Individual Project: Microarray Based Tumor Classification

Author: Italo Duran (Programmer & Biologist)

Introduction

Colorectal cancer is the third most common type of cancer and the fourth most common cause of death [6] globally. Pathological staining is the most widely used method to determine the presence of CC; however, it does not accurately predict the recurrence of CRC [7]. The goal of this project was to reproduce the programmer and biologist results from the comparison of the C3 and C4 tumor subtypes from Marisa et al. data. [1] This study established a classification of the CC subtypes based on their molecular features by exploiting “genome-wide mRNA expression analysis” using microarray analysis. 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 CEL data files will be used for the programmer part to create comma-separated files. Containing the RMA normalized, ComBat adjusted gene expression values, a histogram of median RLE scores, a histogram of median NUSE scores, and a plot of PC1 vs. PC2 with the percent variability attributed to these principal components shown on the x and y axes labels. For the biologist part, instead of re-creating the whole analyst part for one sample data, the sample data provided in the shared computing cluster, in "~/differential_expression_results.csv"; the gene expression statistics will be used, comparable to the file obtained in 5.6.

Methods

To perform the normalization and quality control of a large dataset of microarray samples and the employment of noise filtering techniques to reduce data dimensionality, Rstudio (version 1.4, R version 4.0.2) was used, with two significant repositories: CRAN and BiocManager. The packages used for this analysis were: affy, affyPLM, sva, AnnotationDbi, and hgu133plus2.db.

The CEL files were read using the ReadAffy function, then using the robust multiarray analysis (RMA) function to normalize all the CEL files together. Using the Bioconductor package affyPLM, Relative Log Expression (RLE) and Normalized Unscaled Standard Error (NUSE) scores of the microarray samples were computed. All points in the data were centered around zero based on the metrics; to suffice the quality of the data. NUSE has standardized across arrays such that the median standard error for those genes is one across arrays [4].

The ComBat function from the SVA library was used to correct for batch effects [5]. The RMA normalized data was provided as the input and the metadata file from the Marisa et al. study. The 'normalizationcombatbatch' variable gave the model matrix for the output of interest. The result was written out to a set object and was exported as a CSV file.

Principal Component Analysis was executed to reduce the data dimensionality. Using the scale function helps preserve as much of the data's variation as possible. The 'prcomp' function was used to perform the PCA on the data matrix. Each principal component generated was accessible through the rotation attribute. Once prcomp has been executed, the values for each of the principal components can be viewed by accessing the $rotation attribute of the prcomp object. The critical attribute of the summary function of the prcomp output provided us with various measures to understand each component.

For the biologist role, the gene set collections were obtained from MSigDB and accessed via Bioconductor package GSEABase. The top ten upregulated and downregulated differentially expressed probe sets; Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Hallmark gene sets. They were analyzed to understand the biological significance of the different gene expressions. By using thehgu133plus2.db. library and function; the differentially expressed probe sets with the same matching gene symbol; were selected based on T-test statistics and mapped. To discover the most differentially expressed probe sets. A Fisher Exact test was performed on the filtered probe sets. Finding they contain 50, 186, and 10402 genes, respectively. A comparison between upregulated and downregulated genes of 1000 probes was conducted. Later, a finer comparison was made, and significantly enriched sets of $p < 0.05$ were counted, and the top 3 probe sets were taken from each comparison for reflection to Marisa et al. [1].

Results

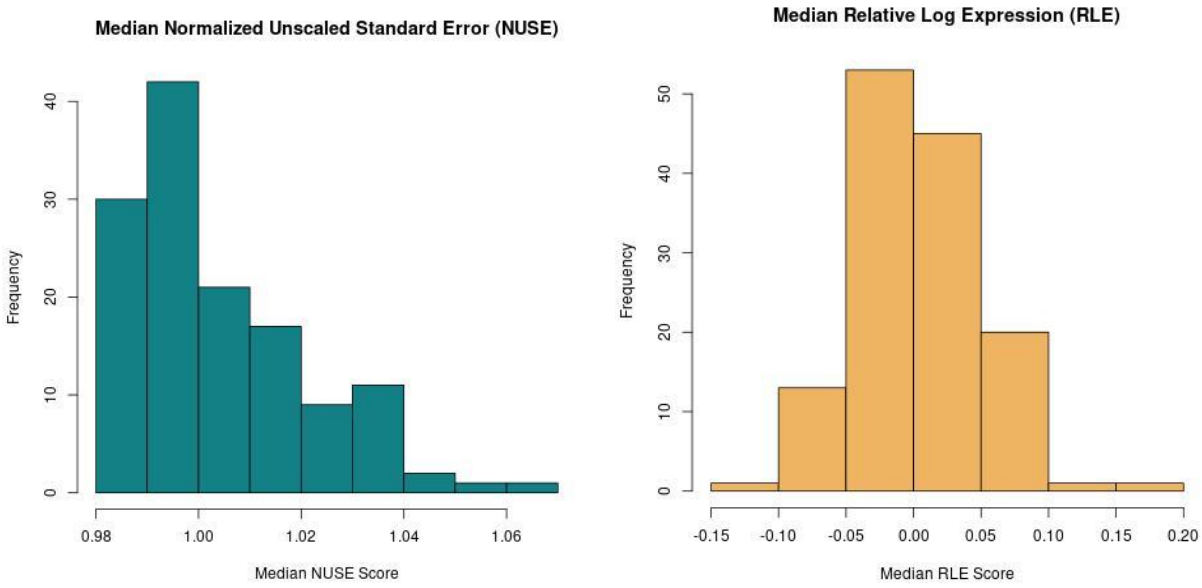

Fig 1: *(Left) Histogram of median NUSE scores for 134 CC samples. (Right) Histogram of median RLE scores for 134 CC samples.*

The quality of the data collected through the microarrays was assessed by calculating the median RLE and NUSE for each chip. Figure 1 represents the raw data's median RLE (right) and median NUSE (left) values. The highest frequency of values belonged to 0 in the RLE plot and 1 in the NUSE plot, which fell into accordance with the nature of the quality assessment metrics. The distributions of these values indicated the quality of the samples was high and sufficient for the rest of the analysis to be carried out. There were 2 samples above 0.10 in the median RLE plot (GSM971993_JS_71_U133_2, GSM972390_VB_156T_U133_2) and 2 samples above 1.05 in the median NUSE plot (GSM972113_070123.15, GSM972269_AD_436_U133_2), however, they were not drastically different enough to justify removal. All samples were included in the subsequent analysis.

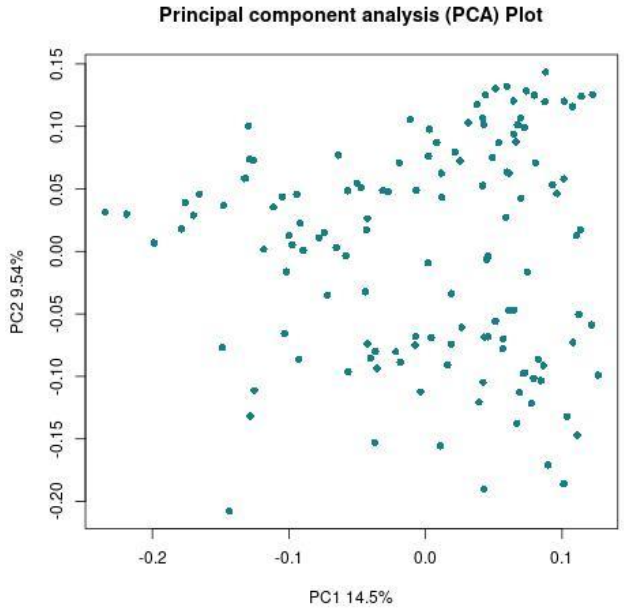

Fig 2: *PCA plot of the first and second principal components across the CC subtypes.*

Figure 2 shows the first and second principal components (PCAs) plotted against each other. The first two components captured roughly 20% of the data variance. The graph shows that the C3 and C4 cancer subtypes separated distinctly into two different clusters with a few outliers. Based on the plot, it can be concluded that the gene expression patterns were distinct in subtypes C3 and C4.

+ Up	PROBEID	SYMBOL	t	p	padj
1	223122_s_at	SFRP2	23.3067212	1.35E-48	3.08E-44
2	207266_x_at	RBMS1	22.6544687	2.57E-47	2.94E-43
3	204457_s_at	GAS1	22.167177	6.43E-45	2.94E-41
4	225242_s_at	CCDC80	21.2792505	2.01E-43	5.90E-40
5	213413_at	STON1	21.035535	3.83E-40	4.39E-37
6	202363_at	SPOCK1	20.9774484	3.86E-43	8.85E-40
7	226930_at	FNDC1	20.9556456	2.55E-43	6.48E-40
8	202291_s_at	MGP	20.9040764	2.06E-43	5.90E-40
9	227059_at	GPC6	20.8854379	1.14E-42	2.38E-39
10	219778_at	ZFPM2	20.5966764	1.42E-35	3.11E-33

Table 1: *Top 10 upregulated genes. The results of matching gene symbols to their probeIDs, sorted by t-statistic and adjusted p-value. Showing each probe ID to its corresponding t-statistic, p-value, adjusted p-value, and gene symbol.*

-Down	PROBEID	SYMBOL	t	p	padj
1	234008_s_at	CES3	-12.588712	2.43E-24	8.83E-23
2	235350_at	C4orf19	-12.608365	1.75E-22	5.03E-21
3	222764_at	ASRGL1	-12.60942	1.62E-23	5.37E-22
4	218189_s_at	NANS	-12.687304	5.39E-24	1.88E-22
5	205489_at	CRYM	-12.803856	1.15E-24	4.31E-23
6	214106_s_at	GMDS	-12.80637	1.14E-21	3.00E-20
7	227725_at	ST6GALNAC1	-13.137294	1.15E-22	3.40E-21
8	211715_s_at	BDH1	-13.417874	2.63E-25	1.07E-23
9	220622_at	LRRC31	-13.543142	1.54E-26	7.18E-25
10	203240_at	FCGBP	-13.788116	2.69E-25	1.09E-23

Table 2: *Top 10 downregulated genes. The results of matching gene symbols to their probeIDs, sorted by t-statistic and adjusted p-value. Showing each probe ID to its corresponding t-statistic, p-value, adjusted p-value, and gene symbol.*

Gene Set Database	Number of Gene Sets
Hallmark (h.all.v7.5.1.symbols.gmt.txt)	50
Kegg (c2.cp.kegg.v7.5.1.symbols.gmt.txt)	186
Go (c5.go.v7.5.1.symbols.gmt.txt)	10402

Table 3: *description of the gene set databases used, specifying the number of gene sets considered in each gene database set. The gene set database was obtained from MSigDB and accessed via Bioconductor package GSEABase.*

	Setname	p-value	estimate	exp	BH
1	KEGG_ECM_RECEPTOR_INTERACTION	1.32E-15	9.11616124686	UP	4.91E-13
2	KEGG_FOCAL_ADHESION	4.57E-14	4.50646540604	UP	8.50E-12
3	KEGG_COMPLEMENT_AND_COAGULATION_CASCADES	2.66E-10	8.17121869099	UP	3.30E-08
4	GOCC_COLLAGEN_CONTAINING_EXTRACELLULAR_MATRIX	4.24E-57	8.67909235374	UP	8.83E-53
5	GOCC_EXTERNAL_ENCAPSULATING_STRUCTURE	4.13E-56	7.18532185883	UP	4.30 E-52
6	GOMF_EXTRACELLULAR_MATRIX_STRUCTURAL_CONSTITUENT	8.41E-42	13.7905967075	UP	5.83E-38
7	HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	1.09E-70	16.5471636797	UP	1.09E-68
8	HALLMARK_MYOGENESIS	1.38E-13	4.63158217711	UP	6.91E-12
9	HALLMARK_UV_RESPONSE_DN	6.96E-12	4.49134265521	UP	2.32E-10

Table 4: *Top 3 Enriched KEGG (yellow), GO (green), and Hallmark (red) Gene Sets. Fisher test performed on all three gene set collection. The results showing each's probes p-value, adjusted p-values, exp, Benjamin & Hochberg p-values.*

SFRP2, RBMS1, and GAS1 were the top three most upregulated enriched probes. They share proteins that control the Wnt signaling, DNA and RNA binding protein, and cellular growth in this list. These proteins encode for RNA Binding Motif Single-Stranded Interacting Protein. For CES3, C4orf19 and ASRGL1 were the top three most downregulated enriched probes. They share proteins that participate in the colon and neural metabolism, a protein-coding gene, and a protein that could be involved in the production of L-aspartate, which can behave as an excitatory neurotransmitter in some brain regions. There were parallel similarities to the original study, from relative upregulated pathways found in the gene set enrichment analysis. As in epithelial-mesenchymal and myogenesis found in the hallmark pathway. As well as focal adhesion and ECM receptor interaction in the KEGG pathway. The errors encountered were that no gene probes had a significant p-value compared to the original study by Marissa et al. [1]. One major component that could significantly differ in error is the partial data analysis from this project, unlike the in-depth analysis with more sample data from Marissa et al. [1].

Discussion

This project focuses on the analyses to reproduce the programmer and biologist results from comparing the C3 and C4 tumor subtypes from Marisa et al. data. [1] for the 134 samples by using normalizing microarray data and computing quality control measures. Cutoffs were used to reduce the dimensionality of the data set, and PCA plots were used to visualize outliers in the data sets by analyzing the clusters. Relative Log Expression (RLE) and Normalized Unscaled Standard Error (NUSE) scores demonstrated the quality of the median distribution. This comparative analysis of C3 and C4 tumor subtypes detected a standard gene set. This analysis allows us to characterize the different gene expression patterns of colon cancer tumors and gain insight into developing potential therapeutic targets for each different subtype. One note to consider is the population metrics since they mainly consisted of a particular cohort and were not diverse enough to consider a diverse cohort study sample that could be more accurate for the study being taken place.

Conclusion

It was an outstanding replication of some of the results of the Marissa et al. [1]. The analysis demonstrated that C3 and C4 subtypes of colon cancer have distinct differences in their gene expression profiles. Thus, it is possible to cluster them separately using principal component analysis. These results establish the belief that colorectal cancer has molecular heterogeneity, and each subtype has a unique gene signature reflected in its gene expression. This categorization provides a means for a more straightforward diagnosis of colorectal cancer subtypes. An insight that was seen was the discovery of many differentially expressed genes between the two clusters. This observation could potentially be that the genes could be targets for treatment and biomarkers for classification. This could benefit the treatment or even provide a means for preventative care in high-risk patients.

References

1. Marisa et al. Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value. PLoS Medicine, May 2013. PMID: 23700391
2. 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.
3. Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., & Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics, 4(2), 249-264.
4. Bolstad, BM (2004) Low Level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization. Dissertation. University of California, Berkeley.
5. W.E. Johnson, C. Li, and A. Rabinovic. Adjusting batch effects in microarray data using empirical bayes methods. Biostatistics, 8(1):118–127, 2007
6. 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.
7. Maguire, A., & Sheahan, K. (2014). Controversies in the pathological assessment of colorectal cancer. World journal of gastroenterology: WJG, 20(29), 9850.