Microarray Based Tumor Classification

Introduction

Colorectal cancer, also known as colon cancer (CC), is the third most common cancer and the fourth-leading cause of cancer death worldwide [1]. Pathological staging, the only prognostic classification used in clinical practice, cannot predict recurrence accurately in many patients undergoing curative surgery [2]. Thus, researchers have been working hard to find a signature that can predict CC prognosis in clinical practice.

Microarray technologies allow for a quick and accurate profile of the expression of tens of thousands of genes at a time at a relatively low cost and many studies have exploited microarray technology to investigate gene expression profiles (GEPs) in CC in recent years. Among those studies, Marisa, L. et al. refined the previously commonly used molecular classification of CC that was based on common DNA markers and established a robust molecular classification by exploiting and analyzing the genome-wide mRNA expression CC samples data generated by microarray. Marisa, L. et al assessed the associations between 6 molecular subtypes but this reproduced microarray based tumor classification analysis of Marisa, L. et al. focuses only on the comparison of the C3 and C4 tumor subtypes.

Methods

To begin, the AffyMetrix microarrays underwent a multitude of processes via the utilization of RStudio and a series of BioConductor packages: sva, affy, affyPLM, AnnotationDbi, hgu133plus2.db, and bladderbatch. After the raw data was retrieved from the Gene Expression Omnibus (GEO) and uploaded to SCC, steps of preprocessing and quality control were performed. In order to read in the CEL file that contains all of the important probeset data information, the ReadAffy function was used. After the CEL file was read in, the Robust Multi-Array average (rma) function was used in order to properly normalize all of the CEL files in conjunction. There are three critical steps that RStudio takes in order to normalize the CEL files. First, the CEL files undergo a background correction, which is responsible for removing any noise or artifacts. Then, the files go through data normalization, which adjusts probes along a multitude of arrays so that they can be compared effectively. Third, they undergo a probe summarization, where probe intensities are grouped together and expressed in a gene level.

Next, two critical quality control methodologies were used on the dataset in order to properly evaluate and interpret the microarray data: Relative Log Expression (RLE) and Normalized Unscaled Standard Error (NUSE). In order to calculate the median RLE values, RStudio median intensities are subtracted from the probes. Generally speaking, median RLE values that are close to 0 imply quality data. On the other hand, in order to calculate the median NUSE values, the standard error is divided by the product of the median expression and standard error. Typically, median NUSE values that are less than 1.00 are considered to be quality data. The hist() function was utilized in order to produce histograms of the median RLE and NUSE of the data.

Furthermore, the ComBat function, which is a part of the BioConductor package, was utilized in order to fix any batch effects that were located within the normalized datasets that were read in from the CEL files. The characteristics that were specifically focused upon were tumor and MMR status. These characteristics were grouped together into a single variable: normalizationcombatmod. On the other hand, batch effects such as RNA extraction and Center methods were combined into a single variable: normalizationcombatbatch. The ComBat function eliminated batch effects and preserved information of interest in the expression data. The expression data was then written out into a CSV file via the write.csv function.

The expression data was properly transposed via the t() function, scaled via the scale() function, and transposed in order to get the original orientation of the expression data. The variance of the expression data was found in order to determine the degree of spread of the values from the mean. The variance was converted into a percentage and then plotted as a barplot in order to visualize the variance percentage.

Onwards, in order to properly visualize the batchless expression data, a Principal Component Analysis (PCA) was conducted. A PCA plot containing outliers was produced via the ggplot function that is located within the ggplot2 library. This PCA plot was produced in order to effectively visualize PC1 and PC2 data as well as their outliers. PC1 and PC2 were individually plotted as boxplots via the ggplot function in order to better support the conclusion that outliers were located roughly 3 standard deviations away from the mean. The outliers that were located within both the PC1 and PC2 plots were removed and the prcomp function, with both scale and center set to FALSE, was then used in order to observe the values of both PC1 and PC2. The filtered expression data with no outliers was then plotted via the ggplot function in order to visualize the data.

Three filters were then applied to the pre-processed data, with these filters being the ones suggested by Marisa, et al. The filters were applied so that probe sets were expressed in at least 20% of the samples, had a significant difference in variance from the median variance of all probe sets (computed using a one-tailed Chi-squared test at $p<0.01$), and had a coefficient of variance greater than 0.186. This resulted in expression matrices, one which passed all three filters and one which only passed the second filter. These filters were necessary in order to reduce noise and reduce dimensionality, in order to ensure that multivariate analysis did not yield meaningless results.

Hierarchical clustering was then performed on the expression matrix which passed all three filters, with two patient clusters being formed. The accuracy of the clustering was then verified by a heatmap of patients with and without the “C3” subtype of colon cancer. The number of genes differentially expressed between the two clusters were then found using a Welch t-test, with the most representative genes from each cluster identified. Hierarchical clustering was chosen due to computational constraints, with the goal of identifying the true number of clusters in the sample.

The Welch t-test results were used to select the top 1000 up and down regulated genes, with the purpose of implementing a fisher test. Three gene set collections, KEGG, GO, and Hallmark sets, were downloaded to compare overlap with the Welch results for the Fisher Test. For the Fisher test, the number of differentially expressed genes needed to be determined as well. A table of the top 10 regulated genes were also made.

Results

First, it is important to understand that Relative Log Expression (RLE) and Normalized Unscaled Standard Error (NUSE) are key quality control methodologists of bioinformaticians in order to evaluate and interpret microarray data. In general sense, RLE values that are close to 0 indicate quality data. Based on Figure 1, it is clear that a majority of the median RLE values hover around or are equal to 0.00, thus indicating the experiment has quality samples. On the other hand, median NUSE values that are less than 1.00 are considered quality data, while NUSE values that are greater than 1.00 may be considered poor data. Figure 2 illustrates that a majority of the median NUSE values are located near or below 1.00, meaning that the arrays are high in quality.

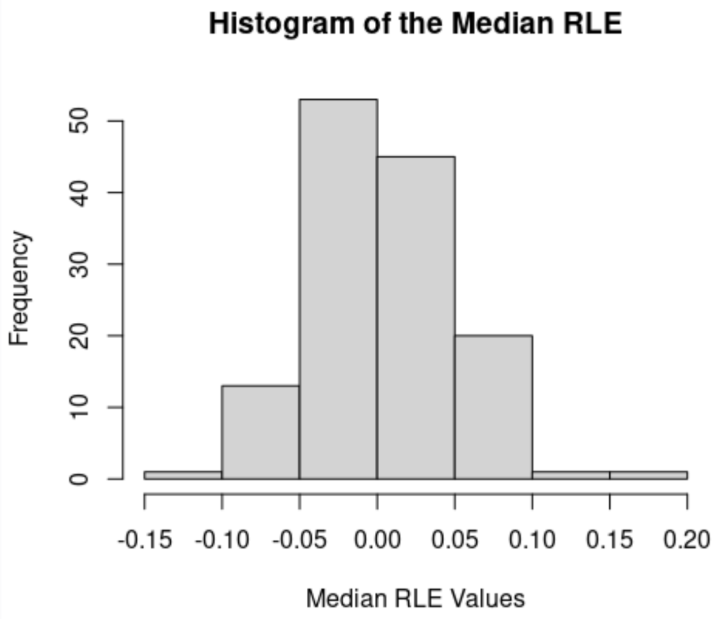


Figure 1. Histogram Depicting the Median Relative Log Expression (RLE) of the samples. The image illustrates the median RLE Values from -0.15 to 0.20 for the samples in the experiment.

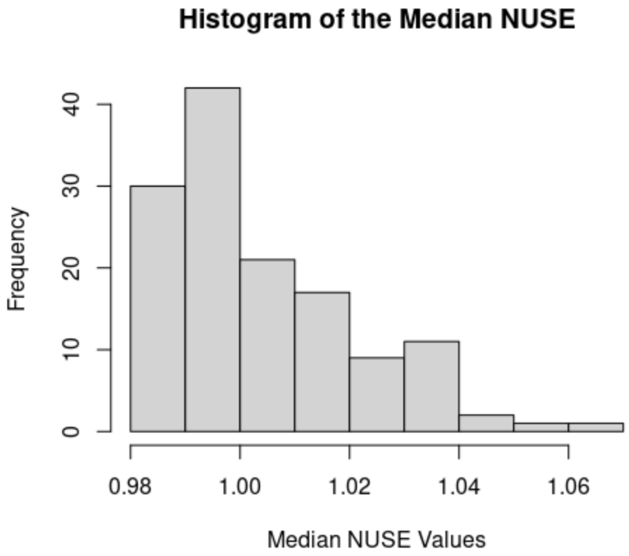


Figure 2. Histogram Depicting the median Normalized Unscaled Standard Error (NUSE) of the samples. The image illustrates the median NUSE values from 0.98 to 1.06 for the samples in the experiment.

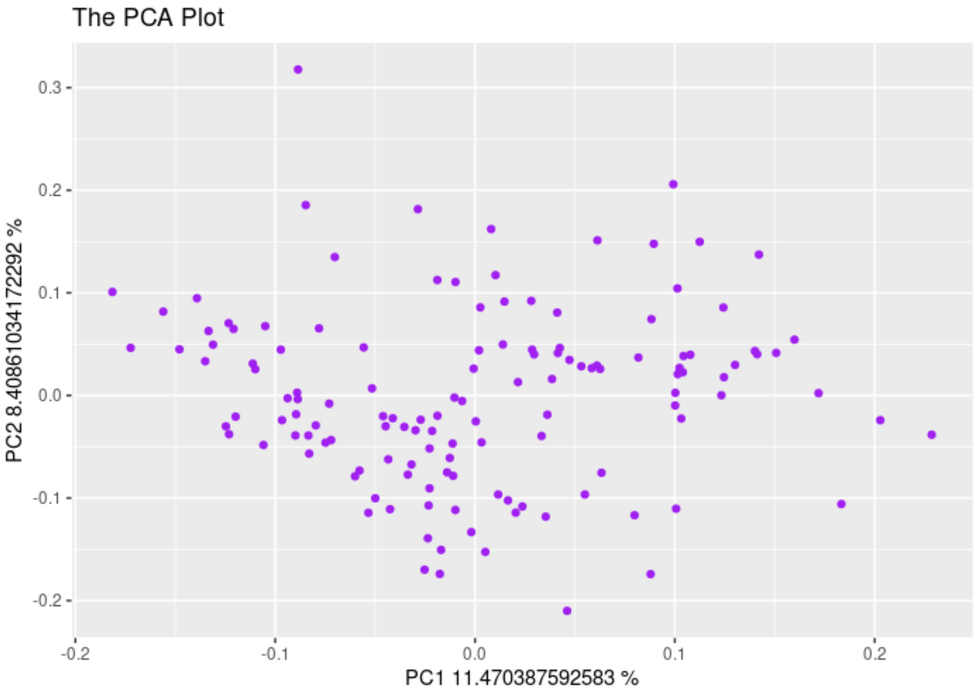


Figure 3. PCA Plot with PC1 and PC2. The PCA plot illustrates the variability of both PC1 andPC2, however, includes outliers.

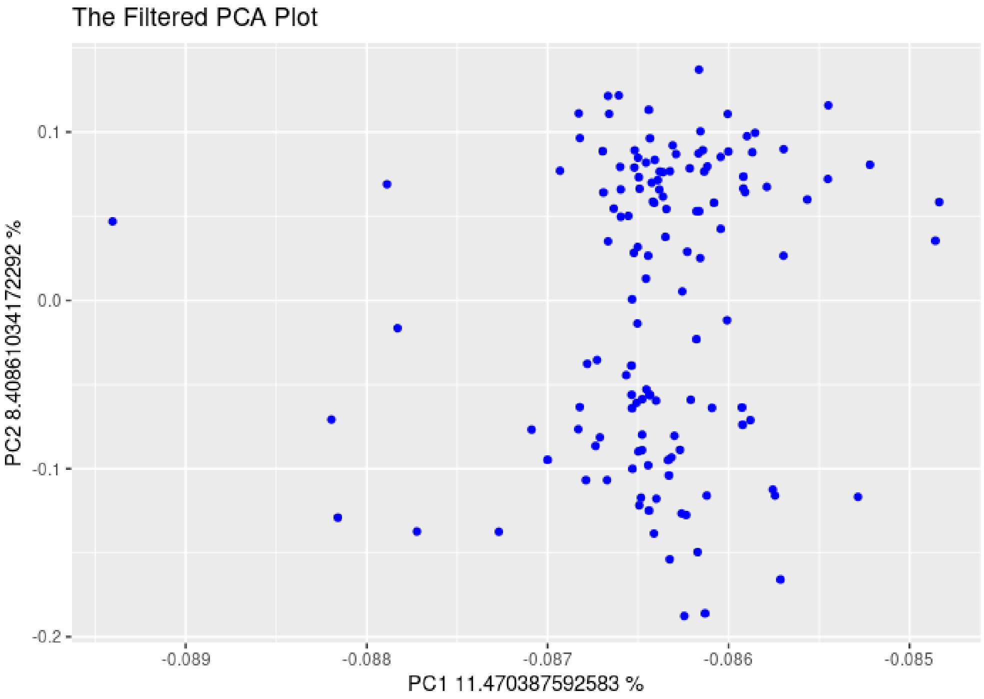


Figure 4. PCA Plot with PC1 and PC2. The PCA plot visualizes the variability of the principal components, however, does not include outliers.

Principal Component Analysis was conducted in order to effectively view the variation amongst The key difference between Figure 3 and Figure 4 is that Figure 5 shows the outliers present in the expression data, while Figure 6 does not. After the outliers were eliminated, another PCA plot was generated in order to show the variance amongst PC1 and PC2. PC1 and PC2. PC1(11.47%) and PC2(8.41%) only account for 20% of the data variance,but in figure 6, after the outliers are eliminated, the points started to show a trend of grouped into two subtypes along PC2.

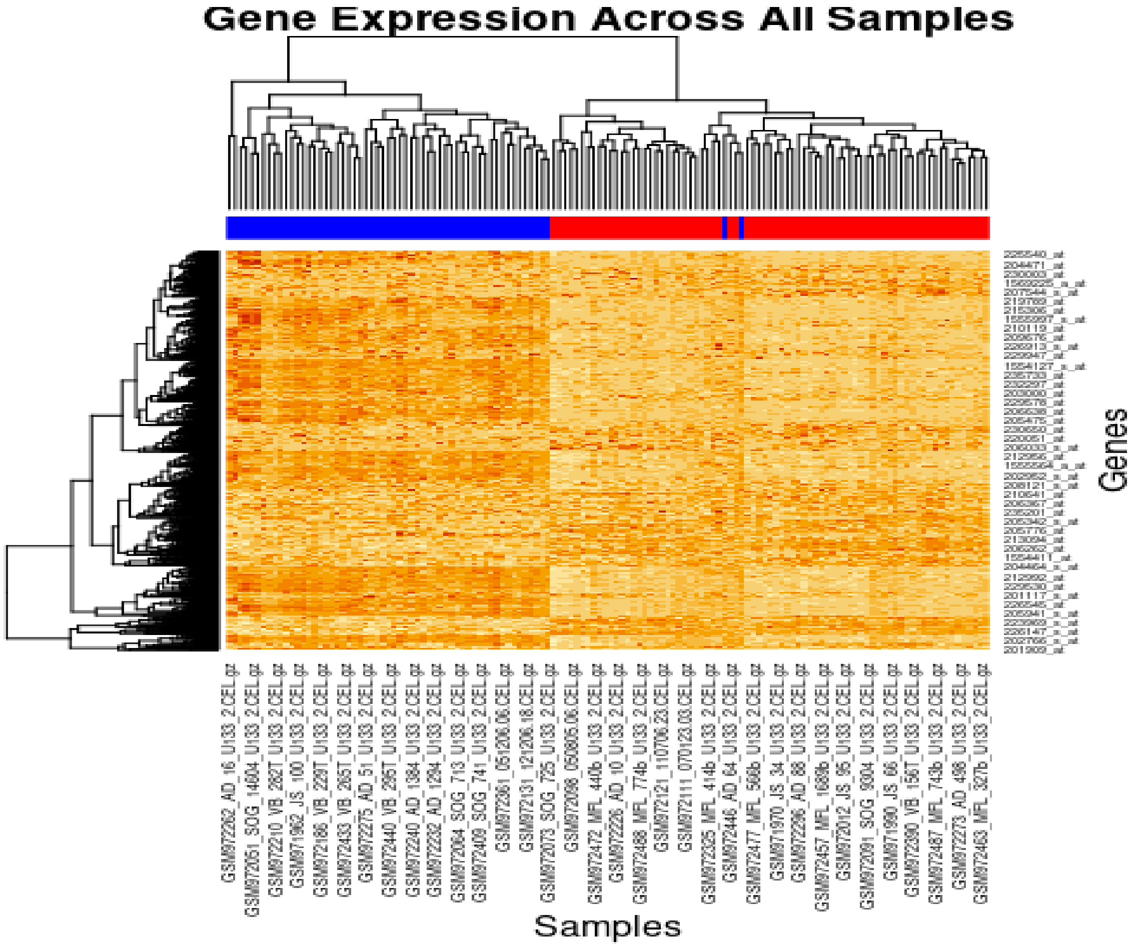
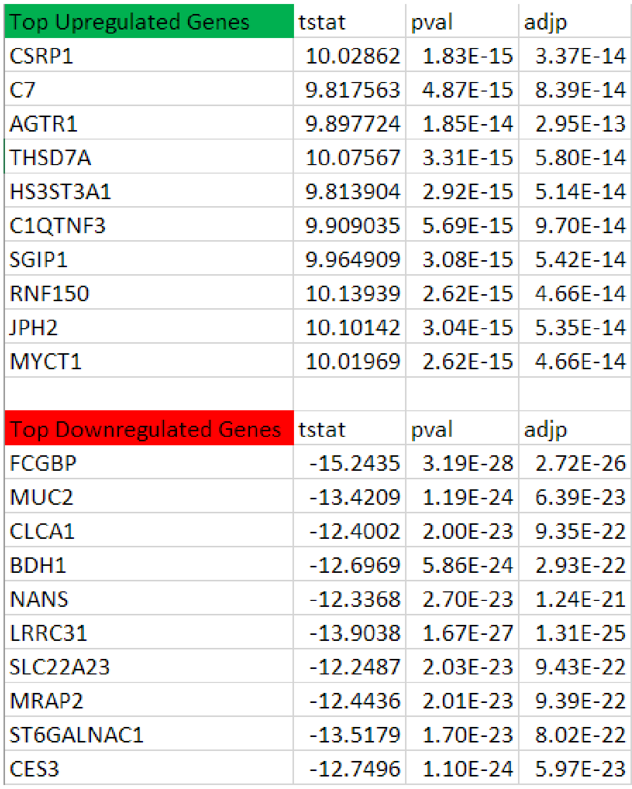


Figure 5. Gene Expression Analysis Across All Samples. This heat map illustrates the results of hierarchical clustering, displaying the differing levels of gene expression when comparing samples of C3 subtype to those not of the C3 subtype.

The RMA normalized gene expression matrix, comprising ~55,000 genes and 134 samples was filtered, with a significant reduction in dimensionality. The resultant number of genes after the first filter was 39562, then 34912 after Chi-squared testing, then finally 1522 genes were identified to have passed all three filters. The variability in the results from both the paper, and other similar methodologies stems from the lack of specification as to whether the Chi-squared test should be lower, upper, or two-tailed. In this case, the filter used was a lower-tailed test which may produce results differing from other tests.

The resulting number of samples from hierarchical clustering was 56 in one cluster and 78 in the other, but more interestingly, the heatmap indicates that there were two samples which were misidentified. This error was most likely due to the two misidentified samples containing gene expression values similar enough to the C3 samples that there was an agglomeration error. Hierarchical clustering is a method known to be sensitive to both noise and outliers, so the two misidentified samples could be considered outliers in this case.



Top Upregulated Genes	tstat	pval	adjp
CSRP1	10.02862	1.83E-15	3.37E-14
C7	9.817563	4.87E-15	8.39E-14
AGTR1	9.897724	1.85E-14	2.95E-13
THSD7A	10.07567	3.31E-15	5.80E-14
HS3ST3A1	9.813904	2.92E-15	5.14E-14
C1QTNF3	9.909035	5.69E-15	9.70E-14
SGIP1	9.964909	3.08E-15	5.42E-14
RNF150	10.13939	2.62E-15	4.66E-14
JPH2	10.10142	3.04E-15	5.35E-14
MYCT1	10.01969	2.62E-15	4.66E-14
Top Downregulated Genes	tstat	pval	adjp
FCGBP	-15.2435	3.19E-28	2.72E-26
MUC2	-13.4209	1.19E-24	6.39E-23
CLCA1	-12.4002	2.00E-23	9.35E-22
BDH1	-12.6969	5.86E-24	2.93E-22
NANS	-12.3368	2.70E-23	1.24E-21
LRRC31	-13.9038	1.67E-27	1.31E-25
SLC22A23	-12.2487	2.03E-23	9.43E-22
MRAP2	-12.4436	2.01E-23	9.39E-22
ST6GALNAC1	-13.5179	1.70E-23	8.02E-22
CES3	-12.7496	1.10E-24	5.97E-23

Figure 6. Top Ten Up- and Down-regulated genes. This table displays the top regulated 10 genes sorted by positive and negative log2 fold change, respective of up and down regulated genes.

The final step of the analysis was identifying the number of differentially expressed genes for those which passed all three aforementioned filters versus those which only passed the first two filters, which were 1,235 and 20,123 respectively.

204457_s_at, 209868_s_at, 223122_s_at, 226930_at, and 218694_at were the five most differentially expressed genes with the lowest p-values while 205518_s_at, 235740_at, and 204042_at were the most representative of the clusters with the highest absolute t-values.

Discussion

Throughout the reenactment of the *Marisa et al. Gene Expression Classification of Color Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value* experiment, the group read in the CEL files garnered from the Gene Expression Omnibus, preprocessed and quality checked the expression data, implemented a series of filters on the combat altered and rma normalized expression data, utilized a hierarchical clustering algorithm in order to visualize the grouping of samples, and finally utilized gene enrichment analysis in order to determine the biological importance of the gene expression profiles. 1,522 genes were identified to have passed all filters, with 132 out of 134 samples correctly classified in accordance with the subtype of colon cancer exhibited.

The implication of the relative success of this methodology means that different subtypes of cancer exhibit different gene expression patterns, something which can be used in the future in order to improve diagnosis. This could also mean that different types of cancer could be potentially identified by gene expression pattern alone, which could lead to faster, and more reliable methods of detection.

The biological interpretations closely align with the main implications of this methodology, which is that different subtypes of CRC exhibit different gene expression patterns, or in this case, the C3 subtype of CRC produced expression data unique enough for a 98.5% correct prediction through hierarchical clustering. The fact that these accurate predictions were made using only 1,522 out of the original ~55,000 genes suggests that the expression of these 1,522 genes could be the identifying factor of the C3 cluster. If put in a broader context, this would suggest that only 2% of the genes linked to CRC would affect the subtype of the cancer, and further studying these genes could help insight into the molecular mechanisms of CRC.

The results from the original paper were not completely reproduced mainly due to the ambiguity of the Chi-squared used in the filtering step and use of a different clustering technique. Marisa et. al did not identify which Chi-squared test was used in order to filter out genes with significant deviations in variance from the overall variance of the dataset, and the use of a lower-tailed test in this reproduction would lead to potential false positives and negatives. Furthermore, due to computational constraints the clustering technique used in this study was not as accurate as the model used by

Marisa et. al, resulting in the two misclassified samples. Overall, this reproduction was able to obtain results which were within the margin of error of the original study (2-3%).

Conclusion

Finally, after looking back at the results and experiment as a whole, our group was not able to completely replicate the results of the *Marisa et al. Gene Expression Classification of Color Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value* experiment due to a plethora of reasons such as: filtering errors with the Chi-Squared test as well as errors in the techniques used in order to cluster our data. Although there were errors in our replication of the experiment, our group concluded that gene expression patterns vary based on the type of CC cancer. The hierarchical clustering supports the case that the C3 subtype of the CC cancer produced a 98.5% correct recurrence prediction. In order to bolster the information that was conducted in this experiment, more samples from different parts of the world should be collected and different subtypes within different types of cancers should be analyzed in order to garner a more holistic analysis.

Our group as a whole faced challenges that ranged anywhere from bugs in code to understanding biological concepts that were key in analyzing the expression data. These problems were overcome by having in-depth discussions over the biological concepts and also conversing with Dr. Labadorf and TA's in order to solve any difficulties we encountered.

References

1. Greenlee RT, Murray T, Bolden S, Wingo PA (2000) Cancer statistics, 2000. *CA Cancer J Clin* 50: 7–33. [PubMed] [Google Scholar]

2. 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