Adalee Koshiol

BF591

4/29/2022

Final Project Write Up: Project 1 Analyst Role

**Introduction**

The first goal of the analyst role for this project was to reduce the number of genes/features in the data based on three metrics. The first metric reduced the data to include genes that had 20% of gene expression values greater than log2(15). The second included reducing the data to include genes that had a variance different from the median with a p-value less than 0.01. And the third reduced the data to those that had a coefficient of variation greater than 0.186. Filtering the data using these restrictions can “clean” the data and remove observations that may contain errors or observations that are not wanted in the analysis since the number of features is much greater than the number of samples.

The second goal of the analyst role for this project was to perform clustering analysis. The clustering analysis is useful for grouping similar data to look at relationships that are similar between them.

**Methods**

All methods executed for this procedure were done using R Statistical Analysis Software (version 4.1.2) using an RMA normalized, ComBat adjusted expression matrix. The data were first filtered for genes that had at least 20% of their expression values greater than log2(15). Then, a chi-squared test was performed for each gene and the resulting test-statistic was compared to the median. Those that varied significantly at a threshold of p less than 0.01 were kept. The last filter calculated the coefficient of variation for each gene and kept only those that had a coefficient of variation greater than 0.186.

Next, hierarchical cluster was done on the filtered data using Euclidean and “averaging” methods. The produced dendrogram was then cut into two clusters for further use. A clustered heatmap was then made using the filtered data. And a Welch t-test was done comparing each gene and expression values between the previously cut clusters. P-values were compared at a threshold of 0.05 to determine significant genes.

**Results**

The original expression matrix after RMA normalization and ComBat adjusting had 54,675 samples. After the first filter for expression values, there were 39,750 genes left. The second filtering with the chi-squared test left 19,926 genes in the dataset. And finally, the last filter using coefficient of variation resulted in 1,658 genes being left in the expression matrix.

The heatmap produced shows two clusters different from the others. These two clusters consist of 76 samples and 58 samples, as seen from the hierarchical clustering performed. The Welch t-test showed there are genes that are statistically significant in differential expression between the two clusters. The top genes for the C4 cluster were GAS1 and SFRP2, whereas the C3 cluster had FAP, ADH6, and XIST as the top genes.

A picture containing text

Description automatically generated

Figure 1. Heatmap of Gene Expression. This heatmap shows the differences between the clusters found after hierarchical clustering.

**Discussion**

The goal of the analyst role was to filter the expression data following three criteria: 1) genes must have expression greater than log2(15) in 20% of samples, 2) the variance must be statistically significant from the median at a p-value of less than 0.01, and 3) the genes must have a coefficient of variation greater than 0.186. After filtering, the goal of the role was to also perform hierarchical clustering and produce a heatmap of the gene expression to aid in visualization of the clusters.

Overall, this analysis showed that the filtering methods were successful. In addition, it found there to be a difference in expression between the genes in the two clusters. The original paper named these clusters ‘C3 ‘and ‘C4’. Genes that were found to be differentially expressed in C4 were GAS1 -- a tumor suppressor gene -- and SFRP2. Genes that were found to be differentially expressed in C3 were FAP, ADH6, and XIST. Further research is needed to explore the Gene Ontology terms related to these genes.