**Project 1 Write-up**

Team: Frazzled

Contributors: Srija Chillamcherla, Kyra Griffin, Adalee Koshiol, Vridhi Vinaykiya

**Introduction**

Colorectal cancer is the third most common cancer, and therefore, a large contributor to death worldwide. So far, there are no reliable ways of predicting recurrence of colorectal cancer in patients after they have undergone curative methods even though recurrence is somewhat common. Gene expression profiling could be a viable method for prognosis of colorectal cancer, though the distinct molecular pathways of CRC make it difficult for studies examining this to be reproduced. More recent studies have shown there to be three molecular subtypes of colorectal cancer, which could maybe cause the difficulties in research reproducibility. It also shows that the molecular classification of colorectal cancer needs revising. The authors of Marisa et al sought to dive deeper into the associations between the found molecular subtypes in order to refine how colorectal cancer is thought of on a molecular level. They did this through a genome-wide mRNA expression analysis. The findings of Marisa et al. could provide insights that would lead to more reliable methods of prognosis of recurrence of colorectal cancer. Here, we sought to reproduce the results of their study.

**Data**

The data were collected and frozen at various institutions from one cohort of 750 patients with colorectal cancer. The patients’ stage of cancer ranged anywhere from I to IV. None of the patients had received chemotherapy or radiation therapy, nor were they patients with primary rectal cancer. Only 566 samples from the cohort fit requirements for gene expression profile analysis, and these samples were then split into two sets: a discovery set with 443 samples, and a validation set with 123 samples. Within the validation set, there were 906 more colorectal samples that were from seven public datasets (GSE13067, GSE13294, GSE14333, GSE17536/17537, GSE18088, GSE26682, and GSE33113). All data were collected with a similar chip platform (Affymetrix U133 Plus 2.0 chips) and came from raw CEL files. Data from The Cancer Genome Atlas was analyzed separately as it was obtained with a non-Affymetrix platform. In addition, within the discovery set, there were 359 patients with stage II or III cancer that were available for survival analysis. For the validation set, there were 416.

**Methods**

All the methods from normalization to visualizing the data were done in Rstudio (version 1.4, R version 4.0.2). R repositories, CRAN and Biomanager 3.14, packages like affy, affyPLM, sva, AnnotationDbi and hgu133plus2.db were installed in R. Affymetrix DNA microarray image analysis was performed on an Affymetrix genechip that contains probes, and the data was stored in the CEL files. CEL files were read using the ReadAffy function into an affybatch object along with a phenodata object that contained the experimental factors for these chips. The affybatch object created in the previous step was preprocessed in two steps: 1) quality control of the data (to check for probe and read anomalies), and 2) adjustment of the data for further analysis, including normalization and correcting for batch effects. The data collected was normalized using different methods to bring a uniformity and remove the differences in expression.

QC of the affybatch object was done using fitPLM function that converted it into a PLMset. This was done on the probeset-by-probeset basis and fit robust Probe Level linear Models to all the probe sets in the object (normalize is set to TRUE). Quality assessment was done on the PLMset using RLE and NUSE. RLE values were calculated for each probe set by comparing the value expressed on each array against the median expression value for that probe set across all the arrays. NUSE standardized after computing the standard error estimates obtained for each gene on each array. The median standard error for that gene was one across the arrays. The outputs of RLE and NUSE were set to statistics. RLE and NUSE histograms were plotted from the outputs, and we see that RLE medians are near zero and NUSE medians are near one.

Affymetrix Genechips are made up of probes, to measure the expression levels of a particular genomic sequence. The presence of noise was attributed mostly to technological differences rather than the biological differences between the probes and is common in large databases. To correct this, normalization of all the CEL files was done using the RMA function in the affy package. This function converts the affybatch object to an ‘expressionset’ object using the robust multi-array average expression measure in the log2-base scale. Summarization was done in the form of combining probe intensities across arrays, so the final measurements represent gene expression levels. Computing the average helps in identifying the outliers.

ComBat was used to adjust for batch effects in datasets where the batch covariate is known. This methodology is described in Johnson et al. (2007). The output is an expression matrix after the data was adjusted for batch effects with empirical Bayes frameworks. Here, a file with batch effects (Center and RNA extraction methods are merged into a single variable) and the features of interest according to Marisa et. Al (Tumor and MMR status) is the input. The features of interest were in a single variable and converted into a model matrix for the ComBat function. The batch effect variable was denoted for the extraction methods. The output file was written to a CSV file with rows as probe sets and columns as samples.

Principal Component Analysis (PCA) was performed on the normalized data using the prcomp function. The data was transposed and scaled before the PCA analysis. It was transposed so that the data is scaled on each gene and not on samples. This scaled data was transposed again to return to the original state and passed to the prcomp function. PCA was plotted to visualize the data by extracting the first and second principal components.

Noise filtering and dimensionality reduction were used to reduce the gene set of features. Three filters were performed sequentially and allowed for the selected probe sets for the genes to meet a certain criteria. The first filter consisted of eliminating the probe sets that had less than twenty percent of their gene expression levels greater than log2(15). The second filter performed was used to eliminate the probe sets that had a variance that was not significantly different from the median variance of all of the probe sets. This filter was conducted utilizing the chi-squared test. The test statistic for each gene was computed and compared to the chi-squared distribution with the degrees of freedom being the number of columns minus one (133). The probes that passed this filter had a test statistic that was greater than the upper limit of the chi-squared test. The third and final filter selected probes where the coefficient of variation was greater than 0.186.

Hierarchical clustering was performed to group together the gene expression data and probe sets for genes that are identified as either C3 or C4 tumor subtypes. Clustering does not rely on classes and can be used for relationship analysis. The filtered data was used to calculate a distance matrix of all of the data values. Then the distance matrix was fed to a clustering function *hclust* to group the data into two groups based on their distance values. The subtype information was found in an annotation file and compared to the filtered data to identify which genes were classified into which subtype. Once the clusters were identified a heatmap was formed to depict expression levels of the probe sets.

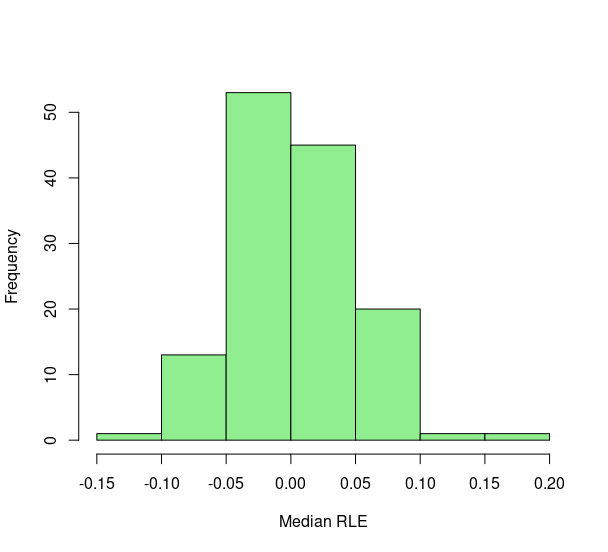
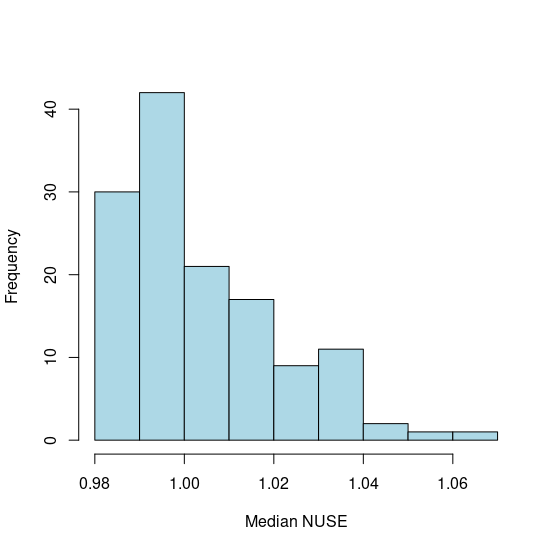
Finally, a t-test was conducted on the clusters to calculate their p-values, and which were then adjusted using the FDR, or false discovery rate, method. The results of the t-test were then compared to a threshold of 0.05, and if the adjusted p-value was less than the threshold, it was deemed as significant and represents differentially expressed genes and probe sets.

The top ten up- and downregulated differentially expressed probe sets from the pseudo data file were further analyzed to understand the biological significance of the different gene expression. The differentially expressed probe-sets, with matching gene symbols, were selected based on T-test statistics to determine the most differentially expressed probe sets were mapped to gene symbols using the hgu133plus2.db.

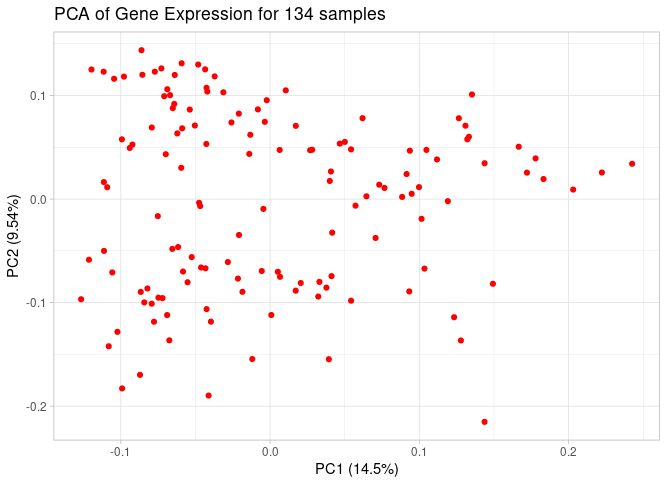
A Fisher Exact test was performed on the pseudo data of differentially expressed gene list. These included Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) and Hallmark gene sets, containing 50, 186 and 10402 genes respectively. A comparison was done between both the up- as well as the downregulated genes of 1000 probes and these gene sets. Genes were compared to the three mentioned gene set collections using the ‘fisher.test’ function. Each gene-set collection was downloaded from the Molecular Signatures Database (MSigDB).

**Results**

The data obtained were normalized using three different approaches: RLE (run length encoding), NUSE (Normalized Unscaled Standard Errors), and RMA (Robust Multi-Array Average expression measure). After the data was normalized, it was corrected for batch effects using the ComBat function.

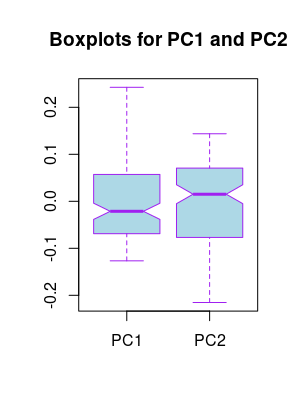
*Figure 1. Histograms of median (a) RLE and (b) NUSE*

Quality control measures of the data were done using RLE and NUSE. (a) RLE function of all the samples have their medians close to zero, indicating that these are good data. (b) NUSE of all the samples median near one indicates that the quality control was a success and the data could be used for further processing.



*Figure 2. PCA Plot of Gene Expression within the Samples. This PCA plot is used to visualize the variance between samples.*

Principal Component Analysis of the samples was done using the prcomp function. The first and second principal components have a variance ratio of 14.5% and 9.54% respectively. The data contained 134 samples in total, and the first two principal components take up to 24.04% variance. Within the plot above, the samples are spread out, and clustering at a single location is not evidently seen. Since it is difficult to identify the outliers from a principal component analysis plot, a box plot of the first two principal components was used to visualize them.



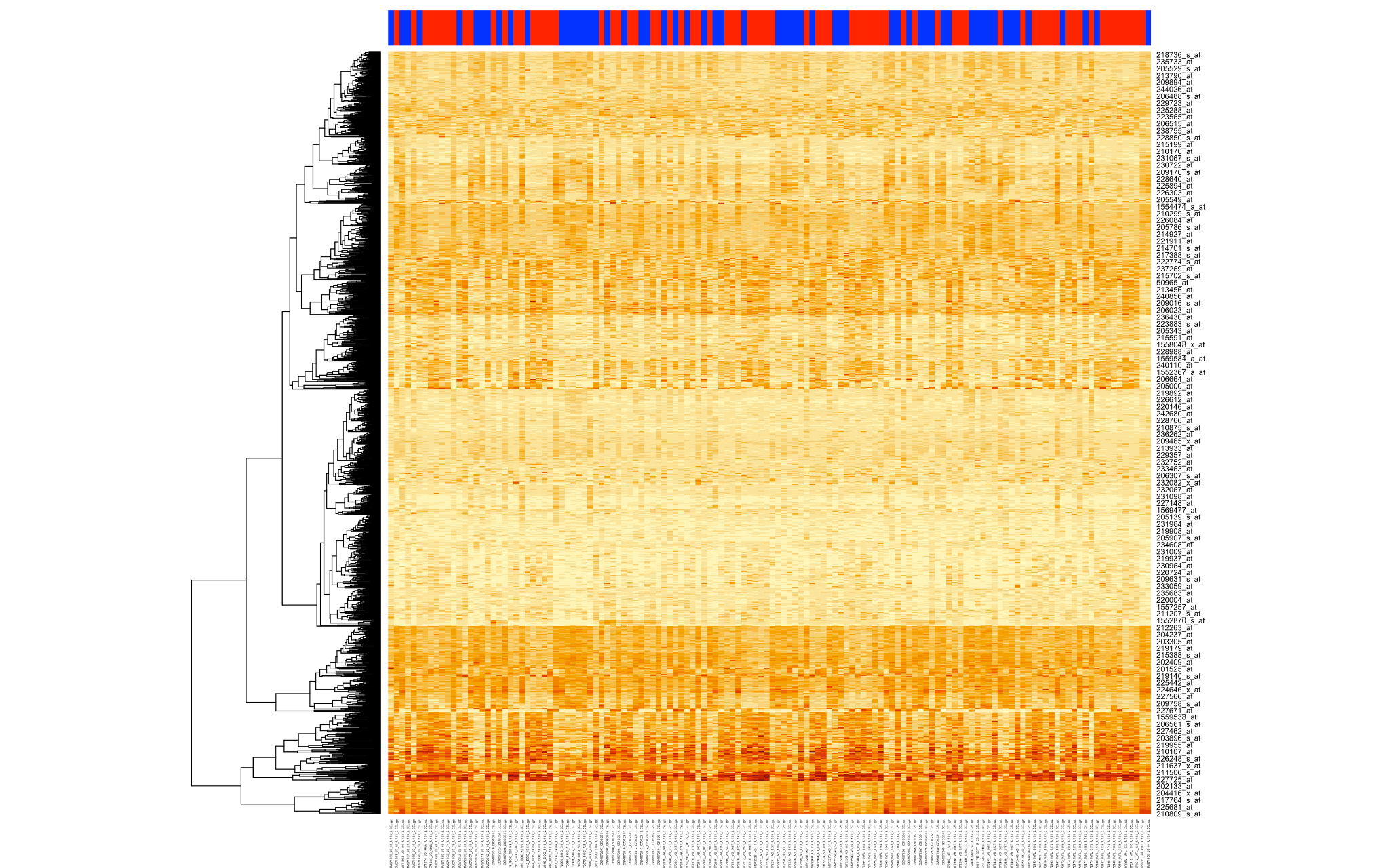
*Figure 3. Outliers within the Principal Components. This box plot was used to visualize the outliers from the PCA in Figure 2.*

The box plot from Figure 3 shows that there are no outliers in the first and second principal components of the analysis.

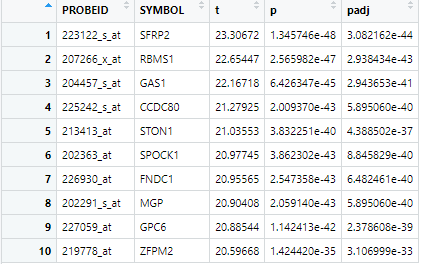
Once the gene expression data was normalized, filtering methods were used to perform noise filtering and dimensionality reduction. The first filter on the ComBat selected the probes where at least 20 percent of the expression values were greater than log2(15). This resulted in 39750 probes out of the original 54675 probes. The second filter utilized the chi-squared test and resulted in a total of 19926 probes that have a variance significantly different from the median variance of all probe sets using a threshold of p < 0.01. The third and final filter examined whether the coefficient of variation was greater than 0.186. This filter resulted in 1658 probes remaining. These three filters were performed sequentially, where the next one utilized the results of the one before it. These results are slightly greater than the filter results of Marisa et. al. where they had a result of 1459 probe-sets after concluding filtering.

After filtering the data to meet the desired requirements, hierarchical clustering was performed. Clustering is an unsupervised method for grouping similar data usually based on a distance function. In the clustering analysis, the euclidean distance function was utilized. This resulted in two clusters being generated, consisting of 76 and 58 samples.

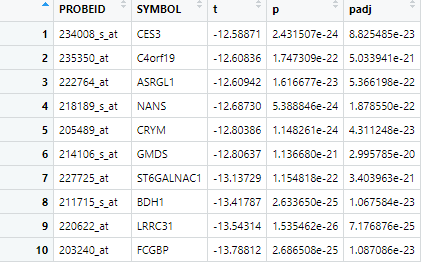
Furthermore, a Welch t-test was performed utilizing these cluster formations. This statistical analysis obtained p-values that were then adjusted. The adjusted values that were statistically significant and less than a threshold of 0.05 were then used to estimate the number of truly significant probe sets. This resulted in 1361 significant probe sets. The probes that best represented the C4 cluster were “204457\_s\_at”, “223121\_s\_at”, “223122\_s\_at” as they were significantly upregulated compared to the C3 cluster and had a significant p-value. “209955\_s\_at”, “214261\_s\_at”, “224588\_at” were some of the genes that represented the C3 cluster the best as they had a higher t-test statistic value compared to the other cluster.



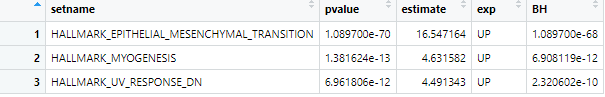
*Figure 4. Heat map of the gene-expression of the 1658 probesetes (y-axis) across 134 samples (x-axis). The heatmap visualizes the clustering of the samples between the two cancer subtypes (C3 and C4). The color bar on the top of the heatmap identifies which of the 134 samples belong to either the C3 (red) or C4 (blue) cluster. The darker colors depict lower gene expression, yellow for downregulation, and red or dark orange for upregulated genes.*



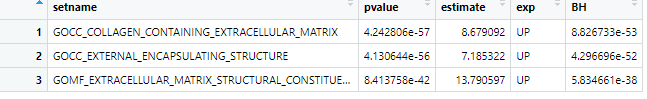
*Figure 5. Top 10 up-regulated genes.**The results of matching gene symbols to their probeIDs, sorted by t-statistic and adjusted p-values to get the top 10 up- and down-regulated genes. Showing each probe ID to its corresponding t-statistic, p-value, adjusted p-value and gene symbol.*



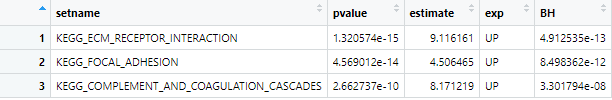
*Figure 6. Top 10 up- and down-regulated genes.**The results of matching gene symbols to their probeIDs, sorted by t-statistic and adjusted p-values to get the top 10 up- and down-regulated genes. Showing each probe ID to its corresponding t-statistic, p-value, adjusted p-value and gene symbol.*



*Table 1. Top 3 Enriched Hallmark Gene Sets. Fisher test performed on Hallmark gene set collection.*



*Table 2. Top 3 Enriched GO Gene Sets. Fisher test performed on GO gene set collection.*



*Table 3. Top 3 Enriched KEGG Gene Sets. Fisher test performed on KEGG gene set collection.*

**Discussion**

Analysis was conducted on CEL files. RLE and NUSE scores were obtained as a part of normalization to determine the quality of the data. The outputs of these two functions show that the data is of good quality. PCA was performed on the normalized data to reduce the dimensionality of the dataset. The first and second principal components constitute about 24.04% variance together. A box plot to identify the outliers is performed on the normalized data, this plot shows that there are no outliers, hence no samples were discarded.

The original researchers in Marisa et al. used consensus clustering to identify the true number of clusters present in their data. The method is highly computational so we used hierarchical clustering to identify the clusters in our analysis. This difference in cluster method may have led to the differences in cluster formations, as well as identifying the probe sets with significant gene expression.

The analysis of the 1000 most up- and down-regulated genes in the enrichment samples found that the top three up-regulated enriched probes were: SFRP2, RBMS1 and GAS1. In the up-regulated list, there is a protein that acts as a soluble modulator of Wnt signaling, a protein that binds to single stranded RNA/DNA, and a protein involved in cellular growth. These proteins encode for RNA Binding Motif Single-Stranded Interacting Protein. The top three down-regulated probes were: CES3, C4orf19 and ASRGL1. These proteins include a protein that participates in colon and neural drug metabolism, a protein-coding gene, and a protein that may be involved in the production of L-aspartate, which can act as an excitatory neurotransmitter in some brain regions.

No genes were found to have a significant p-value, with as much power as the original paper (p < 0.05). This could be possibly due to analyzing only a part of the data, which acted more like pseudo data. Though, there was still a correlation and possible enrichment in similar pathways. In the GSEA (Gene Set Enrichment Analysis), there was a similar pathway upregulated as seen in Marisa et al. (2013). Other pathways seen in the GSEA are also involved in the cell signaling pathways and metabolism in a similar way as seen in the original data classification. These include myogenesis and epithelial mesenchymal transition in the Hallmark pathway, and focal adhesion and ECM receptor interaction in the KEGG pathway.

**Conclusion**

There were upregulated genes in the C3 and C4 tumor subtype clusters, with the genes that were upregulated being different between the two clusters. Differences in upregulated and downregulated genes can be used to characterize colorectal cancer for diagnostic purposes. Furthermore, distinct expression signatures between tumor subtypes further propels the theory of colorectal cancer subtypes, where each subtype can have varying prognostic classifiers. Lastly, more study on the genes that are specifically upregulated or downregulated is needed in order to determine their molecular relationship to colorectal cancer.

One challenge that arose in our analysis was learning how to code the statistical tests and sifting through the data in R. Most of the members of our group have not yet coded in R or done this type of analysis in R, and it provided a challenge. We overcame this by looking up documentation on the specific R functions we were to use. Another problem that we encountered was that we seemed to have different filtering results than those of Marisa et al. We had a greater number of probes after our filtering methods than the published study, which may have skewed the resulting upregulated and downregulated genes found. This may have been in part due to the difference in clustering methods between our analysis and that of Maris et al.

**References**

1. Marisa L., de Reynies A., Duval A., Selves J., 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

<https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001453>

1. Durinck, S., Spellman, P.T., Birney, E. and Huber, W., 2009. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nature protocols, 4(8), pp.1184-1191.

<https://www.nature.com/articles/nprot.2009.97>

1. Bolstad, B., 2011. affyPLM: model based QC assessment of Affymetrix GeneChips.

<https://mirrors.nju.edu.cn/bioconductor/bioc/2.0/vignettes/affyPLM/inst/doc/QualityAssess.pdf>

1. Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E. and Storey, J.D., 2012. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics, 28(6), pp.882-883.

<https://academic.oup.com/bioinformatics/article/28/6/882/311263?login=true>

1. Database, G. C. H. G. (n.d.). *ASRGL1*. GeneCards is a searchable, integrative database that provides comprehensive, user-friendly information on all annotated and predicted human genes. Retrieved February 18, 2022, from <https://www.genecards.org/cgi-bin/carddisp.pl?gene=ASRGL1&keywords=ASRGL1>
2. Database, G. C. H. G. (n.d.). *C4orf19*. GeneCards is a searchable, integrative database that provides comprehensive, user-friendly information on all annotated and predicted human genes. Retrieved February 18, 2022, from <https://www.genecards.org/cgi-bin/carddisp.pl?gene=C4orf19&keywords=C4orf9>
3. Database, G. C. H. G. (n.d.). *CES3*. GeneCards is a searchable, integrative database that provides comprehensive, user-friendly information on all annotated and predicted human genes. Retrieved February 18, 2022, from <https://www.genecards.org/cgi-bin/carddisp.pl?gene=CES3&keywords=CES3>
4. Database, G. C. H. G. (n.d.). *GAS1*. GeneCards is a searchable, integrative database that provides comprehensive, user-friendly information on all annotated and predicted human genes. Retrieved February 18, 2022, from <https://www.genecards.org/cgi-bin/carddisp.pl?gene=GAS1>
5. Database, G. C. H. G. (n.d.). *RBMS1*. GeneCards is a searchable, integrative database that provides comprehensive, user-friendly information on all annotated and predicted human genes. Retrieved February 18, 2022, from <https://www.genecards.org/cgi-bin/carddisp.pl?gene=RBMS1>
6. Database, G. C. H. G. (n.d.). *SFRP2*. GeneCards is a searchable, integrative database that provides comprehensive, user-friendly information on all annotated and predicted human genes. Retrieved February 18, 2022, from <https://www.genecards.org/cgi-bin/carddisp.pl?gene=SFRP2>