUSE Medians. A) The right left histogram shows distribution of the RLE scores. B) The left histogram shows distribution of the NUSE scores. The medians of Relative Log Expression (RLE) and Normalized Unscaled Standard Error (NUSE) scores were computed on normalized microarray samples.

**Chart, scatter chart

Description automatically generated**

Figure 2. First and Second Principal Components. The first and second principal components are compared against each other, showing the percent variance explained by each.