Microarray-Based Tumor Classification

David Lenci, Nikita Tomar, and Daniel Gealow

Introduction

Over their lifetime, 1 in 23 men and 1 in 25 women in the United States develop colorectal cancer [1]. The TNM staging system is the only classification currently used in clinical practice for treatment decisions and prognoses [2]. However, these stages are a poor predictor of recurrence [3]. Marisa et al. posit that attempts to find prognostic biomarkers may be ineffective due to variation in the molecular activity of different colon cancers. Therefore, in *Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value*, they attempt to expand upon and refine previous classifications of molecular subtypes. While previous studies had classified colon cancer based on a few specific markers, such as specific genes [4] or DNA methylation [5], the authors decided to use genome-wide mRNA expression analysis to obtain a more holistic classification system and attempt to determine prognostic markers for each category [3].

Data

The data used in our analysis consists of RNA expression levels in tissue collected from colon cancer tumors and non-tumoral colorectal mucosa. This dataset was generated by Marisa et al. in *Gene expression classification of colon cancer into molecular subtypes* [3] and made public as GSE39582 [6]. Tumor samples were collected by the Cartes d'Identité des Tumeurs program of the Ligue Nationale Contre le Cancer [7] from 750 patients with stage I to IV colon cancer who received surgery in any of seven hospitals across France between 1987 and 2007. These samples were preserved by freezing at -80ºC. Patients who received chemotherapy or radiation

therapy before surgery were excluded from this dataset. mRNA expression was measured using the Affymetrix U133Plus2 chip, using standard procedures. Samples with poor RNA quality were also excluded, leaving 566 colon cancer samples [3]. The original study assigned each sample to one of six molecular subtypes; their categorization was included in the metadata [6]. Of these, we used the 134 samples in subtypes C3 and C4 for our analyses.

Methods

ReadAffy function in the affy R package was used to read probe level information of 134 CC patient samples stored in the CEL files to create an AffyBatch object. These samples were normalized by using the robust multi-array average (RMA) expression measure function found in the affy package. The rma function transformed the probe level data to gene level data by converting the AffyBatch object to an ExpressionSet object. Normalization of the files involved a three-step approach in order to correct for variation between arrays:1) quantile normalization such that arrays can be compared to each other, 2) background correction to remove background noise and artifacts, and 3) summarization which combines probe set intensities across all of arrays into one gene expression measure [8].
The batch effects in the RMA normalized expression dataset were corrected using the ComBat function in the sva package. A file containing clinical and batching annotations, used by Marisa et al. [3], was used to set the batch and mode parameters.The batch effects correction included center and RNA extraction method while preserving two features of interest - tumor and MMR status.

To check the quality of the samples, the Bioconducter package AffyPLM was used to convert the AffyBatch object into a PLM set using the fitPLM functionality. Two quality control metrics were computed for each sample - the Relative Log Expressions (RLE) medians and Normalized Unscaled Standard Error (NUSE) medians.RLE calculates the log expression of each gene in

each array and then subtracts the median gene expression across all arrays for each gene [9]. NUSE normalized 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.

To reduce the dimensionality of the expressed data, Principal Component Analysis (PCA) was performed on the RMA normalized, ComBat adjusted gene expression data. Using the scale function the data was scaled and centered with respect to each gene before being analyzed with PCA. The prcomp function was then used to obtain principal component values. Here the scale and center parameters were set to false. The variations in the principal component were observed and the first two principal components which represented about 20% of the data variability were selected and plotted. Outliers of the first two principal components were determined as being three standard deviations away from the mean.

With the ComBat adjusted RMA normalized gene expression data generated, the matrix was then run through a series of filters according to Marisa et al: (1) Probe Sets were checked to see if they were significantly expressed in at least 20% of samples, (2) using a chi-squared test probeset were checked to ensure they had a variance significantly different from the median variance across all probesets with a p-value < 0.01, and (3) have a coefficient of variation greater than 0.186.

For the first filter each probeset was checked to ensure that at least 20% of its expression values across all samples was greater than log2(15). This was done by summing the number of gene expression values that achieved this threshold and dividing this sum by the number of samples. Any probeset that returned a value less than 20% was filtered out. Next a chi-squared test was performed to find probe sets that have a variance greater than the medium of all

probesets. To do this, the qchisq() function was used to calculate the upper and lower bounds of the test by setting the upper bound to qchisq(p/2,df) and the lower bound to qchisq(1-p/2,df), where p=0.01 and df=133 (Samples-1). Next, the T-statistic was calculated where T = (df*v/p2) for each probeset and was then checked to determine if lower bound < T < upper bound. If the calculated T fell into this range the probeset was filtered out. Lastly, the coefficient of variation(CV) was calculated for each probeset, and probesets with a CV less than 0.186 were filtered from the data set.

Once filtered, hierarchical clustering was then performed on the remaining gene expression data. First, using the agnes function from the cluster module the agglomerative coefficient (AC) was determined for different forms of hierarchical clustering available through the *hclust* function. AC values closer to 1 represent better clustering of our data. According to *agnes* the *ward* method would be the best form of clustering with an AC of 0.72. This also aligns with what the researchers in Marisa et al did when they performed their consensus clustering. Clustering was then performed using the *hclust* function with “ward.D” option.

A Welch T-test was then performed on the gene expression data across the two clusters in order to identify genes differentially expressed with a p-adjusted value < 0.05. This was done by first placing the data in *long* format to work with the *t_test* function. Then the p-adjusted value was calculated using the “fdr” method. The most differentially expressed genes were then further investigated.

Results

Based on the results of RLE, there are two samples with median scores which fall outside of the general distribution with scores greater than 0.10. However all of the RLE median values are close to zero indicating that the samples are of good quality (Fig.1). Similarly, based on the

results of NUSE, two samples with median scores greater than 1.05 were observed. However all median values are close to one, further indicating that the samples are of high quality (Fig. 2).

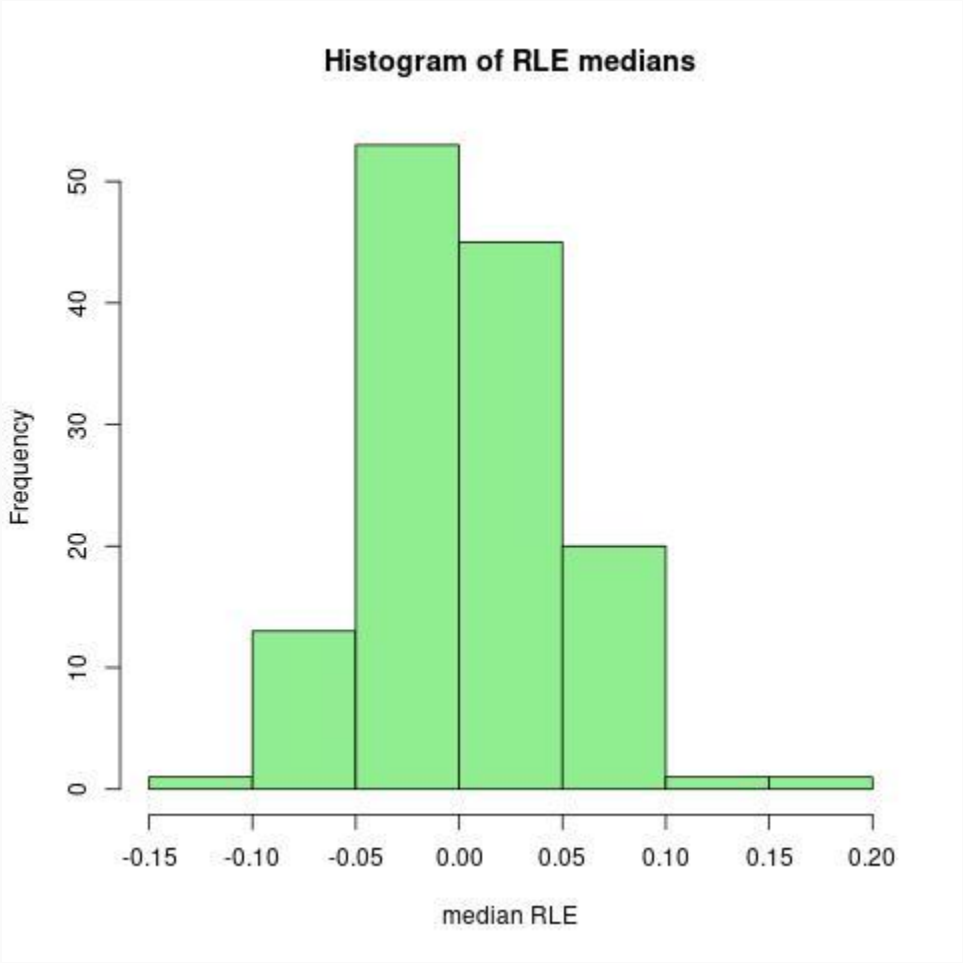

Figure 1. Histogram of median RLE. This histogram represents the distribution of median RLE scores across all samples.

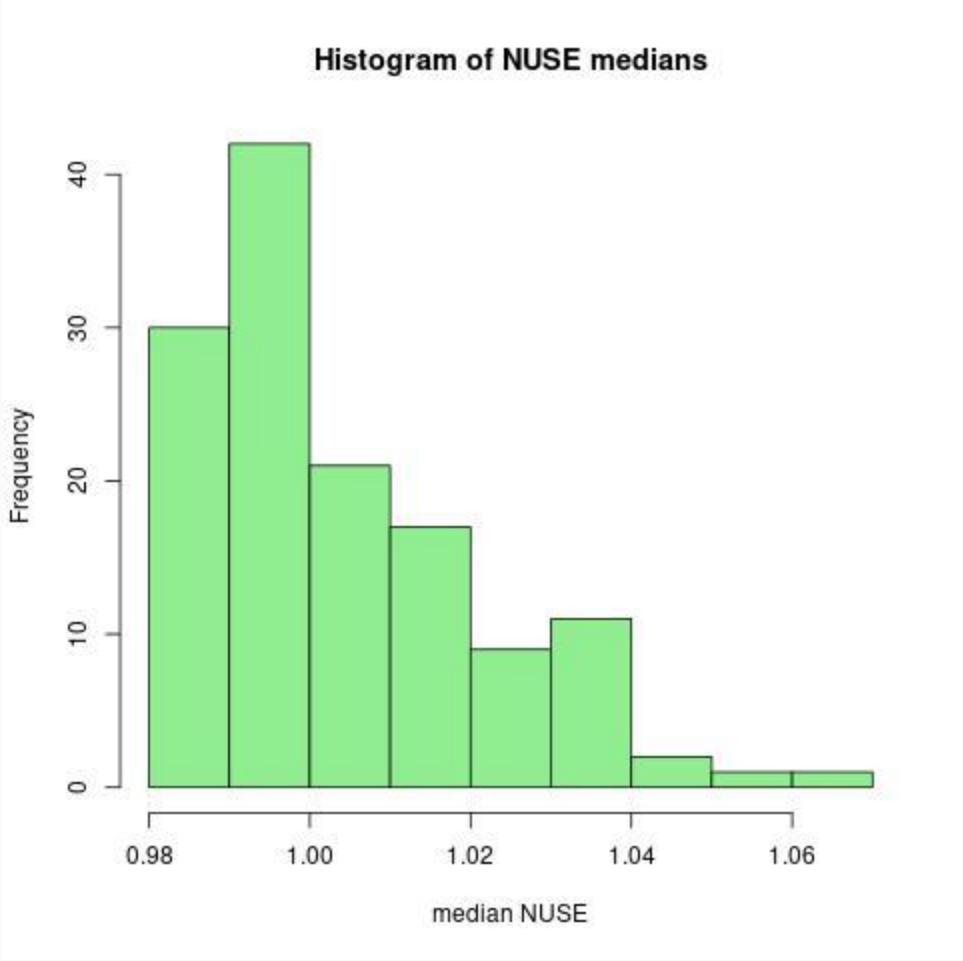

Figure 2. Histogram of median NUSE. This histogram represents the distribution of median NUSE scores across all samples.

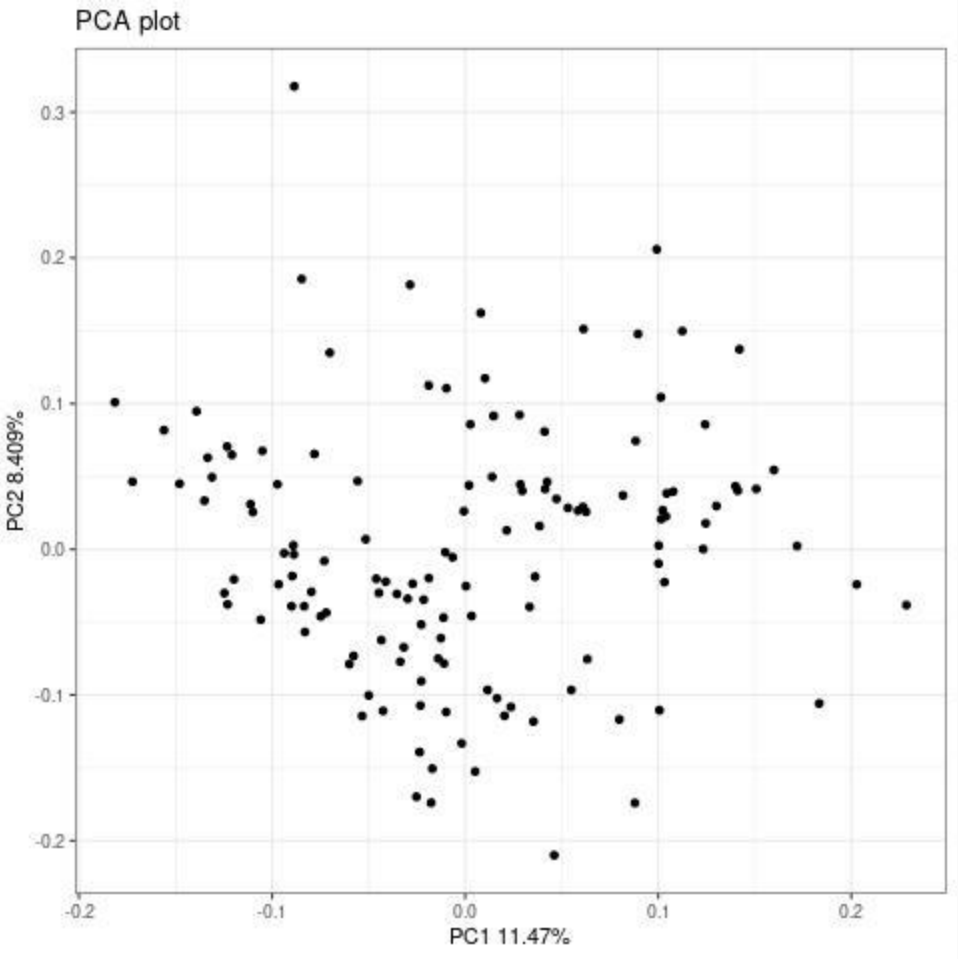

Figure 3. Plot of principal components 1 and 2. PC1 and PC2 represent 11.47% and 8.41% of the variance, respectively.

Fig. 3 shows the plot of the two principal components that displayed the highest variance contributing to the PCA result.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, GSM972350_MFL_036b_U133_2.CEL.gz, GSM972467_MFL_400b_U133_2.CEL.gz files. After the removal of this sample, the PCA was performed again to characterize the effects of the removal on the variance (Fig. 4C).

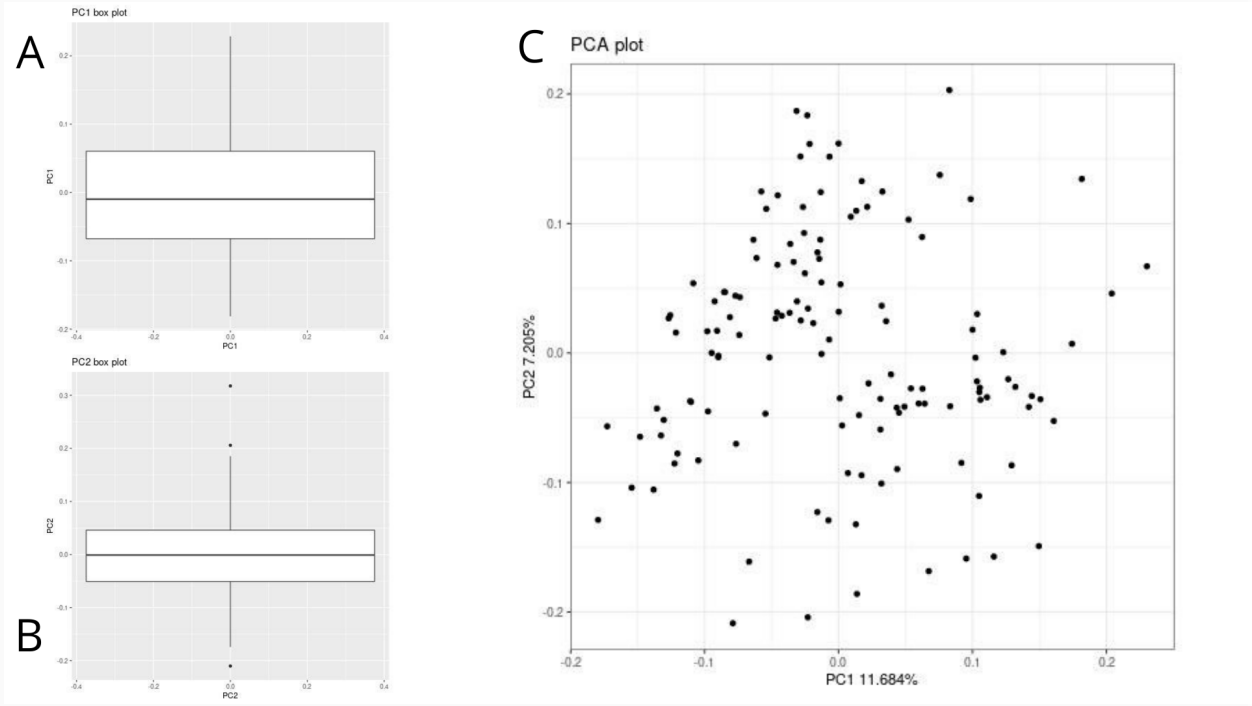

Figure 4. A. Box Plots of PC1. There are no outliers in the plot of PC1. **B. Box Plot of PC2.** There are 3 outliers of PC2 outside of the IQR. **C. PCA of principal components 1 and 2 after removing the samples contributing to the outliers.** PC1 and PC2 represent 11.68% and 7.21% of the variance, respectively.

The RMA normalized ComBat adjusted expression file was first filtered for noise using three filters. First, the gene expression file was filtered to only the genes where the expression value of those genes was at least log2(15) for at least 20% of the samples. This reduced our gene expression matrix to 39,661 genes. Then, a two-tailed chi-squared test was performed on the remaining data using a p-value cutoff of 0.01. This reduced our number of genes down to 29,646. Lastly, genes with a CV less than 0.186 were filtered out. This resulted in a final gene count of 1,151 genes in the expression matrix. Compared to Marisa et al, our filtering resulted in less probesets used to perform clustering, and to generate the heatmap in Figure (5). This could be from an unknown error in the calculations in the code, or due to the fact that we are only

looking at the C3 and C4 subtypes resulting in more data being filtered than would have if the data from the other subtypes was present.

Using the fully filtered expression matrix containing 1,151 genes, hierarchical clustering was performed in order to identify two main clusters. Using the *hclust()* function with the “ward.D" option the data was split into two clusters. This differs from the original paper, which defined six clusters within its data. This is due to our analysis only including the C3 and C4 colorectal cancer subtypes, and not the C1, C2, C5, and C6 subtypes that were included in the paper. From this analysis two clusters were identified with one cluster containing 57 samples and the other containing 77 samples, and from the heatmap in Figure (5) we can see that for the most part these clusters correlate with the C3 and C4 subtype that the samples come from. Although, there are a few samples from subtype C4 that are clustered with subtype C3. These samples were GSM972019, and GSM972412.

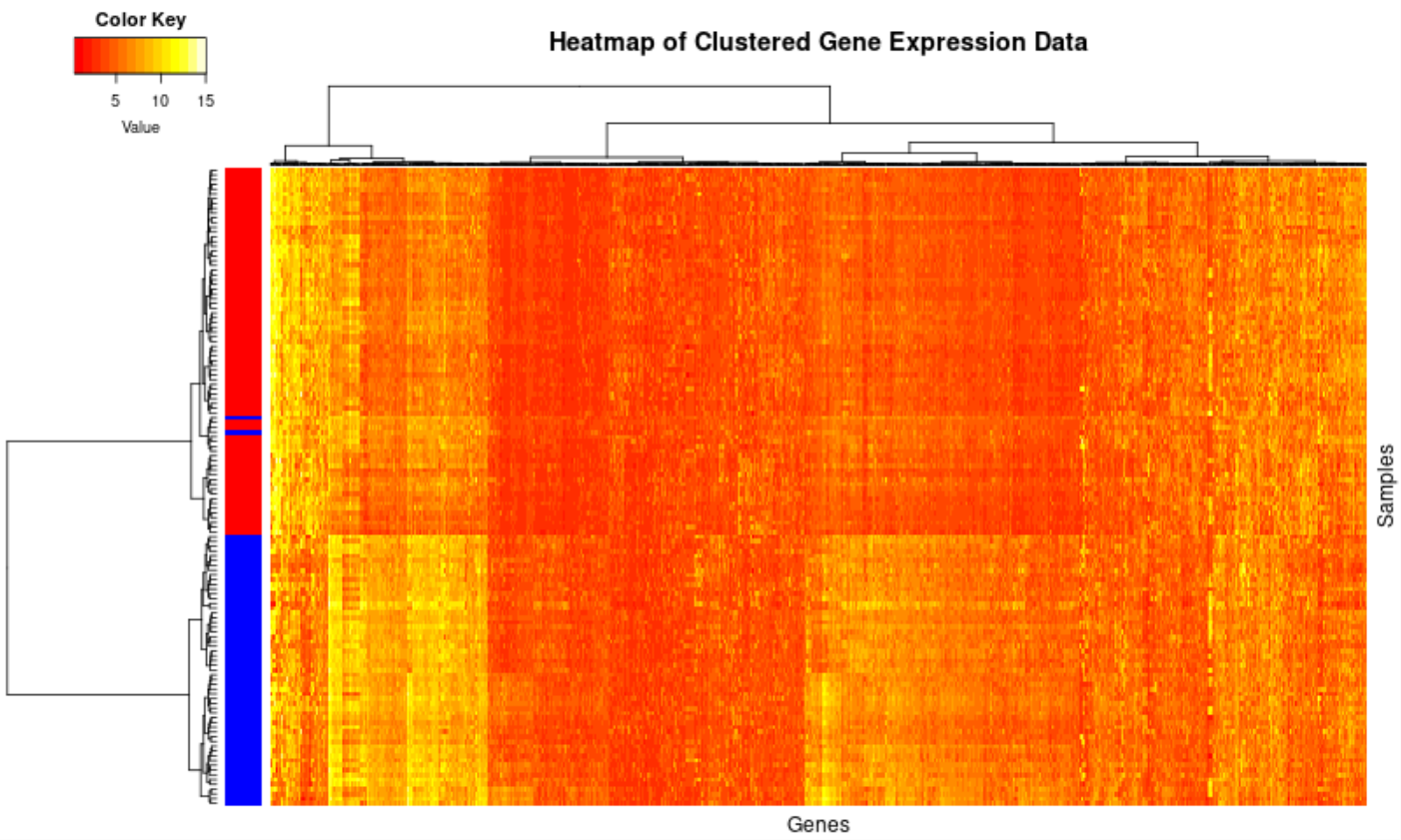

Figure 5: This heatmap shows the expression of the filtered genes across samples clustered using hierarchical clustering. Color bar on the left side of the figure shows how samples are clustered with blue indicating C4 subtype and red indicating C3 subtype.

Using the generated clusters a Welch t-test was performed on the gene expression data between the two clusters with a p-adjusted threshold of 0.05 in order to identify significantly differentially expressed genes. This resulted with 957 differentially expressed genes across the two clusters. From this result we determined that 225242_s_at, 209868_s_at, 218694_at, 227059_at, and 202291_s_at best represented cluster 1 and 227725_at, 204673_at, 210107_at, 238750_at, and 242601_at best represented cluster 2. This was determined from the magnitude of the T-statistic calculated in the Welch T-test, which represents the difference in expression between the two clusters for each gene. The 10 most upregulated genes for each

cluster can be seen in Table 1 along with the T statistic for the gene, p-value and adjusted p-value.

Most Differentially Expressed Genes

Welch T-test performed between clusters 1 and 2 (C4 and C3) with P-adjusted < 0.05

	t	p	padj	hgnc_symbol
Cluster 1 (C4)				
225242_s_at	22.68809	3.10e-46	1.790250e-43	CCDC80
209868_s_at	22.66264	3.84e-47	4.435200e-44	RBMS1P1
218694_at	22.51416	4.99e-45	1.152690e-42	ARMCX1
227059_at	22.12721	4.68e-45	1.152690e-42	GPC6
202291_s_at	22.09310	9.09e-46	3.499650e-43	MGP
223121_s_at	21.97948	1.21e-41	1.466850e-39	SFRP2
226930_at	21.74076	6.48e-45	1.247400e-42	FNDC1
213413_at	21.32879	4.02e-41	3.571615e-39	STON1
238478_at	21.30755	5.94e-37	2.017853e-35	BNC2
227061_at	21.20784	1.71e-38	9.405000e-37	CCDC80
Cluster 2 (C3)				
227725_at	13.49234	2.82e-23	1.467162e-22	ST6GALNAC1
204673_at	13.40697	1.77e-24	1.007069e-23	
210107_at	13.27445	2.90e-25	1.753665e-24	CLCA1
238750_at	12.72529	1.09e-24	6.294750e-24	CCL28
242601_at	12.58634	5.03e-24	2.753389e-23	HEPACAM2
1553828_at	12.51860	8.62e-24	4.588065e-23	NXPE1
230615_at	12.29262	5.89e-23	2.957804e-22	DUOXA2
207214_at	12.21225	7.67e-22	3.487736e-21	SPINK4
227226_at	12.06944	1.90e-22	9.030864e-22	MRAP2
1561387_a_at	11.95321	4.34e-22	2.013133e-21	NXPE1

T shows magnitude of T statistic calculated

Table 1: This table shows the most differentially expressed genes across the two clusters with cluster 1 labeled as C4 and cluster 2 labeled as C3. Note: some of the samples from C4 did cluster

with C3 when hierarchical clustering was performed as shown in Figure (5). Additionally, probe id 204673_at did not return an associated HGNC gene symbol.

Discussion

Reproduction of the analyses performed in Marisa et al demonstrated clear clustering of the C3 and C4 colorectal cancer subtype, and identified significantly differentially expressed genes across the two subtypes. That being said, differences in results were found between our analysis and the analysis performed in the original paper. Firstly, the amount of probes filtered in our data did not match the filtering results in Marisa et al with 1,157 used in our clustering analysis and 1,459 used in Marisa et al. It is unclear which aspect of our filter is removing too many probes, but this did not seem to affect our ability to properly cluster the samples based on our filtered expression data. That being said, two samples from the C4 subtype did cluster with the C3 subtype, which could be a result of over filtering and losing important information. Although, this could also be due to the fact that we used hierarchical clustering in place of the consensus clustering approach used in Marisa et al.

Differential expression analysis through a Welch T-test identified 957 differentially expressed genes with a p-adjusted < 0.05. The three most differentially expressed genes in the C4 subtype included CCDC80, RBSM1, and ARMCXP1. CCDC80 promotes cell adhesion and has been identified as a tumor suppressor in specific cancer types [10], RBMS1 has been identified as a tumor suppressor in colon cancer and is associated with DNA binding [11], and ARMCXP1 is may be associated with tumorigenesis due to the presence of Armadillo repeats [12]. The most differentially expressed genes in the C3 subtype include ST6GALNAC1, CLCA1, and CCL28. ST6GALNAC1 has been identified to be highly expressed in ovarian cancer and correlates with tumor progression [13]. CLCA1 is associated with ion channel transport and transport of glucose, and has been found to suppress colorectal cancer [14]. CCL28 is associated with

inflammatory and immune response [15]. A more robust analysis would attempt to identify gene oncology terms associated with the most highly differentially expressed genes in each cluster in order to identify distinct molecular pathways associated with each subtype.

Conclusion

From this analysis we have demonstrated a semi-accurate recreation of the analyses in Marisa et al, even when overfiltering may have been applied resulting in the loss of information used in the original consensus clustering, and that colorectal cancer can be accurately clustered into distinct subtypes due to the expression profiles of the individual subtypes. The results of these analyses demonstrate the legitimacy of utilizing more robust prognoses based on colorectal cancer subtype biomarkers. Further analysis would follow the same line of thinking in Marisa et al, and identify specific molecular pathways present in each subtype.

References

[1] American Cancer Society. Key Statistics for Colorectal Cancer.

[2] American Joint Committee on Cancer (1997) AJCC cancer staging manual, 5th edition. Philadelphia: Lippincott-Raven.

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

[4] Shen L, Toyota M, Kondo Y, Lin E, Zhang L, et al. (2007) Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer. *Proc Natl Acad Sci U S A* 104: 18654–18659. [PMC free article]

[5] Hinoue T, Weisenberger DJ, Lange CP, Shen H, Byun HM, et al. (2012) Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome Res* 22: 271–282. [PMC free article]

[6] GEO. GSE39582.

[7] Cartes d'Identité des Tumeurs (CIT). *Ligue contre le Cancer.*

[8] Gautier, Laurent et al. “affy--analysis of Affymetrix GeneChip data at the probe level.” *Bioinformatics (Oxford, England)* vol. 20,3 (2004): 307-15. doi:10.1093/bioinformatics/btg405

[9] McCall, M.N., Murakami, P.N., Lukk, M. *et al.* Assessing affymetrix GeneChip microarray quality. *BMC Bioinformatics* 12, 137 (2011). https://doi.org/10.1186/1471-2105-12-137

[10] Grill, J.I., Kolligs, F.T. *DRO1/CCDC80*: a Novel Tumor Suppressor of Colorectal Carcinogenesis. *Curr Colorectal Cancer Rep* 11, 200–208 (2015). https://doi.org/10.1007/s11888-015-0276-3

[11] Yu, Johnny et al. “RBMS1 Suppresses Colon Cancer Metastasis through Targeted Stabilization of Its mRNA Regulon.” *Cancer discovery* vol. 10,9 (2020): 1410-1423. doi:10.1158/2159-8290.CD-19-1375

[12] NCBI. ARMCX1 armadillo repeat containing X-linked 1 (human).

[13] 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

[14] Li, Xiaofen et al. “CLCA1 suppresses colorectal cancer aggressiveness via inhibition of the Wnt/beta-catenin signaling pathway.” *Cell communication and signaling : CCS* vol. 15,1 38. 3 Oct. 2017, doi:10.1186/s12964-017-0192-z

[15] NCBI. CCL28 CC motif chemokine ligand 28 (human).