**Microarray Tumor Classification**

**Introduction:**

Colorectal Cancer, also referred to as colon cancer, is the third most common cancer and fourth leading cause of cancer death worldwide (Marisa et. al). The colon is an important and fundamental part of our digestive tract and its principal functions include absorption of water, minerals, and nutrients. Environmental and hereditary factors contribute to the development of colon cancer which is shown through the build up of mutations found on the oncogenes. Oncogenes are genes that suppress and repair DNA that signal the existence of various pathways where tumors may occur (Arvelo et. al 2015). Previous studies have investigated whether there are any possible gene expression profiles that can help predict prognosis of colorectal cancer but have thus far not produced any firm conclusions with any clinical significance. Colorectal cancer is a disease with several root causes, which is one of the reasons why current colorectal cancer pathological staging protocols fail to accurately predict recurrence. The main purpose of this study was to create a comprehensive molecular classification of colorectal cancer based on mRNA expression profile analyses, which could then be used for more effective prognosis stratification. Colorectal cancer is the third most common cancer in men and second most common cancer in women worldwide, claiming 608,000 lives each year. Prognosis and treatment plans vary widely depending on which of the five pathological stages the patient falls into, so the need for accurate staging and recurrence prediction is very clear.

Bioinformatics techniques were employed by the authors to generate information about the molecular profile of colon cancer for use in disease classification. Whole genome and transcriptome screening arrays were employed by the authors. These arrays allowed for the capture of DNA and RNA information featured in the samples. Based on this information, further analyses may be used to build a molecular profile of colon cancer and its biologically distinct subtypes. Common DNA alterations were characterized in the samples using bioinformatics techniques that were applied to the whole genome analysis results. These characterizations were pertinent, because they associated the genomes of samples with well studied mutations and may lead to driver gene identification (Chakravarthi et al.).

Transcriptome array data was normalized to address experimental biases affecting the expression measurements. Furthermore, batch effect correction was used to address residual measurement biases. These effects on the data would have confounded the detection of signals of interest in downstream analysis. Genes in the expression data were filtered out if they did not meet a quantitative threshold. These genes would unnecessarily reduce the statistical power of downstream methods in differentially expressed gene detection. Additionally, genes were filtered out of the dataset based on variance and coefficients of variation. This subset the expression data for genes that were more likely to be meaningfully differentially expressed among the samples. Colon cancer subtypes were determined by consensus clustering. This technique allowed the authors to build molecular classifications for the samples using gene expression data. Consensus clustering is a relatively robust method of classification, because it features many iterations of clustering (Lancichinetti and Fortunato). Samples were classified as belonging to a subtype in order to associate their molecular characteristics with a group of samples. This technique was facilitated by associating the differential expression of probe sets with a subtype and applying these associations to a centroid-based predictor of the sample subtype.

The gene signatures of the defined subtypes were associated with gene signatures defined in literature to elucidate molecular subtype characteristics based on previously published findings. KEGG pathways and some Gene Ontology gene sets were tested against the top differentially expressed genes for each subtype in order to add functional detail to each molecular profile. Array comparative genomic hybridization analysis added genomic detail to the subtype classifications. Comparative genomic hybridization analysis provided information about the DNA alterations featured in the samples of the defined subtypes. These analyses were applied to additional datasets featuring colon cancer samples and collected under different conditions in order to validate the robustness of defined classifications. Survival analyses of samples with stage II-III identity provided prognostic information to the molecular classifications. Stage I samples were exempt from the survival analysis, because these tumors were not likely to feature recurrence after detection and surgical intervention. Additionally, recurrence risk scores were computed to enhance the prognostic value of the defined subtypes.

**Data:**

Frozen colorectal tumor samples were taken from 750 patients in varying stages of prognosis. Samples were prepared using the following protocol: prior to RNA extraction, the sample was stored at -80C. total RNA extraction was achieved with Trizol. Biotinylated cRNA targets were prepared from 3.0 mg of total RNA. Following fragmentation,7 10 mg of cRNAs were hybridized for 16 hours at 45 C on human GeneChip HG-U133 Plus 2.0 Affymetrix arrays.

Of the 750 tumor samples, 566 fulfilled RNA quality requirements for GEP analysis. 443 of these 566 samples were used in a discovery set to create the classification, while the remainder were used in a validation set to test the classification. The validation set also included 906 colorectal tumor samples from seven public datasets (GSE13067, GSE13294, GSE14333, GSE17536/17537, GSE18088, GSE26682, and GSE33113).

All the CEL files aside from one were acquired from the SCC. The remaining CEL file was obtained by accessing the Gene Expression Omnibus (Sample GSM971958). Data in raw CEL files were normalized independently in batches using the robust multi-array average method (RMA), which is implemented in the Bioconductor R package - affy. For the CIT dataset, technical batch effects were corrected with the ComBat method implemented in the SVA R package.

**Methods:**

The following packages from Bioconductor were used on the CEL files of the data: affy, affyPLM, sva, AnnotationDbi, and hgu133plus2.db. The CEL files were read in using readAffy() and then in order to normalize the data the Robust Multiarray Averaging (rma) algorithm was used. The rma() function converts the Affybatch object into an expression set object.

The relative log expression (RLE) and normalized unscaled standard error (NUSE) was computed in order to check the quality control of the raw data using the Bioconductor package affyPLM. RLE plots are useful to visualize unwanted variation in high dimensional data and NUSE plots help to show the standard error estimates of probe level models for each probe set standardized across all chips.

ComBat, an sva package was used on the normalized data in order to control for batch effects. An annotation file which contained a host of clinical and batch annotations from Marisa et al. (2013) was used. In the annotation file normalizationcombatbatch was a variable that contained the batch effects from both Center and RNA extraction method and normalizationcombatmod was a variable containing features of interest from both tumor and MMR status. Both variables were used as arguments in the ComBat() function in order to correct the batch effects. The expression data was then written out in a CSV file where the probe sets were the rows, and the samples were the columns.

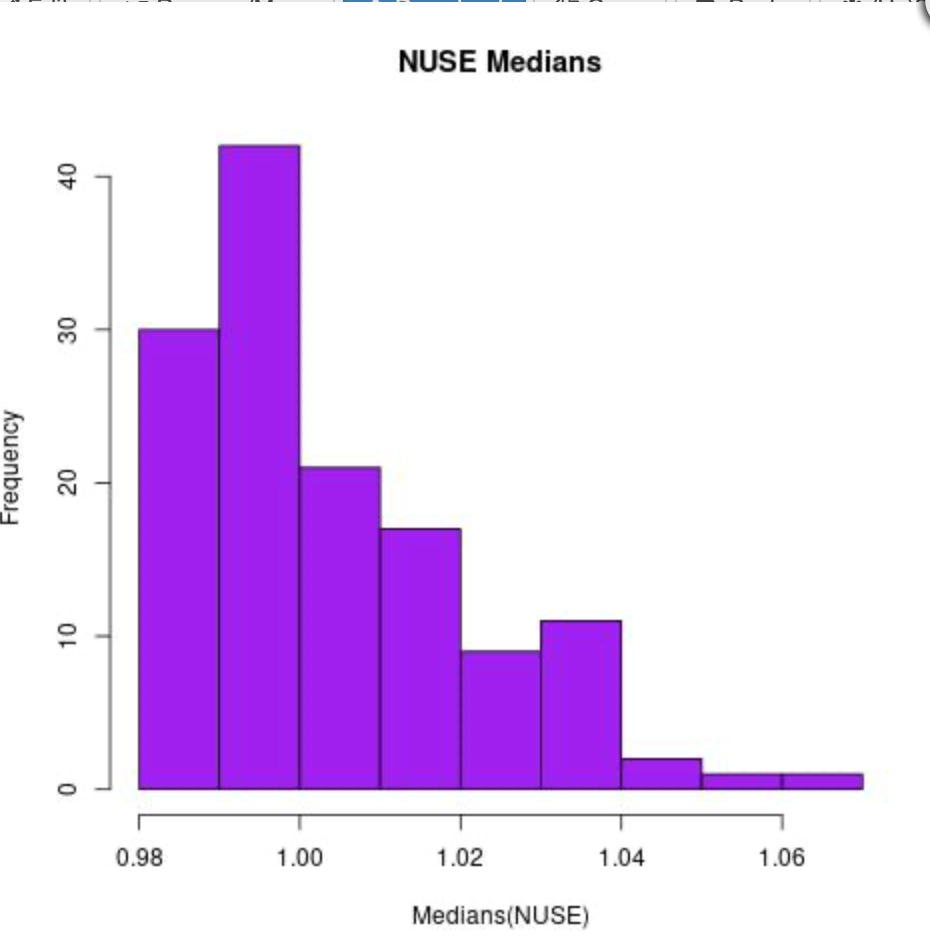
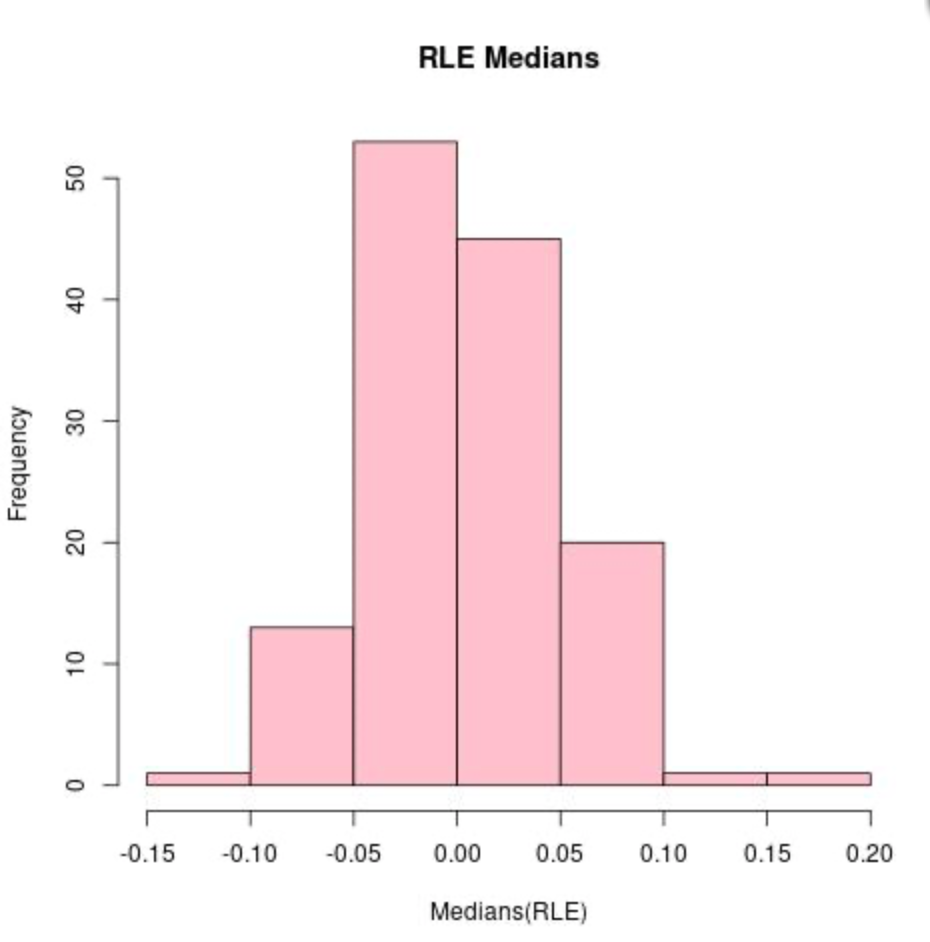
A principal component analysis (PCA) was performed on the normalized ComBat effect data to reduce the data dimensionality and extract the signal from data. Scaling was performed on each gene rather than within each sample by transposing the expression data then using the scale() function and finally re-transposing the data to its original orientation. The function prcomp() was used in order to perform the PCA analysis where scale and center variables were equal to false, since the data was already scaled. The first two principal components were plotted using the plot() function.

Genes featured in the expression matrix were filtered based on three criteria. The general purpose of this filtering was to increase the resolution of biological signals in the dataset and decrease the computational resources required for downstream analysis. Firstly, genes were filtered for a threshold expression level of > 𝑙𝑜𝑔2(15) across at least 20% of samples. This filter removed genes whose captured expression was insufficient for differential expression analysis downstream. These genes were undesirable for downstream analysis, because the presence of many lowly expressed genes in a dataset may reduce the power of statistical tests to identify differentially expressed genes (Sha et al.). Secondly, genes were filtered for expression variance across samples by comparing their variance to the median variance of all probes in the dataset using a chi-squared test. This filter removed genes whose variance across samples was not significant enough to likely represent interesting biological expression heterogeneity. The removal of these genes increased the likelihood that the more of the variance in the dataset represented meaningful biological expression heterogeneity (Hackstadt and Hess). Lastly, the genes were filtered for the coefficient of variation filter of > 0.186. This filter additionally increased the efficiency of downstream statistical testing for differentially expressed genes by removing genes whose variance was less likely representative of interesting biological expression heterogeneity.

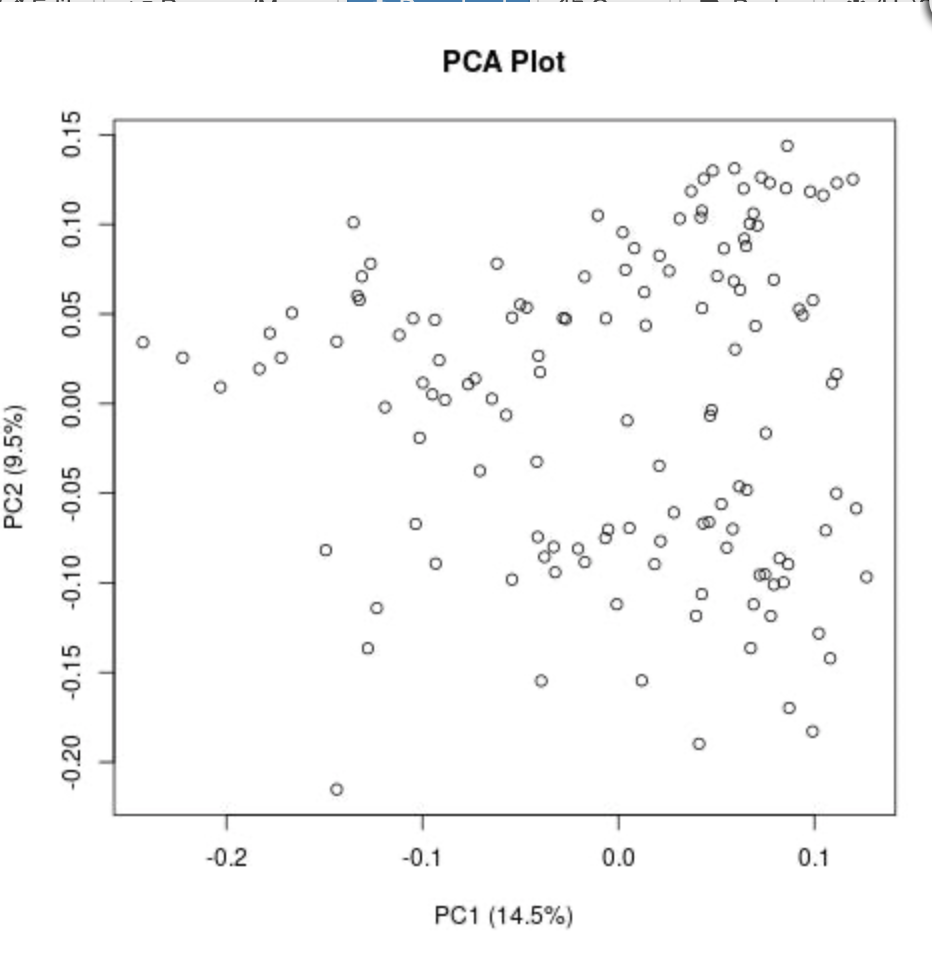
Sample clusters were identified and analyzed in the filtered expression matrix using functions from the Stats R package. Clusters were defined by computing a distance matrix from the expression data and applying the distance matrix to a hierarchical clustering analysis. The clustering analysis results were visualized in a dendrogram and cut such that each sample in the expression matrix was assigned to one of two clusters. A heatmap featuring a color bar that represented sample cancer subtype identity was used to visualize a comparison of gene expression and patient sample. This heatmap allowed for the visual inspection of relationships between sample clusters, sample cancer subtype identity, and gene expression. A Welch t-test between the two identified clusters was used to perform differential gene expression analysis with the filtered expression matrix. P-values resulting from the differential expression analysis were adjusted using the false discovery rate method. Adjusting the p-value was warranted, because multiple testing featured in the statistical analysis confounded the results with false positive inflation (Jafari and Ansari-Pour). The top ten percent of genes in the t-test results that featured a statistically significant adjusted p-value (< 0.05) were subset from the expression matrix to generate a list of genes whose expression was representative of the previously identified sample clusters.

**Results:**

Quality control measures were run on the raw data to look into the RLE and NUSE scores. In RLE, it subtracts the median intensity across all arrays from each probe. The histogram is centered around a median of 0.0 which signifies that the number of up and down regulated are the same, thus indicating that the samples could be of good quality (Figure 1 left). In the histogram that shows the median NUSE scores it is centered around a median of 1.0, which indicates that it could be that the sample is of high quality. After performing the RLE and NUSE quality controls it shows that there are no drastically different genes or outliers in the data so there were no genes removed. In the PCA that was performed, the first two principal components were plotted (Figure 2). Principal components 1 and 2 accounted for around 24% of the total variability. The PCA was performed in order to reduce dimensionality in the data.

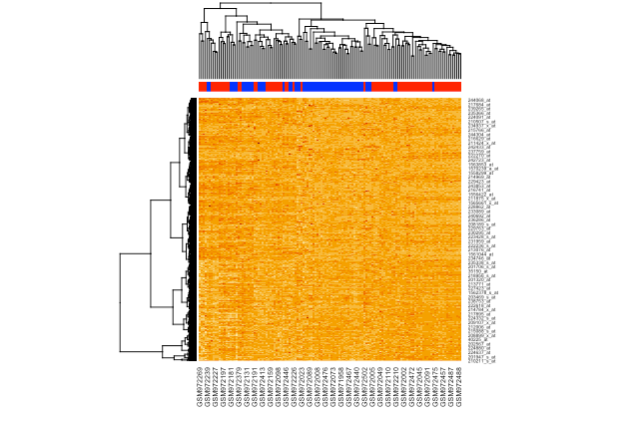


**Figure 1: (Left) Histogram of median RLE scores (Right) Histogram of median NUSE scores**



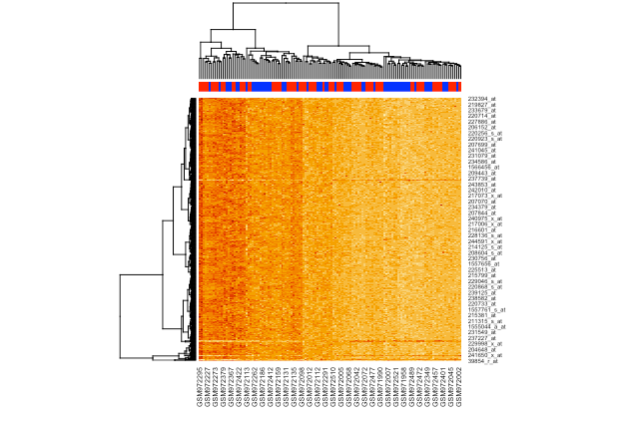
**Figure 2: Principal Component Analysis for Principal Component 1 and 2.** PC1 and PC2 account for 14.5% and 9.5% of the variance respectively.

Overall, 9649 genes passed the three filtering criteria. A filtered expression matrix was generated from the normalized and batch effect adjusted dataset using this list of genes. Classifications from the hierarchical clustering analysis were selected at a level that grouped samples into a cluster of 84 and a cluster of 50. A heatmap of gene expression across samples that featured color labeling of sample cancer subtype identity was generated from the filtered expression matrix (Fig. 3). A Welch t-test identified 6009 genes to be differentially expressed with an adjusted p-value < 0.05 between the two sample clusters. A list of 600 genes was subset from the 6009 differentially expressed genes to represent the gene expression heterogeneity between the two clusters. An additional heatmap was generated using the expression data of these 600 genes to qualitatively observe the gene expression differences among these genes across all samples (Fig. 4).



**Figure 3: Heatmap of Genes and Samples in Filtered Matrix.**

Gene expression does not clearly distinguish cancer by molecular subtype or hierarchical clusters. Hierarchical clusters are heterogeneous for molecular subtype.



**Figure 4: Heatmap of Genes and Samples in Top Ten Percent Differentially Expressed Genes.**

Gene expression shows clear differences between samples based on cluster identity. Samples do not show grouping by molecular subtype based on gene expression.

**Discussion:**

Two sample clusters were found to classify the gene expression profiles of 134 colon cancer samples. 6009 genes were found to be significantly differentially expressed between the clusters. The top ten percent of these differentially expressed genes were qualitatively observed feature expression profiles that distinguish the two sample clusters. These 600 genes were subset as a list that defined the clusters. The expression of the genes that best defines the two clusters was not homogenous within samples of the same molecular subtype identity. Therefore, the results of this analysis imply that the defined hierarchical clusters represented a distinct colon cancer classification from the molecular subtypes featured in sample metadata.

The difference of 1459 probe sets passed by the authors’ filters and 9649 probe sets passed by the filters described in this analysis may have been due to coding error or differences in experimental design. The authors described a discovery set of 443 samples and validation set of 1029 samples. 134 samples were analyzed in this experiment. Differences in filtering results may have arisen from this difference in sample size, because calculations of gene expression variance in samples under similar conditions may change with an increase in sample size (Stretch et al.). However, it was also possible that the integrity of the gene expression filtering was defeated by errors in the coding of filtering calculation. Although the author of the analysis script featured in this experiment did not detect errors in the code, the difficulty of the logical expressions relative to the authors skill may have introduced unidentified errors in filter calculations. In either of the cases described, the probe sets passed by the authors of the previous experiment would be interpreted to be more reliable for downstream molecular characterizations of colon cancer samples.

Removing a criteria from probe set filtering or removing probe set filtering entirely would have had the most impactful effect on the analysis results of this experiment. Altering probe set filtering would have introduced tests into the analysis that did not exist in the analysis that was performed. This would have allowed for results that were not a possibility in the experiment that occurred. Furthermore, the increase in probe sets featured in downstream analyses would have negatively impacted the performance of statistical tests. Therefore, analysis results would have featured probe sets whose expression was less representative of biological differences between samples.

**Conclusion:**

Two molecular subtypes of colon cancer were identified from differential expression analysis of tumor gene expression profiles. Hierarchical sample clustering did not reflect molecular subtype metadata annotation, and therefore sample clusters represented novel molecular classifications. A list of genes whose expression characterized the clusters was generated for future profiling and identification of the tumor classifications.

**References:**

1. Arvelo, Francisco et al. “Biology of colorectal cancer.” *Ecancermedicalscience* vol. 9 520. 9 Apr. 2015, doi:10.3332/ecancer.2015.520
2. Chakravarthi, Balabhadrapatruni V. S. K., et al. “Genomic and Epigenomic Alterations in Cancer.” *The American Journal of Pathology*, vol. 186, no. 7, July 2016, pp. 1724–35. *PubMed*, https://doi.org/10.1016/j.ajpath.2016.02.023.
3. Hackstadt, Amber J., and Ann M. Hess. “Filtering for Increased Power for Microarray Data Analysis.” *BMC Bioinformatics*, vol. 10, Jan. 2009, p. 11. *PubMed*, https://doi.org/10.1186/1471-2105-10-11.
4. Jafari, Mohieddin, and Naser Ansari-Pour. “Why, When and How to Adjust Your P Values?” *Cell Journal*, vol. 20, no. 4, Jan. 2019, pp. 604–07. *PubMed*, https://doi.org/10.22074/cellj.2019.5992.
5. Lancichinetti, Andrea, and Santo Fortunato. “Consensus Clustering in Complex Networks.” *Scientific Reports*, vol. 2, 2012, p. 336. *PubMed*, <https://doi.org/10.1038/srep00336>.
6. Marisa L, de Reyniès A, Duval A, Selves J, Gaub MP, Vescovo L, Etienne-Grimaldi MC, Schiappa R, Guenot D, Ayadi M, Kirzin S, Chazal M, Fléjou JF, Benchimol D, Berger A, Lagarde A, Pencreach E, Piard F, Elias D, Parc Y, Olschwang S, Milano G, Laurent-Puig P, Boige V. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. PLoS Med. 2013;10(5):e1001453. doi: 10.1371/journal.pmed.1001453. Epub 2013 May 21. PMID: 23700391; PMCID: PMC3660251.
7. Sha, Ying, et al. “Effect of Low-Expression Gene Filtering on Detection of Differentially Expressed Genes in RNA-Seq Data.” *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2015, 2015, pp. 6461–64. *PubMed*, https://doi.org/10.1109/EMBC.2015.7319872.
8. Stretch, Cynthia, et al. “Effects of Sample Size on Differential Gene Expression, Rank Order and Prediction Accuracy of a Gene Signature.” *PloS One*, vol. 8, no. 6, 2013, p. e65380. *PubMed*, https://doi.org/10.1371/journal.pone.0065380.