Microarray Based Tumor-Classification

Rachel Thomas[1], Yichi Zhang [2], Varun Raghuraman[3], Neha Gupta[4], and Jackie Turcinovic [5]

[1]Data Curator
[2]Programmer
[3]Analyst
[4]Biologist
[5]TA

February 24, 2021

1 Introduction

Colorectal cancer (CRC) is the third most common cancer worldwide, and is the fourth leading cause of death by cancer (1). Pathological staging is the only method used in clinical practice to select for adjuvant chemotherapy for CRC patients. This approach is not effective and has failed to accurately predict recurrence; 10-20 percent of patients with stage II CRC, and 30-40 percent of patients with stage III CRC develop recurrence. Microsatellite Instability(MSI) is currently the only molecular marker that is reproducibly found to play a role in prognostic classification [1]. Previously scientists have utilized microarray technology to investigate the possibility of a gene expression profile that could be helpful in predicting prognosis of CRC, currently there are none. The goal of this study was to identify a standard reproducible molecular classification of CRC based on mRNA expression profile analyses and to assess whether there were any associations between identified molecular subtypes and clinical or pathological factors, common DNA alterations and prognosis [1]. mRNA expression profiling was used to help determine if CC cells expressed higher levels of certain genes. If so, then further analysis could be done to investigate if these genes code for a specific protein receptor and its potential role in cancer pathogenesis. This technique was a way to classify the varying types of CC into subgroups to aid in better treatment and prognosis.

2 Data

2.0.1 Data Description

We used data previously described [1] and collected from the Gene Expression Omnibus (GEO) database (2) . The data sets were analyzed and prepared using mRNA expression profiles using Affymetrix U133plus2 chip and DNA alteration profiles using the CGH Array [1]. The discovery and validation sample sets were combined into a single dataset that consisted of 134 samples of C3 and C4 human colon cancer subtypes.

2.0.2 Data Quality Assessment

The data was first normalized using Robust Multiarray Averaging (RMA) algorithm to create an expression matrix from Affymetrix data (3) collected from the public database. Then Relative Log Expression (RLE) and Normalized Unscaled Standard Error (NUSE) analysis were employed to determine if batch effects and lower quality arrays existed. ComBat via the sva package is used to correct batch effects using an empirical Bayesian framework. Finally, Principal Component Analysis (PCA) was used to identify any noise or outliers in the datasets. No samples were eliminated due to low quality or contamination

3 Methods

3.0.1 Tools for the Workflow

This workflow is much dependent on an R script integrated with multiple popular libraries in this field. For data processing and analysis, affy, affyPLM, sva, AnnotationDbi, and hgu133plus2.db from Bioconductor are used and we applied libraries based on ggplot for visualization. Alternatively, a python script using numpy, pandas, and matplotlib could be used for the analysis after generating a normalized expression matrix file. We recommend leaving more than 8GB RAM space to render and process data in this analysis.

3.0.2 Normalization of Microarray Data

Raw data of expression intensities with more than 50000 probes among 134 samples is stored in Affymetrix CEL Data File format and is directly readable via the affy package.

The primitive normalization of expression levels is done by the RMA algorithm. RMA is used because of two major nosies: measured expression intensity from background increases with concentration and the existence of the probe effect. RMA applies the quantile normalization, an approach to normalize each array against all others to see if probe intensities have the same distribution. For large data used in this analysis,

this normalization will not eliminate meaningful differences in gene expression. We used the built-in rma function to obtain the background-adjusted, quantile-normalized, and log-transformed intensities matrix.

RMA uses Median Polish for summarization. It is robust for outliers in large datasets and meanwhile maintains certain computational efficiency. However, this method takes more interest on column effects (per sample) instead of row effects (per probe) and thus it is better for datasets which arrays are more than probes.

3.0.3 Quality Control

Using the Bioconductor package affyPLM, we fit a probe level model to Affymetrix Genechip Datawe (fitPLM) and compute the Relative Log Expression (RLE) and the Normalized Unscaled Standard Error (NUSE) scores of the microarray samples after data normalized using quantile normalization and using RMA to correct background. The motive of using RLE is to visualize unwanted variation in high dimensional data. There are many causes of such variation. For example, batches of samples may be processed in different laboratories which operate at different temperatures leading to variation between the batches, i.e. a so-called batch effect. Moreover, the temperature of a particular laboratory may be quite variable throughout the day leading to additional variation within a batch. In any particular study, however, the physical causes of such variation will typically be unknown. This unwanted variation can be so large that comparing gene expression values between samples, often the main objective of such a study, can no longer be sensibly done(4). RLE plots are particularly useful for assessing whether a procedure aimed at removing unwanted variation, i.e. a normalization procedure, has been successful.

An RLE plot is constructed by calculating the median expression across all samples for each gene. After calculating the deviations from this median, a boxplot of all the deviations for that sample is generated.

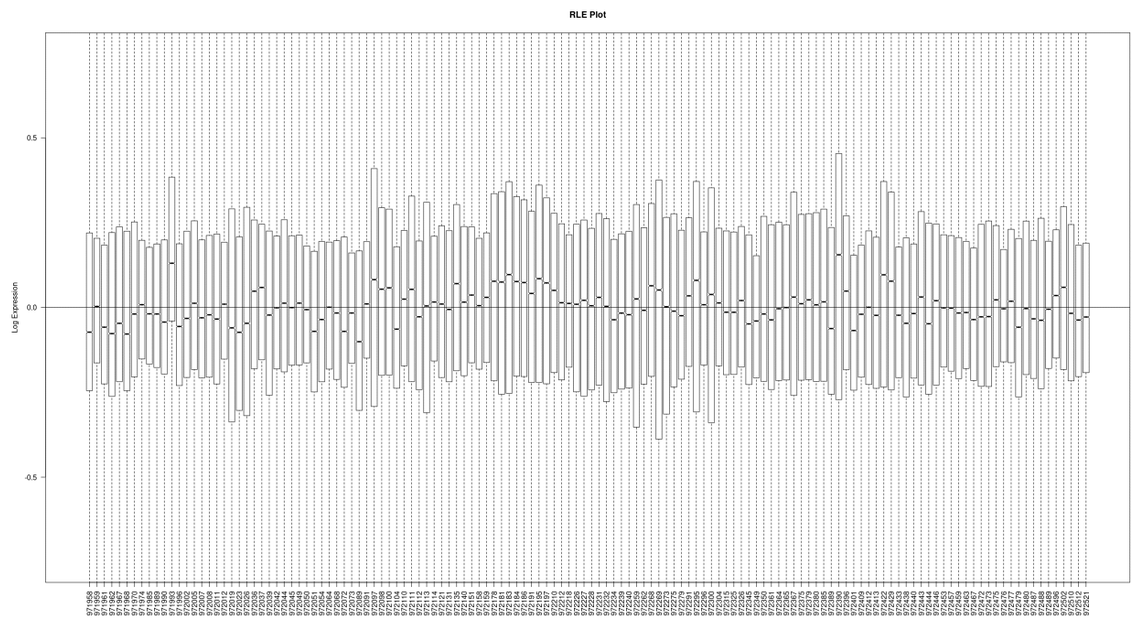

Figure 1: RLE plot of patient data

The aim of this study was to establish a comprehensive molecular classification of CC based on mRNA expression profile analyses[1]. We assume expression levels of a majority of genes are unaffected by the biological factors of interest. In ideal circumstances, an RLE plot would only display box plots roughly centred on zero and would roughly be the same size. We see unwanted variation both between and within batches as indicated by the varying position and heights of the boxplots (gene variation has been removed). Considering the range of difference among sample variations is less than 0.5, experimental data is reliable and acceptable after proper normalization.

Another assessment is NUSE plot. Compared to RLE values, NUSE values are not comparable across dataset, therefore in the NUSE plot low quality arrays are significantly elevated or more spread out, relative to the other arrays. We are interested in standard error estimates obtained by the PLM fit because variability differs between genes, we standardize these errors so that the median standard errors across arrays is 1 for each gene(5).

When examining a NUSE plot, we are looking for arrays with boxes that are significantly elevated or spread out than other arrays. In Fig 2, arrays bounded by red boxes may indicate lower quality arrays due to sample heterogeneity. Those arrays will still be kept for later analysis because there is no median intensity greater than 1.1.

RLE and NUSE analysis have revealed that batch effects and lower quality arrays exist in the dataset, ComBat via sva package is used to correct batch effects using an empirical Bayesian framework. We use an annotation file containing a host of clinical and batching annotations used by the authors for their analysis to correct for batch effects and other unwanted variations while preserving features of interest. Batch effects include both Center and RNA extraction methods have been merged into a single variable called normalizationcombatbatch in the annotation file indicating known batches. Features of interest include both tumor and MMR status as per Marisa et. al. and have been merged into a single variable called

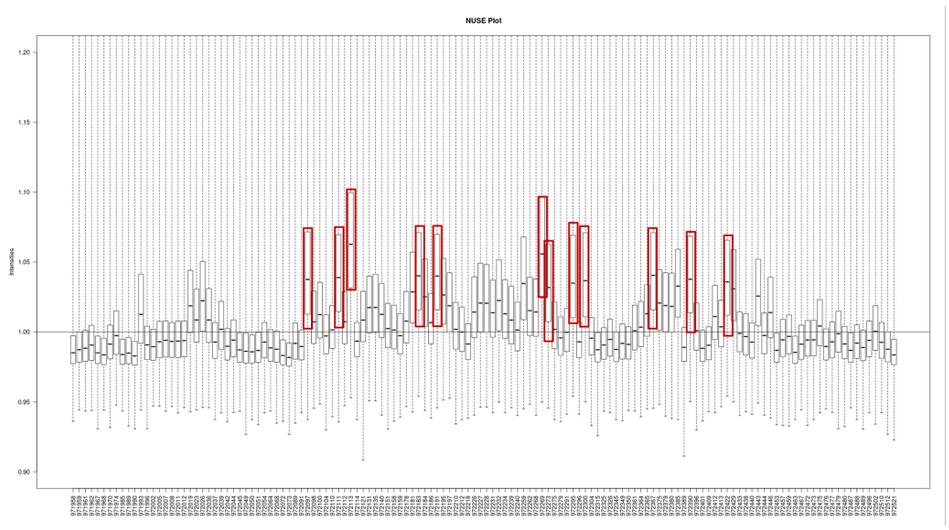

Figure 2: NUSE plot of patient data. Red boxes indicate arrays that are elevated more and spread out more than others

normalizationcombatmod to construct a model matrix for outcome of interest and other covariates besides batch.

3.0.4 Principal Component Analysis and Outlier Detection

The main motive to implement PCA is to figure out unique patterns and correlations in the given dataset. When a strong correlation is found between data sets and variables, a final decision is made about reducing the data so that the final data can be constructed with the significant data still retained.

Normalized expression matrix from the ComBat function needs to be standardized into a comparable range by subtracting each value in the data from the mean and dividing it by the overall deviation in the dataset. Correlation between the different variables in the data set is expressed through a Covariance Matrix. It is important to identify heavily dependent variables which contain redundancy and biased information, which can alter with the outcome and reduce the performance of any specific system. Principal components are computed by the prcomp function and ensure that data is centered and scaled within each gene.

We choose to plot PC1 against PC2, sets of variables possess the most important and useful information that was scattered in the initial stage amongst the initial variables.

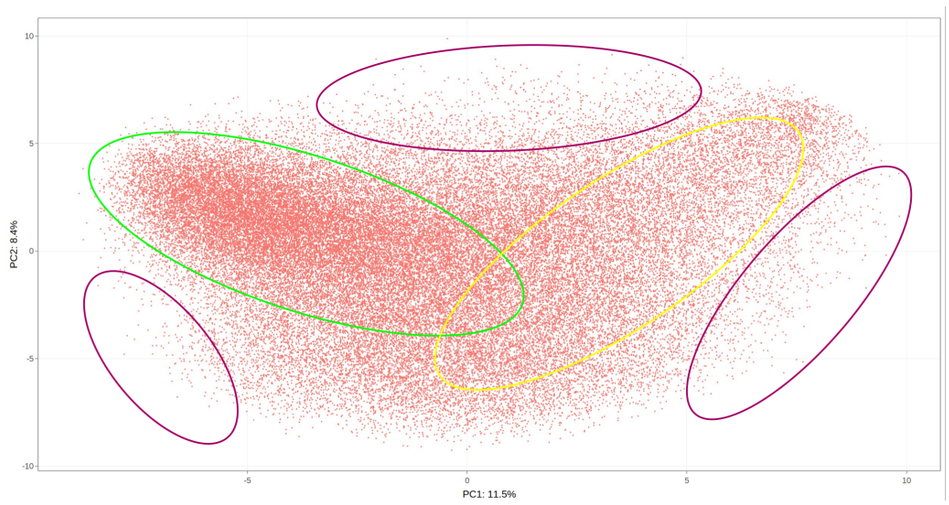

Figure 3: PC1 vs. PC2. Each dot represents a probe in the dataset, green and yellow circles indicate major trends and purple circles indicate noises/outliers. The percentage numbers on legends indicate the percentage variance of each PC among all PCs.

Imagine an ideal PCA plot: suppose an analysis only involves 10 genes. 5 genes only express in situation A and the rest only express in situation B. We will have a linear decision boundary that 10 dots will split into 2 clusters in the plot. Since the purpose of the workflow is to explore a biomarker for Colon Cancer, we assume critical genes would express differently between lesion and healthy tissue, so that data points . A clear decision boundary in hyperplane might be tricky for real-life data even after variations due to experimental factors have been normalized since the dataset contains data that is completely irrelevant to the analysis, “noise”. The PC plot provides initials on how to reduce noises and what features extracted later would be considered as great features.

3.0.5 Gene set Enrichment

Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states. Using the Hgu133plus2.db package from Bioconductor, we mapped the probe IDs to their appropriate gene symbols. A challenge faced is duplicate probe IDs for the same gene symbol. In that case, we filtered the probe IDs with minimum p-adjusted value(with minimum fold change values) and discarded the others. The genes with the highest significance level were selected for further analysis. After sorting the genes in decreasing order based on t statistics, we extracted top 1000 up(highest values) and down(lowest values) regulated genes.

The KEGG, GO and Hallmarks gene sets collections were downloaded from MSigDB. The gene sets were imported into the R environment using the GSEABase package of Bioconductor(6). By performing the Fisher Test, the top 1000 differentially expressed genes were enriched with genes from each set's collections(7). The p values were adjusted using the Benjamini-Hochberg method.

4 Results

4.0.1 Primary Findings

Three filters were applied to the datasets post their normalization and adjustment as detailed in the Methods section. The starting dataset contains 134 patients with 54,674 probe sets that each correspond to a single gene. The first filter removed all probe sets with expression intensities less than log2(15) for 20 percent of the patients examined (Project1-4-1.csv). This reduced set of 39,661 rows was then filtered a second time through a chi-squared test for variance–slightly different in nature to that of the one detailed by the paper. The second filter took the median variance of the initial dataset and compared it against the observed variances for the rows filtered by the dataset. This produced a much-reduced table with 17,721 rows. The test statistic used was (N-1)*(s/u)2, with N being the patients/CEL files, "s" the variance of a gene/probeset, and "u" the median variance. The point of the test was to examine whether the variance of the data was close to the "expected" median result. The final filter examined the coefficient of variance. The coefficient of variance is a ratio of the standard deviation to the mean of a dataset–thus its value can describe how the data varies about its mean. The final filter only took data with a coefficient of variation greater than 0.186, leaving just 1,691 rows in the table (output-4-4.csv). Hierarchical clustering using the ward function (ward.D in R syntax)(8) was used to produce two clusters of patients—58 in one group and 76 in another. The agglomerative coefficient was used to evaluate the best clustering algorithm

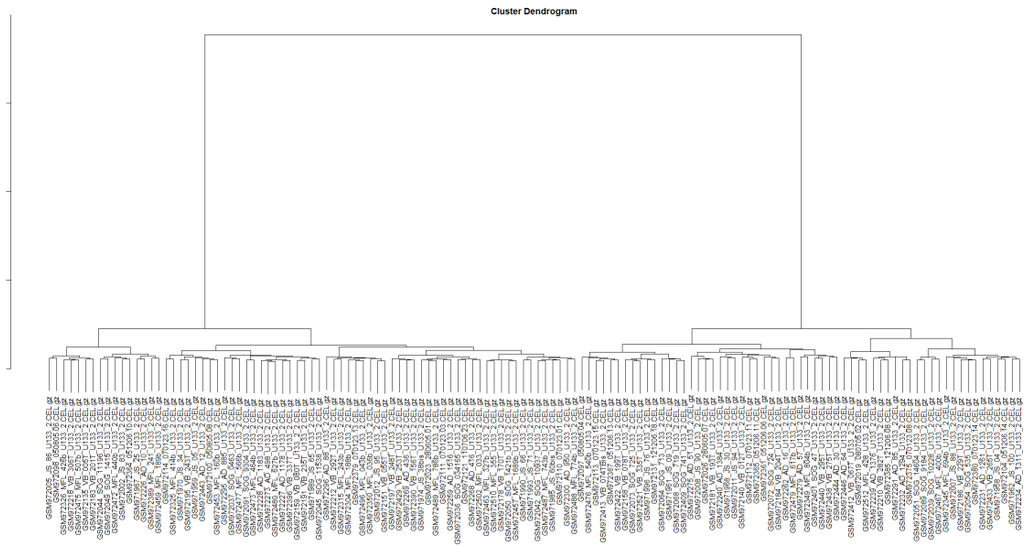

Figure 4: hierarchical clustering of gene expression based on samples

These two subtypes were subjected to Welch T-tests to evaluate whether there was a specific difference in the expression of genes between the two cluster populations, despite differences in the numbers of patients in each cluster. It was found that, using a threshold of p¡0.05 (after adjusting the p-value for differences between group sizes), 1,166 genes were notably differentially expressed out of the initial 1,691–suggesting that the unsupervised clustering did actually produce genes that were different from each other (differential-expression-results.csv). The T-statistic of the two genes was used to determine the genes that most define each set. The t-statistic's signage does not indicate level of difference, merely direction. When the genes that had passed the p-value threshold (both the experimental and adjusted) were compared it was noted that the maximum t-statistic was 24.07, and the least -15.62, with a range of values in between them. Thus while different genes varied in different ways between the clusters, the genes with the greatest magnitude of change just so happened to have positive t-statistics. The top 50 genes with t statistics of greatest magnitude were collected (filter-results-5-5.csv) as they best represent the greatest level of difference in expression between the two cluster populations.

A common method of visualising exceptional gene expression data is to display it as a heatmap, but unfortunately the heatmap generated based on hierarchical clustering does not give us notable information

name	t.statistic	p.value	p.adjusted
207266_x_at	24.07237	5.81E-50	9.82E-47
203748_x_at	23.56101	4.03E-49	6.80E-46
209868_s_at	23.38609	9.17E-49	1.55E-45
202291_s_at	22.23936	1.14E-45	1.93E-42
202363_at	22.04136	9.13E-45	1.54E-41
226930_at	22.01966	5.07E-45	8.55E-42
212607_at	21.42487	2.12E-41	3.52E-38
225782_at	21.30052	4.09E-44	6.89E-41
215127_s_at	21.2313	2.15E-44	3.63E-41
209210_s_at	21.10012	4.44E-44	7.47E-41

Table 1: Top 10 genes from the list of top 50 most differential genes

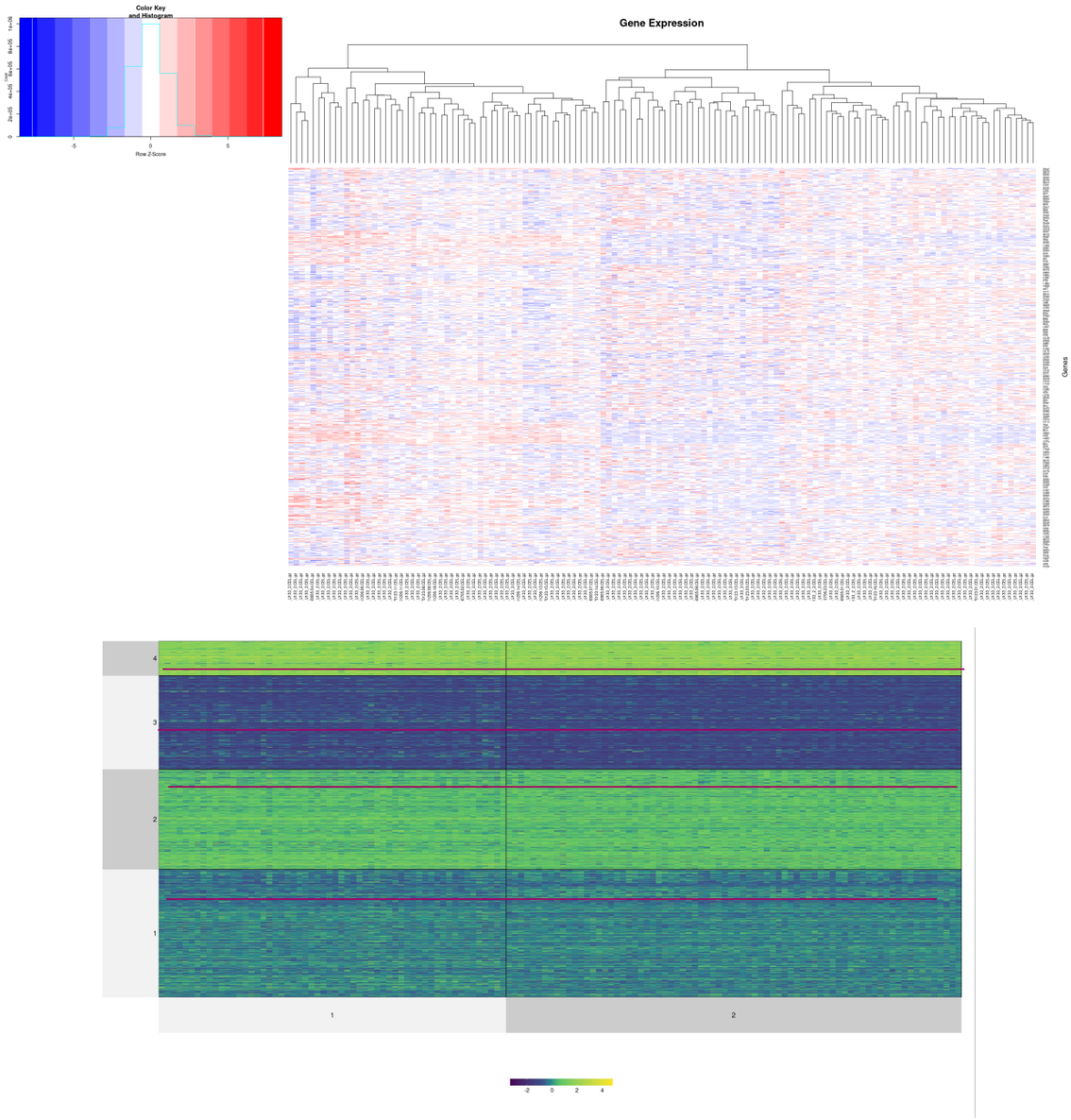

Figure 5: Heatmaps of expression and filtered genes. Generated based on sample clusters and gene sets

about biomarkers that distinguish between healthy tissue and lesion. Based on sample clusters, we manually set a number for genes to generate gene clusters. On Fig 5 bottom, visually we are interested in genes that express high in one sample cluster but low in another. We are particularly interested if gene sets are statistically significant via GSEA fit features we find on the heatmap.

4.0.2 Biological significance of expressed genes

The enriched genes for top 1000 up and down regulated genes were obtained using 3 gene set collections from MSigDB database (9). Out of those, top 10 up and down regulated genes are shown in table 2 and table 3. The GO(Gene Ontology) gene set is divided into three categories: BP(biological processes), CO(cellular components and MF(molecular functions. This collection had 10271 gene sets. The KEGG(Kyoto Encyclopedia of Genes and Genomes) is a canonical pathways gene set derived from the KEGG pathway database. This collection consisted of 186 gene sets. The Hallmark collection summarizes and represents specific well-defined biological states or processes and displays coherent expression. There are 50 gene sets in this. We

got 1580 enriched gene sets with p-value ¡ 0.05.

Probe-id	t-statistics	P-value	p_adj	Symbol
204457_s_at	22.69869	1.27E-45	2.25E-41	GAS1
207266_x_at	22.66197	2.50E-47	4.44E-43	RBMS1
218694_at	21.86439	9.69E-44	1.72E-39	ARMCX1
223122_s_at	21.65212	2.94E-45	5.21E-41	SFRP2
226930_at	21.60388	1.22E-44	2.15E-40	FNDC1
213413_at	21.35635	5.86E-41	1.04E-36	STON1
202363_at	21.22895	1.57E-43	2.77E-39	SPOCK1
202291_s_at	21.04658	1.25E-43	2.21E-39	MGP
217764_s_at	21.03102	5.92E-44	1.05E-39	RAB31
225242_s_at	20.80849	3.16E-42	5.59E-38	CCDC80

Table 2: Top 10 up regulated genes

Probe-id	t-statistics	P-value	p_adj	Symbol
210107_at	-12.4002	2.00E-23	3.40E-19	CLCA1
227226_at	-12.4436	2.01E-23	3.42E-19	MRAP2
207214_at	-12.4581	2.72E-22	4.60E-18	SPINK4
211715_s_at	-12.6969	5.86E-24	9.97E-20	BDH1
214106_s_at	-12.7099	1.05E-21	1.77E-17	GMDS
234008_s_at	-12.7496	1.10E-24	1.88E-20	CES3
204673_at	-13.4209	1.19E-24	2.03E-20	MUC2
227725_at	-13.5179	1.70E-23	2.89E-19	ST6GALNAC1
220622_at	-13.9038	1.67E-27	2.88E-23	LRRC31
203240_at	-15.2435	3.19E-28	5.51E-24	FCGBP

Table 3: Top 10 down regulated genes

For GO, top 3 significantly expressed gene sets are listed in table 4. The extracellular matrix which provides structural support, biochemical pr biomechanical cues for cells or tissues. The collagen contains extracellular matrix secreted by cells in the vicinity and form a sheet underlying or overlying cells such as endothelial and epithelial cells. Extracellular structure organization is a process that is carried out at the cellular level which results in the assembly, arrangement of constituent parts, or disassembly of structures in the space external to the outermost structure of a cell. The sets are all up-regulated list.

Name	pvalue	estimate	expression	BH
GO Extracellular matrix	5.59E-59	8.296428	UP	5.90E-55
GO Collagen containing extracellular matrix	5.74E-59	9.7625	UP	5.90E-55
GO Extracellular structure organization	1.39E-49	8.372985	UP	9.51E-46

Table 4: Enriched data sets on GO collection based on nominal p-value

The top 3 enriched gene sets in the KEGG collection are listed in table 5. The extracellular matrix (ECM) consists of a complex mixture of structural and functional macromolecules and serves an important role in tissue and organ morphogenesis and in the maintenance of cell and tissue structure and function. Cell-matrix adhesions play essential roles in important biological processes including cell motility, cell proliferation, cell differentiation, regulation of gene expression and cell survival.

Name	p-value	Estimate	Expression	BH
KEGG ECM receptor interaction	1.01E-18	13.1601	UP	3.75E-16
KEGG FOCAL Adhesion	4.98E-18	6.078024	UP	9.26E-16
KEGG valine leucine and isoleucine degradation	6.59E-10	9.818737	Down	7.28E-08

Table 5: Enriched data sets on KEGG collection based on nominal p-value

The top 3 enriched gene sets from the Hallmark collection are presented in table 6. Epithelial mesenchymal Genes defining epithelial-mesenchymal transition, as in wound healing, fibrosis and metastasis. Myogenesis gene set describes as Genes involved in development of skeletal muscle (myogenesis). The other gene set is described as genes in response to ultraviolet (UV) radiation. All the gene sets are up regulated and enriched.

Name	p-value	Estimate	Expression	BH
Hallmark epithelia mesenchymal transition	1.07E-73	18.28073	UP	1.07E-71
Hallmark myogenesis	2.25E-16	5.78861	UP	1.12E-14
Hallmark UV Response DN	8.18E-13	5.005041	UP	2.73E-11

Table 6: Enriched data sets on Hallmark collection based on nominal p-value

5 Discussion

The biological relevance of our results can be interpreted by the gene set enrichment data. The top 10 up regulated differentially expressed genes we found had GAS1 (growth arrest specific 1) protein that plays a role in growth suppression. GAS1 blocks entry to the S phase and prevents cycling of normal and transformed cells. Gas1 is a putative tumor suppressor gene. These genes may therefore work as markers of the initiation and growth of Colon Cancer cells and may constitute potential therapeutic targets. This gene was also reported in the reference paper. The most down regulated gene from our study is CLCA1 (Chloride Channel Accessory 1) protein coding gene that is involved in the regulation of tissue inflammation in the innate immune response and may also play a role as a tumor suppressor gene. The reference paper also exhibits epithelial mesenchymal transition that is also should by our analysis in the up regulated form.

6 Conclusion

In this project, we carried out microarray analysis on 134 colon cancer samples. Among all top 10 up-regulated differentially expressed genes we got, one was GAS1 protein which functions in suppressing cell growth. This protein was included in the poor prognosis cluster signature [1]. Additionally the most down-regulated gene, CLCA, could potentially play a role in tumor suppression. Therefore, these genes may be a potential target for future therapeutic strategies for CC patients, and classification of these markers could help to improve clinical prognosis. The major challenge encountered is to extract interests from tons of data after eliminating statistically biased factors from data. To address this problem, a good initial would be toggling on a hypothesis before processing the data. In this analysis, we expect certain genes to be expressed differently in the clinical importance for Colon Cancer and thus deciding which chi squared test results to use and how to make data more sensitive to our criteria became the focus. Generating a heatmap that was easily readable to display useful information from our analysis might not be feasible for the high-dimensional data. Even after narrowing down data scales to focus on genes expressed differentially, we still have noises left. Genes having high expression levels might be responsible for normal metabolic reactions, in other words, they are active among all regions of the human body. Therefore, we are more interested in genes that are exceptionally low among samples or certain samples based on our clustering results.

References

[1] L. Marisa, A. de Reyniès, A. Duval, J. Selves, M. P. Gaub, L. Vescovo, M.-C. Etienne-Grimaldi, R. Schiappa, D. Guenot, M. Ayadi, S. Kirzin, M. Chazal, J.-F. Fléjou, D. Benchimol, A. Berger, A. Lagarde, E. Pencreach, F. Piard, D. Elias, Y. Parc, S. Olschwang, G. Milano, P. Laurent-Puig, and V. Boige, "Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value," *PLoS Medicine*, vol. 10, no. 5, p. e1001453, May 2013. [Online]. Available: https://dx.plos.org/10.1371/journal.pmed.1001453

[2] "https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi."

[3] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, no. 2, pp. 249–264, Apr. 2003.

[4] L. C. Gandolfo and T. P. Speed, "RLE plots: Visualizing unwanted variation in high dimensional data," *PLOS ONE*, vol. 13, no. 2, pp. 1–9, 2018, publisher: Public Library of Science. [Online]. Available: https://doi.org/10.1371/journal.pone.0191629

[5] "https://math.usu.edu/ jrstevens/stat5570/Bowles."

[6] M. Morgan, S. Falcon, and R. Gentleman, "Gseabase: Gene set enrichment data structures and methods," 2020, r package version 1.52.1.

[7] "https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/fisher.test."

[8] "https://www.datacamp.com/community/tutorials/hierarchical-clustering-r."

[9] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov, "Molecular signatures database (MSigDB) 3.0," *Bioinformatics*, vol. 27, no. 12, pp. 1739–1740, 05 2011. [Online]. Available: https://doi.org/10.1093/bioinformatics/btr260