**Project 1 Report**

Cody Webb, Yueh-ting Wang, and Chris Lin

**Introduction**

Colorectal cancer is one of the top five most common cancers in the world, both in occurrence and in total mortality. The current method used to classify colorectal cancer classifies its severity in a stage system, with Stage I Colorectal Cancer (CRC) being the least severe and Stage IV being the most severe. This classification is then used to select patients for chemotherapy. However, this staging system does not seem to predict recurrence risk accurately, as 10-20% of Stage II patients and 30-40 of Stage III patients develop recurrence.

This study was carried out in an attempt to find markers that would better classify CRC cases. Most previous Gene Expression Profile studies had not produced the result of finding better biomarkers for predicting CRC severity, and often were poorly reproducible. However, studies that used unsupervised hierarchical clustering had identified at least 3 distinct molecular subtypes of CRC. These markers had not yet been standardized and reproducible, so this study set out to create a system using unsupervised hierarchical clustering to refine these data and provide a potential standard for CRC classification.

**Data**

The Cartes d'Identité des Tumeurs program collected fresh-frozen primary tissue samples from 750 patients with Stage I-IV colon cancer from seven different French hospitals between 1987 and 2007. These were characterized for common DNA alterations and were also screened with whole genome and transcriptome arrays using Affymetrix U133 Plus 2.0 microarrays. Of those 750 samples, 566 were determined to be of high enough RNA quality that they could be used for gene expression profiling. These 566 samples were split into a discovery set of size 443 and a validation set of size 123.

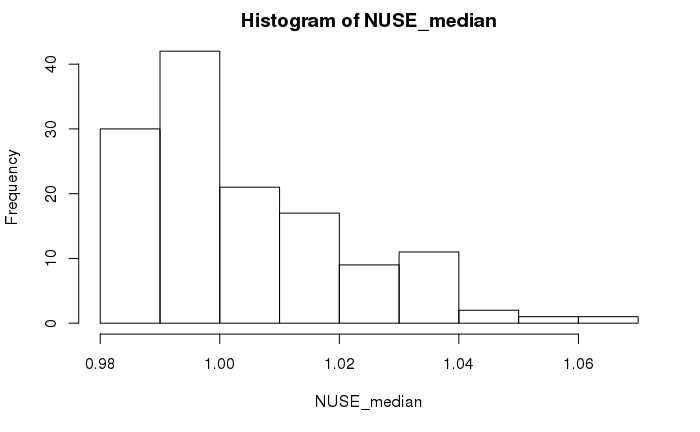
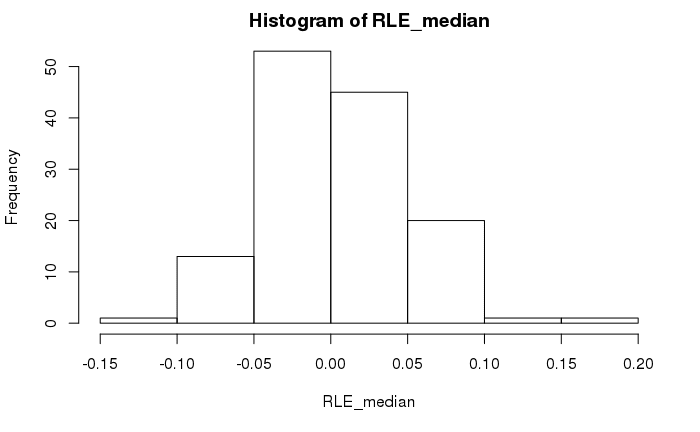
To have a more robust experiment, the validation set included samples from the following datasets: GSE13067, GSE13294, GSE14333, GSE17536/17537, GSE18088, GSE26682, and GSE33113. These datasets were chosen because they were profiled using a similar microarray, there was tumor location information, and there was either patient outcome data available or common DNA alteration data available. The dataset from the Cancer Genome Atlas was also included despite the fact that its screening took place on a non-Affymetrix platform because it included extensive DNA alteration annotation, and thus was analyzed separately.

After the inclusion of these datasets, the validation set was of size 1,029. In the discovery set and the validation set, 359 and 416 patients, respectively, had stage II or stage III CRC and had documented relapse free survival information available, which made them suitable for a survivability analysis. Finally, the data were normalized using the R package affy, and batch effects were removed via the ComBat method in the R package sva.

**Methods**

Data preprocessing and normalization was performed using R version 3.5.1 on the SCC RStudio interactive server. Although the instructions suggest the use of R version 3.1.1, this specific version was not available on the SCC. First, the compressed CEL files were read into R using the ReadAffy() function. The data files were then normalized using the rma() function, ensuring that variation between arrays are corrected for.

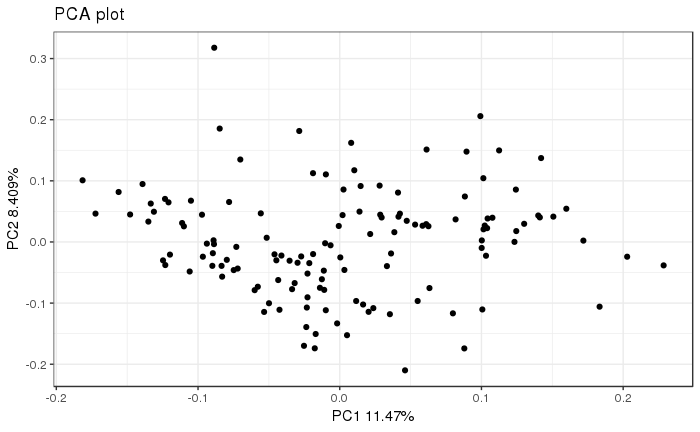
Next, the Bioconductor package affyPLM was used, converting our original data in the form of an AffyBatch into a PLMset. This PLMset was then used to visualize the quality of our data. Both the Relative Log Expression (RLE) and Normalized Unscaled Standard Error (NUSE) scores of the dataset were calculated, and the median was computed (Fig 1A, B). RLE is obtained by subtracting the median intensity across all arrays from each probe, and NUSE is obtained from normalized standard error estimates from the probe level model. The median NUSE score was generally distributed around 1.00, skewed right, while the median RLE score was generally normally distributed around 0. This suggests that the majority of our samples are of high quality, with very few samples having a median RLE > 0.10, or a median NUSE > 1.05.



**Fig. 1 Histogram of median NUSE and RLE scores.** Histograms show the distribution of median RLE and NUSE scores.

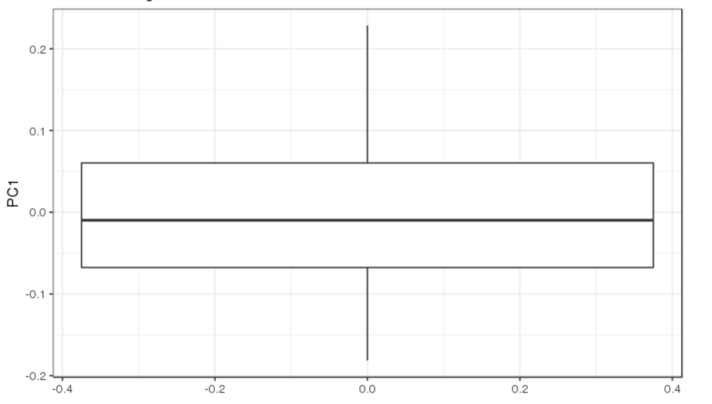
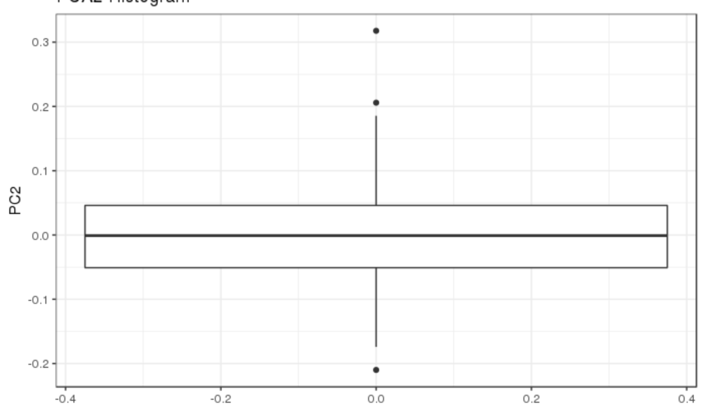
We then used the ComBat package to correct for batch effects. A metadata csv file was provided to us, containing many clinical and batching annotations. The specific columns of interest to us include ‘normalizedcombatbatch’ and ‘normalizedcombatmod’. The features tumor status and MMR status have been merged into ‘normalizedcombatmod’, and center and RNA extraction method have been merged into ‘normalizedcombatbatch’. These two columns were extracted and used as input variables to the ComBat function, along with our normalized expression data from RMA to correct for the stated batch effects. The results of the ComBat were then written to a CSV file.

Principal Component Analysis was then performed on our ComBat output as a way of visualizing and quality controlling our data via dimensionality reduction. Our data was first transposed, then scaled and centered, and transposed back to its original orientation in preparation for PCA. The prcomp function was then used to perform PCA on our scaled and centered data, with the options ‘scale’ and ‘center’ set to false in prcomp(). When our data was plotted along PC1 and PC2, 11.47% and 8.409% of the variance was explained by these two components respectively (Fig 2). Although no clear clustering can be seen, overall an 19.88% explanation of variance in two principal components for 134 variables is a promising result.



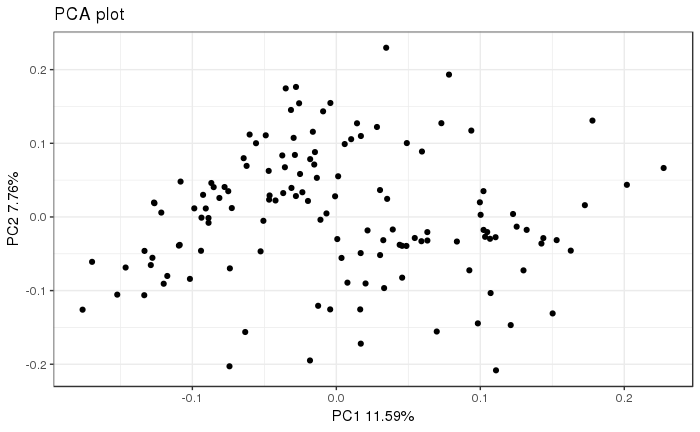
**Fig. 2 Principle component analysis.** Each dot represents the expression profile of 134 genes.

Some outliers can be seen in this PCA plot, so a boxplot was created for both PC1 and PC2 to more clearly identify any outliers (Fig 3).



**Fig. 3 Boxplots of PC1 and PC2.** Points outside of the interquartile range for both PC1 and PC2 are plotted in each boxplot.

PC1 seems to have no outliers outside of the IQR, while PC2 contains 3 outliers. To remove any outliers that may affect our analysis, all values that were outside of 3 standard deviations of the mean in both PC1 and PC2 were removed. This resulted in one sample being removed from our dataset. PCA was then performed again as to characterize the effect of removal of this outlier (Fig 4).



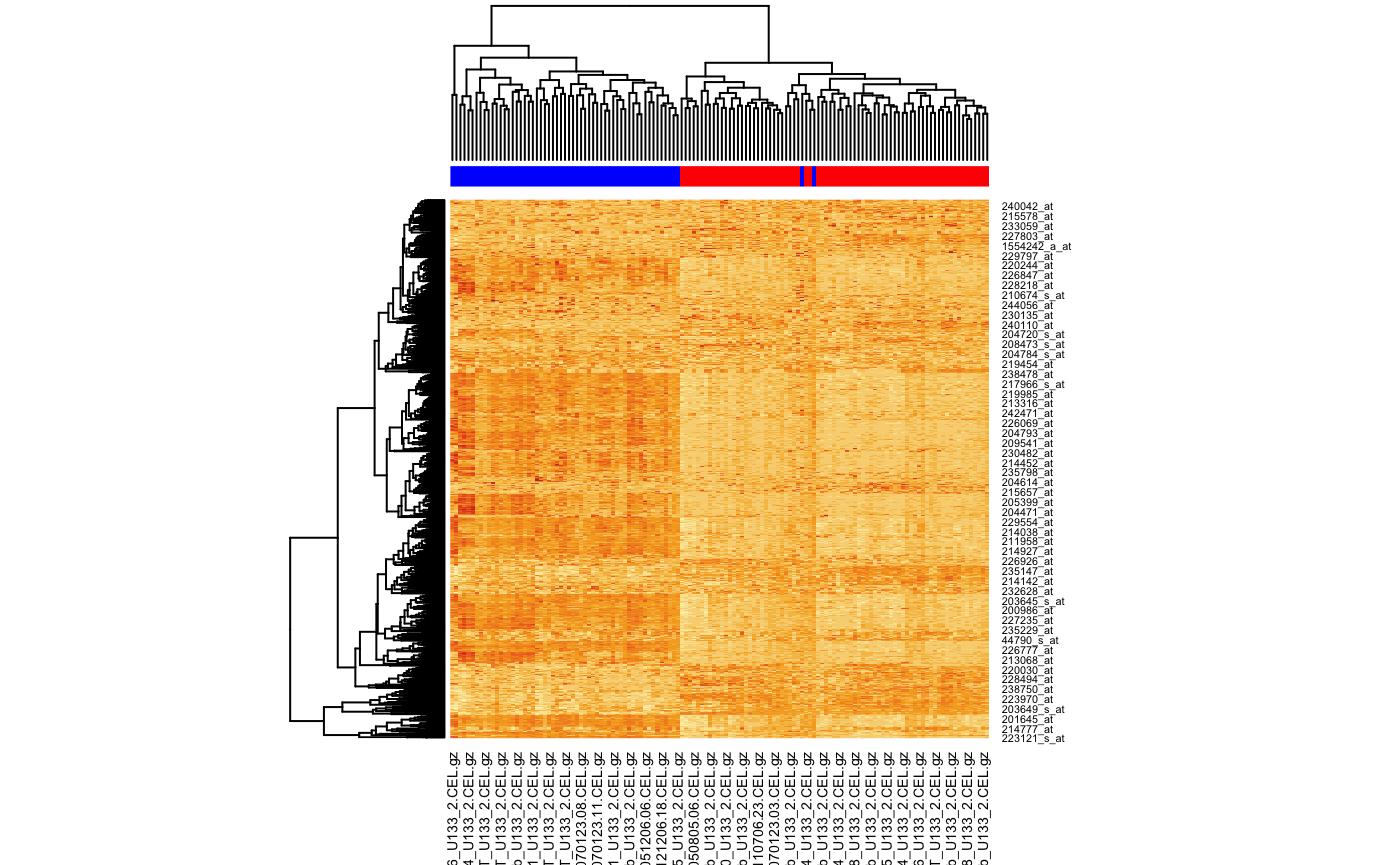
**Fig. 4 Principal component analysis of all points within 3 standard deviations of the mean.** Each dot represents the expression profile of 133 genes.

PC1 showed a slight improvement in variance explained, increasing to 11.59% from 11.47%. PC2 showed a decrease in variance explained, decreasing from 8.409 % to 7.76%. The presence of this outlier seems to have not made a significant change in our data, however, a CSV file with the outlier removed was also generated. Both CSV files, containing 134 and 133 genes, respectively were sent to the data analyst.

**Results**

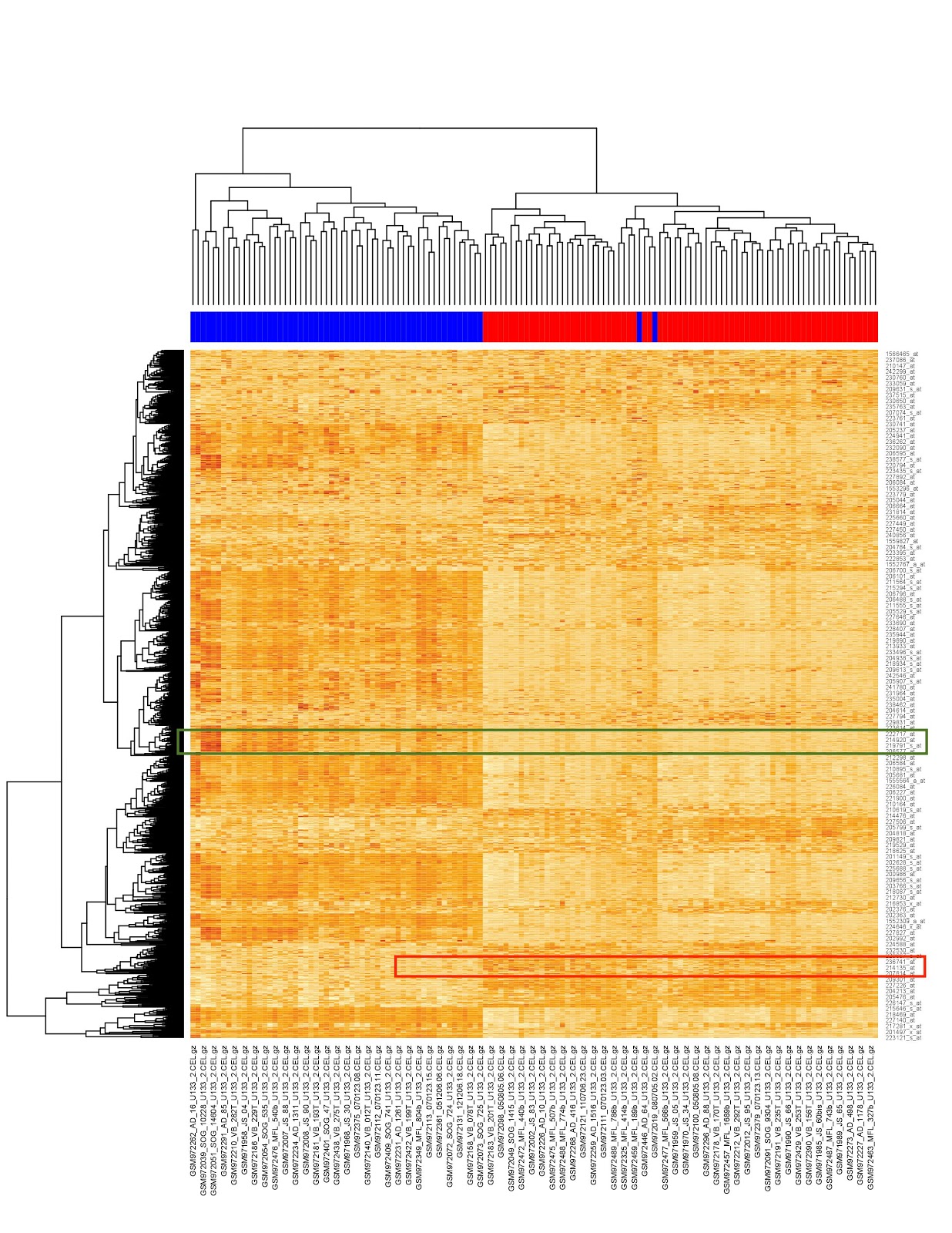
After receiving the data from the programmer, there were 54,675 genes and 134 samples available for analysis. After searching for genes that were expressed in at least 20% of samples after using the substrate function, there were 39,661 genes left in our data set. Then, we performed a one-tailed chi-squared test in order to find the genes that were differentially expressed. After this test, there were now 15,508 genes in total. Finally, we took the genes that had a coefficient of variation that was greater than 0.186, leaving us with 1,531 genes to do our final analysis on. All the filtered results are saved as “filtered\_results.csv”.

Hierarchical clustering was done using the Euclidean and Ward method. The samples were separated into two clusters, with 57 in one group and 77 in the other. The heat map below shows the clusters, with red indicating that the sample is from the C3 subtype of the original paper, and blue indicating that it is from some other subtype.



**Fig. 5** **Heat map of the gene expression data.** This heat map shows the results of the hierarchical clustering analysis. The samples classified as C3 subtype in the original paper are marked in blue, and those classified as not being the C3 subtype are classified in red.

In order to identify genes that were differentially expressed between the two clusters, a Welch t-test was used, with the p-value cut-off being 0.05. There were 1,012 genes that were differentially expressed at a p<0.05 level between the two clusters. All the results of the Welch t-test for all genes are saved as “welch\_t-test.csv”. We used the adjusted p-value and sorted from minimum to maximum. Using this sorted list we were then able to select the three most differentially expressed genes between clusters. Here are the top ten genes that are the most differentially expressed, 204457\_s\_at, 209868\_s\_at, 223122\_s\_at, 225242\_s\_at, 202291\_s\_at, 227059\_at, 218694\_at, 225782\_at, 226930\_at, 202363\_at.



**Fig 6 Heat map of the selected region** This heat map shows the results of genes that represent the clusters. The selected area in green represents the genes that are most differentially expressed in the C3 subtype cluster, and the selected area in red represents the genes for the non-C3 subtype cluster.

 Lastly, based on the hierarchical clustering heatmap, we choose the genes in the color that appear to be the most distinct as the genes that represent the cluster.

These genes are significant since these will be the genes that are the least similar to the other cluster, and will likely be able to be used as a benchmark for what we want our clusters to be. The genes that were the most differentially expressed in the C3 cluster were 222717\_at, 214920\_at, 219791\_s\_at, and 206577\_at. The genes that were the most differentially expressed in the non-C3 cluster were 236741\_at, 214135\_at, and 207814\_at.

**Discussion**

We found that 98.5% (132 of 134) of the samples were clustered correctly, with all samples that were not in C3 being classified as such and all but 2 of the C3 subtype samples being classified together. This seems to corroborate the original paper’s claim about the clustering in regards to C3. However, the fact that two of the C3 subtype samples were classified as not being in C3 is somewhat strange, and should be looked into further.

This discrepancy may be a result of differences in the filtering of outliers, as we used a subjective method involving PCA and standard deviation ranges to determine whether points were outliers. Other ways of determining outliers in the initial data normalization and quality control steps, such as RLE, NUSE score, and RIN may result in different downstream analysis results. In a similar vein, the clustering method we used--Euclidean and Ward-- is one of many ways to cluster data. Since we used hierarchical clustering instead of consensus clustering, this may explain our discrepancy.

**Conclusion**

Given the gene expression data, we can confidently say that the C3 subtype from the original paper is a valid classification. Both our analysis and the original paper have limited applicability, as all subjects were from France. Differences in lifestyle and eating habits may play a strong role in gene expression, which could impact the data and the overall conclusion. That being said, our results do confirm the results from the original paper, which shows that even when using different clustering methods, it is possible to arrive at very similar results.

**References**

[1] Marisa L, de Reynie`s A, Duval A, Selves J, Gaub MP, et al. (2013) Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value. PLoS Med 10(5): e1001453. doi:10.1371/journal.pmed.1001453