Project 1: Microarray Based Tumor Classification

Authors: Preshita Dave (Programer), Italo Duran (Analyst), Monica Roberts (Data Curator)

Introduction

Colorectal cancer (CRC) is the third most common type of cancer and the fourth most common cause of death [6] in the world. Pathological staining is the most widely used method to determine the presence of CC, however, it does not accurately predict recurrence of CRC [7]. About 20% of patients who are diagnosed with stage II or III CRC develop recurrence. 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 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. After RNA quality control criteria was applied, only 566 samples were kept in the study. These RNA underwent hybridization to Affymetrix U133 Plus 2.0 chips where their DNA alterations were characterized. The samples were split into discovery (n=443) and validation (n=123) groups. Also included in the validation group was 906 samples from seven publicly available datasets (GSE13067, GSE13294, GSE14333, GSE17536/17537, GSE18088, GSE26682, and GSE33113) that had also used Affymetrix U133 Plus 2.0 chips. The total size of the validation group was 1,029 samples. Patients with stage II or III CC and relapse-free survival (n = 359 for discovery and n = 416 for validation) were used in survival analysis. Lastly, despite the use of a non-Affymetrix chip, a dataset of 152 CC samples from The Cancer Genome Atlas (TCGA) was included in the validation set and analyzed separately due to its available DNA alteration annotations [1].

The final dataset of CEL files was stored in a central location on the Shared Computing Cluster with one sample missing, GSM971958. This sample was located in the paper’s repository on Gene Expression Omnibus (GEO) using the ascension number GSE39582. It was then downloaded and securely transferred to the team's 'samples' repository for the project on the Shared Computing Cluster (SCC). In order to save storage space, the rest of the CEL files were not transferred to this folder. Instead, symbolic links were created for each file. 134 samples in total were available for analysis.

Methods

Essential packages were downloaded with the help of the BiocManager package, which installs and manages other packages. The packages in Bioconductor that were used were: affy(1.72.0), affyPLM(1.70.0), sva(3.42.0), AnnotationDbi(1.56.2), hgu133plus2.db(3.13.0) and ggplot2(3.3.5). The input data format was a CEL file, a data file created by the Affymetrix DNA microarray image analysis program. It stores the expression levels of each probe and whether it was a perfect match (PM) or a mismatch (MM). To read in the CEL files, the ReadAffy() function came in handy to store the data of the CEL files as a batch. The function ReadAffy is a wrapper for the functions read.affybatch, tkSample-Names, read.AnnotatedDataFrame, and read.MIAME [2].

Before any preprocessing, we assessed the quality of the data. This was done through two performance 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 [4].

The rma() function was used to read in the AffyBatch data and perform the robust multiarray analysis (RMA) preprocessing algorithm. This converted the data into an ExpressionSet object. RMA performed background correction, normalization, and summarization in a modular way. Normalizing the arrays was necessary because each array displayed its own variation and if not appropriated properly, could have given misleading results when comparing across multiple arrays. Quantile normalization was employed here as well, which normalized the distribution of intensities across arrays as the same [3].

The ComBat() function from the sva library was employed to correct for batch effects. This function used a Bayesian framework since the batch covariate was known [5]. The rma normalized data was provided 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 Central and the RNA Extraction method. The model matrix for the outcome of interest was given by the 'normalizationcombatbatch' variable which combined the tumor and MMR status features. The result was written out to an Expression Set object and was exported as a csv file for further analysis.

Principal Component Analysis (PCA) was carried out to reduce the data dimensionality. This helped in the analysis 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 within each gene to understand its contribution to the variance. 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 determined this way.

To filter noise in the data and perform dimensionality reduction, three filters were applied to ComBat normalized data based on the Marisa et al. study [1]. Genes with higher than log2(15) expression values in at least 20% of samples were filtered to obtain the genes that were commonly expressed. The second filter applied was the two-tailed chi- square test. This ensured that there was a significant difference between a gene's variance and an overall probe's median variance. Its purpose was to discard genes that had low variance in expression whose threshold was lower than $p<0.01$. For the third filter, only genes that had a coefficient of variation of 0.186 were retained. This was done by calculating a high coefficient of variation across every probe set in all the samples. The formula used was CV= $\boldsymbol{\sigma}/\boldsymbol{\mu}$. $\boldsymbol{\sigma}$ was the standard deviation and $\boldsymbol{\mu}$ was the mean.

Hierarchical clustering and subtype discovery was performed on the filtered dataset to create cluster classification between C3 and C4 tumor subtypes. The hclust function, which measured the euclidean distance, was used to create the two clusters. Then, the clusters were cut using the function cutree. The samples were separated into the two subtype clusters and a Welch t-test was performed with the t.test function. The p-values for the test were adjusted using FDR corrections with the p.adjust function. Then, a heat map of the cluster data and expression levels was generated with the function heatmap.2. C3 tumor subtypes were color-coded red and C4 tumor subtypes were color coded blue. The intensity and color change from yellow to red was shown in the color key in the map. Red signified a gene being highly downregulated and yellow signified being highly upregulated.

All of the data, procedures, analyses, and figures were done using Rstudio via SCC by Boston University. Github was used to maintain the data processing and control.

Results

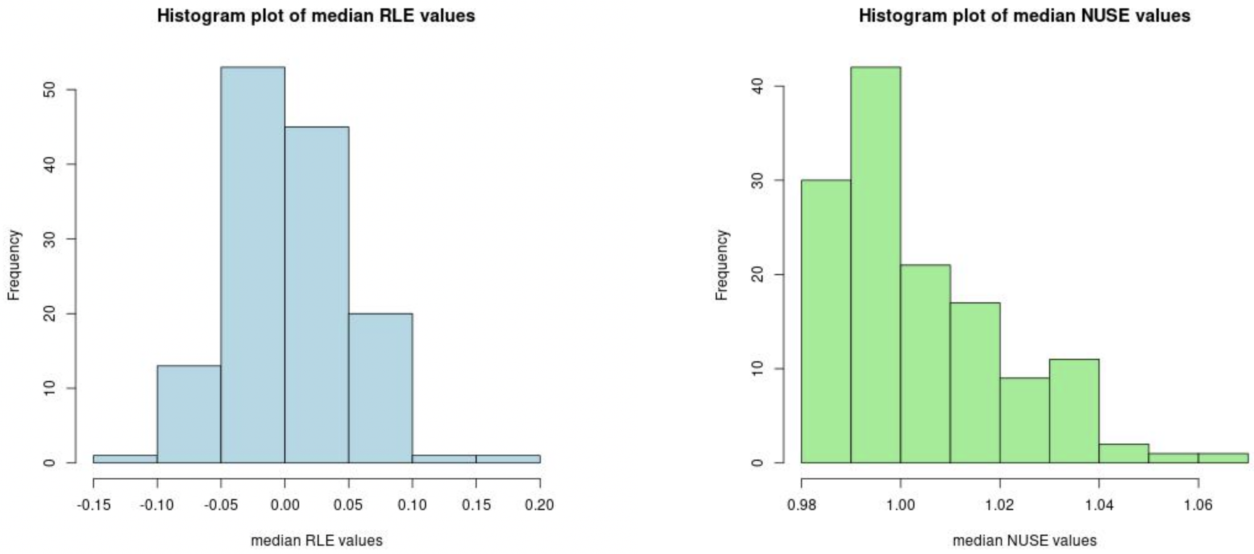

Fig 1: (Left) Histogram of median RLE scores for 134 CC samples. (Right) Histogram of median NUSE scores for 134 CC samples.

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 median RLE (left) and median NUSE (right) values of the raw data. 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 enough 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.

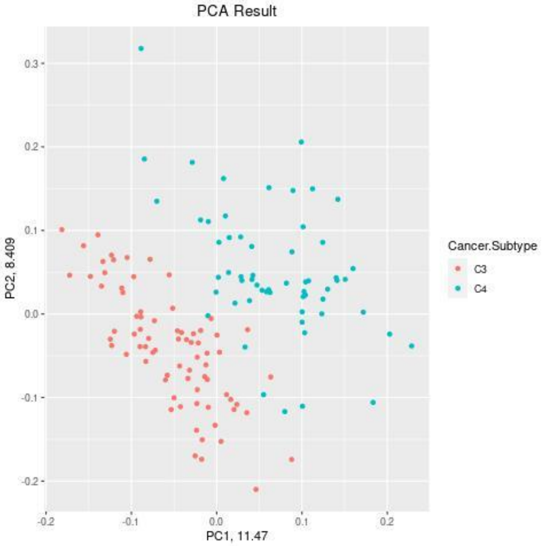

Fig 2: PCA plot of the first and second principal components across the CC subtypes.

Figure 2 shows the first and second principal components plotted against each other. The first two components captured roughly 20% of the variance of the data. As exemplified in the plot, 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.

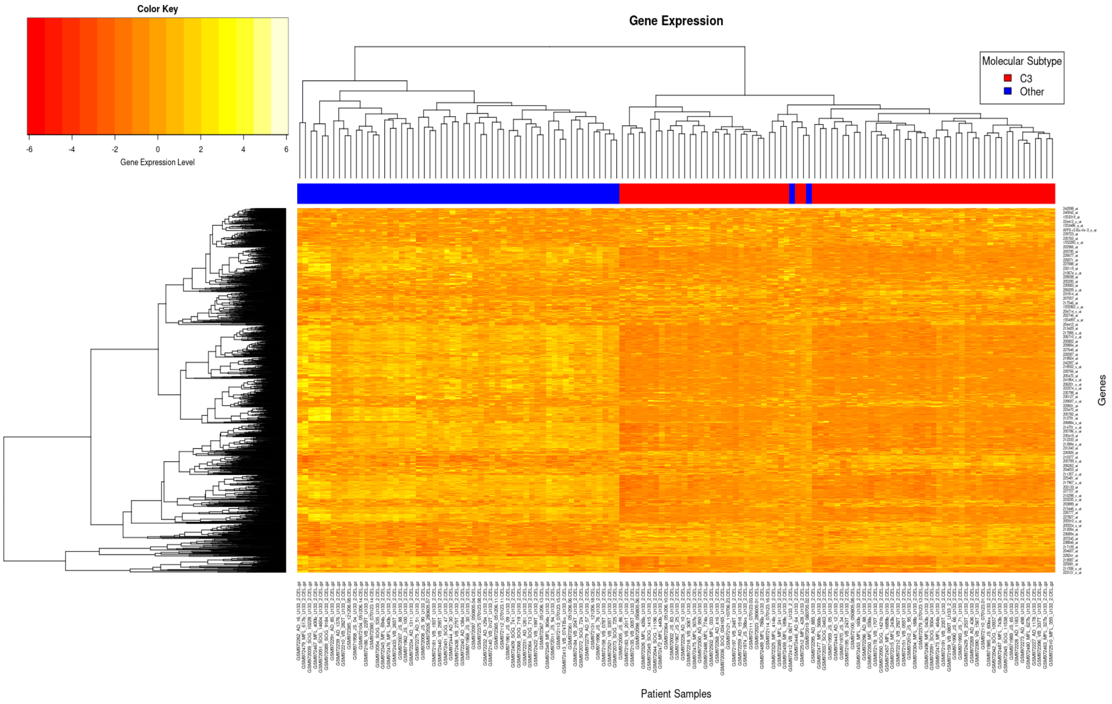

Figure 3. Heatmap of gene expression of the 1531 genes (rows) against the 134 samples (columns). The results show two different cluster groups and their expression levels. C3 shown in red with 77 samples and C4 subtypes shown in blue with 57 samples. The color key represents the gene expression levels with lighter (white/yellow) indicating low expression and darker (red) indicating higher expression.

After noise filtering and dimensionality reduction was conducted, results showed 1531 probes passed all the filters. After running the test statistics and chi square test, 29,645 probes passed the test. Hierarchical clustering and subtype discovery was then performed on the patient samples. The total number of samples was 13. 57 were in cluster one and 77 were in cluster two. The number of differentially expressed genes at $p < 0.05$ between the two clusters was 1,236. The top genes with the largest difference in the clusters were: 204457_s_at, 225242_s_at, 209868_s_at, 218694_s_at, 223122_s_at, 227059_s_at. For the t-test the expression matrix for probe-sets that passed the expression threshold, the number of samples in cluster 1 was 55 and for cluster 2 was 79. The number of differentially expressed genes at $p < 0.05$ between the two clusters was 17,350. The top genes with the largest difference in the clusters were 204457_s_at, 213413_at, 223121_s_at, 223122_s_at, 209356_x_at, 207266_x_at, as presented in figure 3.

Discussion

This project focuses on the analyses to reproduce a comparison of C3 and C4 tumor subtypes for the given 134 samples. By means of normalizing microarray data and compute quality control measures in reference to Marisa et al. study [1].Cutoffs were used to reduce that data set, PCA plots were used to analyze and visualize outliers in the data sets as well as clustering. NUSE and RLE scores demonstrated the quality of the median distribution. To test the quality of these medians, we applied noise reduction and dimensionality reduction, by three filters, to reduce the overall noise in the data. As well as Hierarchical clustering, to divide our reduced feature set into two subtypes clusters. This comparative analysis in both C3 and C4 tumor subtypes resulted in the detection of a common gene set. Four hundred thirty-four genes were expressed distinctively discovered between the C3 and C4 tumor subtypes. This analysis allows us to differentiate different gene expression patterns of colon cancer tumors and in time to develop potential therapeutic targets for each subtype.

Another thing to think about is that the population studied mainly consist of French nationals and hence, doesn't represent a diverse population. It is possible that the gene expression studies are influenced by a certain set of environmental and physical factors which may include diet, weather conditions et cetera. Although validated with external public datasets, more studies should be conducted on a diverse population to verify if the same findings hold true with greater confidence.

Conclusion

Overall, we were able to successfully replicate some of the results of the Marissa et al. study. Our analysis clearly showed the C3 and C4 subtypes of colon cancer have distinct differences in their gene expression profiles that make it possible to cluster them separately using PCA. These results confirm the belief that CC has molecular heterogeneity and each subtype has a distinct gene signature that is reflected in its gene expression. This classification provides a means for easier subtype diagnosis of CC. In addition, we found a large number of differentially expressed genes between the two clusters. These genes could be potential targets for treatment as well as biomarkers for classification.

Areas for further exploration would involve the pathways of our identified gene targets. Understanding and identifying the molecular function of these genes would provide greater insight into the pathogenesis of C3 and C4 colon cancer. This insight could aid in treatment or even provide a means for preventative care in susceptible patients.

The greatest challenge of the project was making sense of and connecting the different components of the analysis while also collaborating with other members of the group. It was difficult at first to correctly carry out the functions of a role without understanding the other roles on the team as well. In the future, this could be improved by more discussion amongst team members.

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.