Data curator: Aravind
Programmer: Dhatri
Analyst: Caroline
Biologist: Eliza

BF528 Project 1
2/17/23

Validating Unsupervised Methods to Classify Colon Cancer Tumors Using Gene Expression Microarrays

Introduction

Colorectal cancer (CRC) is the third most commonly diagnosed cancer and the fourth-leading cause of cancer death worldwide in both sexes (Lotfollahzadeh et al., 2022). In the United States, CRC is the second leading cause of death among all cancer types, with an estimated 50,260 deaths per year when colon and rectum are combined (Lotfollahzadeh et al., 2022). Although new cases and mortality have steadily declined in recent years, CRC remains a global threat to the population. The progression of CRC from colorectal adenoma to carcinoma results from three major pathways: microsatellite instability, chromosomal instability, and CpG island methylator phenotype (Nguyen & Duong, 2018). More recently, evidence has shown that CRC is a heterogeneous disease, suggesting that genetic characteristics of the tumors determine their prognostic outcome and response to drug therapies (Nguyen & Duong, 2018). Despite recent advances in CRC screening and diagnosis, pathological and clinical factors are still the main tools used to select patients to receive chemotherapy (Auclin et al., 2017). While these factors are useful, they do not provide enough information to decide which patients will benefit from adjuvant treatment, and they fail to accurately predict recurrence (Marisa et al., 2013). Due to the heterogeneity of CRC and the lack of predictors for recurrence, there is a clear need for a comprehensive molecular classification to be established for determining prognosis in clinical practice.

The aim of the current study was to create a standard and reproducible molecular classification of colon cancer (CC) using mRNA expression profile analyses (Marisa et al., 2013). To do this, the authors used a multicenter series of CC samples, including tumor samples from 750 patients with stage I-IV CC who underwent surgery. With the patient samples, the authors performed a genome-wide mRNA expression analysis using gene expression microarrays. Using this dataset, further assessment was done to explore the associations between molecular subtypes and clinicopathological factors, DNA alterations, and prognosis. The results from this study provide the first robust transcriptome-based classification of CC into six molecular subtypes that account for clinicopathological variables and common DNA markers. This analysis could improve prognostic models and therapeutic strategies to combat CC in the future. For our analysis, we aimed to reproduce differential gene expression analysis as seen in Marisa et al. for the C3 and C4 subtypes and uncover the biological significance of the resulting gene expression patterns.

Data

The two data sources used in this study were collected from the French Cartes d'Identité des Tumeurs (CIT) program and publicly available datasets. There were 750 CC samples collected from CIT suitable for common DNA alteration characterization as mentioned above. Out of these 750 samples, utilizing quality control criteria, 566 tumor RNA samples were selected. The discovery set consisted of 443 tumors from the CIT cohort, and the remaining were used in the validation set. Included in the validation set were the remaining CIT CC samples, CC samples from seven Affymetrix (Affymetrix U133P2) publicly available datasets (GSE13067, GSE13294, GSE14333, GSE17536/17537, GSE18088, GSE26682, and GSE33113), and CC samples from a non-Affymetrix TCGA program which amounted to a total of n=1181 samples. The data obtained from the publicly available data were used in the validation dataset only. Only stage II and III patients were considered for survival analyses, and a subset of the data used for validation and discovery data was selected. Thus, 359 cases in the discovery set and 416 in the validation were selected for survival analysis. In our analysis, we utilized 134 samples to conduct subsequent studies.

The protocol for preparing and reading the RNA libraries was as follows (de Reyniès, A et al., 2009). Tumor samples were extracted from patients, snap-frozen, and stored under liquid nitrogen at -80 C until they were used in this study. The steps involved in preparing the sample were powdering with liquid nitrogen. Samples were extracted using 6 different methods depending on the sample type. The methods used to extract RNA were given in the project_metadata.csv file. They are:

E1: total RNA was extracted using Trizol (Invitrogen, Carlsbad, CA),
E2: total RNA was extracted using Manual-Cesium Chloride,
E3: total RNA was extracted using Manual-Rneasy Micro (Qiagen),
E4: total RNA was extracted using Manual-Rneasy Mini (Qiagen),
E5: total RNA was extracted using RNA NOW,
E6: total RNA was extracted using TriReagent & Manual-Rneasy Mini (Qiagen)

Quality control measures were performed on samples analyzed by electrophoresis and quantified with Nano-Drop. One of the quality control metrics applied to the samples was a 28s/18s ratio of above 1.8g for microarray data. Other criteria for exclusion of selection of data for patients were those diagnosed with primary rectal cancer or who received radiation therapy and/or preoperative chemotherapy. Microarray analyses were carried out with 3 µg of total RNA of each sample as starting material and 10 µg cRNA per hybridization. The total RNA was amplified according to the manufacturer-provided one-cycle target labeling protocol. The labeling was done with biotinylated cRNA targets. After labeling, 10 μg of cRNAs were hybridized for 16 hours at 45°C, 60 rpm on Human GeneChip HG-U133 Plus 2.0 arrays. The chips were then scanned with a GCOS 1.4 with a wavelength of 570 nm and a size pixel selection of 1.56 micrometers. The reference genome used in this study is the human genome (Taxid: 9606) (Genome Reference Consortium GRCh37).

Methods

Normalization and Batch Correction

Microarray data must be normalized prior to downstream processing to account for technical variation between samples. To carry out this analysis, Bioconductor packages such as affy (1.74.0), affyPLM (1.72.0), sva (3.44.0), AnnotationDbi (1.58.0), and hgu133plus2.db (3.13.0) were utilized. Raw data was in the form of CEL files containing information extracted from probes on an Affymetrix GeneChip. This data was read using the ReadAffy function and normalized using the robust multi-array (RMA) average expression measure implemented in the R package affy (Gautier, Cope, Bolstad, & Irizarry, 2004). Using the median polish algorithm, the RMA function performs background correcting, normalization, and summarization.

Before processing the data, the quality was assessed using two tools: Relative Log Expression (RLE) and Normalized Unscaled Standard Error (NUSE). RLE and NUSE scores were computed using the fitPLM function in the R package affyPLM; normalize and background was set to true as additional arguments (Bolstad et al., 2005). RLE values are computed for each probeset by comparing the expression value on each array against the median expression value for that probeset across all samples. Similarly, for NUSE, the standard error estimates obtained for each sample from fitPLM are taken and standardized across all samples so that the median standard error for that sample is 1 (Bolstad et al., 2005). These processes account for differences in variability between samples. The median RLE and NUSE of each sample were examined by plotting the distribution of the medians in a histogram (**Figure 1**).

The batch effects were corrected using the ComBat method implemented in the sva R package (Johnson, Li, & Rabinovic, 2006). Sva was used with the ComBat function to remove known batch effects and other potential latent sources of variation (Leek, Johnson, Jaffe, Parker, & Storey, 2022). Using an empirical Bayesian framework, the ComBat function adjusts for known batches (Johnson, Li, & Rabinovic, 2006). The batch effects of this study included the center and RNA extraction methods. Features of interest, such as tumor and MMR status, were considered per Marisa et al. and included in the batch correction. Using these two variables when running ComBat to correct for batch effects helped preserve features of interest in the data. The resulting output was written out to a CSV file.

Principal Component Analysis (PCA)

Creating the PCA plot required the ggplot2 (3.3.6) (Wickham, H. 2016) and ggfortify (0.4.15) packages (Tang and Li, 2016). The PCA plot was made using the built-in R function prcomp, and the parameters scale and center were set to false. Scaling was performed within each gene to ensure that all the variables were on the same scale and the variance explained by each variable was comparable. Centering was another important step in ensuring the resulting components were only considering the variance within the dataset and not capturing the overall mean of the dataset as an important variable. The values of each principal component were accessed using the rotation attribute. Percent variability explained by each principal component was conveyed with the importance attribute. The resulting output was a PCA plot of principal component 1 versus principal component 2 of all samples (**Figure. 2**).

Noise Filtering

To reduce noise in the normalized gene expression dataset, three expression filters were applied using established methods from Marisa et al. To isolate genes consistently expressed across samples, the dataset was filtered for genes expressed in at least 20% of subjects as determined by the threshold RMA-normalized expression value > $\log_2(15)$. To filter for genes with variable expression levels across subjects, we also selected probes with a variance significantly greater than the median probe variance using a two-tailed chi-square test ($p < 0.01$). Finally, we filtered for genes with a robust coefficient of variation (CV) by selecting genes with CV > 0.186, the threshold determined by Marisa et al. using a Gaussian mixture model clustering approach.

Hierarchical Clustering

In order to stratify patients according to gene expression, hierarchical clustering was performed after applying all three probe filters to the normalized dataset. Clustering was performed using the 1 - Pearson correlation method for distance and Ward's method for linkage. The resulting dendrogram was cut into two clusters for further analysis (Marisa et al., 2013).

Subtype Discovery

The Welch t-test was performed for each probe to determine differential gene expression between the two clusters of patients. The false discovery rate (FDR) was calculated to account for multiple comparisons. This process was repeated using the original dataset with the variance expression filter only.

Gene Set Enrichment Analysis

Three gene sets were retrieved from the Molecular Signatures Database. The cancer hallmark gene set collection contained 50 total genesets, the Kyoto Encyclopedia of Genes and Genomes (KEGG) gene set collection contained 186 total gene sets, and the gene ontology (GO) pathway gene set collection contained 10,561 total gene sets. The geneset files from the database used gene symbol identifiers. Using the differential expression matrix, these genesets were compared with the top 1000 up-and-down-regulated genes of each subtype from the gene expression profiles. Fisher's exact test was used to compare the data sets, and the enriched gene sets were found using the Benjamini-Hochberg (FDR) procedure.

Results

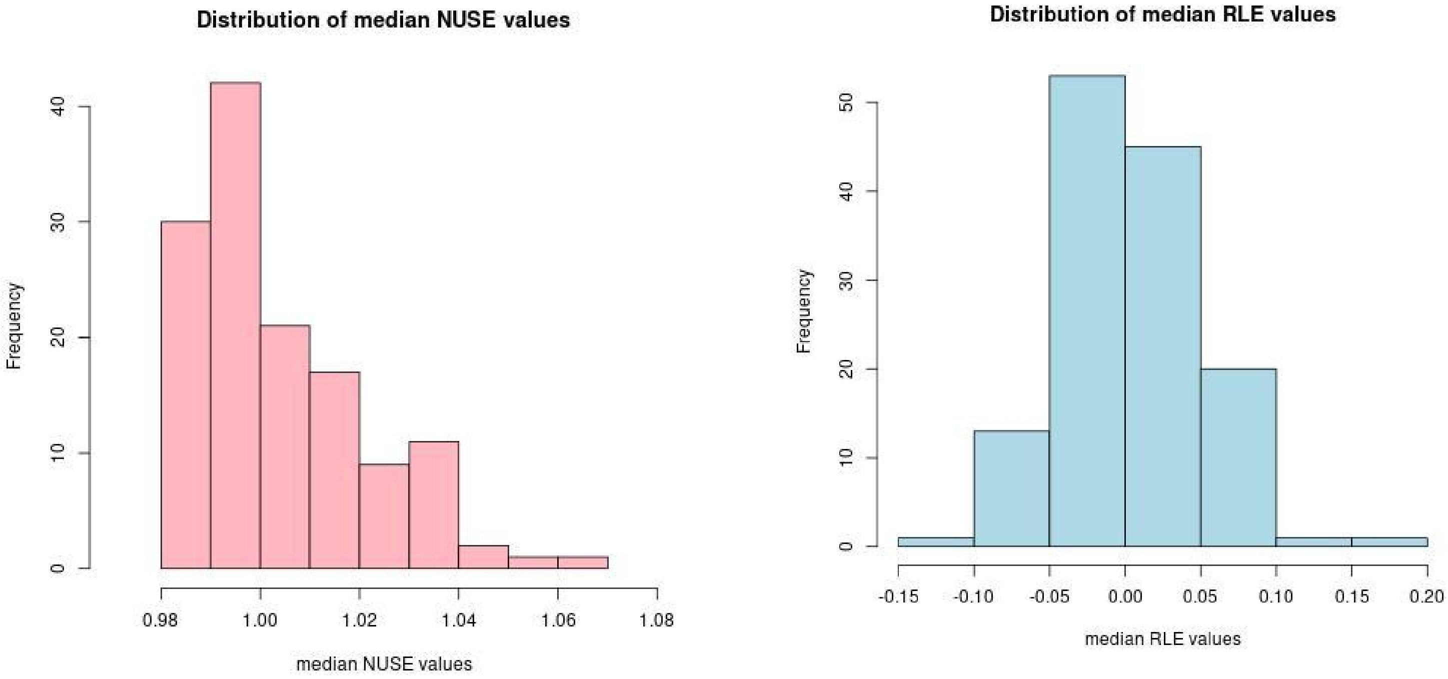


Figure 1. (A) Histogram on the left depicts the median NUSE scores of the 134 samples. (B) Histogram on the right depicts the median RLE scores of the same 134 samples.

The median NUSE and RLE were computed to assess the microarray samples' quality (**Fig. 1**). The NUSE scores are from 0.98 to approximately 1.07. However, most of the distribution is centered around 1, an acceptable standard metric, as the median NUSE error was computed to be 1 across all samples. RLE scores are from -0.15 to 0.20, with most distributions centered around 0. Typically, poorer-quality samples show up with samples not centered about 0 and more spread out. Since the distributions of the NUSE and RLE values fell in the acceptable standard metric, the samples were considered high-quality for downstream analyses.

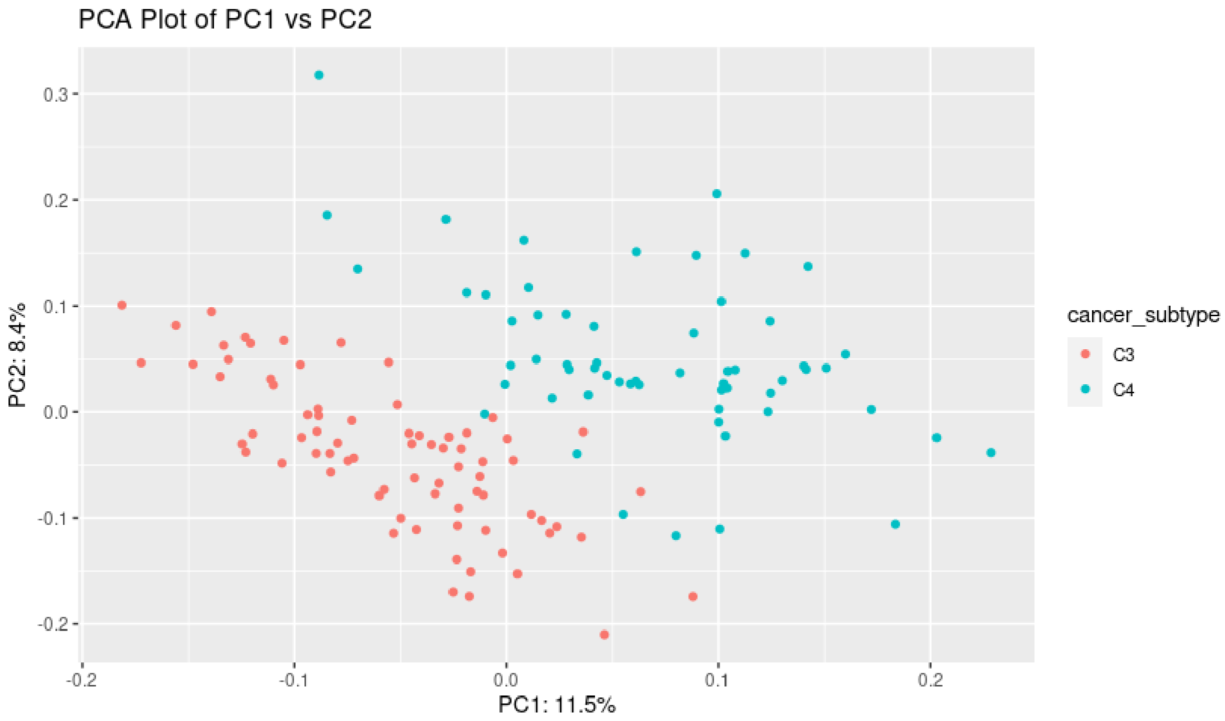


Figure 2. Principal component 1 vs. principal component 2 of all the samples clustered according to cancer subtype.

The plot shows that the samples are loosely clustered into 2 different groups with a handful of outliers. Although the principal components had a total variability of ~20%, the plot does not explain much variance. The sample points are low, and we are unable to conclude anything which will contribute to our analysis. We could extrapolate information from the plot and see actual clustering if there were more sample points.

Hierarchical Clustering and Subtype Discovery

After applying three expression filters to the normalized, quality-controlled dataset, 1,531 probes were left for further analysis. Unsupervised clustering resulted in two groups of patients, a C4 cluster containing 59 subjects and a C3 cluster containing 75 patients (**Fig. 3**). The two patient clusters correlated exactly with their reported cancer subtypes. These cluster assignments were then used for differential expression analysis.

Differential expression was assessed for the probes that passed the chi-squared test for variance alone. From this set of probes, 23,781 passed the significance threshold for differential expression between C3 and C4 patients (FDR < 0.05). Of the 1,531 probes that passed all three filtering conditions, 1,251 passed the significance threshold for differential expression between

C3 and C4 patients (FDR < 0.05). From this list of filtered probes, the 10 with the lowest FDR were chosen as those best representing the C3 and C4 clusters. All 10 genes except GPC6 were found to exist in the authors' 1108 subtype discriminant probeset, which was generated using a supervised method in their discovery patient subset, confirming our top 10 genes as those likely to be distinct in C3 or C4 (Marisa et al., 2013).

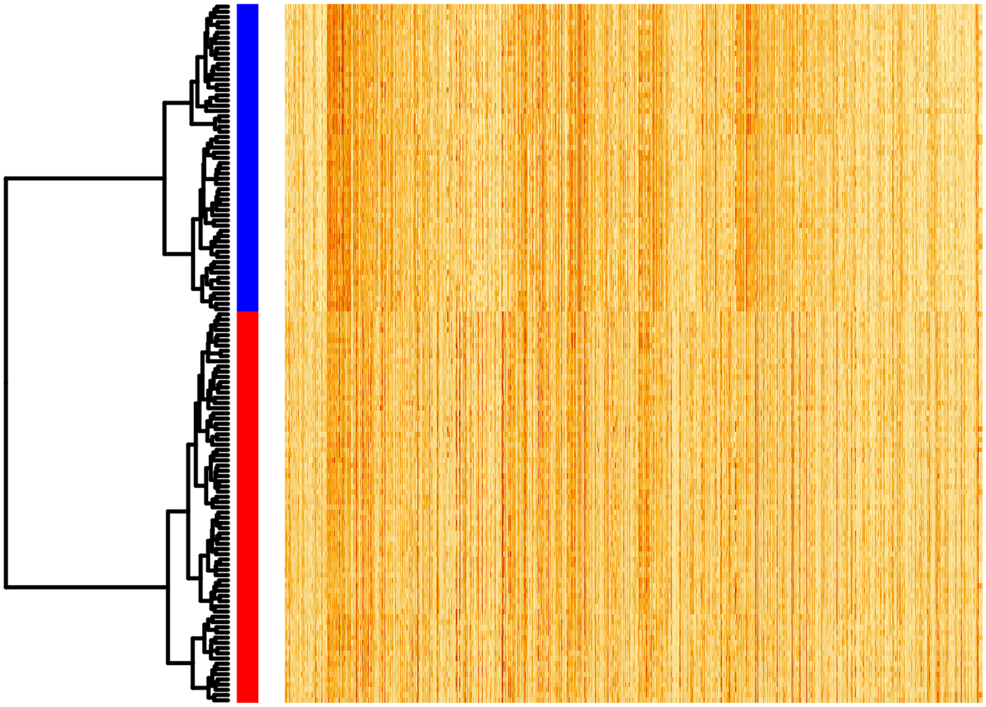


Figure 3. Heatmap showing gene expression (columns) across samples (rows) with color intensity corresponding to increased expression. 134 subjects were stratified into two groups via hierarchical clustering and colored by subtype (C3 in red, C4 in blue).

In addition to selecting the 10 probes with the lowest FDR for our characteristic gene set, we also compared our resulting 1,251 genes with those that Marisa et al. found to be characteristic of C3 and C4. The authors validated their six-subtype classification system by building a centroid classifier based on their discovery sample set containing 57 genes (Marisa et al., 2013). We compared our list of 1,251 filtered, differentially expressed genes to the authors' 57-gene classifier and found 9 matches out of the 20 probes in the C3 and C4 centroids, representing 5 upregulated and 5 downregulated genes specific to each subtype. A partial overlap was expected, given that the authors' classifier was used to distinguish between all 6 subtypes and this analysis only included C3 and C4. One member of the 57-gene classifier, MGP, was also found in our top 10 differentially expressed genes list. After combining the top 10 differentially expressed genes from our analysis with the 9 genes overlapping between our analysis and the authors' 57-gene classifier, we generated a list of 18 genes representative of each cluster summarized in **Table 1**. The most strongly associated subtype was chosen by determining which subtype showed the largest |log(fold change)| compared with other subtypes, according to Marisa et al.

Table 1. Summary of genes characteristic of C3 and C4 subtypes generated after noise reduction and differential expression analysis with the Welch T-test. The associated subtype was determined by selecting the subtype with the highest |log(fold change)| compared to other subtypes based on the authors' findings (Marisa et al., 2013).

Probe Set Id	**Gene Symbol**	**Associated Subtype**
227059_at	GPC6	--
203240_at	FCGBP	C3
211959_at	IGFBP5	C3
209868_s_at	RBMS1, RBMS1P1	C3
223970_at	RETNLB	C3
201147_s_at	TIMP3	C3
210517_s_at	AKAP12	C4
225242_s_at	CCDC80	C4
221019_s_at	COLEC12	C4
209210_s_at	FERMT2	C4
226930_at	FNDC1	C4
225464_at	FRMD6	C4
204457_s_at	GAS1	C4
202291_s_at	MGP*	C4
225782_at	MSRB3	C4
226069_at	PRICKLE1	C4
223122_s_at	SFRP2	C4

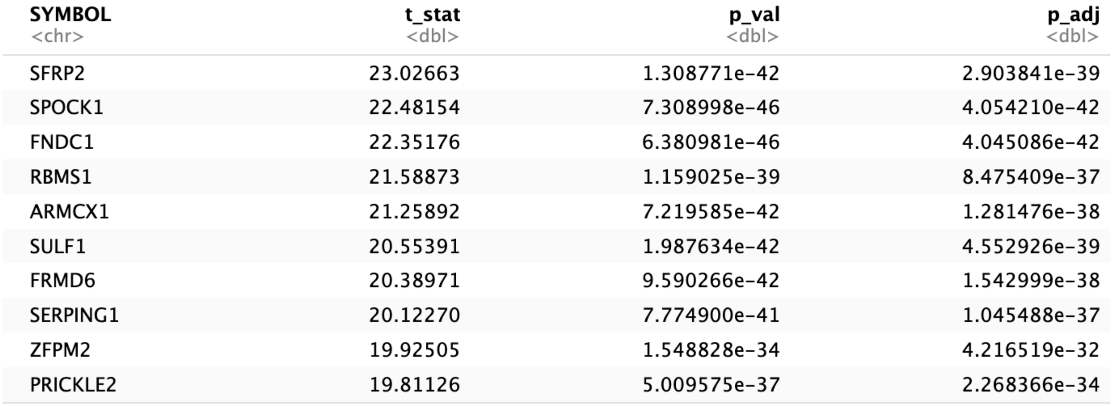
202363_at	SPOCK1	C4

*found in top 10 differentially expressed genes and Marisa et al. 57-gene classifier.

Table 2. The number of significantly enriched gene sets in each gene set collection with adjusted p-value < 0.05.

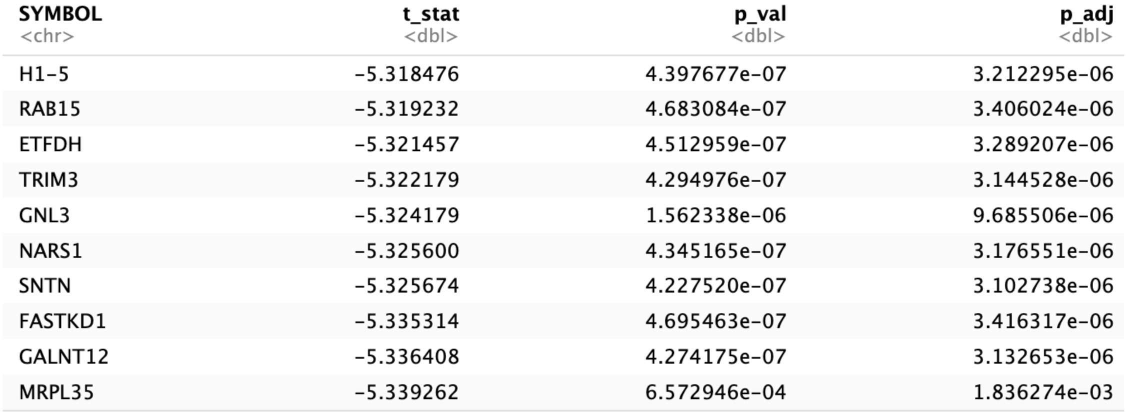
Gene Set Collection	Up-Regulated	Down-Regulated
Hallmark	37	19
GO	2,841	910
KEGG	57	56

Table 3. Top 10 Up-Regulated Probesets.



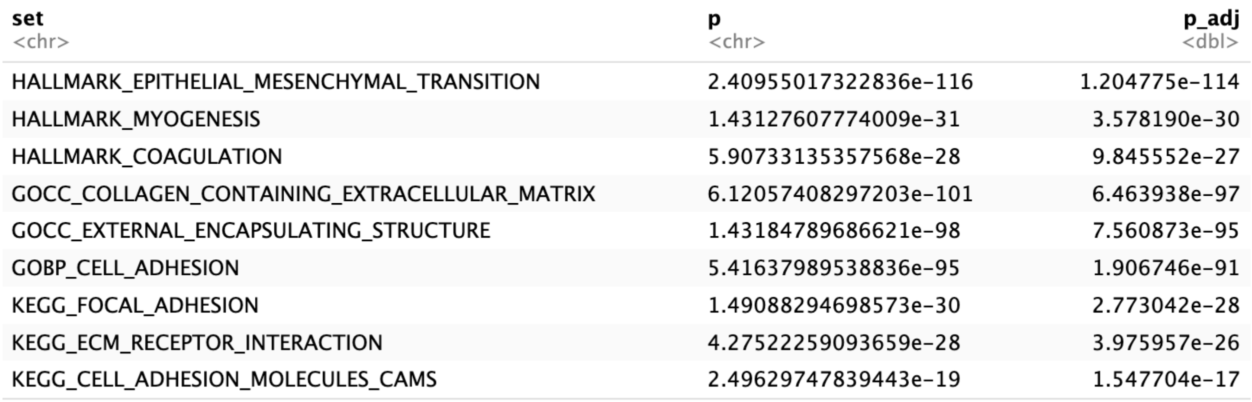
SYMBOL	t_stat	p_val	p_adj
<chr>	<dbl>	<dbl>	<dbl>
SFRP2	23.02663	1.308771e-42	2.903841e-39
SPOCK1	22.48154	7.308998e-46	4.054210e-42
FNDC1	22.35176	6.380981e-46	4.045086e-42
RBMS1	21.58873	1.159025e-39	8.475409e-37
ARMCX1	21.25892	7.219585e-42	1.281476e-38
SULF1	20.55391	1.987634e-42	4.552926e-39
FRMD6	20.38971	9.590266e-42	1.542999e-38
SERPING1	20.12270	7.774900e-41	1.045488e-37
ZFPM2	19.92505	1.548828e-34	4.216519e-32
PRICKLE2	19.81126	5.009575e-37	2.268366e-34

Table 4. Top 10 Down-Regulated Probesets.



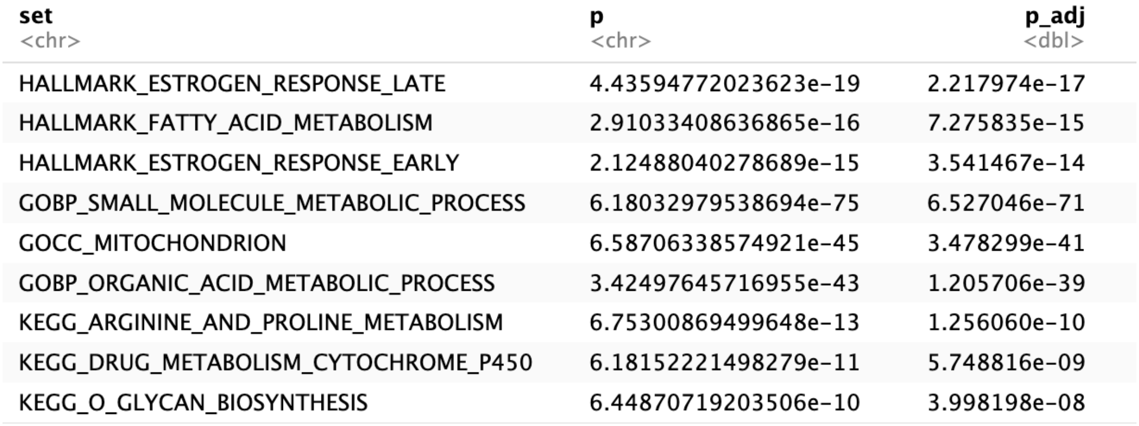
SYMBOL	t_stat	p_val	p_adj
<chr>	<dbl>	<dbl>	<dbl>
H1-5	-5.318476	4.397677e-07	3.212295e-06
RAB15	-5.319232	4.683084e-07	3.406024e-06
ETFDH	-5.321457	4.512959e-07	3.289207e-06
TRIM3	-5.322179	4.294976e-07	3.144528e-06
GNL3	-5.324179	1.562338e-06	9.685506e-06
NARS1	-5.325600	4.345165e-07	3.176551e-06
SNTN	-5.325674	4.227520e-07	3.102738e-06
FASTKD1	-5.335314	4.695463e-07	3.416317e-06
GALNT12	-5.336408	4.274175e-07	3.132653e-06
MRPL35	-5.339262	6.572946e-04	1.836274e-03

Table 5. Top 3 enriched gene sets for the upregulated genes



set <chr>	p <chr>	p_adj <dbl>
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	2.40955017322836e-116	1.204775e-114
HALLMARK_MYOGENESIS	1.43127607774009e-31	3.578190e-30
HALLMARK_COAGULATION	5.90733135357568e-28	9.845552e-27
GOCC_COLLAGEN_CONTAINING_EXTRACELLULAR_MATRIX	6.12057408297203e-101	6.463938e-97
GOCC_EXTERNAL_ENCAPSULATING_STRUCTURE	1.43184789686621e-98	7.560873e-95
GOBP_CELL_ADHESION	5.41637989538836e-95	1.906746e-91
KEGG_FOCAL_ADHESION	1.49088294698573e-30	2.773042e-28
KEGG_ECM_RECEPTOR_INTERACTION	4.27522259093659e-28	3.975957e-26
KEGG_CELL_ADHESION_MOLECULES_CAMS	2.49629747839443e-19	1.547704e-17

Table 6. Top 3 Enriched Gene Sets for the Downregulated Genes



set <chr>	p <chr>	p_adj <dbl>
HALLMARK_ESTROGEN_RESPONSE_LATE	4.43594772023623e-19	2.217974e-17
HALLMARK_FATTY_ACID_METABOLISM	2.91033408636865e-16	7.275835e-15
HALLMARK_ESTROGEN_RESPONSE_EARLY	2.12488040278689e-15	3.541467e-14
GOBP_SMALL_MOLECULE_METABOLIC_PROCESS	6.18032979538694e-75	6.527046e-71
GOCC_MITOCHONDRION	6.58706338574921e-45	3.478299e-41
GOBP_ORGANIC_ACID_METABOLIC_PROCESS	3.42497645716955e-43	1.205706e-39
KEGG_ARGININE_AND_PROLINE_METABOLISM	6.75300869499648e-13	1.256060e-10
KEGG_DRUG_METABOLISM_CYTOCHROME_P450	6.18152221498279e-11	5.748816e-09
KEGG_O_GLYCAN_BIOSYNTHESIS	6.44870719203506e-10	3.998198e-08

Discussion

In this study, unsupervised clustering was applied to gene expression microarray data from C3 and C4 tumors derived from colon cancer patients to stratify samples based on molecular characteristics and determine gene pathways involved in specific tumor subtypes. Hierarchical clustering resulted in C3 cluster and C4 clusters consistent with reported subtype classifications from Marisa et al. Differential expression analysis was performed based on the two subtype clusters, resulting in a list of probes characteristic of C3 and C4 tumors. From this group of probes, a set of 18 genes specific to C3 and C4 was determined.

Our set of 18 C3 and C4 genes are consistent with subtype characteristics established by Marisa et al. (**Table 1**). Marisa et al. found that the C3 group was characterized by the down-regulation of signaling pathways as well as pathways related to immunity and cell motility. C4, the cancer stem cell subtype, is characterized by the down-regulation of cell growth and death pathways and the up-regulation of pathways associated with the epithelial-mesenchymal transition (EMT) and cell motility (Marisa et al., 2013). Marisa et al. previously found C4 to be the most distinct subtype and one of 2 subtypes associated with poor patient prognosis. We affirmed Marisa et al.'s finding that several of the genes most associated with C4, including SFRP2 and GAS1, are found in the poor prognosis gene signature established by Oh et al., which suggests a potential link between the stem cell phenotype and disease aggressiveness (Oh et al., 2012).

Several other genes in our 18-gene list are also consistent with previously established cancer stem cell signatures. CCDC80 plays a role in extracellular matrix (ECM) organization and cell adhesion (Stelzer et al., 2016). Removal of junctions and reorganization of ECM are both processes critical to the EMT, suggesting a potential mechanism for CCDC80 to promote cancer cell stemness (Dongre & Weinberg, 2019; Stelzer et al., 2016). Indeed, Wang et al. recently published a 15-gene signature for stemness in colon adenocarcinoma containing CCDC80 and COLEC12, both genes we found to be specific to C4 (Wang et al., 2021). COLEC12 is known to be a receptor associated with host defense (Stelzer et al., 2016). Paralogs of IGFBP5, a gene involved in cAMP signaling and cell migration, are also included in stem cell-like signatures for colon cancer (de Sousa E Melo et al., 2011; Kosinski et al., 2007; Merlos-Suárez et al., 2011). IGFBP5 is upregulated in C4 relative to C3 (Marisa et al, 2013). PRICKLE2, a paralog of PRICKLE1, a negative regulator of the Wnt/beta-catenin pathway, was also found in the 187-gene signature for colon cancer stem cells published by de Sousa E Melo et al. The only gene in our 10 most significant genes list which was not included in the authors' set of 1,108 most discriminant genes was GPC6. Interestingly, GPC6 is a closely related paralog of GPC4, a cell surface protein and candidate stem cell marker previously found in colon cancer crypts, areas that may contribute to the colon cancer stem cell niche (Kosinski et al., 2007; Paine-Saunders et al., 1999; Stelzer et al., 2016).

Other genes in our top-18 list which have not been formally recognized as contributing to a stem cell signature may also play a role in stemness, cell motility, or EMT. MGP and SPOCK1 are involved in the extracellular matrix, and FERMT2 has been linked to cell differentiation, adhesion, and signaling (Stelzer et al., 2016). TIMP3, a gene up-regulated in C4 relative to C3, is involved in extracellular matrix degradation and has been shown to be highly expressed in colon cancer basal crypts (Kosinski et al., 2007; Stelzer et al., 2016). FRMD6 is active in the cytoskeleton and may be located in the apical junction, suggesting possible involvement in cell adhesion (Stelzer et al., 2016). RBMS1 contributes to cell cycle progression and apoptosis and may also play a role in stem cell status, according to the signature established by de Sousa E Melo et al.

Other genes in our top 18 list may play a more general role in cancer progression or have an otherwise undetermined function. For example, SFRP2 modulates Wnt signaling and is a marker for colon cancer when methylated (Stelzer et al., 2016). MSRB3 is involved in protein metabolism and repair, and RETNLB is involved in epithelial cell proliferation (Stelzer et al., 2016). FCGBP is located in the extracellular exosome, but it is unclear if it could link to the EMT (Stelzer et al., 2016). This is unlikely given that RETNLB and FCGBP were the only two genes in our signature predominantly expressed in C3 rather than C4 (Marisa et al., 2013). AKAP12 is related to cell growth and serves as a scaffold in signal transduction, which may explain why we found it to be strongly upregulated in C4 compared to C3 (Stelzer et al., 2016). Similarly, FNDC1, involved in G protein signaling, is down-regulated in C3 compared to C4 (Stelzer et al., 2016).

The most highly enriched genes from our analysis follow the same general trends found in the paper by Marisa et al. For the up-regulated genes, the gene sets that were most enriched are involved in cell adhesion and cell structure. Alternatively, in the down-regulated genes, the gene sets that were most enriched function in metabolic processes. More importantly, these results are consistent with Marisa et al.'s findings compared to the molecular subtypes. More specifically, Marisa et al. found that the immune system and cell growth pathways were up-regulated in the C2 subtype, and C4 and C6 both showed down-regulation of cell growth pathways. Although there were several genes that were not stated in the paper, many of the up and down-regulated genes that we reported overlap with functions of common cancer-causing genes. For example, SFRP2, a protein that regulates cell growth and differentiation, is upregulated according to our data. Additionally, FRMD6 is a protein linked to prostate cancer that was also found to be upregulated in our analysis (Haldrup et al., 2021). In general, the results from our gene set enrichment analysis are consistent with what we'd expect and the findings from the work of Marisa et al.

Conclusion

We confirmed that unsupervised clustering methods could be used to stratify tumors based on molecular characteristics, including microarray expression data. The resulting patient subtypes were found to be associated with distinctive genes and gene pathways, consistent with Marisa et al.'s findings. Like Marisa et al., we found C4 to be the more distinctive subtype compared to C3, with genes related to its stem cell and EMT features dominating the list of most significantly differentially expressed genes between C3 and C4. It is possible that the stem cell characteristics of the C4 subtype may also be linked to poor patient prognosis; the EMT can contribute to greater tumor-initiation and metastatic potential in cancer cells as well as resistance to some therapeutics (de Sousa E Melo et al., 2011; Dongre & Weinberg, 2019; Marisa et al., 2013; Merlos-Suárez et al., 2011; Oh et al., 2012). Thus, developing reproducible, robust molecular markers to stratify patients may be critical in developing diagnostic and prognostic clinical tools to help improve patient outcomes.

Due to the dominance of the C4 subtype, we gained less information about the C3 subtype. Only 2 genes in our 18-gene group showed high expression in C3 relative to C4, and their expression did not necessarily align with a specific biological function as many of the up-regulated C4 genes did. We noted signaling genes that were down-regulated in the C3 subtype, consistent with Marisa et al.'s findings.

References

Auclin, E., Zaanan, A., Vernerey, D., Douard, R., Gallois, C., Laurent-Puig, P., Bonnetain, F., & Taieb, J. (2017). Subgroups and prognostication in stage III colon cancer: Future perspectives for adjuvant therapy. *Annals of Oncology*, *28*(5), 958–968. https://doi.org/10.1093/annonc/mdx030

Bolstad, B. M., Collin, F., Brettschneider, J., Simpson, K., Cope, L., Irizarry, R. A., & Speed, T. (2005). Quality Assessment of Affymetrix Genechip Data. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor,* 33-47. doi:10.1007/0-387-29362-0_3

de Reyniès, A., Assié, G., Rickman, D. S., Tissier, F., Groussin, L., René-Corail, F., Dousset, B., Bertagna, X., Clauser, E., &amp; Bertherat, J. (2009). Gene expression profiling reveals a new classification of adrenocortical tumors and identifies molecular predictors of malignancy and survival. Journal of Clinical Oncology, 27(7), 1108–1115. https://doi.org/10.1200/jco.2008.18.5678

de Sousa E Melo, F., Colak, S., Buikhuisen, J., Koster, J., Cameron, K., de Jong, J. H., Tuynman, J. B., Prasetyanti, P. R., Fessler, E., van den Bergh, S. P., Rodermond, H., Dekker, E., van der Loos, C. M., Pals, S. T., van de Vijver, M. J., Versteeg, R., Richel, D. J., Vermeulen, L., & Medema, J. P. (2011). Methylation of Cancer-Stem-Cell-Associated Wnt Target Genes Predicts Poor Prognosis in Colorectal Cancer Patients. *Cell Stem Cell*, *9*(5), 476–485. https://doi.org/10.1016/j.stem.2011.10.008

Dongre, A., & Weinberg, R. A. (2019). New insights into the mechanisms of epithelial–mesenchymal transition and implications for cancer. *Nature Reviews Molecular Cell Biology*, *20*(2), Article 2. https://doi.org/10.1038/s41580-018-0080-4

Gautier, L., Cope, L., Bolstad, B. M., & Irizarry, R. A. (2004). Affy—analysis of *affymetrix genechip* data at the probe level. *Bioinformatics, 20*(3), 307-315. doi:10.1093/bioinformatics/btg405

Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and Graphics. *Journal of Computational and Graphical Statistics, 5*(3), 299-314. doi:10.2307/1390807

Johnson, W. E., Li, C., & Rabinovic, A. (2006). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics, 8*(1), 118-127. doi:10.1093/biostatistics/kxj037

Kosinski, C., Li, V. S. W., Chan, A. S. Y., Zhang, J., Ho, C., Tsui, W. Y., Chan, T. L., Mifflin, R. C., Powell, D. W., Yuen, S. T., Leung, S. Y., & Chen, X. (2007). Gene expression patterns of human colon tops and basal crypts and BMP antagonists as intestinal stem cell niche factors. *Proceedings of the National Academy of Sciences*, *104*(39), 15418–15423. https://doi.org/10.1073/pnas.0707210104

Leek, J., Johnson, W. E., Jaffe, A., Parker, H., & Storey, J. (2022, November 1). The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. Retrieved February 16, 2023, from https://www.bioconductor.org/packages/devel/bioc/vignettes/sva/inst/doc/sva.pdf

Lotfollahzadeh, S., Recio-Boiles, A., & Cagir, B. (2022). Colon Cancer. In *StatPearls*. StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK470380/

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

Merlos-Suárez, A., Barriga, F. M., Jung, P., Iglesias, M., Céspedes, M. V., Rossell, D., Sevillano, M., Hernando-Momblona, X., da Silva-Diz, V., Muñoz, P., Clevers, H., Sancho, E., Mangues, R., & Batlle, E. (2011). The Intestinal Stem Cell Signature Identifies Colorectal Cancer Stem Cells and Predicts Disease Relapse. *Cell Stem Cell*, *8*(5), 511–524. https://doi.org/10.1016/j.stem.2011.02.020

Nguyen, H. T., & Duong, H.-Q. (2018). The molecular characteristics of colorectal cancer: Implications for diagnosis and therapy (Review). *Oncology Letters*, *16*(1), 9–18. https://doi.org/10.3892/ol.2018.8679

Oh, S. C., Park, Y.-Y., Park, E. S., Lim, J. Y., Kim, S. M., Kim, S.-B., Kim, J., Kim, S. C., Chu, I.-S., Smith, J. J., Beauchamp, R. D., Yeatman, T. J., Kopetz, S., & Lee, J.-S. (2012). Prognostic gene expression signature associated with two molecularly distinct subtypes of colorectal cancer. *Gut*, *61*(9), 1291–1298. https://doi.org/10.1136/gutjnl-2011-300812

Paine-Saunders, S., Viviano, B. L., & Saunders, S. (1999). GPC6, a Novel Member of the Glypican Gene Family, Encodes a Product Structurally Related to GPC4 and Is Colocalized withGPC5on Human Chromosome 13. *Genomics*, *57*(3), 455–458. https://doi.org/10.1006/geno.1999.5793

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Stein, T. I., Nudel, R., Lieder, I., Mazor, Y., Kaplan, S., Dahary, D., Warshawsky, D., Guan-Golan, Y., Kohn, A., Rappaport, N., Safran, M., & Lancet, D. (2016). The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Current Protocols in Bioinformatics*, *54*(1), 1.30.1-1.30.33. https://doi.org/10.1002/cpbi.5

Tang Y, Horikoshi M, Li W (2016). “ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages.” *The R Journal*, **8**(2), 474–485. doi:10.32614/RJ-2016-060

Wang, W., Xu, C., Ren, Y., Wang, S., Liao, C., Fu, X., & Hu, H. (2021). A Novel Cancer Stemness-Related Signature for Predicting Prognosis in Patients with Colon Adenocarcinoma. *Stem Cells International*, *2021*, e7036059. https://doi.org/10.1155/2021/7036059

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Haldrup, J., Strand, S.H., Cieza-Borrella, C. et al. FRMD6 has tumor suppressor functions in prostate cancer. Oncogene 40, 763–776 (2021). https://doi.org/10.1038/s41388-020-01548-w