Project 1: 2/18/22
Group: Tinman

Microarray Based Tumor Classification:
An Attempt to Replicate Findings from Marisa et al. [1]

Data Curator: Allison Choy
Programmer: Aneeq Husain
Analyst: Rizky Kafrawi
Biologist: Reshma Ramaiah

INTRODUCTION (Data Curator)

Colorectal Cancer (CRC) is one of the most common malignancies in the world [6] and pathological staging has been the standard practice in determining patient prognosis, but it has limitations: patients still had recurrences despite removing lower grade cancerous growth in cases of localized CRC. Yet despite that, it is not uncommon for this method to serve as a basis for chemotherapy treatment. [2] In as much as 30% - 40% of those with stage III CRC and 10% - 20% for stage II CRC developing recurrences, many studies using microarray technologies have been done on the gene expression profiles (GEP) to identify a unique signature for this disease with an interest in survival rates [3] but have been poorly reproducible. This is possibly due to the different pathways affected from distinct molecular features, resulting in inconsistencies observed across studies. [4] However, microarray studies with unsupervised hierarchical clustering have identified three distinct subtypes of CRC, indicating that this is indeed a heterogenous disease.

In this study, Marisa et al. created a robust method to reproducibly classify CRC samples based on their microarray mRNA expression analysis with unsupervised hierarchical clustering– or more specifically, consensus clustering. The study reports six molecular subtypes for CRC, but this method can potentially provide a new standard in which patient prognosis is determined. Note that due to the computationally intensive nature of consensus clustering, our focus in this report will be to replicate these findings with hierarchical clustering on publicly available microarray datasets.

DATA (Data Curator)

This study examines genome-wise mRNA expressions of tumor samples from a large French multi-center (CIT program) cohort as its source as summarized in Figure 1. Out of a cohort of 750 patient samples, 566 satisfied the RNA quality control requirements. These were then further split up into two sets: a discovery set of 443 and a validation set of 123. The validation set also included 906 CRC samples from seven publicly available Affymetrix datasets

with accession numbers: GSE13067, GSE13294, GSE14333, GSE17536/17537, GSE18088, GSE26682, and GSE33113. Data from all datasets were obtained using the Affymetrix U133 Plus 2.0 chip platform and selected based on data availability, tumor location, and either the presence of common DNA alterations and/or patient survival data availability. The purpose of the latter is for relapse-free survival (RFS) analysis. As such, only pertinent data relating to patients with stage II-III CRC were used; stage I and IV will not be as informative as the data will be skewed one way or the other when compared to II-III. This study also included 152 samples from the Cancer Genome Atlas (TCGA), [5] which were obtained with a different platform (Agilent platform) but were included due to the extensive annotations of DNA alterations of these samples. In total, out of 443 cases for discovery and 1,029 for the validation set, only 359 and 416 samples, respectively, were used in this study. Our data for this study were stored in a centralized repository for analysis, but was missing one dataset, which was subsequently downloaded to the file system from Gene Expression Omnibus (GEO) with accession number GSM971958.

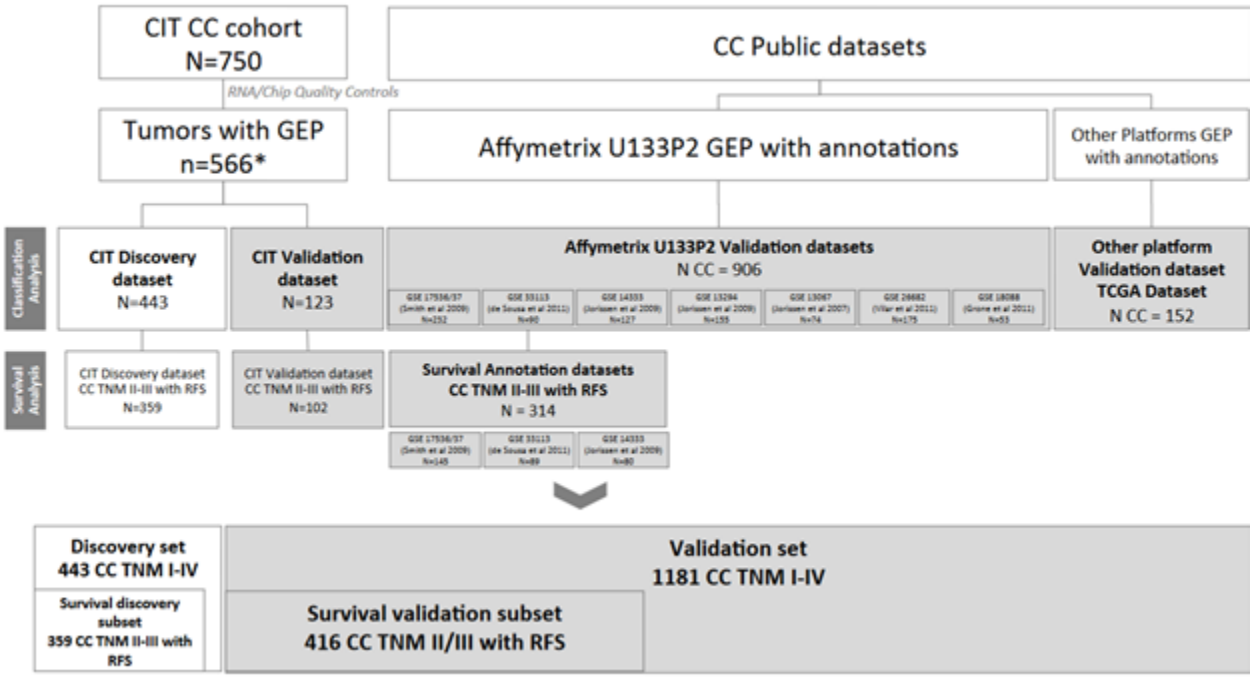


Fig. 1: Summary of Discovery and Validation sets used in study. From Supplementary Figure 1 from Marisa et al. [1]

Project 1: 2/18/22
Group: Tinman

METHODS (Programmer/Analyst)

The raw data downloaded from GEO was pre-processed on R (v4.1.1). First, the data was batch imported using the Affy R package (v1.72.0). Normalization was then carried out with the help of the robust multi-array average (rma) method from the previously mentioned Affy R package. Subsequently, the Relative Log Expression (RLE) and the Normalized Unscaled Standard Error (NUSE) scores were computed with the methods implemented in the AffyPLM R package (v1.70.0). This was done to check the quality of the raw data. The medians of the RLE and NUSE scores were also plotted as histograms to detect outliers. Following this, the data was corrected for batch effects with the use of the metadata file provided by Marisa et al. The ComBat method from the SVA R package (v3.42.0) was used to perform batch effect correction. PCA analysis was also performed on this data to check the quality. Further analysis was then conducted on the normalized data.

A series of filters was then applied to the RMA normalized, ComBat adjusted expression matrix. Firstly, the gene expression values for the starting matrix were filtered such that every gene had an expression value of >log2(15) in at least 20% of samples. The passing probesets were stored in another variable and fed into the next filter, which was to ensure that the variance of samples were significantly different ($p<0.01$) from the median variance of the collective set of probes. To do so, a test statistic for each probe ID was manually generated using the mutate(), ncol(), and rowVars() functions and stored in a new column. The confidence level (CL) and p values (p) were set to 0.99 and 0.01 respectively when setting up the qchisq() function. To perform a two-tailed chi-squared test, the critical values had to be calculated. The lower critical value was calculated as follows: qchisq((1- CL)/2, c-1). Conversely, the upper critical value was calculated as follows:qchisq((CL)/2, c-1, lower.tail = FALSE). Once the chi-squared test critical values had been determined, the data was filtered for t values greater than the upper critical value OR t values less than the lower critical value. The final filter applied to the gene expression matrix was finding samples greater than a coefficient of variation (CV) threshold of 0.186. This was done using the rowSds() and rowMeans() functions. Additionally, as our group had a biologist, an additional expression matrix was generated for probesets from the original RMA normalized, ComBat adjusted expression matrix filtered by just the second filter.

Hierarchical clustering was then performed on the fully filtered gene expression matrices samplewise. This was done by transposing the filtered genes and then feeding the resulting matrix into the hclust() function. Following the cluster visualization, the resulting dendrogram was then split into 2 clusters using the cutree function and specifying k = 2. A heatmap for the gene expression of every gene across all 35 samples was then performed using the heatmap() function. In order to visualize the subtype classifications of samples on the dendrogram generated at the top of the gene expression heatmap, we prepared a matrix containing a column with all the samples present in the filtered geneset. Using cbind(), we affixed the subtype classification of

'C3' or 'C4' for each and every sample in a new column. Determining the aforementioned classifications was done by referencing the proj_metadata.csv file in the scc directory 'doc' under the following filepath: /project/bf528/project_1/doc. The data frame containing the samples and their respective subtype classifications were then transmuted into a new column containing 'red' if a sample was classified as 'C3' and blue if a sample had a non-C3 classification– which was conducted using the transmute() and if_else() functions. A Welch-t test was then conducted on genes passing all 3 filters using the row_t_welch() function. The most differentially expressed genes were then selected based on t-statistic. For the biologist, a t-test analysis was also performed on the additional expression matrix that was only subjected to the chi-square test filter.

RESULTS (Programmer/Analyst)

The results of the median RLE histogram (Fig. 2) indicate that there are two samples which have median scores greater than 0.10. This represents the samples GSM971993 and GSM972390. The median NUSE scores histogram (Fig. 3) also shows similar results wherein two outlier samples (GSM972113, GSM972269) were detected with median scores greater than 1.05.

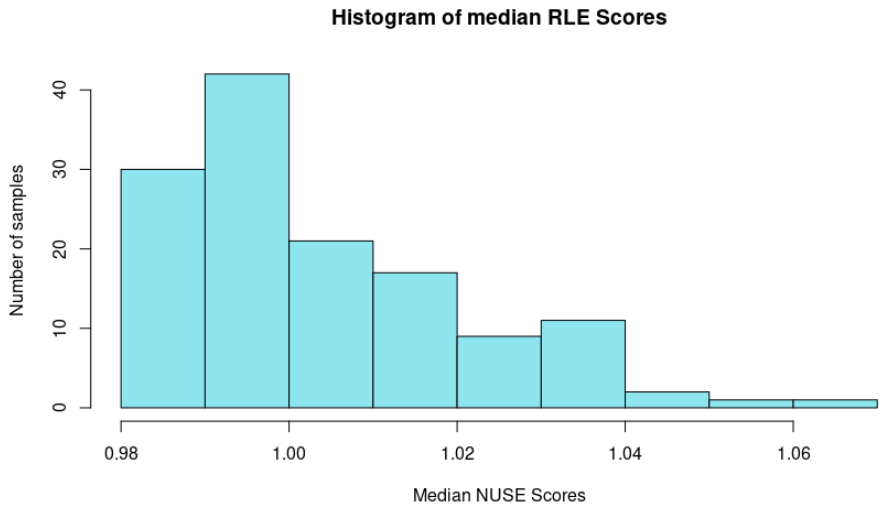
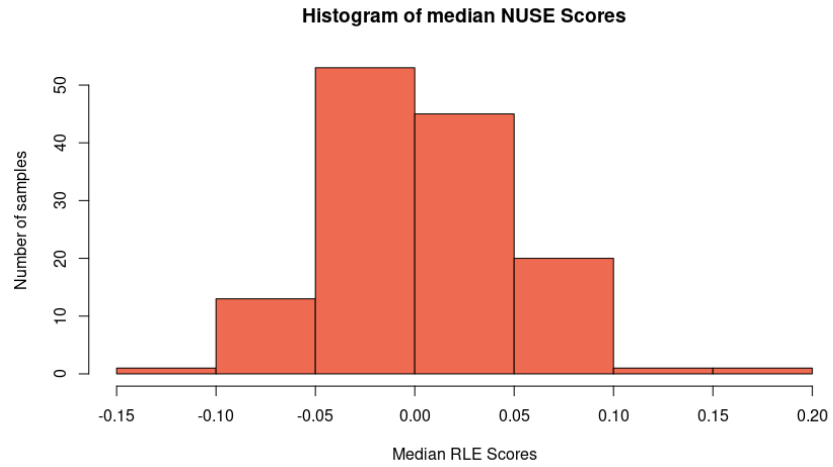


Fig. 2: Histogram showing the median RLE scores for 134 normalized and background corrected samples.



.

Fig.3: Histogram of median NUSE scores for 134 normalized and background corrected samples.

The principal component analysis performed on the normalized data produced a plot (Fig. 4) which accounted for 52.46% of the variance with the first principal component alone accounting for 46% of the variance. Since there is not much clustering it appears that none of the samples are outliers and the dataset is of good quality.

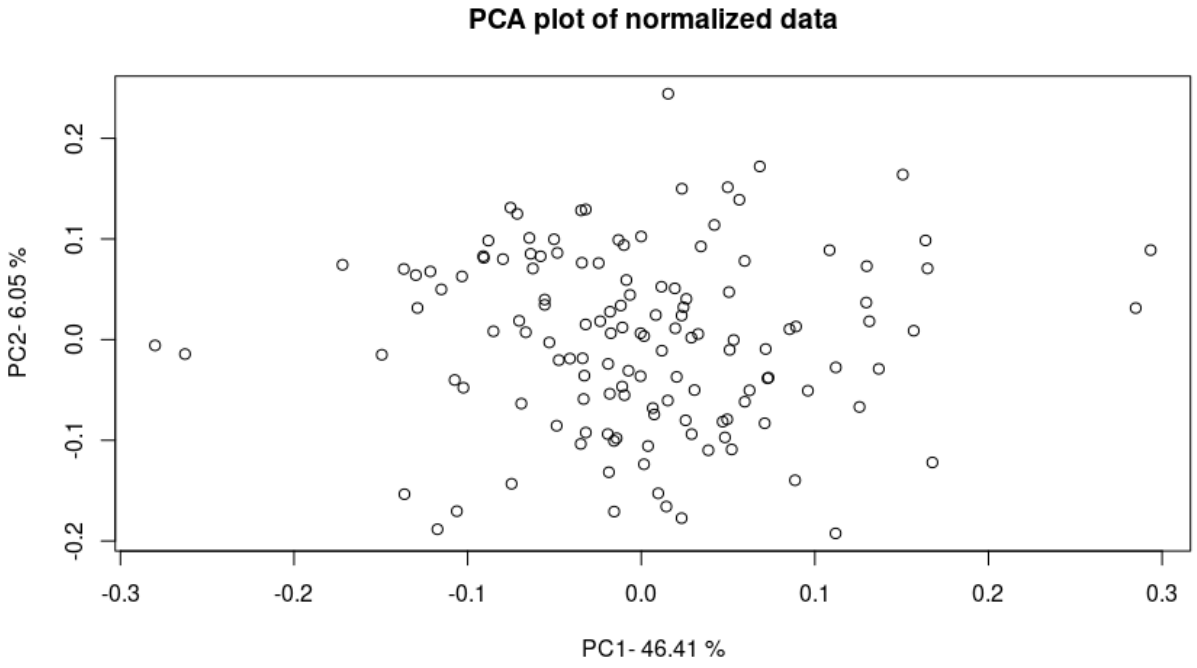


Fig. 4: PCA plot of the 134 normalized CC samples, values on axes indicate the variance.

Project 1: 2/18/22
Group: Tinman

A series of filters was applied to the RMA normalized, ComBat adjusted expression matrix, resulting in 39,354 gene probes passing the first filter, 29,373 gene probes passing the second filter, and 1482 gene probes passing the third filter. When initial hierarchical clustering was performed on these 1482 passing gene probes, 25 genes were designated as cluster 1 while the remaining 11 samples were part of cluster 2. Note that for the biologist gene expression matrix, 43215 genes from the original normalized gene expression matrix passed the chi-square filter.

It appears as though the sample clustering dendrogram depicted in the heatmap generated in Figure 6 is not consistent with the initial clustering I did to predetermine the number of samples per cluster. As seen in the figure below, 13 samples are classified as non-C3 samples, but 12 are clustered on the left-hand branch of the dendrogram and one non-C3 sample is clustered . These clustering inconsistencies could be a byproduct of the unsupervised methodology involved in grouping similar sets of data based on some criterion such as gene expression. When performing the Welch t-test on the gene probesets, the number of genes with a p-value of less than 0.05– i.e. that are differentially expressed– was 943. Furthermore, when accounting for FDR and generating an adjusted p-value, the number of differentially expressed genes drops even further, to 311 out of the original filtered set of 1482 genes.

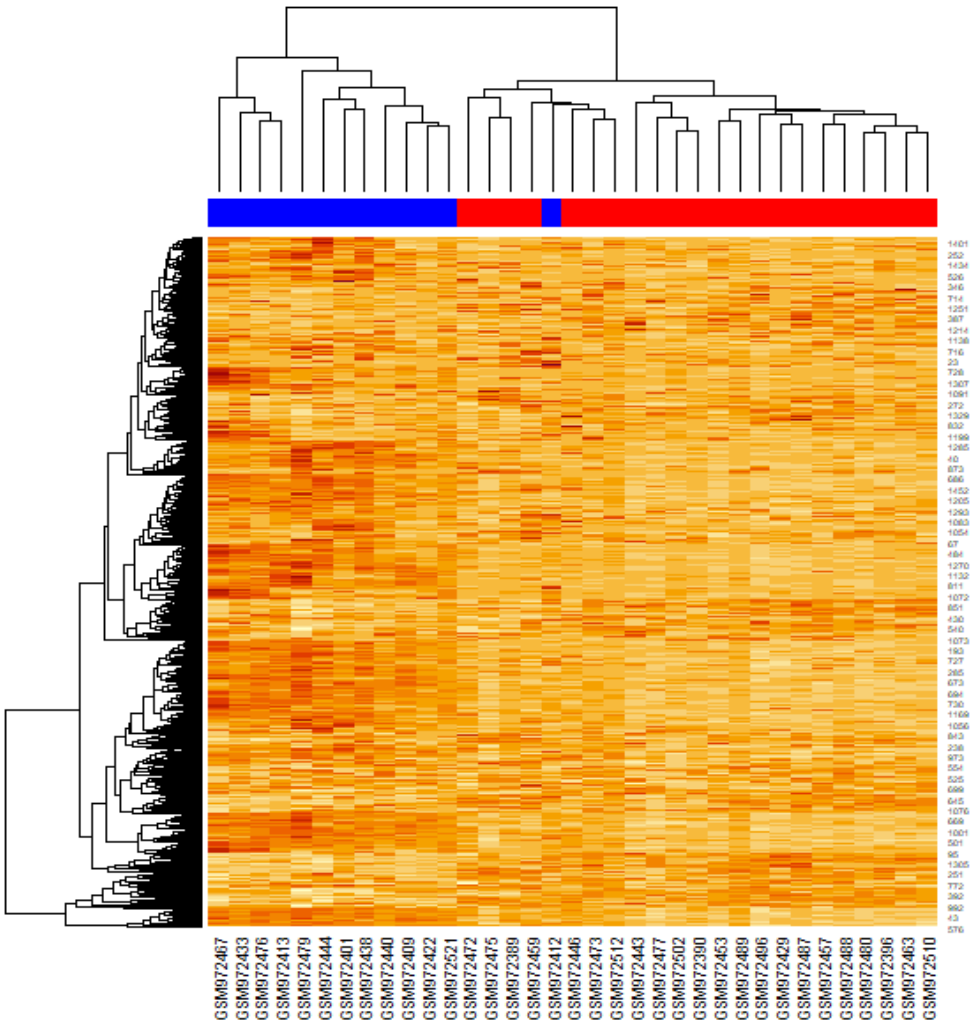


Fig. 5: Heatmap depicting gene expression across all 35 samples in our gene expression matrix dataset. Samples are present on the x-axis and gene probesets are located on the y-axis. The blue/red bar is representative of sample subtypes. If a sample belonged to the C3 subtype it was visualized as red, while blue illustrates non-C3 classified samples. Darker shaded cells are indicative of higher gene expression while lighter cells are representative of lower gene expression.

DISCUSSION (shared)

Top Ten Downregulated Genes				
Gene	**t-statistic**	**p-value**	**adj_p_value**	**GENE SYMBOL**
205100_at	-13.1448119	4.83E-10	7.12E-07	
204457_s_at	-12.31880504	3.01E-11	4.45E-08	GAS1
225242_s_at	-11.61158035	7.26E-11	1.07E-07	CCDC80
205880_at	-10.87314889	1.36E-08	1.98E-05	PRKD1
226930_at	-10.70792668	2.35E-10	3.48E-07	FNDC1
203695_s_at	-10.38639847	2.45E-08	3.53E-05	
214844_s_at	-10.25873899	2.43E-09	3.58E-06	
201843_s_at	-10.16032016	5.93E-09	8.68E-06	EFEMP1
213413_at	-10.14343704	3.82E-09	5.60E-06	STON1
202766_s_at	-10.07033295	2.72E-10	4.02E-07	
209946_at	-9.986585094	7.44E-08	1.06E-04	
Top Ten Upregulated Genes				
Gene	**t-statistic**	**p-value**	**adj_p_value**	**GENE SYMBOL**
210107_at	11.15445596	4.21E-12	6.24E-09	
223597_at	9.760621989	6.17E-11	9.13E-08	
227725_at	7.432790008	1.22E-06	1.65E-03	ST6GALNAC1
228004_at	7.316920327	3.05E-08	4.39E-05	NCRNA00261
232707_at	7.216223476	2.63E-07	3.67E-04	ISX
219955_at	7.067899401	6.09E-08	8.70E-05	L1TD1
236894_at	7.049964318	9.19E-08	1.30E-04	L1TD1
220234_at	6.980726443	6.64E-08	9.46E-05	CA8
203240_at	6.818014095	1.14E-05	1.43E-02	FCGBP
214142_at	6.589254633	3.26E-07	4.55E-04	ZG16
1559538_at	6.390745608	1.11E-06	1.51E-03	

Table 1: Top ten upregulated and downregulated genes in the C3 subtype group. Genes are ranked by the t-test statistic result after filtering for p-values < 0.05 to reflect the highest

deviations from the dataset. Missing gene symbols were unable to be matched in the variant probe set from Supplemental Table 2 of Marisa et al. [1]

In the analysis of the top ten most upregulated and downregulated genes (Table 1) in the enrichment studies of the C3 subtype, we have found 7 upregulated genes (ST6GALNAC1, NCRNA00261, ISX, L1TD1, CA8, FCGBP, and ZG16) and 6 downregulated genes (GAS1, CCDC80, PRKD1, FNDC1, EFEMP1, and STON1) that were identifiable from the probe set from the Supplementary Table S2. Interestingly, while some of these genes, such as NCRNA00261, [7] L1TD1, [8] and ZG16, [9] are specific for colorectal cancer, these biomarkers are also associated with a better prognosis of disease-free survival when upregulated. Others, such as ST6GALNAC1, ISX, and FCGBP, are pro-inflammatory and have been known to promote tumor growth (ISX) [10] or cancer stem cell differentiation into colorectal phenotypes (ST6GALNAC1) [11]. FCGBP, interestingly, is expressed in various tumors, but it is associated with immune infiltration, and thus can participate in tumor immunity.[12][13] The implications of an upregulation of this gene would have a more severe prognosis.

In contrast, the seven genes that were downregulated are involved in growth arrest (GAS1), [14] immune filtration (CCDC80), [15] tumor suppression (EFEMP1), [16] and cellular adhesion of this or other cancers. Downregulation of all these markers would implicate poor prognosis. Due to the key roles these genes play in the regulation of regular cellular processes and maintenance, they would be an appropriate set of markers to helpfully determine a patient's prognosis.

The filtration steps we performed on the RMA normalized, ComBat adjusted expression matrix to generate our filtered gene probe set for unsupervised hierarchical clustering yielded a similar count of variant probe sets. Marisa et al.,[1] generated a list of 1459 variant probe sets, while our analysis yielded 1482 genes. This was more likely to be a result of filtering by different thresholds/ranges than user error in the code considering how close the number of passing probe sets are for both our analysis and that of Marisa et al.,[1] For example, in the Marisa et al.,[1] paper itself, the authors specify a different filter from the log2(15) cutoff we used in favor of the "gene expression log2 fold changes of subtype Cj versus the other subtypes." Additionally, due to time constraints, the data that was filtered and used to generate the data necessary to generate the heatmap and its associated analyses was generated using the provided example gene expression matrix, not the normalized expression matrix put together by the programmer. Performing the same analysis on this matrix would yield notably different data. Since there are nearly 100 more samples included in the programmer normalized expression matrix, there is a lot of data currently missing from our analysis, which could also account for why the number of variant probe sets differ between our analysis and that of Marisa et al.[1] When looking at the second filter, the reason for only looking at genes with variances that are significantly higher than that of the

median variance would be that high variance in gene expression across an array of samples could be indicative of differential expression since samples that are expressed differentially have different quantities of RNA transcription based on an assortment of genomic factors. In a similar vein, genes from part 4.3, with high coefficient of variation values are more likely to be differentially expressed.

With regards to the heatmap, it is interesting how clustering using hclust() and then using cuttree() to assign a cluster to each sample node generates a different clustering pattern than that which is illustrated in Fig. 6. As previously mentioned in the results section, this inconsistency in the pattern of sample clustering could simply be a potential byproduct of the unsupervised nature of the clustering methods used for the analysis of these gene expression matrices. Additionally, had the full dataset of 134 samples been included in our analysis, several key metrics such as the p-value for each gene would have been significantly more robust, which could have possibly resolved the inconsistent cluster grouping we observed.

CONCLUSION (shared)

Perhaps due to the heterogeneity of colorectal disease, there has been no previously known methodology, at the time of publication of Marisa et al.,[1] to accurately predict the recurrence of, and therefore the prognosis of, colorectal cancer (CRC) in patients. The unsupervised hierarchical clustering methodology utilized in the study conducted by Marisa et al.,[1] allowed for informative classifications of colon cancer that would be able to be used for downstream analysis. Our analysis, while modeled on the methodology laid out by Marisa et al.,[1] fell short in two major areas: (1) the full dataset was not incorporated into the analysis due to time constraints and (2) due to a smaller dataset, our calculated data frames for the Welch t-test and chi-square analysis had less robust t-statistic values and p-values.

REFERENCE

1. Marisa L, de Reyniès A, Duval A, Selves J, Gaub MP, Vescovo L, Etienne-Grimaldi MC, Schiappa R, Guenot D, Ayadi M, Kirzin S, Chazal M, Fléjou JF, Benchimol D, Berger A, Lagarde A, Pencreach E, Piard F, Elias D, Parc Y, Olschwang S, Milano G, Laurent-Puig P, Boige V. **Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value**. *PLoS Med*. 2013;**10**(5):e1001453. doi:10.1371/journal.pmed.1001453. Epub 2013 May 21. PMID: 23700391; PMCID: PMC3660251.

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

3. Wang Y, Jatkoe T, Zhang Y, Mutch MG, Talantov D, Jiang J, McLeod HL, Atkins D. **Gene expression profiles and molecular markers to predict recurrence of Dukes' B colon cancer**. *J Clin Oncol.* 2004 May 1;**22**(9):1564-71. doi: 10.1200/JCO.2004.08.186. Epub 2004 Mar 29. PMID: 15051756.

4. Guinney J, Dienstmann R, Wang X, de Reyniès A, Schlicker A, Soneson C, Marisa L, Roepman P, Nyamundanda G, Angelino P, Bot BM, Morris JS, Simon IM, Gerster S, Fessler E, De Sousa E Melo F, Missiaglia E, Ramay H, Barras D, Homicsko K, Maru D, Manyam GC, Broom B, Boige V, Perez-Villamil B, Laderas T, Salazar R, Gray JW, Hanahan D, Tabernero J, Bernards R, Friend SH, Laurent-Puig P, Medema JP, Sadanandam A, Wessels L, Delorenzi M, Kopetz S, Vermeulen L, Tejpar S. **The consensus molecular subtypes of colorectal cancer**. *Nat Med*. 2015 Nov;**21**(11):1350-6. doi:10.1038/nm.3967. Epub 2015 Oct 12. PMID: 26457759; PMCID: PMC4636487.

5. National Human Genome Research Institute (NIH-NHGRI). **The Cancer Genome Atlas (TCGA) homepage**. https://www.genome.gov/Funded-Programs-Projects/Cancer-Genome-Atlas

6. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. **Global cancer statistics, 2012**. *CA Cancer J Clin.* 2015 Mar;**65**(2):87-108. doi: 10.3322/caac.21262. Epub 2015 Feb 4. PMID: 25651787.

7. NCBI GeneID. **LINC00261 long intergenic non-protein coding RNA 261**. https://www.ncbi.nlm.nih.gov/gene/140828 Accessed 02/18/2022.

8. Chakroborty D, Emani MR, Klén R, Böckelman C, Hagström J, Haglund C, Ristimäki A, Lahesmaa R, Elo LL. **L1TD1 - a prognostic marker for colon cancer**. *BMC Cancer*. 2019 Jul 23;**19**(1):727. doi: 10.1186/s12885-019-5952-2. PMID: 31337362; PMCID: PMC6651905.

9. Meng, H., Li, W., Boardman, L.A. *et al.* Loss of ZG16 is associated with molecular and clinicopathological phenotypes of colorectal cancer. *BMC Cancer* **18,** 433 (2018). https://doi.org/10.1186/s12885-018-4337-2

10. Hsu SH, Wang LT, Lee KT, Chen YL, Liu KY, Suen JL, Chai CY, Wang SN. **Proinflammatory homeobox gene, ISX, regulates tumor growth and survival in hepatocellular carcinoma**. *Cancer Res*. 2013 Jan 15;**73**(2):508-18. doi: 10.1158/0008-5472.CAN-12-2795. Epub 2012 Dec 5. PMID: 23221382.

11. Ogawa T, Hirohashi Y, Murai A, Nishidate T, Okita K, Wang L, Ikehara Y, Satoyoshi T, Usui A, Kubo T, Nakastugawa M, Kanaseki T, Tsukahara T, Kutomi G, Furuhata T, Hirata K, Sato N, Mizuguchi T, Takemasa I, Torigoe T. **ST6GALNAC1 plays important roles in enhancing cancer stem phenotypes of colorectal cancer via the Akt pathway**. *Oncotarget*. 2017 Nov 8;**8**(68):112550-112564. doi: 10.18632/oncotarget.22545. PMID: 29348846; PMCID: PMC5762531.

12. Yan T, Tian D, Chen J, Tan Y, Cheng Y, Ye L, Deng G, Liu B, Yuan F, Zhang S, Cai L, Chen Q. **FCGBP Is a Prognostic Biomarker and Associated With Immune Infiltration in Glioma**. *Frontiers in Oncology*. 2022 (**11**) https://doi.org/10.3389/fonc.2021.769033 ISSN:2234-943X

13. Wang K, Guan C, Shang X, Ying X, Mei S, Zhu H, Xia L, Chai Z. **A bioinformatic analysis: the overexpression and clinical significance of FCGBP in ovarian cancer**. *Aging* (Albany NY). 2021 Mar 3;**13**(5):7416-7429. doi: 10.18632/aging.202601. Epub 2021 Mar 3. PMID: 33686968; PMCID: PMC7993703.

14. Li Q, Qin Y, Wei P, Lian P, Li Y, Xu Y, Li X, Li D, Cai S. **Gas1 Inhibits Metastatic and Metabolic Phenotypes in Colorectal Carcinoma**. *Mol Cancer Res*. 2016 Sep;**14**(9):830-40. doi: 10.1158/1541-7786.MCR-16-0032. Epub 2016 Jul 11. PMID: 27401611.

15. Wang WD, Wu GY, Bai KH, Shu LL, Chi PD, He SY, Huang X, Zhang QY, Li L, Wang DW, Dai YJ. **A prognostic stemness biomarker CCDC80 reveals acquired drug resistance and immune infiltration in colorectal cancer**. *Clin Transl Med*. 2020 Oct;**10**(6):e225. doi:10.1002/ctm2.225. PMID: 33135356; PMCID: PMC7603297.

16. Shen H, Zhang L, Zhou J, Chen Z, Yang G, Liao Y, Zhu M. **Epidermal Growth Factor-Containing Fibulin-Like Extracellular Matrix Protein 1 (EFEMP1) Acts as a Potential Diagnostic Biomarker for Prostate Cancer**. *Med Sci Monit*. 2017 Jan 13;**23**:216-222. doi: 10.12659/msm.898809. PMID: 28085790; PMCID: PMC5256367.