Liz Murphy as Data Curator
Meghana Peshwa as Programmer
Go Ogata as Analyst
Haroon Qureshi as Biologist

Microarray Based Tumor Classification

Introduction

Colorectal cancer (CRC) is the third most common cancer and a leading cause of cancer death worldwide. Currently there is only one prognostic classification used in clinical practice, pathological staging, however it is a poor indicator of recurrence for many patients. Molecular markers of CRC have been investigated via gene expression profiles by microarrays, but no signatures have been identified that can reliably predict prognosis. There is some indication that there are at least three molecular subtypes of CRC, which could explain the difficulty in reproducibly identifying molecular signatures, as CRC is likely not a homogenous disease as previously thought. The molecular classification of CRC currently in use is based on DNA markers including microsatellite instability (MSI), CpG Island methylator phenotype (CIMP), chromosomal instability (CIN) as well as *BRAF* and *KRAS* mutations, but is in need of refinement. The purpose of Marisa et al was to establish a standard molecular classification of colon cancer (CC) with clinical and prognostic significance, through the analysis of mRNA expression from a large, multicentered dataset (1). Through unsupervised analysis of gene expression profiles from CC patient samples ranging from stage I to stage IV, the authors identified six distinct clusters, which were then validated and assessed for associations with known clinical markers, DNA alterations and prognostic indicators. The aim of this project was to recreate these results using a subset of available samples, following a similar analysis pipeline.

Data

The dataset used by Marisa et al. came from a multicentered French cohort of 750 patients with stage I to IV colon cancer who had undergone surgery from 1987 to 2007. Patients who received chemotherapy before surgery, and those with primary rectal cancer were not included in this initial study. 566 samples met RNA quality control

requirements for gene expression profiling analysis, which the authors further split into a discovery set (n=443) and validation set (n=123). These samples were profiled using Affymetrix U133 Plus 2.0 gene chips. Of the 566 patients, 19 had normal tissue available for comparison, and those samples were also profiled. The authors used previously published methods for RNA extraction, probe production, hybridization and raw data processing (2). In addition to the 123 samples from the French cohort, Marisa et al. included 906 CC samples from 7 publicly available datasets through the NCBI GEO (GSE13067, GSE13294, GSE14333, GSE17536/17537, GSE18088, GSE26682, and GSE33113) in their validation set. All of these datasets were profiled using a chip similar to Affymetrix U133 Plus 2.0, and had both raw CEL files available and metadata regarding tumor location, DNA alterations, and patient outcome. Data from TCGA was included in the validation set as well due to thorough DNA alteration annotations, although those 152 samples were analyzed separately as they were profiled using a non-Affymetrix platform. For this project's repeat analysis, 134 samples from the discovery set profiled on Affymetrix U133 Plus 2.0 gene chips were reexamined. These samples had been assigned subtypes C3 or C4 by the classification system reported by Marisa et al. All sample data is available via the NCBI GEO using accession number GSE39582 (https://www.ncbi.nlm.nih.gov/geo/).

Methods

The dataset consisting of 134 CC samples underwent normalization and pre-processing. All of the CEL files were read together using the ReadAffy function, which produced an AffyBatch object. A probe level linear model (PLMset) object was made from the AffyBatch object using the fitPLM function. This is a prerequisite for computing and plotting the Relative Log Expression (RLE) and Normalized Unscaled Standard Error (NUSE) scores. These are quality control metrics that are used to assess the quality of microarray data. The AffyBatch object was normalized using the rma (robust multi-array average) function.

ComBat (sva package) was used to correct for batch effects in the normalized data while preserving features of interest. The batch effects requiring correction were Center and RNA extraction method, while the features of interest included tumor and MMR status as per Marisa et. al. The resulting expression data was written to a CSV file using the exprs and write.csv functions, where probe sets are rows and samples are columns.

Principal Component Analysis (PCA) was performed on the batch corrected data using the prcomp() function in order to reduce the dimensionality of the data and examine the percent variability of the principal components. The first two principal components were plotted using the ggplot function to examine for outliers.

The normalized and batch corrected probe set intensity levels from comBat were then filtered to obtain data to allow for biologically relevant results. The first filter looked for probe sets in which at least 20% of samples had expression levels greater than log2(15) intensity value (gene > 3.9) in, and 39,750 probeset passed the first filter. The second filter focused on genes that showed significantly different levels of variance than the rest of the genes ($p < 0.01$) using a chi-squared test, resulting in 15,540 probe sets. The final filter looked for a coefficient of variation greater than 0.186. The three filters left 1,658 probe sets which minimized noise within the data and allowed for an effective analysis using clustering.

The packages hgu133plus2.db and AnnotationDbi were used to map gene symbols to probe set IDs found through t-test analysis of the differential expression matrix. Any probe sets that did not pair with a gene symbol were dropped from the list, as were any probe sets with duplicate gene symbols. This new list of probe sets was ordered by descending t-statistic (t-stat) value, and was filtered for the most significant genes using the adjusted p-value. From this, the top 1000 up and down regulated genes were acquired, respectively.

This study used the hierarchical clustering algorithm provided by R, hclust(). The samples were clustered into two clusters using the 1,658 genes that passed all three filters based on gene intensity. The hierarchical clustering labeled samples that were colon cancer subtype 3 vs. other colon cancer subtypes. A heatmap was generated with columns by the clustering and the expression level of the gene sets using the built-in R function heatmap(). A Welch t-test was performed on the 1,658 genes between the two clusters with a significance level of 0.05.

Using the getGmt() function from the Bioconductor package GSEABase, the Kyoto Encyclopedia of Genes and Genomes (KEGG), Gene Ontology (GO), and Hallmark gene sets, which were downloaded from MSigDB, were analyzed. The number of genesets in each collection was noted and then compared to the top 1000 up and down regulated genes using Fisher's Exact test. Fisher's Exact test computed hypergeometric statistics and p-values comparing overlap between the gene sets and the up and down regulated genes, and found the statistically significantly enriched gene sets for each gene set collection. The p-values for these enriched gene sets were then

adjusted using the Benjamini-Hochberg (FDR) procedure, and the top 3 enriched gene sets from each gene collection were made into a table.

Results

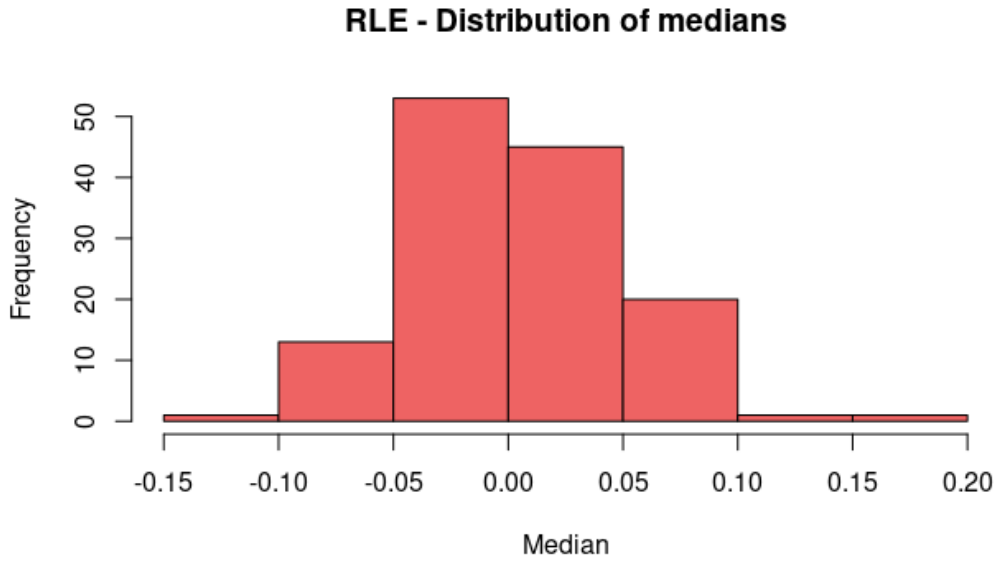

Figure 1. Histogram of median RLE scores. The histogram represents the distribution of median RLE scores across the 134 CC samples.

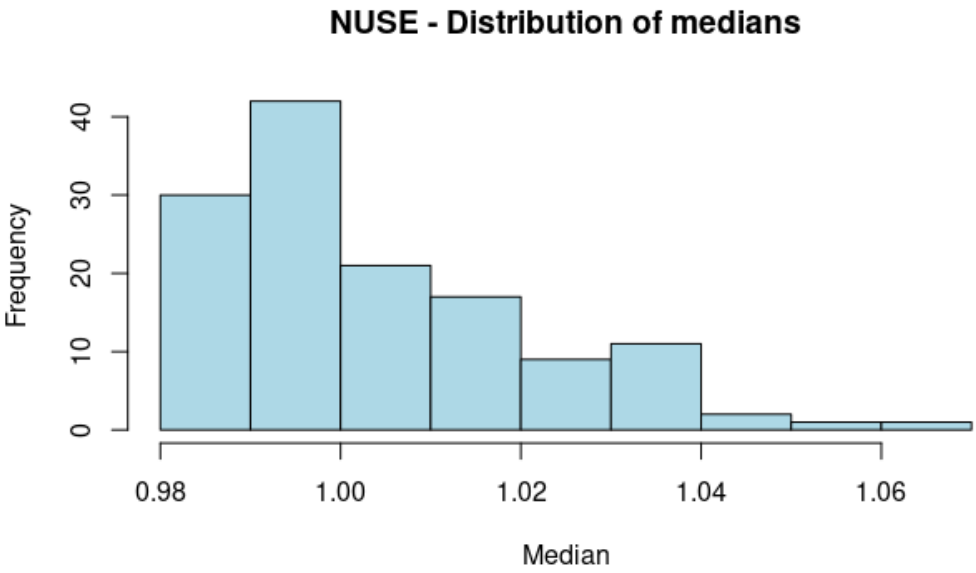

Figure 2. Histogram of median NUSE scores. The histogram represents the distribution of median NUSE scores across the 134 CC samples.

The median RLE scores in Figure 1 are centered near RLE = 0. Figure 2 shows that the median NUSE scores are reasonably centered around 1. Therefore, based on the median RLE and NUSE scores, the samples are of good quality and present no quality control problems.

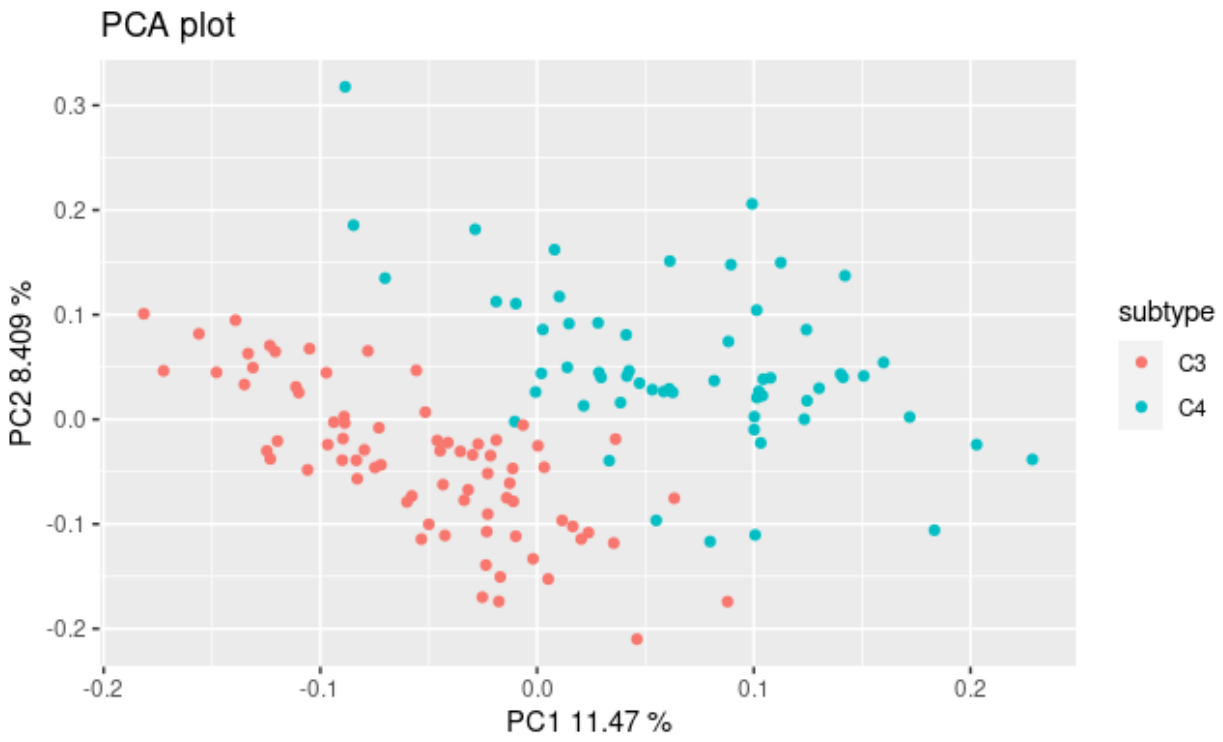

Figure 3. PCA plot of first two principal components. The first two principal components together represent about 20% of the variance. The samples were clustered based on CC subtype present in the metadata.

Figure 3 represents the PCA plot of the first two principal components which capture the highest variance in the dataset. The samples formed distinct clusters based on the CC subtypes provided in the sample metadata. There are no outliers because there appear to be no samples that are far away from other samples. Therefore, the median RLE and NUSE scores, as well as the PCA plot suggest that the dataset is of good quality for further analysis.

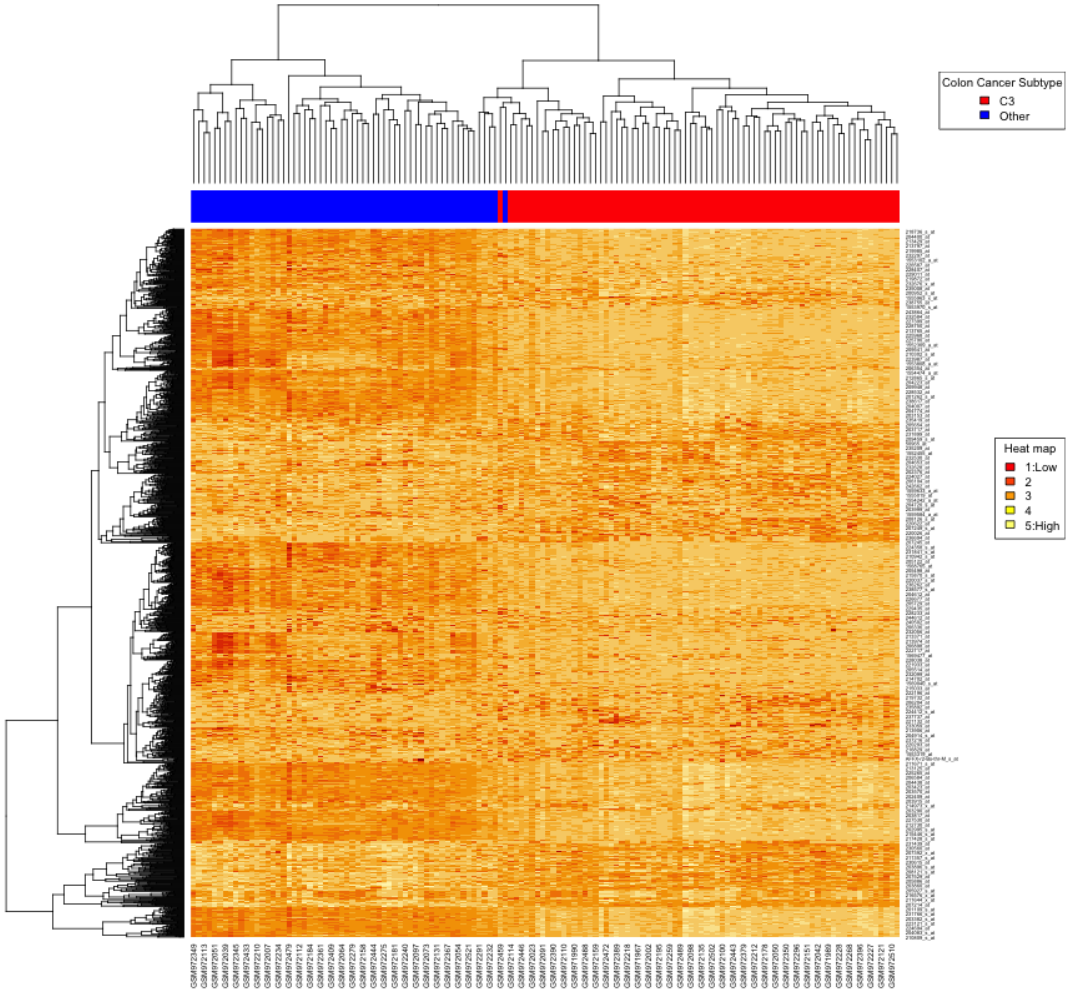

Figure 4. Heatmap of the 1,658 probeset and Patient samples The heatmap depicts the hierarchical clustering of the samples(column) where all but 5 samples were clustered correctly. The heatmap of the probeset shows how the expression levels differ between the two clusters.

The clustering shows two distinct clusters, one composed of non-C3 colon cancer subtypes (n = 58) and another composed of mostly C3 colon cancer subtypes (n = 76) (Fig. 4). There are 5 samples that are labeled as non-C3 subtypes that are clustered with the C3 samples and the inspection of the heatmap supports this clustering decision. Through the clustering of the different subtypes, a clear differentiation of gene expression is observed. With the exception of the 5 non-C3 samples that are clustered with the C3 samples, different probe set clusters show the entire C3 and non-C3 cluster expressing the same expression change.

The Welch t-test was performed between the C3 vs non-C3 and out of the 1,658 probe sets that were used, only 798 probesets were observed to be differentially

expressed. Out of the 798 probesets that were differentially expressed, 43 of them were up-regulated in C3, and 755 of the probe sets were down-regulated.

Probe set IDs were mapped to their respective gene symbols and arranged by descending T-stat values, after removing duplicates and null gene symbol values. The top-1000 up and downregulated genes were then collected from a final pool of 1,281 filtered genes. The top-10 for both expression types are shown in Table 1. According to t-stat value, the most upregulated gene was FCGBP, and the most downregulated gene was MIR100HG, as seen in Table 1.

	Probeset ID	Symbol	T-stat	P-value	P_adj
1	203240_at	FCGBP	7.150764	1.567079e-11	2.598218e-08
2	227725_at	ST6GALNAC1	6.749278	1.552486e-10	2.574022e-07
3	228241_at	AGR3	6.498036	8.447443e-10	1.400586e-06
4	204673_at	MUC2	6.108669	4.947303e-09	8.202628e-06
5	220622_at	LRRC31	5.982796	9.871183e-09	1.636642e-05
6	226248_s_at	ELAPOR1	5.804042	2.423092e-08	4.017486e-05
7	210107_at	CLCA1	5.692932	4.359435e-08	7.227944e-05
8	238846_at	TNFRSF11A	5.680254	4.587661e-08	7.606341e-05
9	207214_at	SPINK4	5.623221	6.167072e-08	1.022501e-04
10	1553828_at	NXPE1	5.586527	7.530412e-08	1.248542e-04

	Probeset ID	Symbol	T-stat	P-value	P_adj
1	225381_at	MIR100HG	-8.565432	8.490885e-15	1.407789e-11
2	208131_s_at	PTGIS	-8.474560	9.535670e-15	1.581014e-11
3	238478_at	BNC2	-8.456452	9.731887e-15	1.613547e-11
4	219778_at	ZFPM2	-8.429968	1.369452e-14	2.270551e-11
5	221019_s_at	COLEC12	-8.411041	1.450093e-14	2.404254e-11
6	202437_s_at	CYP1B1	-8.400291	1.938599e-14	3.214198e-11
7	227061_at	CCDC80	-8.391255	1.336698e-14	2.216245e-11
8	204938_s_at	PLN	-8.331668	2.891306e-14	4.793785e-11
9	225790_at	MSRB3	-8.302132	2.563793e-14	4.250769e-11
10	226065_at	PRICKLE1	-8.281361	3.328029e-14	5.517872e-11

Table 1. Top 10 upregulated and downregulated genes The top 10 most up regulated (top) and down regulated (bottom) genes found through t-score analysis in the differential expression dataset.

Gene set enrichment analysis of the KEGG, GO, and Hallmark gene sets was performed. These collections held 186, 50, and 10,402 gene sets, respectively. After conducting Fisher's Exact Test to each of these gene set collections, and filtering for sets where p-value<0.05, it was found that the KEGG database had a total of 21 significantly enriched genes, whereas GO and Hallmark had 888 and 12 significantly enriched gene sets, respectively. The top-3 significantly enriched pathways from each collection, organized in descending order by Benjamini-Hochberg (FDR) values, are listed in Table 2. Table 2 displays the name of the geneset, the p-value, the statistic estimate (est.), expression type (reg.) and FDR value of each top enriched gene set. The top-3 significantly enriched gene sets in KEGG were found to be downregulated pathways related to the metabolism of starch and sucrose, nitrogen, and retinol. The GO database's most enriched gene set was the supramolecular fiber organization set, but two other sets related to protein binding were also found. All three of these sets

were upregulated. Interestingly, the Hallmark database's most enriched and statistically significant gene sets pertained to epithelial mesenchymal transitions, with both up and downregulated pathways being expressed.

	Geneset Name	P_Value	Estimate	Reg.	FDR
1	KEGG_STARCH_AND_SUCROSE_METABOLISM	2.13E-05	0	Down	0.0079165
2	KEGG_NITROGEN_METABOLISM	9.98E-05	0	Down	0.0103102
3	KEGG_RETINOL_METABOLISM	1.12E-04	0.0898543	Down	0.0103102
1	GOBP_SUPRAMOLECULAR_FIBER_ORGANIZATION	2.16E-09	0.2374535	Up	4.49E-05
2	GOMF_PROTEIN_CONTAINING_COMPLEX_BINDING	1.21E-07	0.3321391	Up	1.25E-03
3	GOMF_CYTOSKELETAL_PROTEIN_BINDING	2.37E-07	0.2876871	Up	1.64E-03
1	HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	1.08E-08	0.2513953	Up	1.08E-06
2	HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	6.56E-08	23.650105	Down	3.28E-06
3	HALLMARK_MYOGENESIS	4.09E-05	0.2531069	Up	1.36E-03

Table 2. Top 3 enriched gene sets The top 3 enriched gene sets found for the KEGG, GO, and hallmark gene set databases, respectively. Ordered by ascending FDR value.

Discussion

The quality of the dataset consisting of 134 CC samples was examined. RLE and NUSE scores were calculated on the normalized dataset. These are quality metrics that assess reproducibility and identify apparent outliers. Ideally, a good quality array has a median RLE around 0. Also, it has a median NUSE around 1 and a small IQR (interquartile range). It was observed in this dataset that the median RLE and NUSE scores are centered at 0 and 1 respectively (Figure 1 and Figure 2). As a result, the dataset was determined as good quality and no samples were removed.

PCA was performed to detect any outliers. The first two principal components represent the highest variance in the dataset (11.47% and 8.409% respectively) (Figure

3). These two principal components capture a large fraction of the variance in the dataset. There were distinct clusters formed on the basis of the CC subtypes. From this plot no samples appeared to be outliers, thus indicating that the dataset is of good quality for further analysis.

The dendrogram of the hierarchical clustering shows two major clusters of samples used in the original study, and are clustered in a similar way. While the algorithm used for clustering was different in the Marisa et al. study, hierarchical clustering was effectively able to cluster most non-C3 colon tumors (Fig. 4). The algorithm did fail to cluster 5 of the non-C3 colon tumors correctly, and by observing the heatmap, the error in clustering appeared to be a difference in the expression levels. The 5 samples that were clustered incorrectly may have some characteristics that may assist for further classification.

The Welch t-test ($p < 0.05$) was used to check which probe sets were differentially expressed in the data and found that 43 genes were up-regulated and 755 genes were down-regulated. The top 10 differentially expressed genes for up and down regulation seen in Table 1 can be used to represent the relationship between the C3 and non-C3 colon cancer subtypes. While it is difficult to represent each cluster with specific genes, the heatmap depicts a greater up regulation and the up regulated genes are most likely best associated with the C3 clusters. In the same logic, it may be reasonable to represent the non-C3 subtypes with the top 10 down regulated genes (Fig. 4). For a greater accuracy of classification and representation of the different subtypes, a deeper study into the data will be necessary.

The results from gene set enrichment were partially replicated in this study. Marisa et al. identified 1,108 probe sets that were filtered through t-stat and adjusted p-value; there were 1,281 appropriately filtered probe sets found in this replicate study. This could have been due to differing sample sizes. For similar reasons, the discrepancies between the top-10 up and downregulated genes reported by our replicated study and by Marisa et al. can be explained. The influx of over 200 probe sets may have skewed the findings, not to mention the original study worked with data regarding more tumor subtypes.

The top two genes that were upregulated, as seen in Table 2, were FCGBP, an antibody binding protein, and ST6GALNAC1. FCGBP was found to have a positive correlation with the levels of B-cells and macrophages combating colorectal tumors. The upregulation of such a gene is no surprise in the presence of cancer growths, as these are a reason for increased immune response (3). This gene was also found by Marisa

et al. to be upregulated in the C3 subtype. ST6GALNAC1 was not reported in the original study, however, it is also of note in the context of cancer, as it has been previously observed to be highly expressed in cancer stem cells, a major factor behind tumor-initiating and self-renewal, and has also been shown to be highly expressed in malignant forms of cancer (4).

Interestingly, the most downregulated gene, MIR100HG, has also been linked to having a strong influence on colorectal cancer metastasis when heavily promoted (5). It is intriguing that this would be so down regulated for a colorectal cancer subtype, and perhaps the relationship between this gene and FCGBP could be researched in a future study, as there may be a correlation between the two.

After conducting Fisher's exact test on the downloaded genesets from KEGG, GO and Hallmark, and adjusting the p-values according to the FDR value, the most enriched gene sets were found for each database. From these enriched sets, the top 3 from each collection were extracted and can be seen in Table 2. The results from the enrichment process held up relatively well to the findings in the original study, although there were differences noted. In the KEGG database, 186 enriched gene sets were identified. From these, the top 3 sets were each shown to be involved in metabolomics. Respectively, these were downregulated gene sets involved in metabolizing starch and sucrose, nitrogen, and retinol. The results for starch and sucrose, and nitrogen metabolism reflects what was found in the original study as well, as both of these genesets were found to be downregulated in the C4 subtype. Retinol metabolism was not found in Figure 2 in Marisa et al. In the GO database, 888 enriched gene sets were found after proper analysis. Of these, the top 3 were the supramolecular fiber organization, protein containing complex binding, and cytoskeletal protein binding genesets, all of which were upregulated. These were not found in Figure 2 in Marisa et al. Again, the cause for this discrepancy may be attributed to methodological differences in generating the differential matrix used for downstream statistical analysis. Finally, in the Hallmark collection, 12 enriched genes were found. The top 3 enriched gene sets were the epithelial mesenchymal transition (upregulated), epithelial mesenchymal transition (downregulated), and myogenesis sets. The epithelial mesenchymal transition set was expressed both with upregulation and downregulation. This somewhat mirrors what is seen in Figure 2 of Marisa et al., given that the same pathway in GO is downregulated in the C3 subtype and upregulated in the C4 subtype. This is perhaps a reason why both expression types are present after applying the FDR procedure. The epithelial mesenchymal transition has been recorded as having a link to cancers,

leading to both the loss of epithelial characteristics in cells and transitions to more mesenchymal characteristics, such as increased motility and probability of metastasis (6).

Conclusion

The expression of the Hallmark epithelial mesenchymal transition geneset being both up and downregulated was of particular note. In Marisa et al., the epithelial mesenchymal transition pathways in GO were noted to be downregulated in the C3 tumor subtype, and upregulated in the C4 subtype. This could be a potential reason for the appearance of both expression types in the top-3 most significantly enriched gene sets found in Hallmark.

FCGBP was the most upregulated gene in the final mapped dataset. Due in part to its previously recorded positive correlation with the levels of macrophages (3), it could be reasoned that the upregulation of FCGBP could be related to an increased immune response to the colorectal tumor. Marisa et al. found this gene to be upregulated in the tumor subtype C3, and this could further indicate a strong correlation between this tumor subtype and gene.

A particular challenge of replicating the original study was coordinating each role in a way that resulted in a streamlined flow of information. Due to each role relying on its predecessor, there needed to be a quick turnaround of data. This was achieved due to communication between each role, and group assistance for any roadblock that would have affected downstream analysis.

References

1. 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
2. de Reyniès, A., Assié, G., Rickman, D. S., Tissier, F., Groussin, L., René-Corail, F., Dousset, B., Bertagna, X., Clauser, E., & 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 : official journal of the American Society of Clinical Oncology*, *27*(7), 1108–1115. https://doi.org/10.1200/JCO.2008.18.5678

3. Zhuang, Q., Shen, A., Liu, L., Wu, M., Shen, Z., Liu, H., Cheng, Y., Lin, X., Wu, X., Lin, W., Li, J., Han, Y., Chen, X., Chen, Q., & Peng, J. (2021). Prognostic and immunological roles of Fc fragment of IgG binding protein in colorectal cancer. *Oncology letters*, *22*(1), 526. https://doi.org/10.3892/ol.2021.12787
4. Wang, WY., Cao, YX., Zhou, X. *et al.* Stimulative role of ST6GALNAC1 in proliferation, migration and invasion of ovarian cancer stem cells via the Akt signaling pathway. *Cancer Cell Int* 19, 86 (2019). https://doi.org/10.1186/s12935-019-0780-7
5. Li, W., Yuan, F., Zhang, X., Chen, W., Tang, X., & Lu, L. (2019). Elevated MIR100HG promotes colorectal cancer metastasis and is associated with poor prognosis. *Oncology letters*, *18*(6), 6483–6490. https://doi.org/10.3892/ol.2019.11060
6. Vu, T., & Datta, P. K. (2017). Regulation of EMT in Colorectal Cancer: A Culprit in Metastasis. *Cancers*, *9*(12), 171. https://doi.org/10.3390/cancers9120171