Haroon Qureshi
Programmer and Analyst Portions of Project 1

Microarray Based Tumor Classification (Abbreviated)

Introduction:

This study was undertaken with the goal of replicating the programmer and analyst portions associated with the microarray-based colorectal tumor classification study conducted by Marisa et al. In this reproduction study, the programmer role produced microarray data that had been normalized to ensure that any downstream analyses were not contaminated or biased due to artifacts other than the sample expression data. This role was vital to make sure that the data produced was accurate and relevant to the expression data. This normalized data underwent quality control measures to ensure the data was not biased due to outside elements, and histograms pertaining to the RLE and NUSE scores, two of these quality control metrics, were made. A PCA plot was also produced after scaling the data, and this visualized the percent variability among principal components, allowing for any outliers to be detected. The analyst role took the expression data produced by the programmer and first performed various filtering methods to clean data, reducing noise. Once cleaned it was possible to find the genes that were strongly expressed from the given gene-pool. Once the significantly expressed genes were found, they were clustered using hclust(), a function built into R. This was used to classify genes based on hierarchical attributes, and identify whether genes were associated with the colon cancer subtype C3, or another subtype, and a heatmap was made using dendrograms made from the clustered data. Finally, a Welch test was performed on the genes of both subtypes to find the genes that were most significantly differentially expressed. Overall, the analyst role was extremely important, as it found the most significant genes, reduced noise, and created associations between genes and the cancer subtypes, allowing for the genes associated with subtype C3 to be identified.

Methods:

To begin the programmer role, the CEL files provided by the data curator were all read into the RStudio session using the ReadAffy(), a function provided by the Bioconductor package affy. Using the function rma() from the same package, the CEL files were all normalized together. Using the fitPLM() function, a probe level linear model was created. This was needed for processing the data for the Relative Log Expression (RLE) and Normalized Unscaled Standard Error (NUSE) values, metrics used for ensuring the quality of the data is sufficient. The Surrogate Variable Analysis (sva) package was then installed and its function ComBat() was utilized on the normalized data for correcting batch effects. Two fields were of particular interest, in accordance to Marisa et. al: Tumor and MMR Status. The expression data was then saved as a .csv file for downstream analysis. Principal Component Analysis was performed, and a PCA plot was created using the function ggplot().

Beginning the analyst role, the expression data produced from the ComBat() analysis were filtered for further analysis. To start, probesets that had 20% of their associated samples

with an expression greater than log2(15) were filtered for. Another filter was applied, searching for genes that displayed greater variance with a p-value less than 0.01 ($p<0.01$). Finally, probsets with a coefficient of variation > 0.186 were found. These filters were implemented to reduce noise, which would be beneficial for clustering. Clustering was achieved by utilizing the function hclust(), using the genes found through filtering and forming two separate clusters, those related to the colon cancer subtype 3, and those related to other subtypes. A heatmap was generated to better visualize the expression of the genes associated with these subtypes, and finally a Welch t-test was performed to find genes between both clusters that had a significance of at least p-value < 0.05.

Results:

To begin the analysis of the colon cancer samples provided by the data curator, quality control measures were first taken. Figure 1 depicts the results of conducting the quality control metrics RLE and NUSE on the normalized sample data. As seen in Figure 1, it can be seen that the scores for the 134 colon cancer samples revolved around an average score of 0, with ~95 samples that had scores of 0 +/- 0.5. The NUSE scores are shown to have a similar centralization around 1, with ~110 samples scoring 1 +/- 0.02. Based on these scores, it could be determined that these samples were of good quality, and would be able to used for further analysis without having to drop any samples.

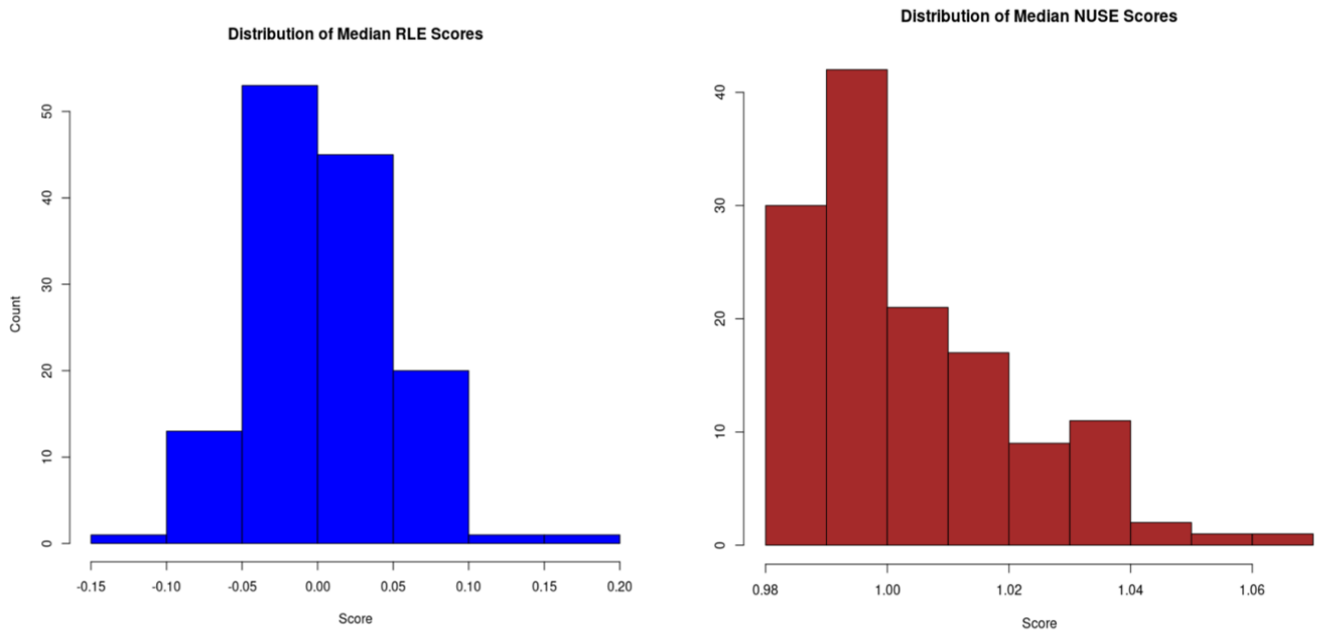

Figure 1. Histograms for RLE (Left) and NUSE (Right), representing the distribution of the median scores across both metrics, respectively.

A PCA plot was then constructed from the two principal components that had the largest variance, as seen in Figure 2. Depending on the subtype that the sample belonged to, a clear visual representation of subtype clusters can be seen. No clear outliers were detected either, therefore, along with the median distribution from Figure 1, it was determined that the data was of good quality, and the samples could all be used further downstream.

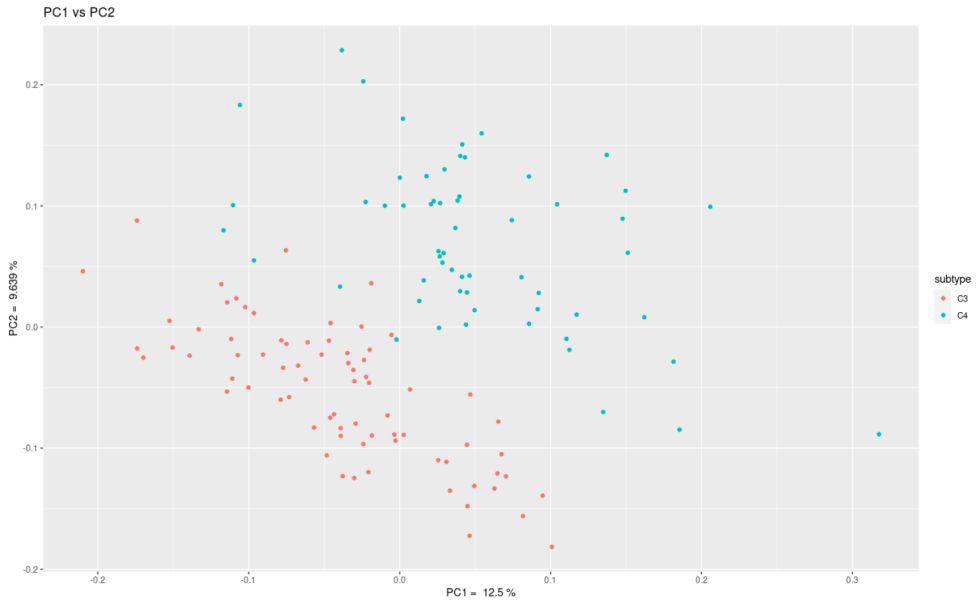

Figure 2. PCA plot depicting the variance between the 134 colon cancer samples.

When filtering the normalized data to reduce noise, there were a few thresholds implemented. To start, probesets that had 20% of their associated samples with an expression greater than log2(15) were filtered for, and then genes with a p-value less than 0.01 ($p<0.01$) and a coefficient of variation > 0.186 were found. Once applying all these filters, it was found that 1,531 probesets passed through. When applying clustering to these samples using Euclidean distance techniques, as seen in Figure 3, it was found that 57 samples were non-C3 subtypes, and 77 were C3 subtypes. A heatmap was created to visualize the expression intensity of each cluster, as can be seen in Figure 4. The expression intensities were well-defined between the two clusters, but there were a couple samples grouped in the non-C3 subtype that were actually C3. The Welch t-test was then performed, finding genes with an adjusted p-value < 0.05, comparing C3 and non-C3 expression between the 1,531 probesets. It was found that 804 probesets were differentially expressed, with 41 being up-regulated in C3 subtypes, and 763 were down-regulated.

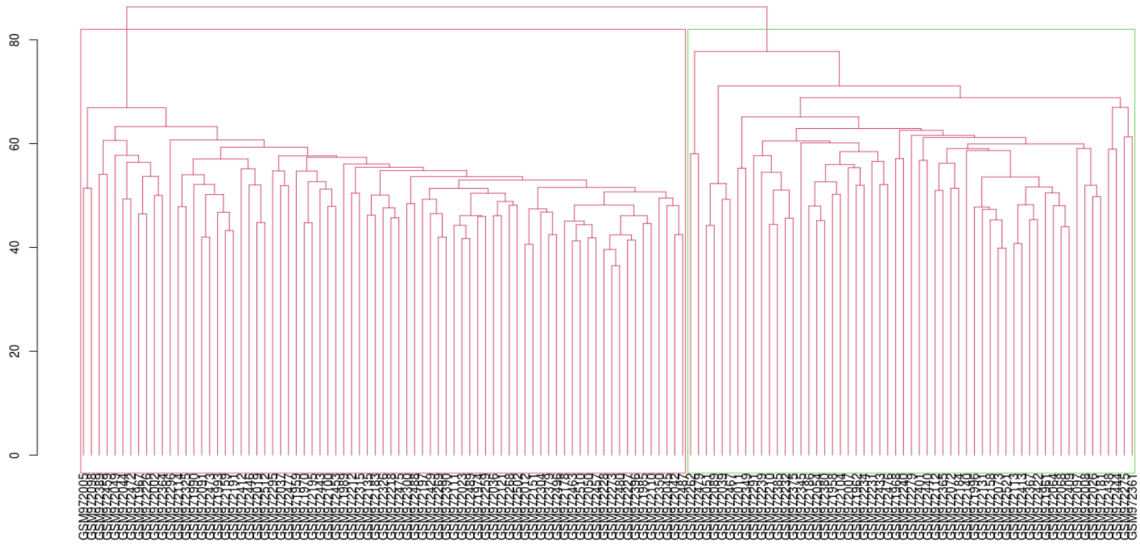

Figure 3. Dendrogram displaying clusters built through Euclidean distance clustering method. Clusters represented C3 and non-C3 subtypes, respectively.

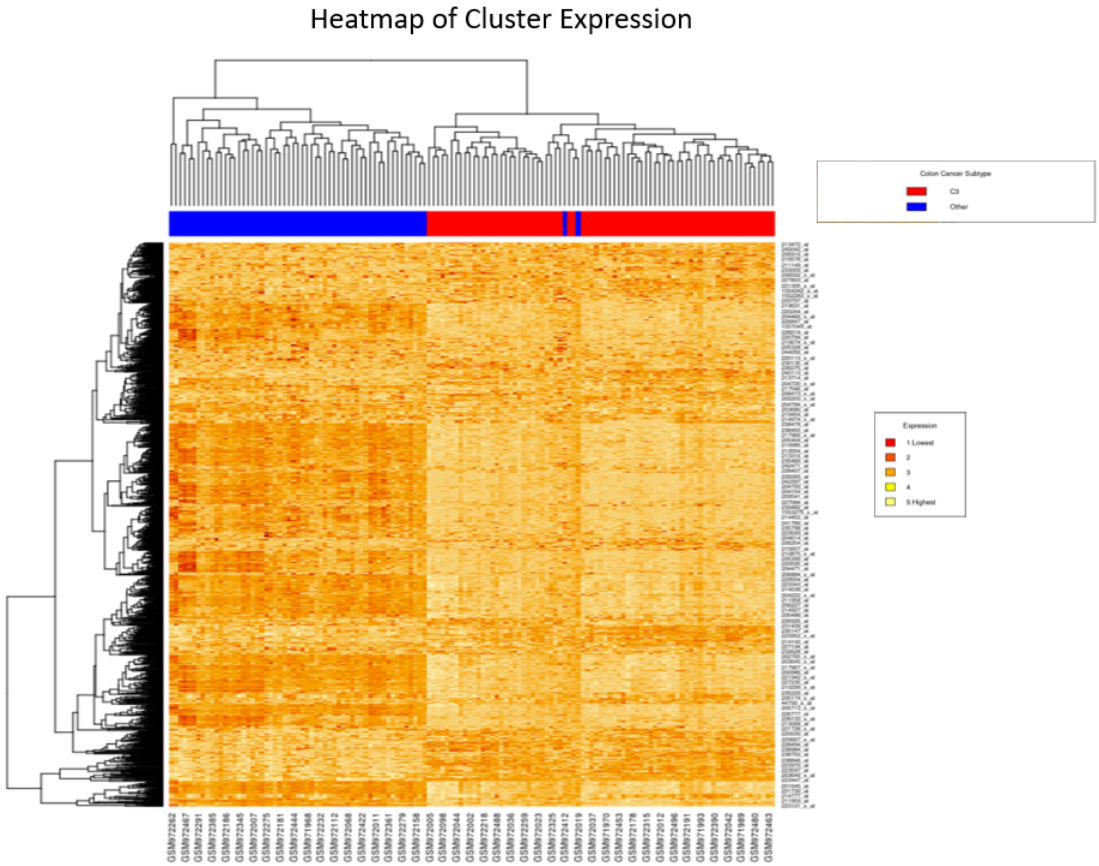

Figure 4. Heatmap indicating gene expression between the C3 and non-C3 subtype clusters. Clear expression differences were shown between the clusters.

Discussion:

The quality of the 134 colon cancer samples supplied by the data curator underwent RLE and NUSE scoring to assess their quality. It was found that the samples were indeed of good quality, as determined by median RLE scores around 0 +/- 0.5 and the median NUSE scores centralized around 1 +/- 0.02, as seen in FIgure 1. Due to the median distributions of these respective scoring methods hovering around 0 and 1 respectively, the samples were determined to be high quality. After performing PCA on these samples, using two principal components, it was found that the PC1 variance was 12.5% and the PC2 variance was 9.639%. The two principal components represent the variance within the sample set. No outliers were found through the PCA, and every sample was viable for use in downstream analysis.

The clustering of the samples into their respective subtypes was similar to what Marisa et al. produced, although they did not use the same clustering method of Euclidean distance. The subtypes were successfully clustered, although a handful of C3 samples were incorrectly identified as non-C3. This is possibly because these samples may have had some attributes dissimilar to the others, as their expression levels are clearly deviant from the other C3 samples. This could be a point of interest, and further research could be done on these samples to understand the reason for this difference in expression intensity a bit better.

After performing the Welch t-test at adjusted p-value < 0.05, there were 41 up-regulated genes of the C3 subtype and 763 down-regulated genes. The probeset determined to best represent the cluster for the C3 subtypes was the FCGBP probeset. This was chosen due to the correlation of these genes and the levels of B-cells and macrophages detected when fighting tumors. It makes sense that the immune response would increase if a tumor was detected. Looking at the heatmap, there was a clear increase of expression amongst the C3 colon cancer subtypes, so those genes are more than likely the genes with the strongest association with C3. Observing the relationship this way, there may also be a correlation between the non-C3 subtypes and the most down-regulated genes. For this reason, the probeset MIR100HG was found to best represent the non-C3 subtypes. This gene is associated with having a strong influence, so it is interesting that it was as down-regulated as it was, but perhaps this was due to other factors. It may be interesting to delve deeper into MIR100HG and FCGBP, as there perhaps is an inhibitory relationship between the two.

References

1. Marisa, L., de Reyniès, A., Duval, A., Selves, J., Gaub, M. P., Vescovo, L., Etienne-Grimaldi, M. C., Schiappa, R., Guenot, D., Ayadi, M., Kirzin, S., Chazal, M., Fléjou, J. F., Benchimol, D., Berger, A., Lagarde, A., Pencreach, E., Piard, F., Elias, D., Parc, Y., … Boige, V. (2013). Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS medicine*, *10*(5), e1001453. https://doi.org/10.1371/journal.pmed.1001453