Introduction

In their 2013 paper, “Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value,” Marisa et. al. sought to better understand color cancer (CC) due to its very common nature amongst men and women. Until this point, colon cancer prognosis and treatment was determined using pathological stage, a classification that was often a poor predictor of treatment outcome and relapse. Previous studies had explored gene expression as a potential predictor of prognosis, but none had yielded conclusive results. In this study, the authors conducted a large multi-center study, using a cohort of 750 CC patients, collecting their genetic information and clinical data. They selected 566 suitable tumor samples to develop a microarray dataset and train a classifier. Using these samples, along with data pulled from public databases, they performed a differential gene expression analysis using microarray technology. They identified six molecular subtypes that correspond with biological and molecular phenotypes. In addition to contributing to a greater understanding of the biology of colon cancer, this study makes significant progress towards more informed diagnosis and clinical decision-making.

Data

In effort to reproduce the paper’s findings, we first had to access their datasets. By using the accession number given in the paper, GSE39582, we were able to locate the study’s public repository. The microarray data was collected from the Gene Expression Omnibus, GEO, database and copied onto a remote server, SCC. A symbolic link was used to connect some of the sample data between multiple groups in the course. This link helped to maintain efficiency throughout the system, by avoiding wasted space on the remote server.

The study’s repository included the microarray data from human tumor tissue samples. A total of 134 samples were downloaded and used for our analysis. Though the original study used 443 discovery samples and 1,029 validations samples, we used a smaller dataset which combined both discovery set samples as well as validation set samples. Discovery samples were used to identify patterns whereas the validation samples were used to test robustness.

The downloaded dataset provided samples in the form of CEL files, which is the typical output form from microarray processing. The microarray technologies were able to investigate the gene expression profiles (GEPs). The samples used in the original study had to meet a RNA quality requirement for the GEP analysis; further description of the data generation can be found in the Marisa et. al.’s paper as well as in the supplementary material (Marisa, et. al., 2013). Overall, the data quality was assessed in the original study, leaving out several samples that did not meet the requirements. Our dataset was also normalized and assessed for quality control, to ensure good quality to continue our analysis.

Methods

All analysis was performed in RStudio (version 1.3.1073, R version 4.0.2) using BU's Shared Computing Cluster. Packages required were affy (1.66.0), affyPLM (1.64.0), and sva (3.36). After reading the data from CEL files and converting to an AffyBatch using the ReadAffy function, preprocessing of the data involved (1) QC of the data (to check for probe and read anomalies) and (2) adjustment of the data for follow-up analysis, including normalization and correcting for batch effects.

The QC of the data was performed as follows: the raw dataset was normalized using fitPLM (normalize set to TRUE), which fits the probe set with a linear model and normalizes the probe data. Histograms of the median NUSE and RLE scores (using those functions on the normalized dataset) showed no outliers, and data was centered on 1 and 0 respectively (Supplementary image 1 & 2).

With this confirmation that the data was of high quality, the raw dataset was then background corrected and normalized by RMA, using the rma function included in the affy package. The function takes in raw AffyMetrix microarray chip reads, and computes a robust average across all arrays in order to discard outlier results that deviate too far from the computed average . This calculation is especially appropriate for microarray data, as it allows for differential binding of probes and removes only true outliers (Irizarry, 2003). The expression results were returned as an array and all values were in $\log_2$ base scale. Next, the data was adjusted for batch effects using ComBat. This package is also able to accommodate multiple features in the adjustment, if provided. In our analysis, we were provided with a table of metadata associated with each sample, including batch number and another feature that combined MMR and tumor status, and those were included as arguments along with the expression dataset. The output from this ComBat adjustment (an expression matrix) was written to a csv for further analysis. Lastly, a principal component analysis was performed to visualize the variability of the data (Supplementary image 3). The first two principal components account for 93.2% and 1.4% of the variance of the dataset.

Due to large file sizes and a large number of files, the initial conversion of CEL files to an AffyBatch ran somewhat slowly. Any normalization steps involving the large AffyBatch dataset (NUSE, RLE, RMA) ran slowly as well, and running these steps unnecessarily should be avoided.

Results

In this section, we analysed the result of noise filtering and dimensionality reduction, and hierarchical clustering and subtype discovery. We presented three noise filtering approaches on RMA normalized sets of genes to adjust gene expression matrix. The first noise filtering approach indicated that 39750 genes expressed in 20% of samples. We then calculated the variance of all probe sets implementing a threshold of $p<0.01$, there were 54391 genes remaining after conducting the Chi-square test for variants. Our last filtering method was conducted to classify the genes that have a coefficient of variation greater than 0.186, our result

indicates 1838 genes qualify this filtering test. A total 1429 genes pass all the three noise filtering and dimensionality reduction requirements.

Clustering is one of the most used methods that divided the dataset into groups, consisting of similar data points. In this experiment, the hierarchical clustering was performed on 134 most variant samples. Clustering dendrogram was divided into two samples based on molecular subtypes, the first cluster has 79 samples and the second has 55 samples. The hierarchical clustering analysis from 134 samples of the discovery revealed two clustering of samples based on the 2 molecular subtypes, C3 molecular subtypes indicated in red and the rest of subtypes indicated in blue. Based on the gene expression matrix result and the two clustering methods, we looked at genes that expressed at adjusted value of $p<0.05$, we found 1190 genes expressed at the adjusted value of $p<0.05$.

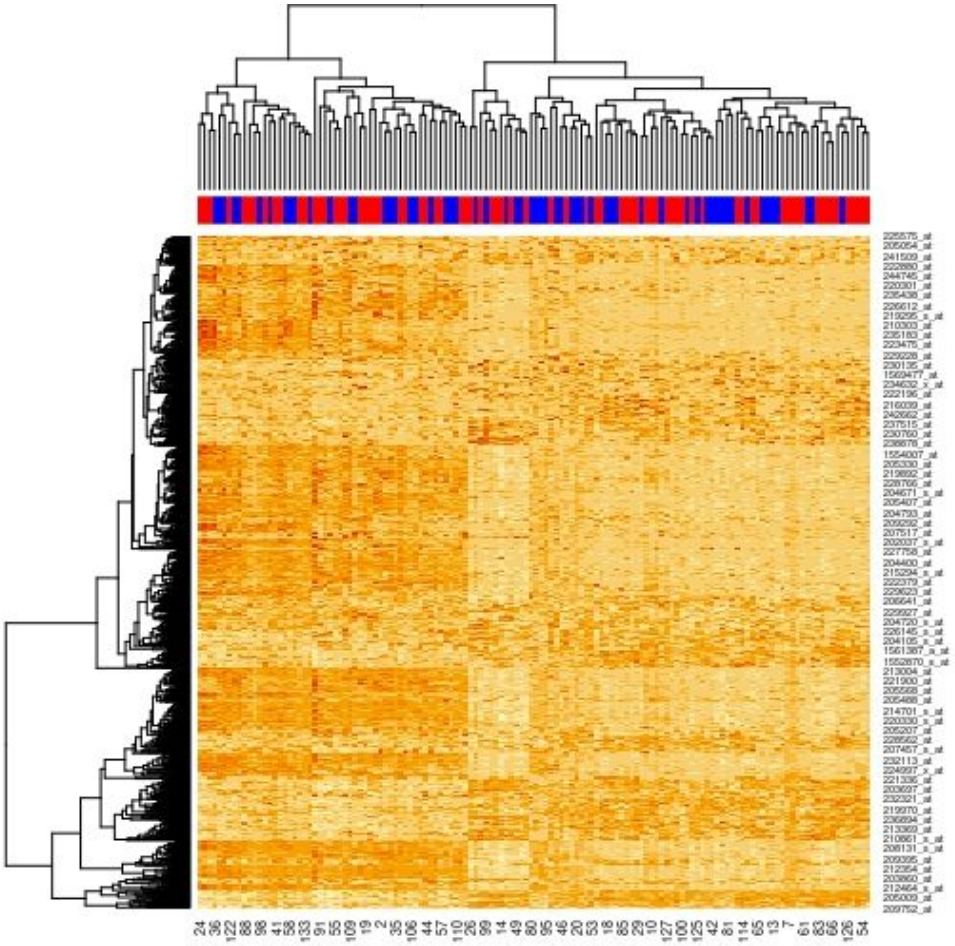


Figure 1: Heatmap of 1190 probes sets ordered by two molecular subtypes

Discussion

Through the use of filters and clustering, we were able to find 1190 genes that expressed at an adjusted value of $p<0.05$. These probes were mapped to the human genome symbols to then perform a gene set enrichment analysis. Within our top 10 differentially expressed genes we found *secreted frizzled-related protein 2* (SFRP2) and *growth arrest-specific 1* (GAS1) which was also found within the paper *Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value* (Figure 2). Furthermore, with our analysis we were able to identify common genes from our dataset and genes within GO, HALL, and KEGG pathways. Using the Fisher's test we found that there was a significant association between our genes and only the KEGG pathways. These pathways involve; metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, human diseases, and drug developments. This significance suggests that the genes differentially expressed can be found within the KEGG pathway gene sets. Unfortunately, due to a few limitations we were unable to identify the top three gene set names that were highly enriched. Overall, from these results we can conclude that there are genes differentially expressed that can be found within the KEGG gene sets. With further analysis we could identify the specific gene pathways that are being affected and their overall impact.

PROBEID	t.stat	p_val	p.adjust	SYMBOL	abso_t_value
213413_at	-23.83783	1.140086e-49	5.903972e-45	STON1	23.83783
225946_at	-23.69502	2.170937e-49	5.903972e-45	RASSF8	23.69502
204457_s_at	-23.33867	1.094387e-48	1.861058e-44	GAS1	23.33867
223121_s_at	-23.28967	1.368651e-48	1.861058e-44	SFRP2	23.28967
219778_at	-23.00211	5.114083e-48	5.563202e-44	ZFPM2	23.00211
238478_at	-22.74931	1.642904e-47	1.489320e-43	BNC2	22.74931
218694_at	-22.59383	3.380879e-47	2.626991e-43	ARMCX1	22.59383
205168_at	-22.50277	5.166254e-47	3.512471e-43	DDR2	22.50277
227061_at	-22.39461	8.560192e-47	5.173304e-43	LINC01279	22.39461
203695_s_at	-22.21176	2.016700e-46	1.096903e-42	GSDME	22.21176

Figure 2: Top 10 differentially expressed genes of data

Conclusion

Ultimately, we were able to reproduce to some degree the results of the paper and identify a subtype of colon cancer through differential expression. A number of differentially expressed genes and their associated KEGG pathways were identified. We were unable to identify top gene sets and specific pathways expressed due to technical difficulties and time constraints.

Throughout our analysis we encountered a few challenges we needed to overcome. To begin, the data was successfully copied and transferred to SCC. Although, once it was time for our group to begin working with the dataset we ran into an issue with access permissions. We were able to overcome this challenge by changing the mode of the directory files. By allowing all users to have reading, writing, and execution abilities, this problem was solved. Additionally, to analyze the gene expression across all genes, we produced a heatmap to visually show the results. Unfortunately, we ran into issues when producing this heatmap and were unable to correctly alter the color settings. With more time and problem solving, we may be able to correctly assign the red and blue colors to distinguish the subtypes that each sample belongs to. Further limitations involved issues with reproducing the top three enriched gene sets and top 10 down regulated genes. Part of this was solved by reproducing analysis results and re-running statistical analysis. The other half of this issue wasn't solved due to limited time to problem solve bugs and coding issues. If there was additional time we would have been able to work out these bugs and produce additional results and their implications.

References

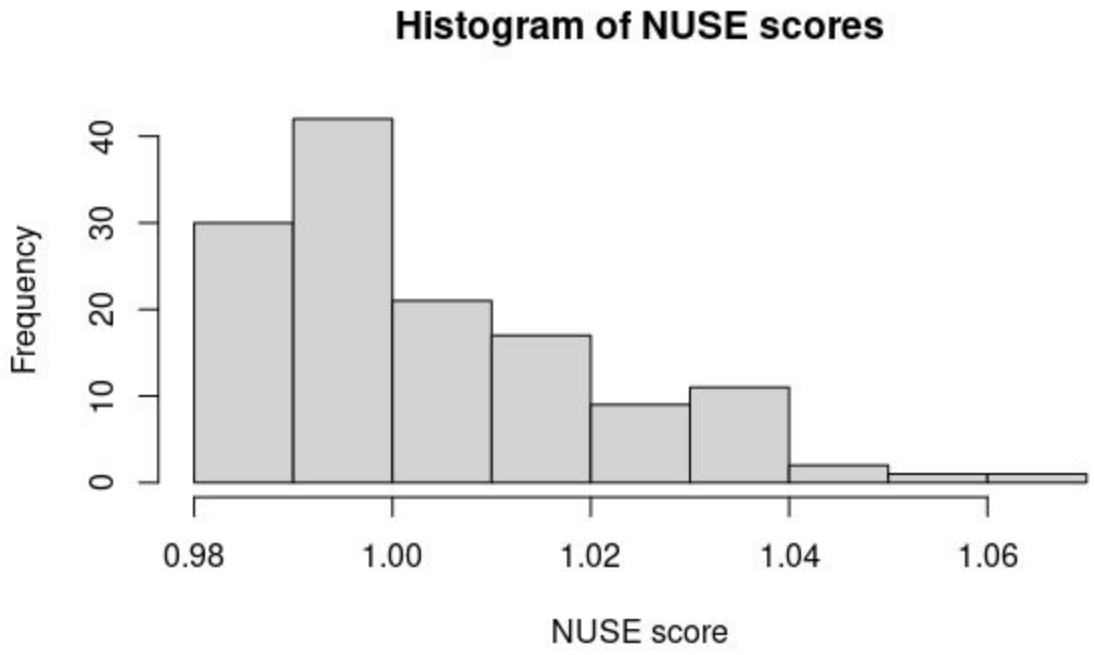
Rafael A. Irizarry, Bridget Hobbs, Francois Collin, Yasmin D. Beazer-Barclay, Kristen J. Antonellis, Uwe Scherf, Terence P. Speed, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data", *Biostatistics*, Volume 4, Issue 2, 1 April 2003, Pages 249–264, https://doi.org/10.1093/biostatistics/4.2.249

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.

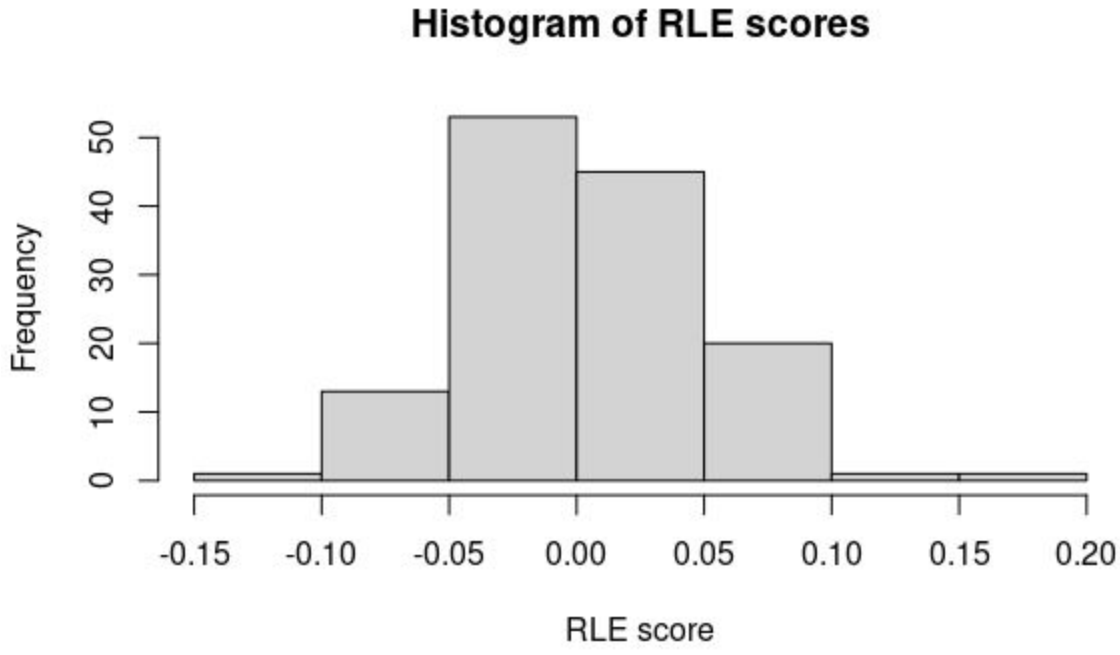
Laurent Gautier, Rafael Irizarry, Leslie Cope, Ben Bolstad (2020). Description of affy. https://www.bioconductor.org/packages/devel/bioc/vignettes/affy/inst/doc/affy.pdf

Supplementary Images

Supplementary image 1: Median NUSE scores of microarray data



Supplementary image 2: Median RLE scores of microarray data



Supplementary image 3: Principal components analysis of normalized and adjusted microarray dataset (after pre-processing).