# **Project 1 Write-up**

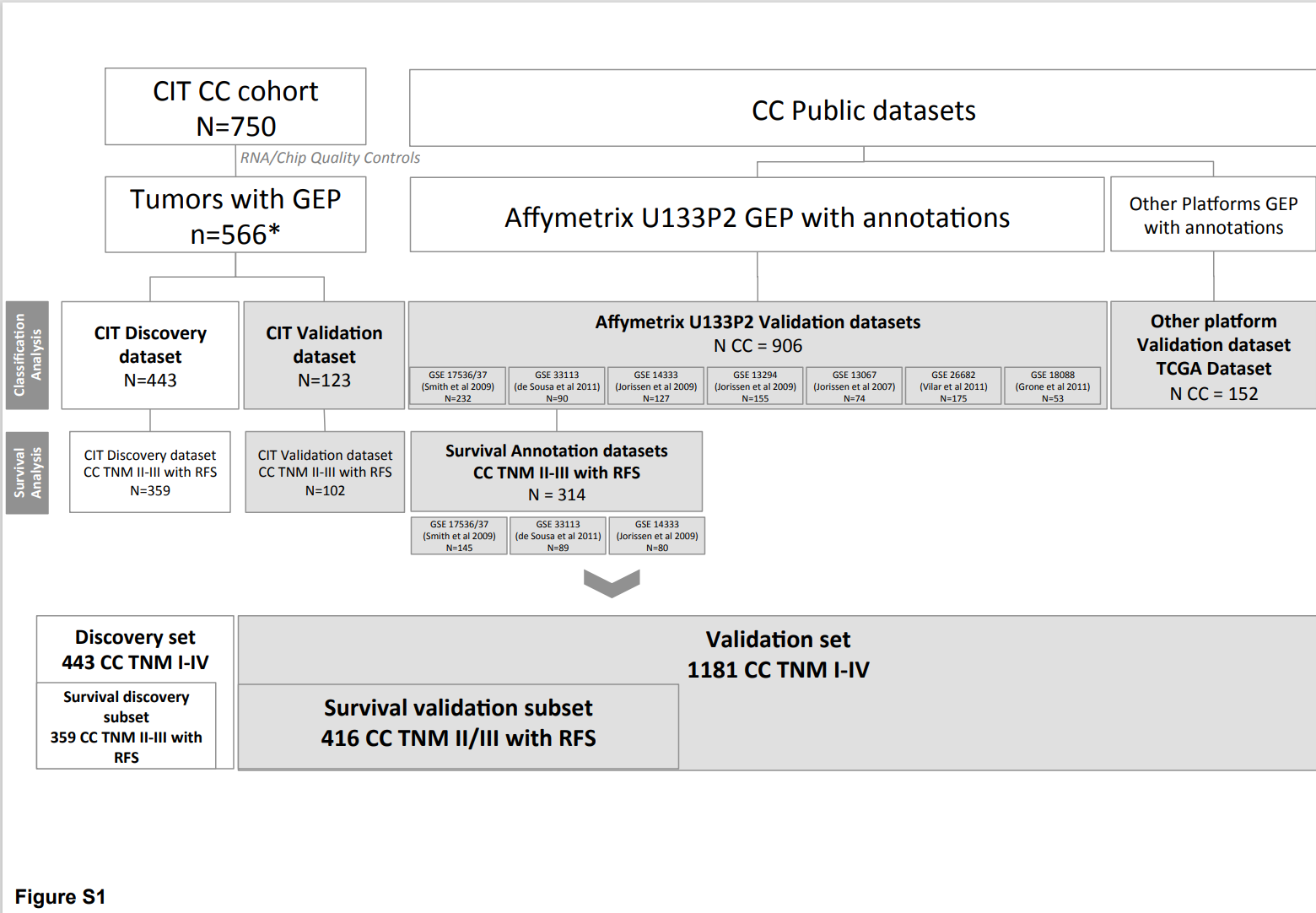
Abhishek Thakar, Allison Nau, Mae Rose Gott, and Sheila Yee

# **Introduction**

Colorectal cancer (CC) remains one of the leading causes of death by cancer despite recent technological discoveries and diagnostic techniques. While pathological staging remains a common tool used to determine a patient’s prognosis, it is evidently insufficient to predict the risk of recurrence for patients who have already undergone surgery to remove malignant tissue. Reportedly 10-20% of stage II CC patients and 30-40% of stage III CC patients develop recurrence (Marisa et al., 2013).

As such, there remains a need to develop better prognostic signatures or markers in order to more effectively characterize CC. Previous techniques used to do so involved microsatellite instability (MSI) and gene expression profile (GEP) studies, both of which require further refinement and better reproducibility (Popat et al, 2005) (Hutchins et al., 2011). CC is a heterogeneous disease as its molecular signatures develop through different molecular pathways (Jass, 2011). GEP studies involving unsupervised hierarchical clustering have identified three molecular subtypes of CC (Salazar et al., 2011). In order to further elucidate the classification of these molecular subtypes, Marisa et al. (2013) used mRNA expression profile analysis. By doing so, they sought to provide robust standard molecular subtypes for CC prognosis. In this project, we aimed to reproduce results from the original study by analyzing the C3 tumor subtype through analysis of publicly available microarray data and hierarchical clustering.

# **Data**



**Figure S1. Datasets flowchart. Discovery and validation sets used in study (Marisa et al. 2013 Supplementary Figure S1).**

Looking at Figure S1 from the study, it can be seen that from the patient sample size of 750, only 566 passed the quality controls check for gene expression analysis. The 566 total were run on Affymetrix U133 Plus 2.0 chips and split into two separate datasets (discovery and validation). More validation datasets were curated from 7 other public datasets which used a similar chip ([GSE13067](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE13067), [GSE13294](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE13294), [GSE14333](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE14333), [GSE17536](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE17536)/17537, [GSE18088](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE18088), [GSE26682](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE26682), and [GSE33113](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE33113)). The Cancer Genome Atlas (TCGA) dataset was obtained through a non-Affymetrix chip and analyzed separately but had useful annotations and extensive information for the colon cancer samples. The data used for analysis was located on a central repository with one missing sample that was obtained by accessing the Gene Expression Omnibus (sample ID: GSM971958).

The CIT CEL files were primarily all normalized using a Robust Multi-array Average (RMA) as described in the text supplementary material (Text S1). This was used to remove any potential batch effects due to the multiple centers as well as using a ComBat method in the SVA R package to aid in residual technical batch effects.

# **Methods**

All data from AffyMetrix microarrays underwent pre-processing, normalization, and statistical analysis using R 4.0.3 and BiocManager 3.12 to download the following Bioconductor packages: affy, affyPLM, sva, AnnotationDbi, hgu133plus2.db, and ggplot2.

Probe-level information (stored in the form of CEL files) was read using the ReadAffy function which created an AffyBatch object. To assess the quality and distribution of the un-normalized data, the Bioconductor package AffyPLM was used to convert the AffyBatch object into a PLMset. Two quality control metrics were measured for the dataset - the median relative log expression (RLE) and median normalized unscaled standard error (NUSE). RLE calculates the log expression of each gene in each array and then subtracts the median gene expression across all arrays for each gene (McCall et al., 2011). NUSE normalizes unscaled standard errors by dividing the standard error for each gene expression estimate by the median gene expression estimate across all of the arrays such that the median standard error for each gene is centered around one across all arrays.

The resulting AffyBatch object was normalized using the robust multi-array average (RMA) expression measure (rma) function found in the affy package. The rma function also transformed probe-level data to gene-level data. Normalization of the files involved a three-pronged approach in order to correct for variation between arrays: 1) background correction to remove background noise and artifacts, 2) quantile normalization such that arrays can be compared to each other, and 3) summarization which combines probe set intensities across all of arrays into one gene expression measure (Gautier et al, 2004).

To correct for batch effects in the RMA normalized expression dataset, the ComBat function in the sva package was used. Batch and mode parameters were set using a file containing clinical and batching annotations used by Marisa et al. (2013). Batch effects included center and RNA extraction method while features of interest that were preserved included tumor and MMR status.

Principal component analysis (PCA) was performed on the normalized, ComBat adjusted data in order to reduce the dimensionality of the data and visualize the variation present in the dataset. Data was scaled and centered within each gene rather than sample. The prcomp function was used to obtain principal component values with the parameters scale and center within the function set to false since the data had already been scaled. The first two principal components were plotted using the ggplot function from the ggplot2 package. Outliers of the first two principal components were determined as being three standard deviations away from the mean. A CSV file containing the ComBat adjusted data with the outliers removed was generated.

3 filters were applied to the ComBat adjusted data using R to filter out insignificant findings via unsupervised probe set selection. The first filter tested for genes whose expression was higher than log2(15) (=3.91) in at least 20% percent of the samples. 39661 genes from the ComBat adjusted matrix passed this filter. The second filter used the median variance of the ComBat matrix to determine which genes had a significantly (p= 0.01) different variance from the median. The measure of variance used was (Var-medianVariance)^2/medianVariance which was then compared to the qchisq() statistic. 54575 genes from the ComBat matrix passed this filter and 39563 genes from filter 1 passed. The final filter ensured that the final gene sets would have a high robust coefficient of variation, greater than 0.186, which was determined by Marisa et al (2013) using a Gaussian mixture model clustering approach via R package mclust. 1690 genes from the ComBat adjusted data passed this threshold and 1435 genes passed all three filters.

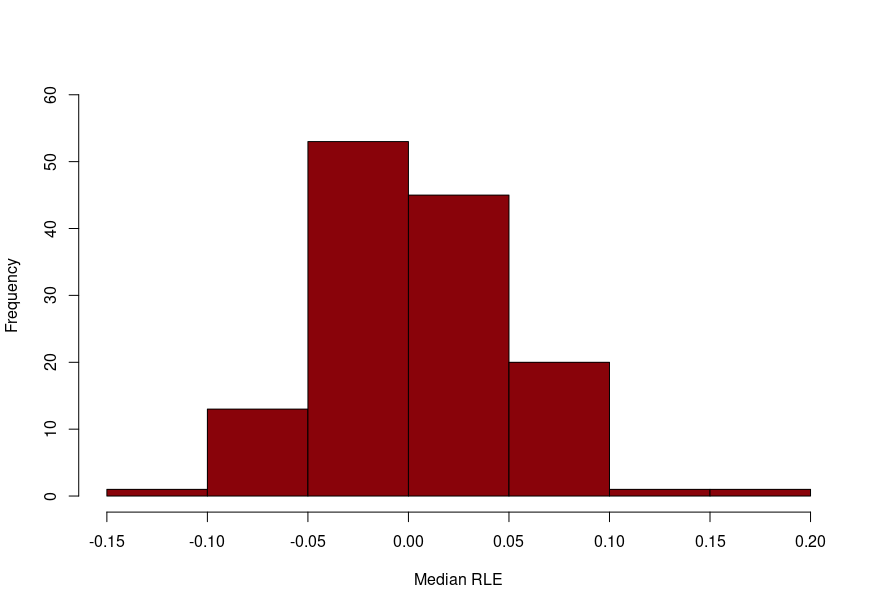
Hierarchical clustering was performed on the 1435 genes outputted from the 3 filters in the unsupervised probe set selection. The created dendrogram was divided into 2 clusters, C3 subtype and Non-C3 subtype. From the clustering and the cut dendrogram, a heatmap was created (Figure 5) to show the relationship of gene expression between the two sample clusters. A paired-sample t-test was then performed between the two sample subtypes using the t.test() function.

Using the R-package hgu133plus2.db, we mapped gene symbols back onto the Affymetrix IDs of the gene probes used (Carlson, 2016). IDs with no mapped gene symbol were dropped. Gene probes used for gene mapping were from data that underwent expression threshold filter and the variance filter for Table 1. That data was not filtered through the coefficient of variance threshold. While Marisa et al kept only the most variant probeset when multiple probes matched to the same gene, we kept the least significant adjusted p-value. This was done to aid choosing the most promising genes for a follow up study and to reduce the chance that there was some bias between sample groups (such as one region of a sequence had a more common gene variant in one group or one region of a sequence was more degraded in diseased group due to sample processing.) The top 1000 up-regulated and the top 1000 down-regulated genes were selected irrespective of significance, using the t-test values as proxy for magnitude of fold change (see Table 1: Top 10 up and down regulated genes). For comparison, gene probes used to produce Table 2 were filtered through all three filters mentioned in the ComBat analysis, including the coefficient of variation threshold not used for Table 1, and were the highest up-regulated and down-regulated genes used for the earlier clustering analysis.

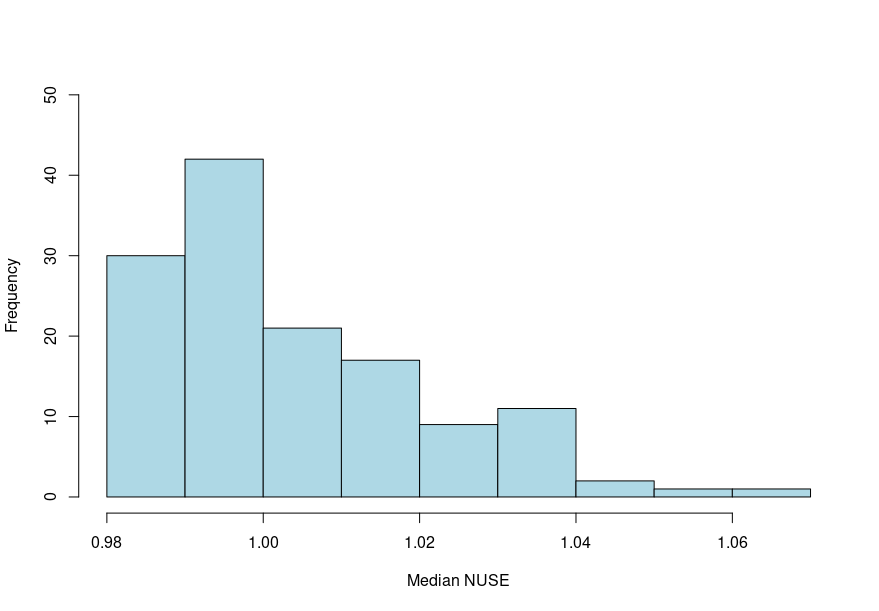
The GSEABase bioconductor package (Morgan, 2020) was then used to compare the differentially expressed genes (padj < 0.05) to the KEGG, GO, and Hallmark gene sets from MSigDB (MIT & UC, 2004-2020). (Note: this is using the gene list that was not filtered by coefficient of variation threshold. This ensured that non-differentially expressed genes remained in the analysis for the gene set enrichment analysis.) GO contains annotated gene sets associated with the same gene ontology term (biological process, cellular component, and molecular function), as derived from the Gene Ontology resource. KEGG contains gene sets derived from the KEGG pathway database. Hallmark contains gene sets that represent well defined biological states or processes, curated using a computational methodology between MSigDB collections. Two GSEABase:GeneSets were created representing the differentially expressed set of genes and the non-differentially expressed genes as determined in our earlier analysis. Each MSigDB gene set was compared (via intersection) with both our differentially expressed GeneSet and our non-differentially expressed gene-set. This comparison was used to create contingency tables for each gene set compared with our gene lists. Fisher tests were performed on each contingency table (fisher.test with default parameters). For each gene collection (KEGG, GO, Hallmark), p-values were then adjusted separately, using the Benjamini-Hochberg method in R’s p.adjust function. The statistically significant gene sets were identified (adjusted p-value < 0.05; see Table 3: Most statistically significant enriched gene sets).

Mapping the gene symbols, identifying the most differentially expressed genes, and identifying the most enriched gene sets took less than 1 minute when run locally on a desktop computer (all Rmarkdown chunks run together via knit within Rstudio). These steps in the analysis did not require computer cluster resources.

# **Results**

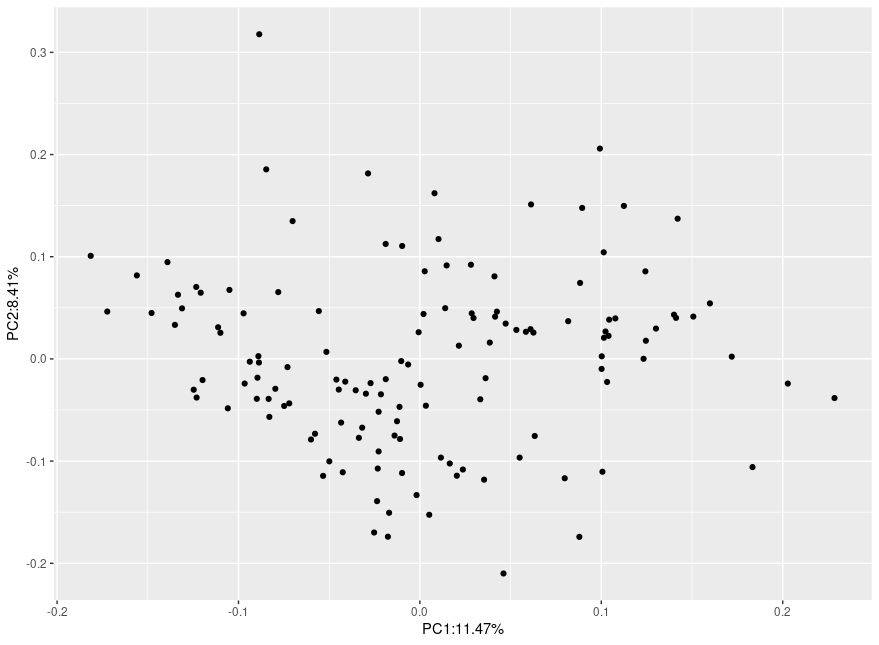


**Figure 1. Histogram of median RLE.** The histogram reflects the distribution of median RLE scores across all samples.

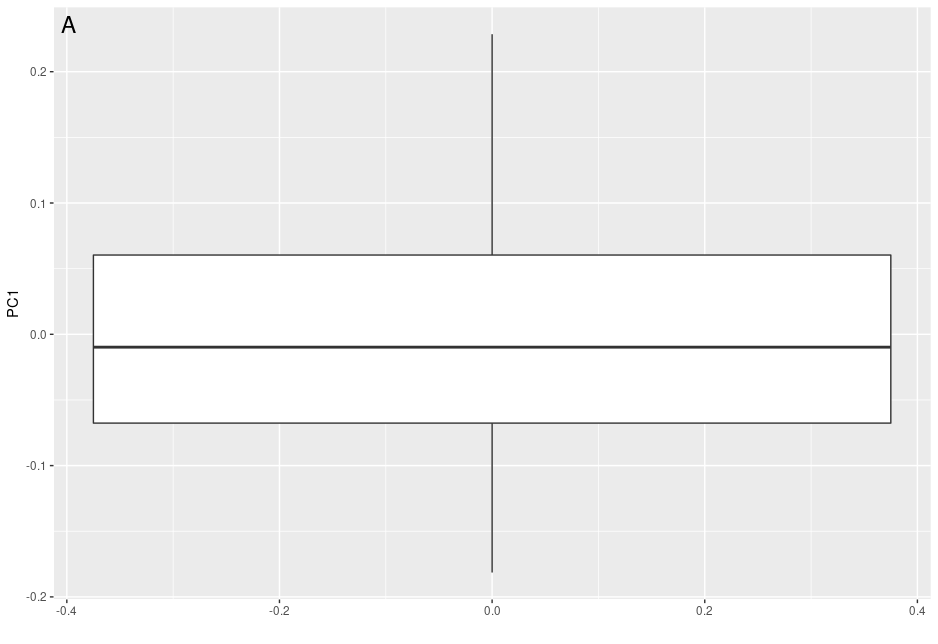
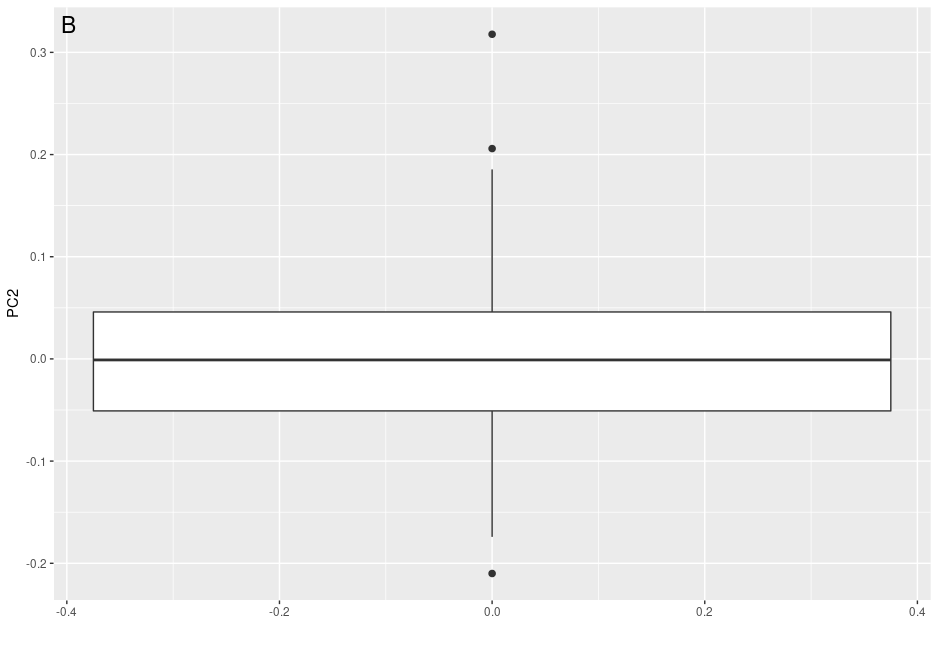


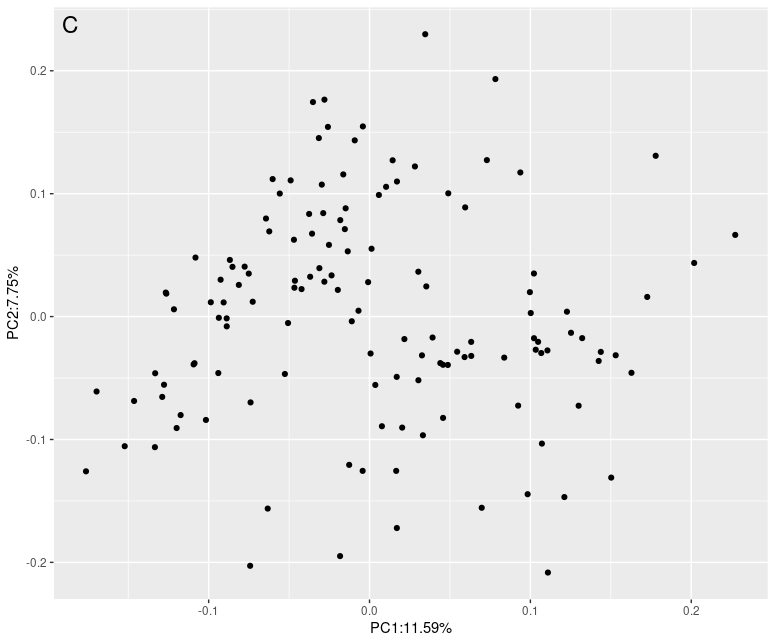
**Figure 2. Histogram of median NUSE.** The histogram reflects the distribution of median NUSE scores across all samples.

Quality control measures were performed to determine the quality of the microarray dataset by calculating median RLE and NUSE for each chip. It is evident that most of the samples have a median RLE close to zero, thus indicating that the samples are of good quality (Fig. 1). The median NUSE score was distributed around one, providing further indication that the samples are high quality (Fig. 2). The resulting distributions from the RLE and NUSE scores for all of the samples suggest that few samples are drastically different relative to the rest of the samples. With such robust quality control measures, no samples were removed from further analysis.



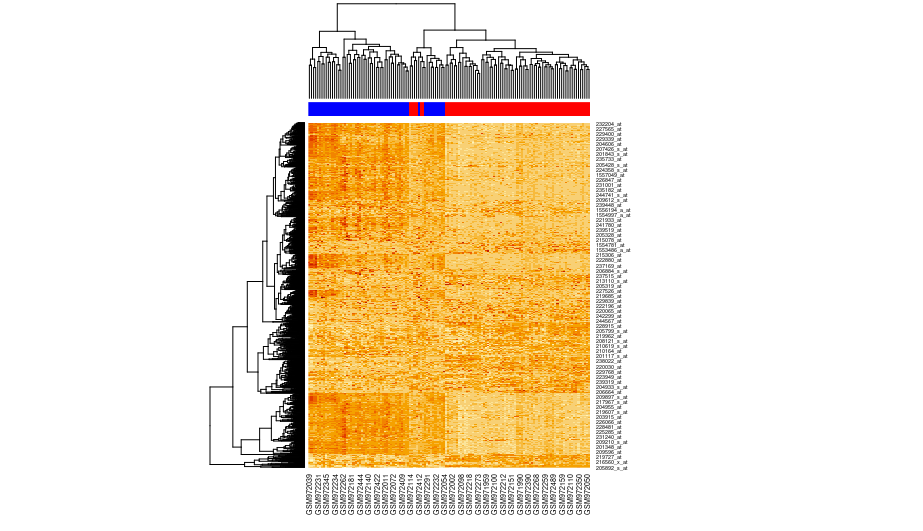
**Figure 3. Principal component analysis of principal components 1 and 2.** PC1 and PC2 explain for 11.47% and 8.41% of the variance, respectively.



**Figure 4. Examination for outliers in PC1 and PC2.** Outliers are evident in PC2 (Fig. 4B), but not in PC1 (Fig. 4A). After removal of the outlier sample, PCA was performed again (Fig. 4C).

PCA was performed to reduce dimensionality in order to better visualize and preserve variation (Fig. 3). No outliers were detected in PC1 data (Fig. 4A) while three outliers were detected in PC2 data (Fig. 4B) outside of the interquartile range (IQR). The three outliers correspond to the GSM972097\_050805-04.CEL.gz file. After the removal of this sample, the PCA was performed again to characterize the effects of the removal on the variance (Fig. 4C).



**Figure 5. Heatmap of gene (rows) expression by sample (columns).** Samples are subdivided into “C3”(58 samples), indicated by red in the column colorbar, and “Non-C3” (76 samples), indicated by blue in the column colorbar. Darker orange/red cells in the heatmap correspond to higher rates of probe set expression in the corresponding samples whereas light orange/yellow cells correspond to lower rates of expression.

Data from the ComBat adjusted matrix was filtered through the three filters for expression, variance, and coefficient of variation. 39661 probe sets passed through the first filter from the ComBat adjust matrix, while 54575 sets passed through the second filter, though 39563 of the filter 1 probe sets passed through the same filter, and filter 3 passed 1690 probes from the original data and 1435 of the probe sets that passed through filters 1 and 2.

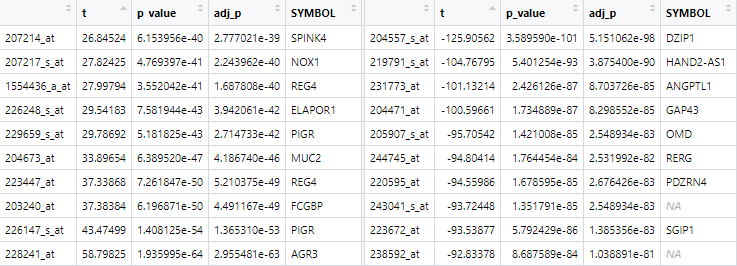
Hierarchical clustering was performed on the set of 134 patient samples. Two clusters were cut, with 58 samples in the first cluster, denoted as blue in Figure 5, and 76 samples were in the second cluster, denoted as red in Figure 5. The second cluster corresponds to the C3 cluster found in Marisa et al, (2013) where the sample data was initially obtained while the first cluster is equivalent to all of the non-C3 clusters from the same. A heatmap was then produced from the data, as shown in Figure 5.

Probe sets underwent paired sample t-test, using values from C3 cluster samples against values from non-C3 cluster samples. After p-value adjusting, there were 1164 of 1435 samples that were differentially expressed with a p-adjusted value of < 0.05 in the data set of the probe sets that passed all three filters. From the data set that underwent filters 1 and 2, 22729 of 39563 probe sets had t-test values with a p-adjusted value of <0.05.



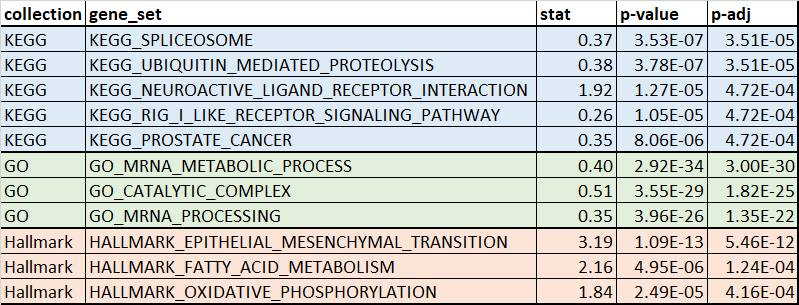
**Table 1: Top 10 up and down regulated genes.** Genes in red are upregulated in the C3 group. Genes in blue are downregulated in the C3 group. Probes are the Affymetrix gene probes, “t” is the t-test statistics, p\_value is the p-value, adj\_p is the adjusted p-value for multiple hypothesis correction, and SYMBOL are the gene symbols mapped to the probes using the hgu133plus2.db bioconductor package.

The most up-regulated genes in the C3 group were IgGFc-binding protein (FCGBP), Leucine rich repeat containing 31 (LRRC31), and Mucin 2 (Oligomeric mucus, MUC2). The most down-regulated genes were Secreted Frizzled Related Protein 2 (SFRP2), Fibronectin Type III Domain Containing 1 (FNDC1), and Armadillo repeat containing X-linked 1 (ARMCX1). (See Table 1: Top 10 up and down regulated genes).



**Table 2. Top 10 genes representative of the C3 (left) and non-C3 (right) subtypes .** These genes were from the data set that passed all three filters from the analysis, differing from those in Table 1 because Table 1 genes were not run through the third filter, the coefficient of variation (CV) filter that filtered for gene CV robustness.

Some of the most representative genes for C3 subtype (Table 2) include Anterior Gradient 3 Homolog (AGR3), Polymeric Immunoglobulin Receptor (PiGR), and Fc Fragment of IgG Binding Protein (FCGBP). Genes that were more representative of the non-C3 subtype include the Zinc Finger Protein (DZIP1), HAND2 Antisense RNA 1 (HAND2-AS1), and Angiopoietin-related protein 1 (ANGPTL1). Probes whose symbol was defined as NA were probes that we could not locate symbols for using the source materials and searching through databases by probe set id.

****

**Table 3: Most statistically significant enriched gene sets.** Top 3 statistically significant gene sets from GO, KEGG, and Hallmark. After Benjamini-Hochberg multiple hypothesis correction, KEGG had 3 pathways that were ranked third for enrichment.

The number of gene sets in GO, KEGG, and Hallmark collections were 10271, 186, and 50 respectively. There were 389, 43, and 9 statistically significant enriched gene sets in the GO, KEGG, and Hallmark gene sets respectively. There were 441 statistically significant enriched gene sets in total. The most enriched GO gene sets were mRNA metabolic process, catalytic complex, and mRNA processing. The most enriched KEGG pathways in our analysis were Spliceosome, Ubiquitin mediated proteolysis, Neuroactive ligand receptor interaction, RIG-I like receptor signaling pathway, and prostate cancer. The most enriched Hallmark gene sets were Epithelial mesenchymal transition, Fatty acid metabolism, and Oxidative phosphorylation. (See Table 3: Most statistically significant enriched gene sets.)

# **Discussion**

Analysis was conducted on 134 CEL files. After normalization of the data, the median RLE and NUSE scores were computed, both serving as quality control metrics to determine the reliability and quality of the data. RLE and NUSE are common but powerful tools for visualizing data heterogeneity. RLE measures a specific probe set’s expression in a microarray and any variation from other microarray samples. The optimal median RLE is centered around zero which assumes that the expression level of most genes are not affected by biological factors and that most genes have the same level of expression across all samples. As such, any deviation from zero is a measure of variation within a microarray. NUSE examines the precision of estimated expression of an array relative to other arrays based on the probe level model fit. The ideal median NUSE score is centered around one, indicating that each probe set on a chip has relatively uniform expression (Mccall et al., 2011). Arrays with standard errors higher than the median standard error (therefore, a NUSE score greater than one) or a larger interquartile range may be indicative of a poor quality array. This dataset had median RLE and median NUSE scores generally centered around zero and one, respectively, thus suggesting that the samples were of high quality (Fig. 1 and 2). Therefore, no samples were discarded.

PCA was performed on batch effect corrected data to reduce dimensionality of the dataset to allow for better visualization of variation and to identify any clusters amongst samples. The first and second principal component explained 11.47% and 8.409% of the variation, respectively (Fig. 3). Together, the first and second principal components explain 19.88% of the variability. These two principal components do not cover a large proportion of the variance in the dataset. Additionally, there is no well-defined clustering in the dataset (Fig. 3). Outliers in both principal components were defined as being three standard deviations away from the mean. In PC1, no outliers were detected while in PC2 three outliers were detected as being outside of the IQR (Fig. 4A and 4B). This resulted in the removal of one sample. PCA was performed again to determine the effect this removal had on the variance explained by principal components one and two (Fig. 4C). PC1 experienced a slight increase in percent variance explained as it had started with 11.47% to now 11.59%. On the other hand, PC2 experienced a slight decrease in variance explained as it had started with 8.41% to now 7.75%.

Filters were applied to the batch effect corrected data to determine which genes were expressed above threshold in 20% or more of the samples, narrowing down the number of probe sets to ultimately reduce the amount of noise within the data. Once genes with low rates of expression across samples were removed from the dataset, the data was filtered to genes whose variance differed significantly from the mean variance of all genes within the data using a chi-squared test and a p-value of 0.01. This removed genes whose gene expression variance differed little across all samples, which could have hindered future clustering and determining up-regulation and down-regulation further down the analysis stream. Finally, the last filter used a threshold of the coefficient of variation (CV) to determine which gene had a high degree of volatility within expression rates across the samples, which would facilitate later analysis and determining later patient clusters.

Once noise and dimensionality reduction finished, we were able to perform hierarchical clustering and subtype discovery on the probe sets that remained and patient samples. Patient samples underwent hierarchical clustering and then the outputted dendrogram was divided into 2 clusters, which correspond to the C3 and non C3 subtypes determined by Marisa et al. Using the subtype groupings found in the previous step, we were able to create the heatmap shown in Figure 5, which shows which patient samples in the columns corresponded to the C3 subtype, denoted with a red mark on the color bar, and the non-c3 subtype corresponding to a blue mark on the same bar. Finally, we were also able to conduct a paired-sample t-test, testing the C3 patient samples against the non-c3 samples of the same probe set. After performing t-tests, tej p-values gleaned were then run through a p-adjust function using “fdr” or Benjamin&Hochburg method to adjust. Of the 1435 probe set p-adjusted values, 1164 samples had p-adjust values less than 0.05. Gene probes regardless of p-adjust value were then mapped to gene symbols, differentially expressed genes were identified, and then gene set enrichment analysis was performed.

10 genes were identified from the t-test scores to represent each subtype, C3 and non-C3. The 10 genes that were most representative of the C3 subtype were for the most part identified by Marisa et al as one of the 1459 most variant probe sets, similar to the 1435 probes that passed the three filters for noise and dimensionality reduction. However, two of the top 10 for C3 did not make it as Marisa et al’s most variant probe set, despite testing in the probe set we filtered for. This could be that the 200+ missing samples that were left out of this analysis greatly affected what was determined as noise in the 134 samples we did use. The two genes were ELAPOR1 (Endosome-Lysosome Associated Apoptosis and Autophagy Regulator) and NOX1 (NADPH Oxidase 1)The 10 genes that were identified as representative of the non-C3 subtype were not on the 1459 most variant probe sets as found by Marisa et al. This could be due to us grouping several subtypes within this category-The group of samples covered by this group in Marisa et al ended up being several different cancer grouping subtypes, C1, C2, C4, C5, and C6.

Genes found to be representative of the C3 subtype included the AGR3 gene, which is a protein coding gene that is mostly expressed in ciliated airway epithelial cells and generally over-expressed in several kinds of cancers, FCGBP gene, which is a protein coding gene that may be involved in the maintenance of the mucosal structure as a component of the mucosa and may be involved in clotting as suggested by its association with Von Willebrand Disease, and PIGR, which is a member of the immunoglobulin superfamily that binds immunoglobulin molecules to the surface of epithelial cells (Stelzer et al, 2016). This could be indicative of C3 subtype cancers specifically affecting epithelial cells and the immune system more so than the other types of cancers involved in the non-C3 group. Genes found to be representative of the non-C3 subtype included the HAND2-AS1 gene, which is affiliated with the IncRNA class and is associated with Endometrial cancer, DZIP1, which is part of the pathways Signaling by GPCR and Signaling by Hedgehog and may participate in spermatogenesis, and ANGPTL1, which is a member of the vascular endothelial growth factor family (Stelzer et al, 2016). This could be indicative that the C3 subtype of cancer could be specifically down-regulating these genes or of several subtypes within the non-C3 subtype up-regulating these genes which would result in a lower strength association when subtypes are grouped into catch-all groupings due to sample size. These genes affect various functions like G-protein signaling between cells and cell growth.

We found many genes to be differentially expressed between C3 and the other colon cancer subtypes in our analysis. MUC2, the glycoprotein mucin 2 related to the formation of oligomeric mucus, was up-regulated in the C3 cancer subtype in our analysis. Marisa et al also found MUC2 to be differentially expressed across colon cancer subtypes and upregulated in C3. MUC2 plays an important role in intestinal function. Loss of MUC2 function promotes the progression of colon cancer and is a potential prognostic marker for colon cancer (Hsu, 2017). MUC2 and FCGBP were two of the top ranked up-regulated genes in the C3 colon cancer subtype whether or not we filtered by coefficient of variation.

FCGBP, Fc fragment of IG binding protein, was upregulated in the C3 subtype in our analysis, and has been previously shown to be related to cancer progression and prognosis. FCGBP may be anti-inflammatory and has been implicated in the chronic inflammatory disease ulcerative colitis. Both FCGBP and ulcerative colitis are associated with colorectal cancer, and FCGBP has been shown to decrease the survival time of colorectal patients (Qi, 2015). Conversely, FCGBP was most recently shown to have decreased expression in disease development in liver metastatic colorectal cancer patients and increased expression was associated with survival (Yuan, 2021). Marisa et al found FCGBP to be variably expressed across the cancer subtypes they looked at. We found this gene to be upregulated in the C3 subtype, and Marisa et al found it to be upregulated in both the C3 and C6 subtypes, which supports the hypothesis that this gene could be either protective or harmful depending on the colon cancer subtype.

Epithelial mesenchymal transition and fatty acid metabolism have been implicated in cancer. These hallmark gene sets overlap with cancer related gene sets (MSigDB). Epithelial mesenchymal transition refers to the process by which epithelial cells lose their cell to cell adhesion and become, in extreme cases, mobile, invasive, and in a stem cell like state (Zhang, 2018). Changes in fatty acid metabolic pathways have been implicated in cancer as a way for cancer cells to help fuel their energy production, to function as secondary messengers, and as important structural components needed for cell division (Koundouros, 2020). Differences in the epithelial mesenchymal transition and fatty acid metabolism gene sets could be associated with cancer severity and prognosis.

The believed relationship between oxidative phosphorylation and cancer has changed over time. It is currently believed that oxidative phosphorylation is required for cancer cells to survive, and an increased dependency characterizes some cancer stem cells; oxidative phosphorylation inhibitors are being considered as therapeutic agents (Sica, 2020). Finding enrichment for the hallmark gene set oxidative phosphorylation suggests that there could be differences in this gene set associated with cancer severity in this dataset.

Generally, the C3 subtype found in Marisa et al was reproducible (see Figure 5), despite the difference in data available. Marisa et al was working with a larger patient sample set, 750 versus the 134 used in this paper, which would account for some of the filtering and clustering discrepancies. The smaller number of patient samples streamlined our process, for instance it cut down on the number of sample subtypes we could find in our clustering, but also could have changed the values we got when we had to take dataset or probeset means, coefficients of variation, or the median variance, all of which would have impacted filter thresholds and selectivity.

We only partially reproduced the results in Marisa et al. Initial filtering yielded similar results, 1435 from our data vs 1459 found in the source material. Clustering yielded similar data as Marisa et al. We were able to differentiate most of C3 samples from the non-C3 samples. Our patient sample size was too small to subdivide the rest of the non-C3 samples into the clusters found by Marisa et al. Once the t-tests were performed between subtypes, we had 1164 genes with p-adjusted values of <0.05, whereas Marisa et al had slightly less at 1108. In either case, the number of probes that we found did not differ much more 5% from the findings of Marisa et al, indicating that the patient samples used to represent Marisa et al’s dataset was fairly well-matched as a subset

Our most up-regulated and most down regulated genes in group C3 did not match by ranking with Marisa et. al’s results as presented in Table S2 of most differentially expressed genes. Several of the genes that made it into our top 10 up or down regulated (see our Table 1, Marisa 2013 Table S2) were not in the author’s list of most differentially expressed genes when searched be either probe id or gene symbol. We found FOXA3, C4orf19, CRYM, MISP3, MIR1199, ASRGL1, ARMCX1, SULF1, FRMD6, SERPING1, and SERPINF1 that did not make it into Marisa et al’s results. However, the list of probes used to produce Table 1 were produced without the coefficient of variation filtration step; this likely is the largest source of the discrepancy. That step, when done on our dataset, reduced the list of probes to 1/27 of the previous size. Genes related to the missing genes for FOX, C4orf, and SERPIN genes did make it into Marisa et al’s most differentially expressed genes. These differences are likely due to several factors, including how we filtered our results, how we defined our clusters, and how we ranked those results. It is worth noting that, unlike the authors, we were focused only on one comparison (group C3 to every other cluster) when determining what genes we believed were most differentially expressed. The genes that we considered most differentially expressed at least agreed with the direction of the change that Marisa et al found.

Our top enriched gene sets did not match the gene sets mentioned in the paper, but there was some qualitative overlap. This difference was not surprising as Marisa et al curated their named enriched gene sets, considered the direction of change for each of their defined subtypes, and did not name Hallmark gene sets. Marisa et al also had a different set of differentially expressed genes that were input into the enrichment analysis. There were some gene sets that qualitatively overlapped. For example, Marisa et al found subtype C3 had decreased expression of epithelial mesenchymal transition and motility, which is a hallmark gene set we found to be significantly enriched in differentially expressed genes. Marisa et al also found several fatty acid metabolic pathway gene sets to be variable between subtypes, which is another hallmark gene set we found to be enriched.

There are several aspects of our analysis that would have a significant aspect on our results and resulted in some deviation from the findings in Marisa et al (2013). Marisa et al used a different gene expression threshold for filter 1; we chose probe sets that were expressed by at least 20% of samples at the same threshold where Marisa et al used probesets expressed above threshold in 5%, potentially resulting in us initially selecting for less samples. Our gene set enrichment analysis was likely impacted by the significance threshold chosen to determine differentially expressed genes. If we had considered the direction of the change (increased vs decreased expression) between treatment groups our gene set enrichment analysis would have likely improved the accuracy and interpretability of our results.

# **Conclusion**

FCGBP was a particularly interesting gene, as it has been implicated for both better and worse prognosis in colorectal cancer. We found this gene to be upregulated in the C3 subtype, and Marisa et al found it to be upregulated in both the C3 and C6 subtypes, which would support the hypothesis that this gene could be either protective or harmful depending on the colon cancer subtype. This gene would be of interest for further follow up studies.

The hallmark gene sets epithelial mesenchymal transition and fatty acid metabolism describe attributes, functions, and processes that have been previously implicated in cancer. The hallmark gene set oxidative phosphorylation points related to a process that has, in different years, been believed to both be necessary and unnecessary for cancer cells to survive. Seeing gene set enrichment may mean that these processes are related to cancer severity and prognosis in colon cancer. Oxidative phosphorylation, in particular, is of current interest in oncology. A better understanding of oxidative phosphorylation’s relationship to the different cancer subtypes described in Marisa et al would be an interesting followup to this analysis.

One challenge of this project was that analysis decisions and bugs were occasionally not found until the next person in the pipeline conducted their analysis and were able to flag something as a possible issue. We overcame this via discussions between the relevant group members and troubleshoot with members of other groups who were having similar issues. We also struggled with how to interpret the results we did have. More time spent on interpreting the results would help in future projects. Data workup and csv files are located under dachshund directory on scc.

# **References**

Broad Institute Massachusetts Institute of Technology; Regents of the University of California. (2004-2020). Molecular signatures database v7.2. Retrieved February 21, 2021, from <https://www.gsea-msigdb.org/gsea/msigdb>

Gautier L, Cope L, Bolstad BM, Irizarry RA (2004). “affy—analysis of Affymetrix GeneChip data at the probe level.” Bioinformatics, 20(3), 307–315.

Hutchins, G., Southward, K., Handley, K., Magill, L., Beaumont, C., Stahlschmidt, J., Richman, S., Chambers, P., Seymour, M., Kerr, D., Gray, R., & Quirke, P. (2011). Value of mismatch repair, KRAS, and BRAF mutations in predicting recurrence and benefits from chemotherapy in colorectal cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, *29*(10), 1261–1270. https://doi.org/10.1200/JCO.2010.30.1366

Hsu, H.-P., Lai, M.-D., Lee, J.-C., Yen, M.-C., Weng, T.-Y., Chen, W.-C., Fang, J.-H., & Chen, Y.-L. (2017). Mucin 2 silencing promotes colon cancer metastasis through interleukin-6 signaling. *Scientific Reports*, *7*(1), 5823.<https://doi.org/10.1038/s41598-017-04952-7>

Jass J. R. (2007). Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. *Histopathology*, *50*(1), 113–130. <https://doi.org/10.1111/j.1365-2559.2006.02549.x>

Koundouros, N., & Poulogiannis, G. (2020). Reprogramming of fatty acid metabolism in cancer. *British Journal of Cancer*, *122*(1), 4–22.<https://doi.org/10.1038/s41416-019-0650-z>

Marc Carlson (2016). hgu133plus2.db: Affymetrix Human Genome U133 Plus 2.0 Array annotation data (chip hgu133plus2). R package version 3.2.3.

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>

Martin Morgan, Seth Falcon and Robert Gentleman (2020). GSEABase: Gene set enrichment data structures and methods. R package version 1.52.1.

McCall MN, Murakami PN, Lukk M, Huber W, Irizarry RA. Assessing affymetrix GeneChip microarray quality. *BMC Bioinformatics*. 2011;12:137. Published 2011 May 7. doi:10.1186/1471-2105-12-137

Popat, S., Hubner, R., & Houlston, R. S. (2005). Systematic review of microsatellite instability and colorectal cancer prognosis. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, *23*(3), 609–618. https://doi.org/10.1200/JCO.2005.01.086

Salazar, R., Roepman, P., Capella, G., Moreno, V., Simon, I., Dreezen, C., Lopez-Doriga, A., Santos, C., Marijnen, C., Westerga, J., Bruin, S., Kerr, D., Kuppen, P., van de Velde, C., Morreau, H., Van Velthuysen, L., Glas, A. M., Van't Veer, L. J., & Tollenaar, R. (2011). Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, *29*(1), 17–24. <https://doi.org/10.1200/JCO.2010.30.1077>

Qi, C., Hong, L., Cheng, Z., & Yin, Q. (2016). Identification of metastasis-associated genes in colorectal cancer using metaDE and survival analysis. *ONCOLOGY LETTERS*, 7.

Sica, V., Bravo‐San Pedro, J. M., Stoll, G., & Kroemer, G. (2020). Oxidative phosphorylation as a potential therapeutic target for cancer therapy. *International Journal of Cancer*, *146*(1), 10–17.<https://doi.org/10.1002/ijc.32616>

Stelzer G, Rosen R, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, Iny Stein T, Nudel R, Lieder I, Mazor Y, Kaplan S, Dahary D, Warshawsky D, Guan - Golan Y, Kohn A, Rappaport N, Safran M, and Lancet D. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analysis , Current Protocols in Bioinformatics(2016), 54:1.30.1 - 1.30.33.doi: 10.1002 / cpbi.5. [PDF]

Yuan, Z., Zhao, Z., Hu, H., Zhu, Y., Zhang, W., Tang, Q., Huang, R., Gao, F., Zou, C., Wang, G., & Wang, X. (2021). IgG Fc Binding Protein (FCGBP) is Down-Regulated in Metastatic Lesions and Predicts Survival in Metastatic Colorectal Cancer Patients. *OncoTargets and Therapy*, *Volume 14*, 967–977.<https://doi.org/10.2147/OTT.S285171>

Zhang, Y., & Weinberg, R. A. (2018). Epithelial-to-mesenchymal transition in cancer: Complexity and opportunities. *Frontiers of Medicine*, *12*(4), 361–373.<https://doi.org/10.1007/s11684-018-0656-6>