Project 1: Microarray-Based Tumor Classification

Alec Jacobsen, Daisy Wenyan Han, Divya Sundaresan, Emmanuel Saake

Programmer Biologist Data Curator Analyst

ENG BF528, Spring 2021

Introduction

Despite the various advancements in diagnostic tools, colorectal cancer has remained the third-most common form of cancer, as well as the fourth-most prominent cause of death worldwide (Greenlee, 2000). In their recent study, *Gene Expression Classification of Colon Cancer into Molecular Subtypes*, Marisa et al. aimed to establish molecular classifications using previously obtained genome-wide mRNA gene expression data from existing colon cancer samples. This data was then used to delineate six new classifications, creating the first known transcriptome-based classification of colon cancer.

The biological significance of these subtypes was corroborated with significant differences in disease prognosis. This was shown by illustrating the specific signalling pathways affected across each of the molecular subtypes, as well as testing each subtype signature against the enrichment of Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) pathways, and gene sets related to cancer hallmarks. These six classifications, in addition to the differential gene expression associated with each one, can allow for better identification of targeted markers of colon cancer subtypes, more individualized treatment, and improved disease prognosis for colorectal cancer patients.

Due to lack of gene expression profiles that helped predict prognosis of colorectal cancer this study was conducted to see if there were any possible gene expression profiles that could be found. This study uses microarray data to determine which genes are differentially expressed, and given these differentially expressed genes can we predict a condition in this case colorectal cancer. Though microarray is not used as much anymore it still is a very common technique used in gene expression analysis.

Each dataset was normalized in batches, and residual technical batch effects were corrected. The analytic techniques used on the data include gene expression analysis, Array-Based Comparative Genomic Hybridization Analysis, Unsupervised Subtype Discovery Based on Gene Expression Analysis, Molecular Subtype Characterization, and validation testing. Further details are mentioned in the methods section below.

Data

Fresh-frozen primary tumor tissue samples from a cohort of 750 patients with stage I to IV CC who underwent surgery collected from various hospitals in France part of the The French national Cartes d'Identité des Tumeurs (CIT) program. The microarray data was generated for the human genome.

Of the 750 tumor samples of the CIT cohort, 566 fulfilled RNA quality requirements for GEP analysis. This is how the high quality data was selected. 184 samples were eliminated due to not meeting the RNA quality standards which was done before the samples reached us (Marisa, 2013). In addition to the 566 CC samples there were 19 non-tumoral colorectal mucosas added. The original paper had the 585 total samples divided into 443 CC discovery set, 123 CC validation set, and the 19 non-tumoral colorectal mucosas set. We had access to the 585 samples for our project via (GEO database). As the data curator I added the one missing sample from GEO into our samples folder on the SCC using wget.

The Affymetrix U133 Plus 2.0 chips for microarray was the instrument used for analysis. The specific type of microarray should be a high density oligonucleotide array based on the chip used though not specifically mentioned in the paper. No sources of contamination were mentioned.

Link to public repository: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE39582

Methods

Data Preprocessing and Quality Control:
CEL files were read into R using the ReadAffy function and normalized with Robust Multiarray Averaging algorithm (RMA), as implemented in the Affy package using the default parameters (Gautier et al. 2004). Batch effects were corrected for with the ComBat function of the SVA package (Leek et al. 2020). Relative log expression (RLE) and normalized unscaled standard error (NUSE) scores were calculated by first fitting probe-level linear models to probe-sets using the fitPLM function of the AffyPLM package with the default model and parameters (Bolstad 2004). RLE and NUSE scores could then be calculated using the RLE and NUSE functions. Medians were calculated with base R. Distributions of medians were plotted for both RLE and NUSE scores to ensure that the samples were within the acceptable ranges (close to zero for RLE and approximately less than one for NUSE). None of the samples were discarded.

Batch effect correction was done with the ComBat function from the SVA package using the medical center and RNA extraction method as covariates. Normalized and batch corrected data were then scaled and used for principal component analysis (PCA) with the prcomp function from the stats package (R Core Team 2013). PCA showed even clustering of samples with no discernable outliers. Again, none of the samples were discarded.

Further analysis consisted of two phases: dimensionality and noise reduction, and hierarchical clustering and subtype discovery. Dimensionality and noise reduction served as a preparatory step for hierarchical clustering, which can produce spurious results when the number of features is far larger than the number of samples, as is often the case in microarray data. By reducing the amount of noise in the data through filtering, hierarchical clustering algorithms are more likely to produce true results. The clustering then served to identify which samples are different from each other, so that the differences in clusters can be probed, revealing the distinguishing biological characters of any cancer subtypes.

Noise filtering and dimensionality reduction:
In this study, noise and data dimensionality reduction took the form of three different tests:

- A logarithmic test involving the selection of genes on the criteria that 20% of their gene-expression value be $> \log2(15)$).
- A second test, a low variance filter with a threshold of $p<0.01$ was undertaken: Two critical values (upper and lower) of the chi-squared distribution was computed to determine the critical region. Test statistics that fall outside the critical region satisfied the condition for rejection of the null hypothesis.
- Finally, variation in samples were measured using the threshold, coefficient of variation > 0.186

Table 4.1 shows the parameters utilized to obtain the unsupervised analysis on the gene expression data. The algorithm is as described below:

Dimension reduction at the gene-level:

```
Testing start: [NB: * user defined function]
        Input:
                Preprocessed gene_exp data, alpha=0.01(for chi square),
        Output:
                Filtered data
        Functions:
                qchisq(), sd(), *execute_logfoldtest(),*execute_chisquarevariancetest(),
        *execute_CoVtest(), mean(), variance(), Intersection()
        Main_run:
               Data: Get  preprocessed data
               Vector genes: Compare on the gene-level whether more than 20% of gene expression >
                 log2(15)
                Test statistics : get_test_statistics() for each gene.
                Upper and lower critical points: qchisq with degree of freedom
                Vector of genes: for each test statistics if value is outside the critical region,
                 reject null hypothesis.
                Filtered list: Compute coefficient of covariance and if >0.186 accept.
                Intersect: Intersection of all three tests.
                Reduced_dim data: Write out Intersection data
End Test
```

Unsupervised subtype discovery based on expression data

Using hierarchical clustering based on a euclidean distance matrix, unsupervised analysis was performed on samples from the gene expression data (*dim* = 1323 *x* 134). The clustering was limited to two (2). The algorithm for this process is as described below:

Clustering algorithm:

```
Clustering start:
        Input:
                Dimension-reduced data
        Output:
                Cluster
        Functions:
                scale(), distance(), hclust(), cutree(), plot()
        Main_run:
                Data: Get  dimension-reduced data
                Data: scale(data)
                Distance : Compute distance matrix of data
                Cluster: Cluster with average method(distance)
                Dendrogram: Plot cluster
                Two_Cluster: Use cutree() to cut cluster into k groups
                Plot: return plot
End clustering
```

Results

At the end of the dimension reduction phase, we had reduced the dimension of input_data from 54675 to 1323. Table 4.1 is a summary of output statistics and parametric configurations.

Unsupervised analysis of gene expression

A sample-wise ($n = 134$) clustering on the aforementioned (output) data resulted in two clusters of samples based on the most variant probe-sets ($n\_row = 1323$): Cluster_1($n = 59$, $\approx 44\%$ *of* 134), and Cluster_2 ($n = 75$, $\approx 56\%$ *of* 134).

Colon cancer subtypes

The analysis focused on the C3 and C4 tumor subtypes. The heatmap of gene-expression of genes across all samples revealed that C3 was most common. There were few outliers from C4 as indicated by the column side color indicator on the heatmap in Figure 4.1. Similar to the clusters from the unsupervised analysis, the following are the number of C3 and C4 tumor subtype : C4 ($n = 59$, $\approx 44\%$ *of samples*), and C3 ($n = 75$, $\approx 56\%$ *of samples*).

Welch t test

The adjustment of the ranked output of the welch test with p_adjust() using the 'fdr' method of Benjamini & Hochberg (1995) revealed that a total of 1064 probe sets for those differentially expressed at $\rho < 0.05$.

Table 4.1 : Estimates from dimensionality and cluster analysis

Dimensionality Reduction		Cluster analysis		Parameters	
	Sub_totals		Sub_totals	Method	Values
#_of probe_set_from_preprocessed	54675	#_of_clusters	2	distance_method	euclidean
#_of_probe_set_from_log_test	39661	#_of_samples_in_Cluster1	59	cluster_method	average
#_of_probe sets_values_outside_critical_region	54416	#_of_samples_in_Cluster2	75	cluster_type	hierarchical
#_of_probes_passed_CoV_greater_0.186	1699	#_of_c3_tumor_subtype	75	pval_method	fdr
#_of_probe_sets_passed_all_three_tests	1323	#_of_c4_tumor_subtype	59		
#_of_differentially_expressed_p<0.05	1064				

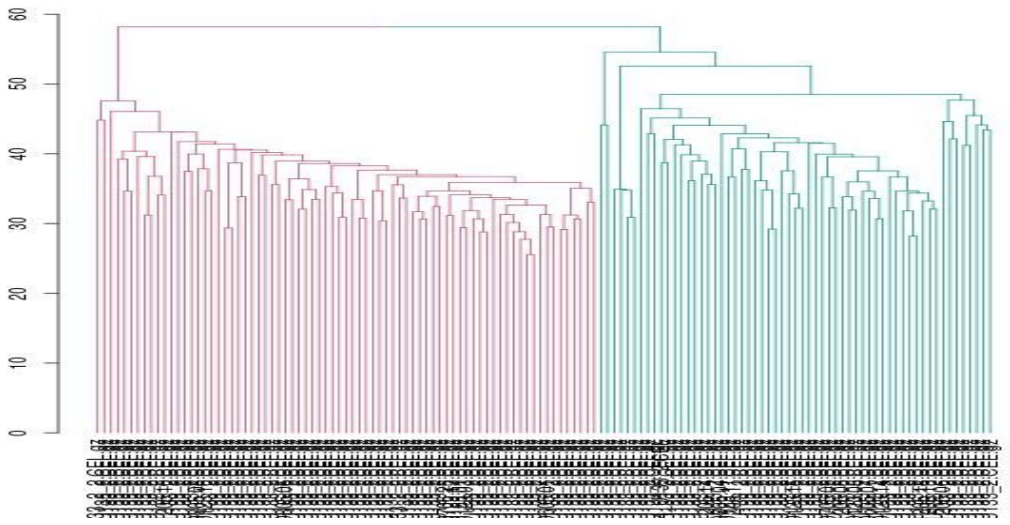

Figure 4.1: A dendrogram showing the hierarchical clustering of patients into two clusters

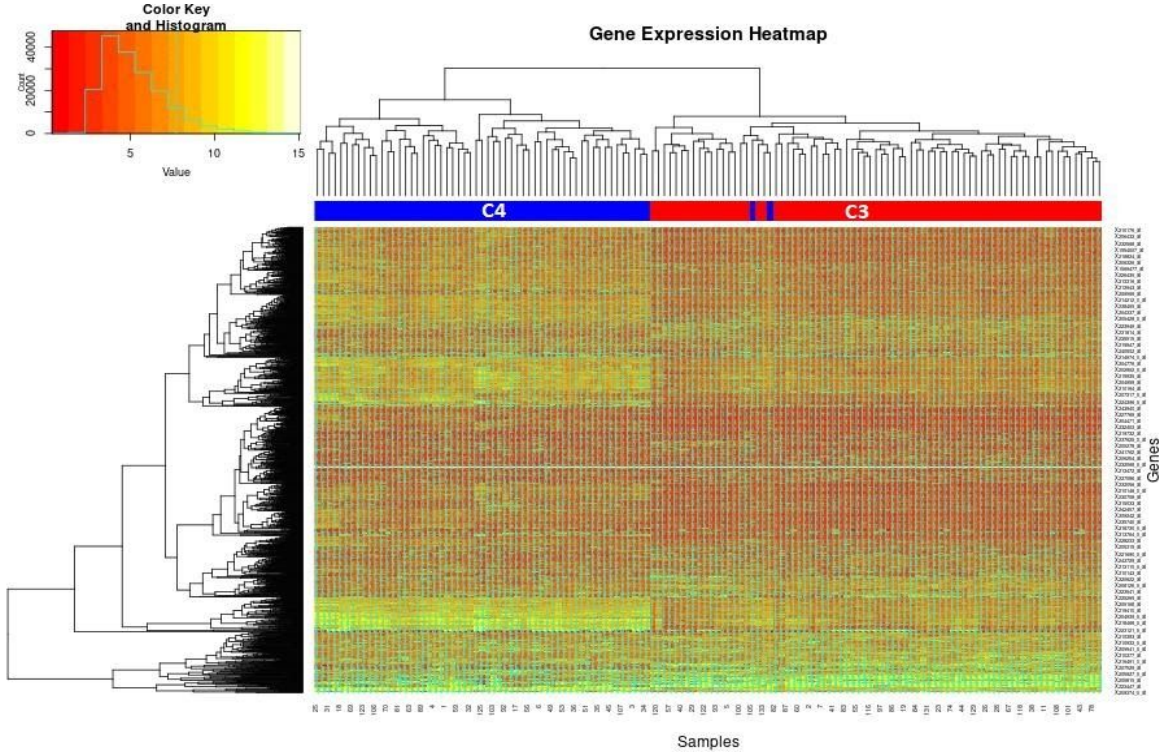

Figure 4.2: Heat map illustrating the expression of genes with colorside bar indicating the colon cancer molecular subtype C3 and C4. Where red indicates C3 and blue indicates C4.

Gene Set Enrichment Analysis

The enrichment of the Kyoto Encyclopedia of Genes and Genomes (KEGG) and GeneOntology (GO) pathways and gene sets related to cancer hallmarks was tested using the set of genes that had passed the quality control parameters for the C3 and C4 CRC subtypes. Table 5 shows the most significantly enriched gene sets in each of the three gene sets, ordered by nominal p-value. These pathways were then compared to those obtained by Marisa et al. for further validation.

The hgu133plus2 database was used to match the Probe sets generated by the microarray data to their respective gene symbols. A total of 1434 Gene Symbol - Probe ID pairs were obtained from the list of 1323 Probe ID's. Duplicated gene symbols, with different Probe ID's were filtered such that only the Probe ID containing the most significant adjusted p-value was retained. This resulted in a total of 999 Gene Symbol - Probe ID pairs to be used for enrichment analysis.

Gene sets were downloaded from MSigDB to be used in the gene set enrichment analysis. In particular, GMTs files of the all hallmarks in cancer, all GeneOntology Pathways, and all KEGG Pathways were obtained for analysis. These contained 50, 10271 and 186 gene sets, respectively, all of which were significantly enriched with an adjusted p-value below 0.05, following FDR correction.

Table 5A. Most highly enriched GeneOntology Gene Sets. The most highly enriched pathways within the GeneOntology gene sets, by smallest nominal p-value.

Geneset	Statistic	P Value	Adjusted P Value
GO_REPRODUCTION	0	0	0
GO_NEGATIVE_REGULATION_OF_TRANSCRIPTION_BY_RNA_POLYMERASE_II	0	0	0
GO_MICROTUBULE_CYTOSKELETON_ORGANIZATION	0	0	0

Table 5B. Most highly enriched Cancer Hallmarks Gene Sets. The most highly enriched pathways within the Hallmark gene sets, by smallest nominal p-value.

Geneset	Statistic	P Value	Adjusted P Value
HALLMARK_MYC_TARGETS_V1	0	4.2685282496816e-248	4.2685282496816e-248
HALLMARK_OXIDATIVE_PHOSPHORYLATION	0	4.2685282496816e-248	4.2685282496816e-248
HALLMARK_E2F_TARGETS	0	5.94605985180604e-245	5.94605985180604e-245

Table 5C. Most highly enriched KEGG Gene Sets. The most highly enriched pathways within the Kyoto Encyclopedia of Genes and Genomes gene sets, by smallest nominal p-value.

Geneset	Statistic	P Value	Adjusted P Value
KEGG_OLFACTORY_TRANSDUCTION	0	0	0
KEGG_PATHWAYS_IN_CANCER	0	3.43852962373845e-298	3.43852962373845e-298
KEGG_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION	0	3.56178888921201e-267	3.56178888921201e-267

This analysis was repeated for the top thousand up-regulated and down-regulated genes following differential expression analysis. The Probe ID's used for this analysis were selected such that their FDR was less than 0.05, and did not undergo the same extensive filtering as the gene list that was used above.

Table 6 shows the most highly upregulated and downregulated genes for each cluster subtype, relative to each other. The "Upregulated" gene sets refer to the clusters in which genes were more highly expressed in Cluster 1 over Cluster 2. For the purpose of our analysis, these refer to subtypes C4 and C3 respectively.

Table 6A. Most highly upregulated enriched gene sets. The most highly enriched Gene Ontology, KEGG and Cancer Hallmarks pathways in the top 1000 genes upregulated in the C4 subtype over the C3 subtype.

Geneset	Statistic	P Value	Adjusted P Value
GO_REPRODUCTION	0	0	0
GO_NEGATIVE_REGULATION_OF_TRANSCRIPTION_BY_RNA_POLYMERASE_II	0	0	0
GO_MICROTUBULE_CYTOSKELETON_ORGANIZATION	0	0	0

Geneset	Statistic	P Value	Adjusted P Value
HALLMARK_E2F_TARGETS	0	4.99695193892528e-234	4.99695193892528e-234
HALLMARK_MYC_TARGETS_V1	0	4.99695193892528e-234	4.99695193892528e-234
HALLMARK_OXIDATIVE_PHOSPHORYLATION	0	5.99634232670954e-231	5.99634232670954e-231

Geneset	Statistic	P Value	Adjusted P Value
KEGG_OLFACTORY_TRANSDUCTION	0	0	0
KEGG_PATHWAYS_IN_CANCER	0	5.49758676947783e-257	5.49758676947783e-257
KEGG_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION	0	1.880304602363e-251	1.880304602363e-251

Table 6B. Most highly downregulated enriched Gene Ontology gene sets. The most highly enriched Gene Ontology, KEGG and Cancer Hallmarks pathways in the top 1000 genes downregulated in the C4 subtype over the C3 subtype.

Geneset	Statistic	P Value	Adjusted P Value
GO_REPRODUCTION	0	0	0
GO_NEGATIVE_REGULATION_OF_TRANSCRIPTION_BY_RNA_POLYMERASE_II	0	0	0
GO_MICROTUBULE_CYTOSKELETON_ORGANIZATION	0	0	0

Geneset	Statistic	P Value	Adjusted P Value
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	0	4.99695193892528e-234	4.99695193892528e-234
HALLMARK_INTERFERON_GAMMA_RESPONSE	0	3.59480722486222e-228	3.59480722486222e-228
HALLMARK_ALLOGRAFT_REJECTION	0	1.02755837577292e-220	1.02755837577292e-220

Geneset	Statistic	P Value	Adjusted P Value
KEGG_OLFACTORY_TRANSDUCTION	0	0	0
KEGG_PATHWAYS_IN_CANCER	0	6.15144047838672e-289	6.15144047838672e-289
KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION	0	2.11595789461939e-265	2.11595789461939e-265

Discussion

Using hierarchical clustering two clusters were obtained C1 of size 59 and C2 of size 79. A careful look at the heat-map in Figure 4.2 reveals a similar grouping of the tumor subtype C3 and C4. This correlation does not only exist in the grouping but in the numbers. Considering the number statistics involved, it seems the hierarchical clustering clustered the patients/samples based on their tumor subtype. Where $Cluster\_1 = patients\ with\ C4\ tumor\ subtype \Rightarrow n = 59$ and $Cluster\ 2 = patients\ with\ C3\ tumor\ subtype \Rightarrow n = 75$. This attests to the strength of unsupervised learning or analysis and significant information hidden in the gene expression data of Microarrays. It appears, using the tumor subtype as a label for the data, we could equally apply supervised learning as a predictive model on the gene expression data. However as observed on the heatmap there were one or two C4 tumor subtype outliers in the C3 region.

Gene Set Enrichment Analysis

While the results of our gene set enrichment analysis did not reproduce those obtained by Marisa et al., that is not to say they did not offer interesting insight. Cell motility, for example, is still an important factor in the C3 and C4 subtypes, as shown by the enrichment of GO Microtubule Organization. Cellular reproduction is also prominent across the enriched gene sets, as seen by the enrichment of the GO_Repoduction and Hallmark_Oxidative_Phosphorylation pathways. The KEGG_Pathways_in_Cancer, Hallmark_MYC_Targets and Hallmark_E2F_Targets pathways were also significantly enriched in our analysis. While this was not noted in the original paper, it was to be expected, as these are typical cancer markers and pathways, and we are working with two sets of cancer microarray data.

Interestingly, when comparing between the gene pathways that were significantly upregulated in C4 tumors in table 6A and those that were significantly downregulated in table 6B, 5 pathways were present in both. This may suggest that different genes in those pathways are being expressed differently between cancer subtypes, which could be useful features to consider when diagnosing between subtypes in the clinical setting. Additionally, the epithelial mesenchymal transition pathway was significantly downregulated in C4 tumors as compared to C3 tumors, which could be indicative of a greater risk of metastasis on C3 tumors as compared to C4. On the other hand, two immune-response pathways

(interferon gamma response and allograft rejection) were significantly downregulated in C4 over C3 tumors, which may suggest that C4 tumors are harder for the body to fight.

We were unable to reproduce the gene set enrichment analysis results from the original paper for a variety of reasons. Most notably, our analysis focused primarily on the differences between C3 and C4 tumor subtypes, while Marisa et al. looked at differences spanning across all six. For example, the researchers noted that subtype C4 showed down-regulation of cell growth and up-regulation of motility pathways, relative to the other five subtypes (Marisa et al., 2013). However, this was not noted in our analysis, possibly because C3 and C4 were more similar in these pathways to each other, than would be noted should the other four subtypes also be accounted for.

Additionally, our filtering appears to have been conducted on a much more conservative basis, leaving fewer genes considered significant, and therefore fewer genes for gene set enrichment analysis. Marisa et al. isolated their top 1000 most upregulated and 1000 most down-regulated genes for this analysis. However, we had a total of 999 significant genes, making differentiating between up and down-regulated genes slightly less feasible. Therefore, the enrichment analysis was done on all 999 genes identified, potentially leading to different results.

This was later rectified by using only the genes that passed the FDR < 0.05 threshold, thereby providing enough genes to select the top thousand up and down-regulated genes for gene set enrichment analysis. A total of 28490 genes were used, decreased to 14362 following removal of duplicated gene symbols. This provided directionality to the gene enrichment, thereby allowing us to infer the genes that are most highly up or down-regulated across the two subtypes. Interestingly, the same or very similar pathways were identified in being both up and down-regulated in the C3/C4 subtypes. This indicates that while particular genes may be significantly up- or down-regulated, the pathways as a whole are influenced in complex ways that require further analysis for elucidation.

The way in which duplicated gene symbols were selected in our analysis also differed from that of Marisa et al.'s group. For gene symbols matching with multiple Probe ID's, they chose the Probe ID with the greatest variance, likely to account for multiple different scenarios. However, as we had few duplicates, making calculating variance redundant, we chose instead to keep the gene symbol with the most significant (smallest) p-value, as this would likely allow for more potentially interesting enrichments.

Conclusion

[*Data Curator*]: As previously mentioned CC is a very common form of cancer that has a high mortality rate. As bioinformatics we are constantly wondering how we use computational techniques to better translational research and in the end benefit personalized healthcare. The major takeaway that I saw from this project was the ability to use gene expression data from microarray and through various bioinformatics analysis offer interesting insights on colon cancer. Understanding how the samples were collected and how they were assessed for quality is very important for further downstream analysis. As I particularly did not have a strong role it was interesting to go through the others scripts and see the data getting transformed into meaningful conclusions.

[*Programmer*]: Data quality control is arguably one of the most important and potentially overlooked steps in the analysis of genomic data, as errors and poor data are impossible to distinguish by looking at the raw data alone. The greatest challenge in preprocessing data is generating metrics through which to judge the quality of the data that are human readable. For this project, that was accomplished using RLE/NUSE scores and PCA, which revealed that the data was of good quality.

[*Analyst: observation & challenges*]: The significance of noise filtering is observed in the elimination of data items that were not significantly variant for analysis. With unsupervised learning we observed the capturing of hidden patterns, categories, as well as underlying indicators of colon cancer subtype. Welch's test helped outline the p-values and test statistics; which allowed for further adjustment. What were my challenges? On my part as the analyst, I must mention I felt very green in the project, especially understanding what I was to do with the data. Initially, I couldn't tell the difference between the samples, the probe set IDs and the genes. I couldn't just tell what I was looking at and what extra steps I needed to take aside the obvious outlines from the problem description on the BF528 website/documentation web page. This affected my ability to effectively apply computational power to the data. And all through the project I felt I could do more. I know there is a lot I still need to understand and learn...and look forward to your reviews and feedback for a more enlightened me.

[Biologist]: Gene set enrichment allows for the results from high throughput experiments to be summarized and analyzed in terms of higher order biological functions. In particular, the specific biological pathways impacted in the C3 and C4 CRC subtypes were elucidated and compared in our analysis, possibly allowing for more targeted treatments in the future. However, as noted above, several of these pathways overlap, and contain genes that allow the pathway to appear both up-regulated and down-regulated simultaneously. In this scenario, we would have to look more closely at the specific genes that are causing these contradictions, and other ways in which they affect the cell, in order to make use of the gene set enrichment analysis results.

References

Bolstad BM (2004). *Low Level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization*. Ph.D. thesis, University of California, Berkeley.

Gautier L, Cope L, Bolstad BM, Irizarry RA (2004). "affy—analysis of Affymetrix GeneChip data at the probe level." *Bioinformatics*, **20**(3), 307–315. ISSN 1367-4803, doi: 10.1093/bioinformatics/btg405.

GEO Accession viewer. (2021). Retrieved 22 February 2021, from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE39582

Greenlee, R T et al. "Cancer statistics, 2000." *CA: a cancer journal for clinicians* vol. 50,1 (2000): 7-33. doi:10.3322/canjclin.50.1.7

Leek JT, Johnson WE, Parker HS, Fertig EJ, Jaffe AE, Zhang Y, Storey JD, Torres LC (2020). *sva: Surrogate Variable Analysis*. R package version 3.38.0.

Marisa, Laetitia et al. "Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value." *PLoS medicine* vol. 10,5 (2013): e1001453. doi:10.1371/journal.pmed.1001453

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.