MICROARRAY BASED TUMOR CLASSIFICATION

Group: Hedgehog
Teaching Assistant: Joey Orofino

Data Curator: Qinrui Wu
Programmer: Dylan Beeber
Analyst: Rojashree Jayakumar
Biologist: Merai Dandouch

Introduction

Colorectal cancer (CRC) has become one of the most predominant cancer types in Western countries, accounting for around 10% of all cancer-related deaths. The prediction of colon cancer recurrence is inaccurate through its pathological staging, even though the only prognostic classification utilized in clinical practice to identify patients for adjuvant treatment is pathological staging (Kuipers et al., 2015). In previous studies, three distinct molecular subtypes of CRC have been identified based on gene expression profiles (GEP) which were determined by microarray technology. A 2013 paper from Marisa et. al. examined different CRC samples in an attempt to learn more about the different kinds of CRC and their influence on disease prognosis. Marisa et al.'s study discovered six subtypes of CRC by unsupervised analysis on gene expression data from a discovery subset of 443 CRC samples. In this study, we examined a subset of these samples to look for further evidence of the differences between the C3 and C4 subtypes.

Data

There were 134 samples included in one dataset that was saved in the ‘samples’ directory for further analysis in terms of this project. In order to avoid using much space, 133 of these provided samples were saved as symbolic links and another missing file was downloaded from GEO by searching for accession number GSE39582 and sample ID GSM971958. This dataset is an Affymetrix microarray that evaluates gene expression of colon cancer. And it was composed of discovery and validation set samples. Each sample file contains the probe ID (54,675) and the corresponding RNA value that is log2 normalized intensity signal. The microarray dataset that is being used originated from Marisa et al. that aims to classify colon cancer based on mRNA expression profile analyses. This paper provided 585 samples on GEO, we only used 134 of these for the current study.

Methods

All processing and analysis steps were performed in R version 4.1.2. Software packages utilized include: affy, affyPLM, sva, AnnotationDbi, hgu133plus2.db, and ggplot2. Initial data processing was completed in under 6 minutes when performed on a single core processor within a remote computing cluster.

Initial Data Processing

Each of the 134 individual CEL files were read into a single AffyBatch object using the affy R package. Samples were then normalized using the robust mutli-array averaging function (rma) from the AffyPLM package. Normalization of the raw expression data is important to account for variation in sample collection and differences between microarray runs. Quality control was then performed on the raw expression data by calculating the median relative log expression (RLE) and normalized unscaled standard error (NUSE). The RLE is a measurement of the relative expression values across all microarray probes in a sample and NUSE is an estimation of standard error normalized across all microarrays. Abnormal RLE and NUSE values may indicate sample issues, but a single abnormal value in either does not necessarily justify exclusion of the sample from the experiment. Both RLE and NUSE calculations were performed via the fitPLM() function within the affyPLM package. No samples were discarded after examination of the RLE and NUSE values.

The normalized expression data was corrected for batch effects using the ComBat() function from the sva package. Preprocessed metadata encoding for batch effects and variables of interest was provided in the original Marisa et. al. study. Batch effects included test site location and RNA extraction method encoded into a single variable passed to the batch argument. Variables of interest included tumor and mismatch repair (MMR) status and were encoded as a single metadata variable passed to the mod argument. The finalized expression data after normalization and batch effect correction was saved for further analysis.

Noise Filtering and Dimensionality Reduction

The RMA normalized and Combat adjusted matrix was passed through three filters. The first filter (expression filter) was used to retain genes, which have an expression greater than log2(15) in at least 20% of the samples. The second filter (variance filter) was used to retain genes with a significant variance greater than the threshold of $p<0.01$. A test statistic was calculated for each gene and significant variants falling in the upper tail of a chi distribution were retained.

Test statistic of a gene = df x Variance of the gene/ Median Variance of the gene set

where df = Number of samples (n) - 1

The third filter was used to retain genes with a covariance > 0.186. The expression data from the variance filter and covariance filter were written out to separate files for further detailed analysis.

Hierarchical Clustering and subtype discovery

Unsupervised hierarchical clustering was performed on the filtered samples using euclidean distance matrix constructed using dist() method and hclust() method in R to cluster samples into subtypes. The clusters were cut into two groups and a heatmap was constructed to visualize the differential expression. A Welch t-test between these clusters was performed using the t.test() method between the clusters for each gene and the adjusted p-value (FDR method) using p.adjust() was calculated. The threshold to determine genes that are differentially expressed from the Welch t-test is a p-adjusted value lesser than 0.05 ($p<0.05$). This t-test analysis was conducted on genes passing filters 1 and 2 and also on genes passing all three filters. The t-test results were written out into csv files for further biological analysis.

MSigDB Gene Set Data Mapping and Fisher's Exact Test

AnnotationDBI and hgu133plus2.db packages were implemented using Bioconductor methods including select(), keys(), and columns() . The hgu133plus2.db was assembled using Affymetrix HG-U133_Plus_2 annotation data gathered from public repositories. The ALIAS object in the hgu133plus2.db package was applied to join commonly used gene symbols. This allowed for sample probe ids and gene symbol mapping. Probe ids had a many to many relationship with the mapped hgnc symbols. Samples that were not mapped to gene symbols or duplicated were removed from the report. To retrieve cancer signaling pathways, KEGG, GO, and Hallmark gene set symbols were downloaded from MSigDB and mapped to the hgnc symbols in the previous step. Expression values were used to filter the top 1000 up and down regulated genes to create contingency tables for Fisher's exact t-test analysis on all signaling pathways for each gene set (KEGG, GO, Hallmark).

Results

Quality control via median RLE calculation flagged three samples for values either greater than 0.1 or less than -0.1 while Median NUSE calculation flagged two other samples for values greater than 1.05. All samples were carried over for further analysis, because no individual samples were flagged for both abnormal RLE and NUSE values.

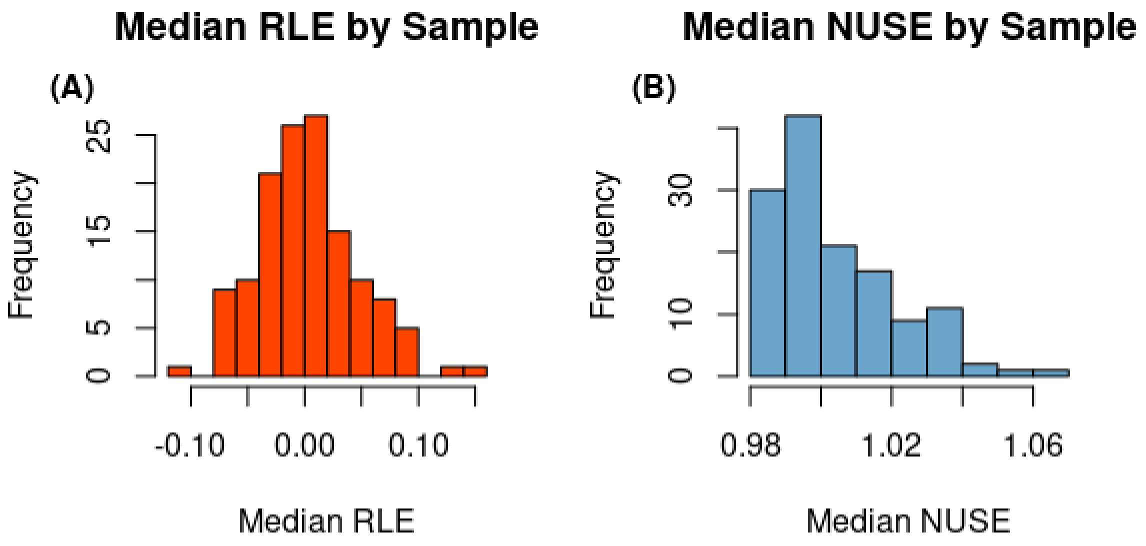


Fig 1: Plot (A) displays the median Relative Log Expression (RLE) for all samples, a measurement of how similar expression was between samples. Samples GSM971993, GSM972089, and GSM972390 were flagged for median RLE values above 0.1 or below -0.1. Plot (B) displays the Normalized Unscaled Standard Error, an estimation of standard error per sample scaled such that the median standard error is equal to 1. Samples GSM972113 and GSM972269 were flagged for median NUSE values greater than 1.05.

Principal component analysis (PCA) on the expression matrix data reveals that the C3 and C4 cancer subtypes are separated into distinct regions when plotting against the first two principal components. The original Marissa et. al. study divided these samples into subtypes using consensus clustering, so it is unsurprising that we find a similar separation between the samples when classifying them by the labels generated by their study, particularly when considering that subtype C4 was found to be the most distinct cluster. The separation of the sample subtypes in principal component space further reinforces the assertion that the C3 and C4 subtypes show separate patterns of gene expression. However, it should be noted that principal components 1 and 2 only explain about 20% of the total variation between samples, with the remaining 80% of variance unaccounted for.

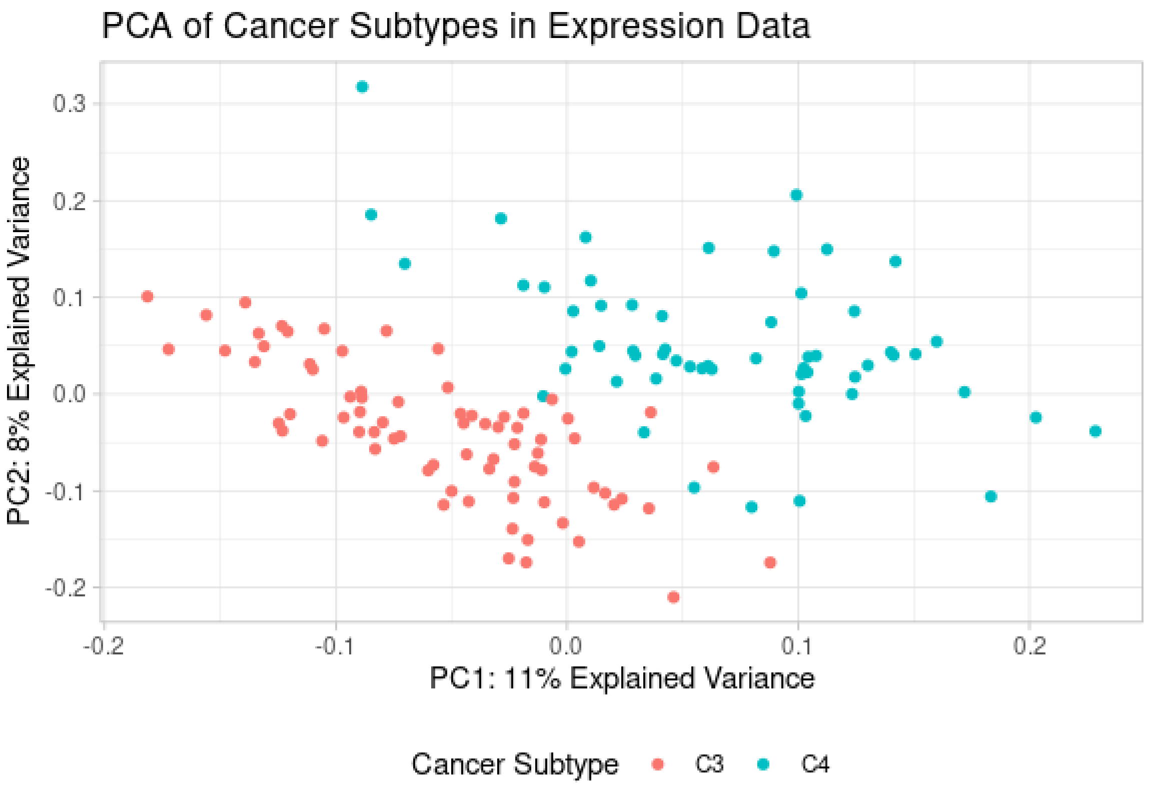


Fig 2: Principal component analysis of expression data of expression matrix values separated into the C3 and C4 cancer subtypes. Plotting samples by PC1 and PC2 separate C3 and C4 samples into distinct regions in the plot. PC1 and PC2 explain about 20% of the total variance between samples.

The expression data, which was RMA normalized and corrected for batch effects, had 54675 genes was filtered to remove noise in the genes (probe ids) and reduce the dimensionality of the data. A similar protocol as Marissa et. al, was adopted with minor differences.

Only the genes, which were expressed with an intensity greater than log2(15) in at least 20% of the samples were selected. Marissa et al., had a less strict filtering criteria where they selected the genes if they were expressed in at least 5% of samples. The number of genes remaining after the first filter was 39661.

These filtered genes were subjected to a variance filter where genes that had a significant difference in their variance from the median variance were retained. A percentile of chi-squared distribution with n-1 (n=134) degrees of freedom and significance level threshold of p<0.01 was used to measure the significance. The upper one tailed test was used, as Marissa et. al. and the

BRB-Array Tools Version 4.6 user manual, had not mentioned whether to use lower or upper one tailed or two tailed tests for analysis (Simon R., 2020). In this analysis, all genes in the lower one tailed test were filtered out in the covariance filter and hence using only the upper one tailed test was sufficient. The resulting number of genes from both the first and second filter was 19896. Additionally, the variance filter on the RMA normalized ComBat adjusted matrix gave a significantly high number of genes (27401 genes), which was extremely high to do further analysis. This is because low sample size reduces the stringency of the filter. Therefore, further analysis was carried out using both the filters.

The final filter was the covariance filter which included genes with a coefficient of variation greater than 0.186. After the covariance filter the expression data had 1531 genes.

The samples (n=134) were clustered using the unsupervised hierarchical clustering, with Euclidean distance matrix, which gave two clusters which were cut to give the clusters belonging to C3 subtype (n=77) and C4 subtype (n = 57). A heat map (Fig 3) was constructed which showed clear distinction between these two subgroups with some samples showing minor differences in their expression levels with respect to their subgroups. Further the colorbar showed that the two samples in C4 clustered with C3 samples and this can be attributed to the type of clustering algorithm utilized and the presence of outliers in C4 and a region of overlap between C3 and C4 which was also visualized in the PCA plot (Fig 2). The C4 subtype had a higher expression level on average (Fig 3).

The mean of expression values of genes in each cluster was calculated to find the genes which represent each cluster best. 226147_s_at (*PIGR*) had the highest mean expression in cluster 2. These genes represent cluster 2 (C3) best. 229271_x_at (*COL11A1*) had the lowest mean expression in cluster 2. 212464_s_at (*FIN1*) had the highest mean expression in cluster 1. 232737_s_at (*ENPP3*) has the lowest mean expression in cluster 1. These genes represent cluster 1 (C4) best.

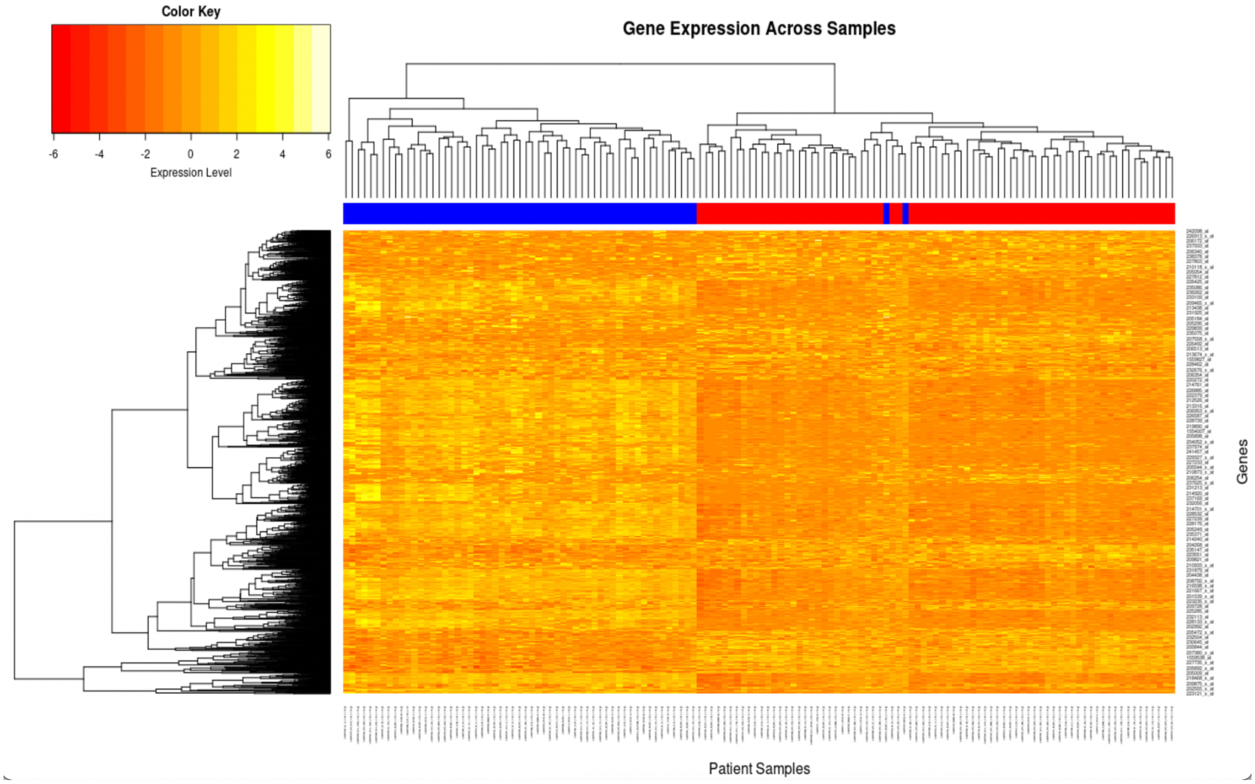


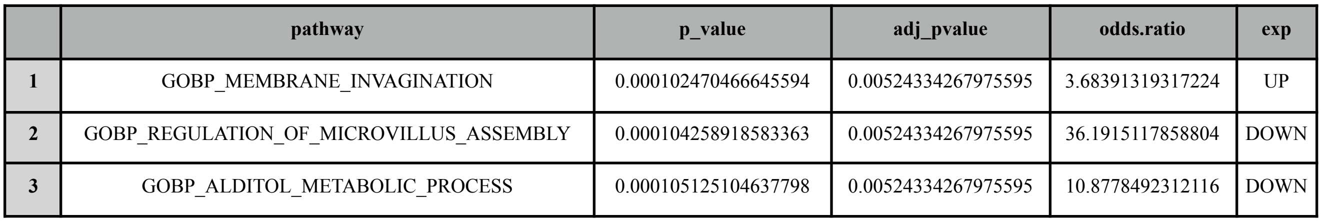
Fig 3. Heat map showing differential expression of the genes across the patient samples. Each row is a gene and each column is a sample. The colorbar in blue represents the C3 subtype and red represents the C4 subtype. Hierarchical clustering was used to cluster patients into subtypes.

However, a simple calculation without any statistical significance cannot be used to assess differential expression and hence further analysis using Welch t-test with a significance level of $p<0.05$ which compares means between two groups was performed. 1236 genes were differentially expressed. The most differentially expressed genes were 204457_s_at (*GAS1*), 223122_s_at (*SFRP2*) and 209868_s_at (*RBMS1P1*) which had a very low adjusted value. These genes are true positives and therefore best define the cluster. For instance., *GAS1* and *SFRP2* best represent the C4 cluster, as this has a significant differential expression and is also reported by Marissa et. al. as top dysregulated genes in C4 and causes aggressiveness of colorectal cancer cells. Since the C4 subtype has a poor prognosis, these genes best represent the C4 cluster.

	PROBEID	SYMBOL	stat_expr	p_val_expr	padjust_expr	reg
1	207266_x_at	RBMS1	20.5784645127299	5.52639924529169E-43	2.19182520467514E-38	up
2	209356_x_at	EFEMP2	20.3321202866381	3.37233989501583E-41	1.91071960823175E-37	up
3	203748_x_at	RBMS1	20.2959974973448	2.81527765785227E-42	5.58283635940394E-38	up
4	1555724_s_at	TAGLN	20.1071686453893	5.6265587136901E-42	7.43849817145544E-38	up
5	200788_s_at	PEA15	19.9729795139428	1.12715030905414E-39	4.47039084073964E-36	up
6	205547_s_at	TAGLN	19.8925122196625	1.54318032157211E-41	1.53010186834679E-37	up
7	209868_s_at	RBMS1	19.8735553307095	2.52059754545202E-41	1.8883573971079E-37	up
8	1555630_a_at	RAB34	19.8073278395449	2.85674702671324E-41	1.8883573971079E-37	up
9	213413_at	STON1	19.6955309165704	8.40737931167118E-39	2.22296713920127E-35	up
10	206580_s_at	EFEMP2	19.6801990862637	9.82316814419265E-38	1.77089396257648E-34	up
11	235350_at	C4orf19	-12.9964393679278	1.38862314490874E-22	8.99904943631137E-21	down
12	236513_at	PRELID2	-12.8255416357119	3.83740274697594E-24	3.08712434782581E-22	down
13	220622_at	LRRC31	-12.7620269107839	4.01798914242922E-24	3.21934277531081E-22	down
14	203240_at	FCGBP	-12.5336464931352	3.67196598495091E-22	2.28984029762796E-20	down
15	227725_at	ST6GALNAC1	-12.2486424560812	4.8684255569258E-20	2.29187645382516E-18	down
16	234008_s_at	CES3	-12.0100402540068	7.45032042636412E-23	5.01638553598356E-21	down
17	214106_s_at	GMDS	-11.9531880694867	3.61183275705403E-19	1.52068894880594E-17	down
18	226302_at	ATP8B1	-11.9453212391497	1.25314911004258E-20	6.31526643626415E-19	down
19	212814_at	AHCYL2	-11.7812618505027	2.74143379212015E-21	1.50592805580716E-19	down
20	219573_at	CARMIL1	-11.6790455159196	1.1134662839615E-20	5.65444126609435E-19	down

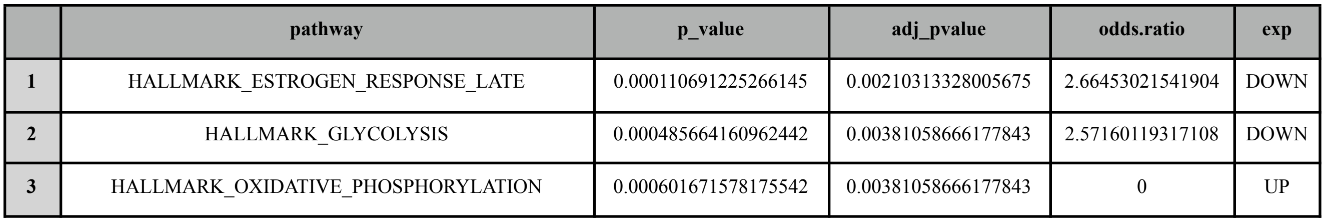
Table 1. Rows (1-10) top 10 up-regulated genes, rows(11-20) top down-regulated genes. The most up-regulated genes were RBMS1, EFEMP2, RBMS1 and TAGLN. Down-regulated genes consisted of C4orf19, PRELID2, LRRC31 and FCGBP.

(1) Top 3 GO



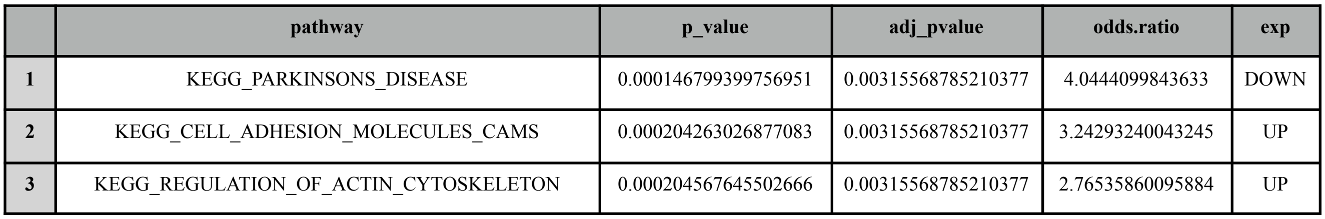
	pathway	p_value	adj_pvalue	odds.ratio	exp
1	GOBP_MEMBRANE_INVAGINATION	0.000102470466645594	0.00524334267975595	3.68391319317224	UP
2	GOBP_REGULATION_OF_MICROVILLUS_ASSEMBLY	0.000104258918583363	0.00524334267975595	36.1915117858804	DOWN
3	GOBP_ALDITOL_METABOLIC_PROCESS	0.000105125104637798	0.00524334267975595	10.8778492312116	DOWN

(2) Top 3 Hall



	pathway	p_value	adj_pvalue	odds.ratio	exp
1	HALLMARK_ESTROGEN_RESPONSE_LATE	0.000110691225266145	0.00210313328005675	2.66453021541904	DOWN
2	HALLMARK_GLYCOLYSIS	0.000485664160962442	0.00381058666177843	2.57160119317108	DOWN
3	HALLMARK_OXIDATIVE_PHOSPHORYLATION	0.000601671578175542	0.00381058666177843	0	UP

(3) Top 3 KEGG



	pathway	p_value	adj_pvalue	odds.ratio	exp
1	KEGG_PARKINSONS_DISEASE	0.000146799399756951	0.00315568785210377	4.0444099843633	DOWN
2	KEGG_CELL_ADHESION_MOLECULES_CAMS	0.000204263026877083	0.00315568785210377	3.24293240043245	UP
3	KEGG_REGULATION_OF_ACTIN_CYTOSKELETON	0.000204567645502666	0.00315568785210377	2.76535860095884	UP

Table 2. (B) Top 3 enriched pathways for GO, Hallmark KEGG gene set types. Hallmark had a total of 19 significantly enriched gene sets. GO had a total of 1689 significantly enriched gene sets. KEGG had a total of 57 significantly enriched gene sets. Hallmark GOBP Membrane Invagination, Hallmark_Oxidative_Phosphorylation, KEGG_CAMS, KEGG_REG_OF_ACTIN _CYTOSKELETON were up-regulated. While KEGG_Parkinsons_Disease, Hallmark_Estrogen_Response_Late, Hallmark_Glyclycosis, GOBP_Regulation_Of_Microvillius_Assembly, GOBP_Alditol_Metabloic_Process were down regulated.

Differential expression results were mapped to HGNC symbols and gene sets using hgu133plus2.db and GSEA packages. In table 1, the top up-regulated genes from patient samples were *RBMS1, EFEMP2, RBMS1, TAGLN, RAB34,* and *STON1*. Along with the most up-regulated genes, the top up regulated pathways were membrane invagination, oxidative phosphorylation, cell adhesion molecules, and regulation of actin cytoskeleton as shown in table 2. Furthermore, upregulated genes remained greatly supported with around a two-fold difference in adjusted p-values compared to down regulated genes. For example, RBMS1 had an adjusted p-value of 2.19E-38 and *C4orf19* had an adjusted p-value of 8.9E-21. In table 1, decreased differentially expressed genes were identified as *C4orf19, PRELID2, LRRC31, FCGBP, PRELID2, ST6GALNACC1, CES3, GMDS, ATP8B1, AHCYL2* and *CARMIL1*. It is shown there are more unique gene probe identifiers in down regulated genes in relation to up-regulated genes.

Discussion

In our attempt to replicate the findings from the Marissa et. al. paper, we processed a subset of 134 samples labeled as C3 or C4 from the original study. Our quality control of these samples suggested that no significantly flawed samples were still present in the data, meaning that we found no issues with the preprocessing performed by Marissa et. al. Our principal component analysis of the C3 and C4 samples illustrate that each of the two subtypes display distinct patterns of gene expression. The number of genes that were obtained after noise reduction was 1531, which was higher (by an order of 100 genes) than the number of genes Marissa et. al. had obtained. This may be due to lower sample size than Marissa et al., Even though Marissa et al, used a threshold of only 5% in the first filter, at a large sample size(n=750), the filters are very stringent as the sample size will push a distribution towards a normal and hence the tails of the distribution will be very less. Therefore, this analysis yielded a higher number of genes after noise filtering due to a lower sample size.

In our analysis hierarchical clustering was used which resulted in two subtypes, C3 and C4. Marissa et. al. had used consensus clustering on a larger set of samples which gave six subtypes, but since in our analysis we used hierarchical clustering on only 134 patient samples, only two distinct subtypes, C3 (n=77) indicated by red and C4 (n=57), indicated by blue in Fig 3. were obtained. A heatmap with a colorbar was constructed to visualize the clustering and expression levels. Marissa et. al had reported that there C4 was the most distinct subtype and our analysis did not entirely replicate this as the colorbar showed that two C4 samples were clustered with C3 sample, which may be because we used hierarchical on a small number of samples instead of consensus clustering on a larger sample size. Differential expression analysis was performed using Welch t-test to determine the upregulated and downregulated genes and only 1236 genes were differentially expressed.

Although we were not able to reproduce exact gene set pathways, high expression of KEGG regulation of actin cytoskeleton pathway was identified. Moreover, there were similarities in pathways between Marissa et al. and our findings. In table 2, our up-regulated KEGG CAMS pathway finding encapsulates KEGG tight junction and T cell receptor signaling pathways found in the Marissa et al. paper. A significant difference was that KEGG Parkinson's Disease pathway showed significantly low expression in our findings. There were little to no connections between GO and Hallmark findings. Nonetheless, a highly expressed Hallmark oxidative phosphorylation pathway was shown in table 2. This response is expected due to extreme physiological stressors in certain diseases(Hon et al., 2021). In Marissa et al., 1,108 discriminant probe sets were obtained. However, we kept probe ids that mapped to the same gene due to disagreeing p-values and adjusted p values, elucidating potential splice variants (Stalteri and Harrison, 2007). In figure 1, it was found that only up-regulated genes, such as *RBMS1, TAGLN, EFEMP2*, contain non-discriminant probe id and hgnc symbol mapping. Importantly, misregulation of splicing could lead to mutations and has been found to progress colorectal cancer through cell proliferation, apoptosis and angiogenesis which is known to involve *RBMS1* (Ward and Cooper, 2010; Chen et al. 2021; Yu et al., 2020). Further analyses on these non-discriminant probe ids and gene symbols could provide scientists more information about targeting splicing processes for therapeutic strategies in treating colorectal cancers (Yu et al., 2020).

Conclusion

Overall, we were generally able to reproduce the findings from Marissa et. al. We observed evidence for distinct patterns of gene expression between the C3 and C4 subtypes through principal component analysis. This supports the original paper's hypothesis that the subtypes of CRC are driven by different pathways and inciting mutations. Since we utilized only 134 samples from the C4 and C3 subsets in our analysis, we had a different number of genes that got filtered and clustering gave only 2 subtypes as expected. Some difficulty was encountered trying to apply the GSEABase package to the top 1000 genes with affecting our exact fisher t-test results in table 1 and 2. The top three gene set pathways were not identical to Marissa et al. because the top 1000 genes were filtered using expression values and not p-values. Our analysis, reaffirmed that mRNA expression analysis of colorectal cancer tissues using microarray are ideal for molecular classification of colorectal cancer

References

Berke Jeffrey T. *et al.*, 2021. sva: Surrogate Variable Analysis.

Bolstad B. M. *et al.*, 2004. Low Level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization. Dissertation. University of California,

Chen, Y. *et al.*, 2021. Alternative splicing of mRNA in colorectal cancer: new strategies for tumor diagnosis and treatment. *Cell death & disease*, *12*(8), pp.1-16.

Durinck S. *et al,* 2009. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt." *Nature Protocols*, **4**, 1184–1191.

Gautier, L. *et al.,* 2004. affy---analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20, 3 (Feb. 2004), 307-315.

Hon, K.W. *et al.,* The Crosstalk Between Signaling Pathways and Cancer Metabolism in Colorectal Cancer. *Frontiers in Pharmacology*, *12*, pp.768861-768861.

Kuipers, E. J. *et al.,* 2015. Colorectal cancer. *Nature reviews. Disease primers*, *1*, 15065. https://doi.org/10.1038/nrdp.2015.65

Marc Carlson *et al.,* 2021. hgu133plus2.db: Affymetrix Affymetrix HG-U133_Plus_2 Array annotation data (chip hgu133plus2). R package version 3.13.0.

Marisa, L. *et al.,* 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

Morgan M *et al.,* 2021. *GSEABase: Gene set enrichment data structures and methods*. R package version 1.56.0.

Pagès H. *et al.,* 2021. *AnnotationDbi:* Manipulation of SQLite-based annotations in Bioconductor. R package version 1.56.2, https://bioconductor.org/packages/AnnotationDbi.

Simon R. *et al*., 2020. BRB-ArrayTools Version 4.6. https://brb.nci.nih.gov/BRB-ArrayTools/Documentation.html

Stalteri M.A. *et al*., Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips. *BMC bioinformatics*, *8*(1), pp.1-15.

Ward A. J. *et al.,* The pathobiology of splicing. The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland. 2010 Jan;220(2):152-63.

Wickham H. *et al.,* ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Yu J. *et al*., 2020. RBMS1 suppresses colon cancer metastasis through targeted stabilization of its mRNA regulon. *Cancer discovery*, *10*(9), pp.1410-1423.