BF528 Project 1: Microarray Based Tumor Classification

Authors: Pooja Savla (Data Curator), Vrinda Jethalia (Programmer), Manasa Rupuru (Analyst), Pragya Rawat (Biologist)

Date: 17th February, 2023

Introduction:

Colorectal cancer (CRC) is the third most common type of cancer and the fourth most common cause of death [1] in the world. Researchers have looked into Gene Expression Profiles (GEPs) through the use of microarrays. Unfortunately, no signature has been established as colon cancer (CC) consists of many molecular entities that can develop through different pathways. Large scale studies need to be conducted to identify different subtypes of CC as well as determine a reproducible molecular signature and classification system. In order to identify molecular subgroups of colon cancer based on patterns of gene expression, the authors employed bioinformatic approaches to examine large-scale gene expression data. They grouped colon cancer samples based on similarity in gene expression patterns using unsupervised clustering techniques. The distinct subtypes were then characterized, and important biological pathways and processes linked with each subtype were identified using differential gene expression analysis and pathway enrichment analysis.

In this project, our goal was to reproduce the results obtained from the Marisa et al. study [1] by utilizing a subset of the data. This study established a classification of the CC subtypes based on their molecular features by exploiting "genome-wide mRNA expression analysis" through the use of microarrays. Initially only three subtypes of CC were identified, but through the Marisa et al study, six subtypes were classified, more accurately reflecting the molecular heterogeneity of CC. To confirm their findings, this was also validated against an independent dataset [1].

Data:

The samples used in this study came from a large multicenter cohort of 750 patients diagnosed with Colon Cancer (CC) and were a part of the French national d'Identité des Tumeurs (CIT) program. Each primary tumor tissue sample was collected during surgery between 1987-2007 and was fresh-frozen. Each sample was also accompanied by clinical and pathological data and staged according to the American Joint Committee on Cancer tumor node metastasis (TNM) system. In this paper the experiments used a microarray instrument to measure gene expression in colon cancer tissue samples. Specifically, the authors used the Affymetrix Human Genome U133 Plus 2.0 Array, which is a type of DNA microarray chip. Next RNA Isolation was performed and RNA Quality Control was undertaken to access the quality of the RNA in which only 566 samples were satisfying the stringent quality criteria. These RNA were hybridized on an Affymetrix chip (asterisk) and used for molecular subtype determinations. The discovery set was composed of 443 tumors from the CIT cohort. The validation set was composed of the remaining CIT cohort CC samples, CC samples from seven Affymetrix publicly available datasets (GSE13067, GSE13294, The Memorial Sloan-Kettering Cancer Center (MSKCC) dataset : GSE14333, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE17536.: GSE17536/The National Cancer Institute (NCI) dataset: GSE17537, GSE18088, GSE26682, and GSE33113), and CC samples from the non- Affymetrix TCGA program. For survival analyses, only stage II and III patients were considered, stage I and IV patients not being informative as almost all survive or die, respectively; there were thus 359 cases in the CIT discovery set and 416 in the CIT validation set and three public datasets included in this analysis. The resulting microarray data was analyzed using various bioinformatics tools and algorithms to classify the colon cancer samples into molecular subtypes based on their gene expression profiles. The study did not use sequencing datasets, but instead used microarray technology to measure gene expression levels in colon cancer tissue samples. Hence, there is no average library size or number of reads (single end and paired end) to report[5].

On the Shared Computing Cluster, a single sample, GSM971958, was missing from the final dataset of CEL files. Using the ascension code https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE39582.: GSE39582, this sample was located in the Gene Expression Omnibus (GEO) repository for the study. It was then downloaded and safely moved to the group's repository for "samples" for the project on the shared computing cluster (SCC). The remainder of the CEL files weren't moved to this folder in order to conserve storage space. Instead, symbolic links were made for every file. There were a total of 134 samples available for examination.

Methods:

Programmer:

Data preprocessing and normalization are essential precursor steps to high quality data analysis. These precursor steps ensure that the data is cleaned, bias has been eliminated and features have been scaled. Bioconductor provides a large collection of R packages that are used to preprocess, analyze, and visualize high-throughput genomic data such as microarrays. First, BiocManager (1.30.19) was installed which accesses the Bioconductor Project Package Repository to install other useful packages. The Bioconductor packages used in this project were - affy(1.76.0), affyPLM(1.74.2), hgu133plus2.db(3.13.0), sva(3.46.0), AnnotationDbi (1.60.0).

The input data of 134 samples were in CEL format on GEO. CEL files are binary files that contain raw intensity data from microarray experiments. The CEL files were read in batch using the ReadAffy() function. The function ReadAffy is a wrapper for the functions read.affybatch, tkSample-Names, read.AnnotatedDataFrame, and read.MIAME [4]. Next, the quality of the data was assessed using two commonly used metrics - Relative Log Expression (RLE) and Normalized Unscaled Standard Errors (NUSE). The quality assessment functions operated on PLMset objects, which were obtained through the fitPLM() function. RLE values were calculated by comparing the expression intensity on each array against the median value across all arrays for that probeset. To ensure the quality of the data was sufficient, all data points were centered around 0 (less variance in expression of the genes across arrays) based on the metric. NUSE standardized across arrays such that the median standard error for those genes is 1 across all arrays, which accounted for variability between genes [5]. Histograms were plotted to see the distribution of RLE and NUSE scores across samples.

The normalization step was performed using the rma() function which read in the affybatch data and applied the Robust Multiarray Averaging (RMA) preprocessing algorithm. We chose RMA since it has been shown to produce more accurate and reproducible results than other normalization methods. Moreover, RMA can be combined with quantile normalization, which adjusts the distribution of intensities across all arrays so that they become similar, thereby reducing the impact of systematic variation across the different arrays. The steps in RMA were background correction, normalization, and summarization performed in a modular way. This converted the data into an ExpressionSet object. To account for the inherent variation present in each microarray, it was necessary to normalize the arrays before comparing gene expression across multiple arrays. Failure to appropriately normalize the arrays could lead to misleading results.

Batch effects refer to technical sources of variation that arise from differences in sample processing or measurement between different experimental batches. Statistical methods are applied to the data to diminish the impact of technical variations on biological signals. We employed the ComBat() function from the sva library to correct for batch effects. The rma normalized data was used as the input, along with the metadata file from the Marisa et al. study. The batch covariate was inferred from the 'normalizationcombatbatch' column by combining the Center and RNA Extraction method. The model matrix for the feature of interest was given by the 'normalizationcombatmod' variable which combined the tumor and MMR status features. The result was written out to an ExpressionSet object which was exported as a CSV file for further analysis.

To reduce the data dimensionality, Principal Component Analysis (PCA) was carried out. It made the analysis easier by reducing the dimensions of the data while preserving as much of the data's variation as possible. Since PCA seeks to maximize the variance of each component, the data was scaled and centered to ensure that each gene (or feature) in the analysis is treated equally. This was done by the scale() function. The prcomp() function was used to perform the principal components analysis on the given data matrix. Each principal component generated was accessible through the rotation attribute. The importance attribute of the summary() of the prcomp output provided us with various measures to understand each component. The variance of each PCA was calculated in this manner. All of the code was written in R (4.2.1).

Analyst:

Due to the large number of genes (54,675) and the small number of patients (134) present in this data, it is important to carefully consider what kind of statistical methods are used to analyze this dataset. Univariate statistical tests can not be used on this size of data due to the low sample to genes ratio, nor can multivariate statistical methods be useful unless noise is filtered out. Hence, a series of noise filtering, hierarchical clustering

In this portion, the metrics that Marisa et al. used to filter their data for noise, were used to obtain the genes they used for their downstream analysis. Marisa et al used three filters. The first filter selected for genes from the normalized data set who had at least 20% of their expression values being greater than log 2(15). The second filter selected for genes that have a variance that is significantly different from the median variance of all probe sets. The variance of each probe set across was calculated and it was tested against the median value of all the variances (median value of the set of variances for each probe). The t-test statistic for each probe was then again compared to a chi squared The probes whose p values are greater than 0.01 are selected for. Lastly, the third filter selected for probes that have a robust coefficient of variation(rCV) that is higher than 0.186. First the rCV for each probe set was calculated by taking the standard deviation of the probe and dividing it by its mean. The purpose of this step is to remove the highest and lowest expression values in each of the probe sets. The value of 0.186 was adapted from Marisa et al. In their analysis using the Gaussian mixture modeling algorithm, they identified four groups of rCV values and the lowest was 0.186. Hence, this value was used as the cutoff.

The next portion of the analysis was to perform hierarchical clustering. Using this unsupervised method, the samples were grouped together in this case for subtype discovery that differentiated patients with “C3” subtype from patients who were not of “C3” subtype. This was performed using the hclust() and dist() functions in R. The dist() function calculated distances and generated a distance matrix of the data that passed all three of the noise filters and the output was given as input into the hclust() function, which calculated the actual clustering. The distance matrix was calculated using the Euclidean method and the clustering with hclust() was calculated using the complete method.A heat map of the result can be found in the results section(Fig 3). Lastly, a welch t-test was performed between the clusters to identify the genes that were differentially expressed between the clusters. Genes with a p-adjust value greater than 0.05 were considered as differentially expressed. This t-test analysis was again performed using the same cluster membership with the dataset of genes that were found to pass only the first two noise filters. All of the code for the noise filtering,clustering and statistical testing was written using R code(4.2.1).

Results:

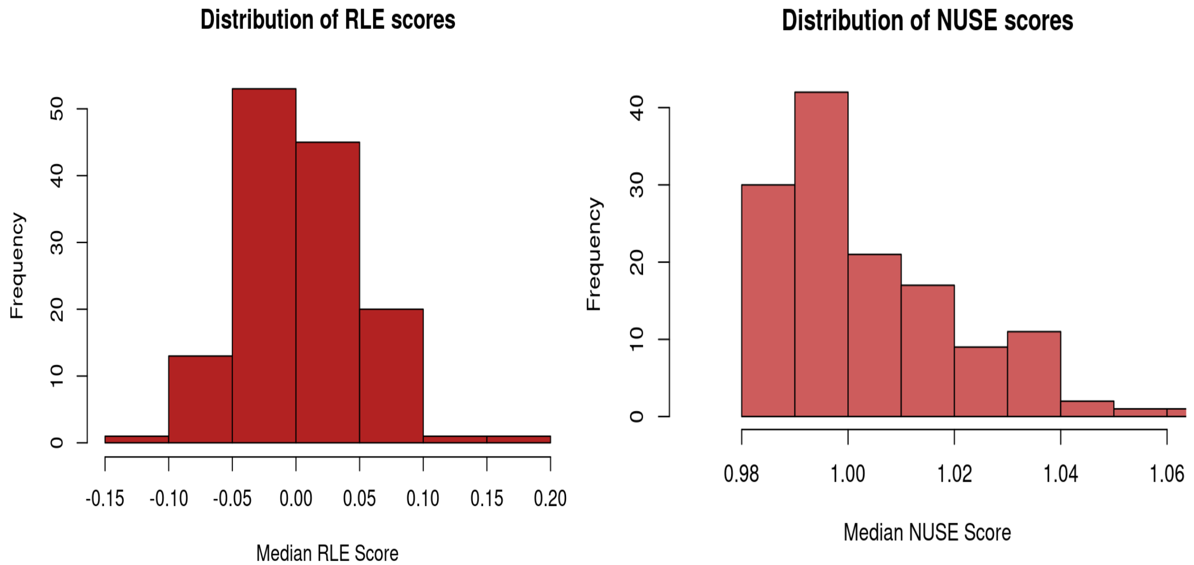

Figure 1: (Left) Histogram displaying the distribution of median RLE scores for 134 CRC samples. (Right) Histogram displaying the distribution of median NUSE scores for 134 CRC samples.

The quality of the data produced using microarrays was assessed by calculating the median RLE and NUSE for each chip. Figure 1, represents the median RLE (left) and median NUSE (right) values of the raw data. The highest frequency of values was associated with 0 in the RLE plot and 1 in the NUSE plot, which is in accordance with the nature of these quality assessment metrics. Through the distributions of these values we could imply that the quality of the samples was high and satisfactory enough for the rest of the analysis to be carried out. There were 2 samples with medians greater than 0.10 in the RLE plot (GSM971993_JS_71_U133_2, GSM972390_VB_156T_U133_2) and 2 samples with medians greater than 1.05 in the NUSE plot (GSM972113_070123.15, GSM972269_AD_436_U133_2), however, they don't have much deviation from the expected values, therefore, it does not justify sample removal. All samples were included in the subsequent analysis.

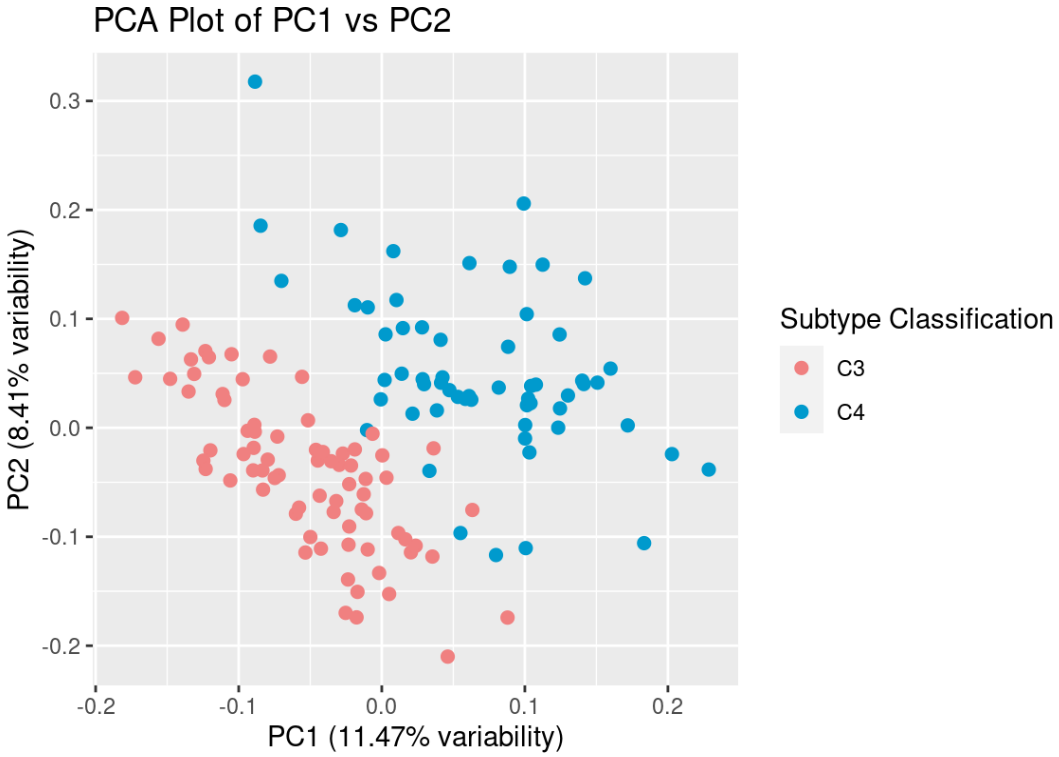

Figure 2: PCA plot of the first and second principal components across CRC subtypes

From figure 2, we can conclude that the first two principal components explain approximately 20% of variance of the data. Two clusters are clearly seen distinguishing the C3 and C4 subtypes of CRC with very few outliers. Based on this PCA plot, we can infer that the gene expression patterns were distinct in C3 and C4 cancer subtypes.

Referring to the noise filtering of the normalized data steps, our analysis found that 39,661 out of the initial 54,675 probe sets had at least 20% of their expression values being greater than log 2(15). 37,517 of probes were found to have variance that is significantly different from the median variance of all probe sets. And lastly, 1,027 probe sets passed all three filters. Once hierarchical clustering was performed on this data, 53 of the samples clustered together and the remaining 81 clustered together. By merging the metadata with this dataset and labeling of the heat map, the analysis showed that the 53 samples were typed as C3 and the remaining 81 as C4.

From the Welch t-test it was found that 712 genes were differential expressed between the clusters. These 712 genes had adjusted p values less than 0.05. The genes with the lowest p values are considered to be the most differences among the clusters. The results of the welch t-test from the data of 37,517 probes (ones that only passed filters one and two) showed 5,758 differentially expressed genes.

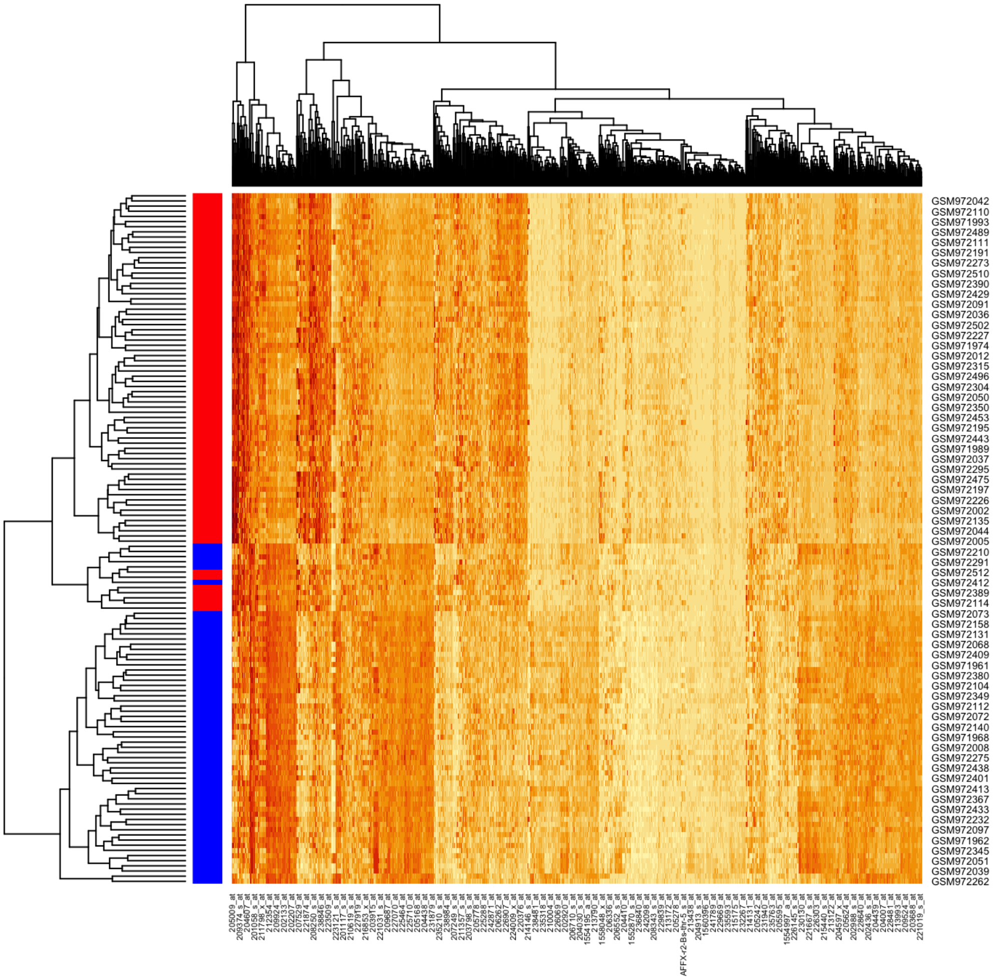

Figure 3: The heat map is a graphical representation of the expression levels of the probe sets for each of the samples. The red labels are representing samples with the C3 subtype and samples with the C4 subtype are in blue.

PROBEID	SYMBOL	p_value	t_statistics	p_adjusted
223121_s_at	SFRP2	1.67E-43	22.44437	6.27E-39
227059_at	GPC6	1.24E-44	21.65425	4.66E-40
213413_at	STON1	3.97E-42	21.31285	1.49E-37
209356_x_at	EFEMP2	4.49E-43	21.19881	1.68E-38
225242_s_at	CCDC80	1.59E-43	21.10503	5.97E-39
203748_x_at	RBMS1	1.84E-43	20.85641	6.91E-39
225464_at	FRMD6	3.93E-42	20.25167	1.47E-37
221019_s_at	COLEC12	1.24E-34	20.14339	4.63E-30
202363_at	SPOCK1	3.98E-41	19.89567	1.49E-36
218694_at	ARMCX1	3.08E-40	19.82895	1.15E-35

Table1: Top 10 up-regulated genes with p-val<0.05 and |t-statistic|>0

PROBEID	SYMBOL	p_value	t_statistics	p_adjusted
203240_at	FCGBP	3.31E-24	-13.31883	1.23E-19
220622_at	LRRC31	8.05E-26	-13.30955	2.99E-21
227725_at	ST6GALNAC1	5.12E-22	-13.13357	1.89E-17
234008_s_at	CES3	2.84E-24	-12.55303	1.06E-19
1553828_at	NXPE1	1.39E-23	-12.5189	5.16E-19
1568598_at	KAZALD1	1.96E-22	-12.0364	7.27E-18
218189_s_at	NANS	9.15E-22	-11.90187	3.38E-17
204130_at	HSD11B2	1.53E-21	-11.90076	5.65E-17
205259_at	NR3C2	2.96E-20	-11.89407	1.09E-15
222764_at	ASRGL1	7.15E-21	-11.57894	2.64E-16

Table2: Top 10 down-regulated genes with p-val<0.05 and |t-statistic|>0

Table 1 and 2 were obtained by using the bioconductor package ‘hgu133plus2.db’ to map probeset IDs to gene symbols. It was assumed that the difference in median expression values is not significant to deal with probe IDs mapping to the same gene symbols. Otherwise further investigation to determine the cause of the difference, such as dynamic range, specificity, and sensitivity of the probes would have to be conducted. For that reason one probe was chosen randomly among the matches in this case.

Set	p-value	adjusted p-value
KEGG_N_GLYCAN_BIOSYNTHESIS	0.0004278119396	0.0004278119
KEGG_OTHER_GLYCAN_DEGRADATION	0.0004278119396	0.0004278119
KEGG_O_GLYCAN_BIOSYNTHESIS	0.0004278119396	0.0004278119
HALLMARK_INTERFERON_ALPHA_RESPONSE	0.0001191179211	1.92E-04
HALLMARK_ALLOGRAFT_REJECTION	0.0001475109442	2.30E-04
HALLMARK_PROTEIN_SECRETION	0.0001718539169	2.60E-04
GOBP_PROTEIN_EXPORT_FROM_NUCLEUS	0.000102283608	0.0004721229
GOBP_POSITIVE_REGULATION_OF_NUCLEAR_DIVISION	0.000102283608	0.0004721229
GOMF_MOLECULAR_FUNCTION_ACTIVATOR_ACTIVITY	0.000102283608	0.0004721229

Table3: Top 3 enriched upregulated KEGG, Hallmark, and GO gene sets with p-value and adjusted p-values link

Set	p-value	Adjusted p-value
KEGG_ARRHYTHMOGENIC_RIGHT_VENTRICULAR_CARDIOMYOPATHY_ARVC	0.0001026753163	2.17E-04
KEGG_GLYCEROPHOSPHOLIPID_METABOLISM	0.0001576278517	3.29E-04
KEGG_PPAR_SIGNALING_PATHWAY	0.0002137267011	4.42E-04

HALLMARK_REACTIVE_OXYGEN_SPECIES_PATHWAY	0.000177670545	2.07E-04
HALLMARK_PANCREAS_BETA_CELLS	0.0002951199577	3.35E-04
HALLMARK_WNT_BETA_CATENIN_SIGNALING	0.0009948799742	1.11E-03
GOBP_PROTEIN_EXPORT_FROM_NUCLEUS	0.000102283608	0.000378625
GOBP_ACTIN_FILAMENT_DEPOLYMERIZATION	0.000102283608	0.000378625
GOBP_RUFFLE_ORGANIZATION	0.000102283608	0.000378625

Table4: Top 3 enriched downregulated KEGG, Hallmark, and GO gene sets with p-value and adjusted p-values

Gene sets were downloaded from the human collection on Molecular Signatures Database or MSigDB. Gene sets used for the purpose of this paper were; 50 Hallmark gene sets - gene sets that summarize and represent specific well-defined biological states or processes and display coherent expression (MSigDB). The 186 KEGG gene sets contain canonical pathway gene sets that are obtained from the KEGG pathway database. And 10561 GO gene sets were derived from gene ontology. These sets represent GO terms belonging to one of each of the three root GO ontologies - biological process, cellular component, and molecular function respectively.

From our analysis, KEGG cell communication pathways is one of the top 3 enriched gene sets - this matches the results from the paper. For instance, the KEGG glycan biosynthesis gene set shows high enrichment. Glycans are known to have many protective, stabilizing, organizational, and barrier functions - similar to the top enriched gene sets included in the paper such as KEGG Focal adhesion gene set, KEGG Tight junction, and KEGG Gap junction.

Discussion

By means of microarray data normalization and quality control measures in reference to Marisa et al. study [5], the PCA plot was used to analyze and visualize outliers in the data sets as well as clustering. NUSE and RLE scores showcased the quality of the median distribution indicating that the samples were good and there was no requirement for discarding outliers. The PCA plot of the first two principal components explained only about 20% variance, however, this can be justified due to the low number of sample points which makes it difficult to obtain proper inference. The plot does show two clusters of C3 and C4 cancer subtypes with few outliers (mostly C4 is more spread out).

The total number of genes that we had from the data that were remaining after all the filter stages were a total of 1,027 genes. However, at this step Marisa et al. resulted in 1,459 genes. This means that there was an error in at least one of the filtering steps. The most likely step in which an error was made could have been at how the second filter was written. The code for this portion was not written out explicitly as its description but rather used a R function called varTest(). It functions to calculate the variance and perform a one sample chi squared test on the variance all together. It was implemented in this analysis with the understanding that it was performing the same statistical calculations as instructed but there is possibility that the calculations performed by varTest() could slightly be different from what was expected. Additionally, there is evidence in Figure 3 that there was some error in the hierarchical clustering step potentially. A source of this error can be either how the dist() or hclust() functions were used. For both of the functions, the defaults were used to generate this data. As can be seen in the heat map, not all the samples of one color clustered together. Perhaps by passing a different method into the dist() or hclus() methods could have resulted in a more distinctly clustered heat map.

Conclusion

The paper does not report any sources of error or contamination that were detected in the study. However, the authors note that the use of different microarray platforms, sample collection and processing methods, and data normalization methods across the different datasets can introduce variability and affect the accuracy of the classification models.

Overall, the authors took several measures to ensure that the data was of high quality, including conducting QC checks on the microarray data and clinical data and filtering out low-quality samples. While the paper does not report any sources of error or contamination, the authors note that the use of different microarray platforms and data normalization methods can introduce variability in the data.

References

1. Mármol, I., Sánchez-de-Diego, C., Pradilla Dieste, A., Cerrada, E., & Rodriguez Yoldi, M. J. (2017). Colorectal carcinoma: a general overview and future perspectives in colorectal cancer. International journal of molecular sciences, 18(1), 197.
2. Maguire, A., & Sheahan, K. (2014). Controversies in the pathological assessment of colorectal cancer. World journal of gastroenterology: WJG, 20(29), 9850.
3. Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. 2004. affy---analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20, 3 (Feb. 2004), 307-315.
4. Bolstad, BM (2004) Low Level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization. Dissertation. University of California, Berkeley.
5. 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.