BF528 Project 1: Group Frizzled

Janvee Patel - Data Curator
Zhuorui Sun - Programmer
Yashrajsinh Jadeja - Analyst
Camilla Belamarich - Biologist

Microarray Based Tumor Classification

Introduction

Marisa et al. (2013) successfully predicted clinical prognostic factors of colon cancer (CC) using mRNA gene expression profiles. However before this study was completed, pathological staging was the only clinical practice used to form a treatment plan for patients with colorectal cancer (CRC), which ultimately failed to account for the heterogeneous manner of the disease and predict recurrence accurately. CC is extremely prevalent among the human population, as well as notorious for its high death rates [2]. There are five stages of CC that must be correctly identified in order to effectively treat this disease. Moving forward, a new clinical practice was essential to hinder the devastating effects CC causes in humans. Previous studies that attempted to take on this challenge used microarray technology to analyze gene expression profiles. Unfortunately, the results of these studies were not significant enough to use in the prognosis of CRC, which is mainly due to the heterogeneous manner in which this disease takes on. This causes there to be several prognostic signatures and molecular subtypes. Later gene expression profile studies used more advanced unsupervised hierarchical clustering and high-throughput methylation to classify molecular subtypes, in which three were found. Common DNA markers of CRC were used in the classification process; however, a more refined and reproducible standard of molecular subtype classification is needed to accurately predict prognosis and recurrence of CRC.

Further classification of these prognostic signatures and molecular subtypes was essential to aid the clinical treatment of CRC/CC moving forward. Researchers used mRNA expression profile analysis in hopes of discovering a complete classification of CC that can be used in clinical practices. The Cartes d'Identité des Tumeurs (CIT) in France collected fresh-frozen primary tumor tissues from patients ranging from stage I to IV who all underwent surgery, which were the primary samples classified in this paper in combination with seven other public datasets. In order to comprehensively classify samples, they used genome-wide mRNA analysis. Specifically, they used an array-based comparative genomic hybridization. Additionally to find common clusters in the data, consensus unsupervised analysis of gene expression profiles revealed six molecular subtypes. Through the classification process, they formed significant associations between these molecular subtypes and clinicopathological factors using the Chi-squared test, logistic regression, and the Kyoto Encyclopedia of Genes and Genomes to observe associated signaling pathways. Lastly, researchers tested the strength and validity of their classified molecular subtypes in a large independent dataset. Compared to previous studies, Marisa et al. (2013) revealed a comprehensive classification process for CC molecular subtypes that can be used in clinical practices to more accurately predict prognosis and recurrence.

Data

From Marisa et al. (2013), there were 750 tumor samples from patients with stage I to IV colon cancer from the French CIT program, of which 566 tumor samples satisfied the RNA quality control measurements for gene expression profiling (GEP) analysis. These samples were further separated into a discovery set which comprised 443 samples and a validation set which comprised the remaining 123 samples. In addition, colon cancer samples from public datasets using an Affymetrix platform and TCGA which did not use an Affymetrix platform were implemented within the validation set. The discovery set was used for determining classification and identifying patterns between those samples, while the validation set was used to test the classification. For the gene expression profiling of the tumor samples, Affymetrix Human Genome U133 Plus 2.0 Array was used [1]. This microarray is of type in situ oligonucleotide array. The data described in this section allowed for further subtype determination and identification of relationships among the samples.

We used the data generated as previously described [1]. In this analysis project, there were a total of 134 samples utilized which were included from the combined discovery and validation datasets. The data for these samples can be accessed at [http://www.ncbi.nlm.nih.gov/geo/; accession number GSE39582]. In the NCBI Gene Expression Omnibus (GEO) database, the repository for the paper was accessed through the accession number GSE39582. From this repository, sample GSM971958 was identified, and the .CEL.gz file for this sample was downloaded and uploaded using an SFTP client. The remaining 133 samples had been downloaded and uploaded previously from the NCBI GEO database into a central location and were made available through the use of symbolic links. These samples were sourced from human tissue of primary colorectal adenocarcinoma, and the sample type was RNA. In addition, from the GEO database, for Sample GSM971958, the TNM staging system criteria was listed, and this sample was sourced from stage IV colon cancer with the size of the primary tumor classified by T3, regional lymph nodes by N2 indicating the number of lymph nodes that had cancer, and metastasis by M1 indicating that the cancer had spread [3,7]. This is an example of the type of tumor samples that were included in this analysis.

From the GEO database, the RNA extraction protocol for Sample GSM971958 mentioned above followed Manual-Trizol [3]. Protocols for extracting total RNA from the remaining 133 samples discussed above included using TRIzol Reagent (Invitrogen), Manual-Cesium Chloride, Manual-RNeasy Micro (Qiagen), Manual-RNeasy Mini (Qiagen), RNA NOW, and TriReagant and Manual-RNeasy Mini (Qiagen). These protocols were implemented for ensuring there was extraction of high-quality total RNA. Then, biotin-labeled cRNA targets were generated using the Affymetrix One-Cycle Target Labeling Assay procedure

[3]. The fragments were hybridized on the Affymetrix Human Genome U133 Plus 2.0 Arrays, and then, the washing and staining procedure were completed using the Affymetrix GeneChip Fluidics Station 450. The Scanner GeneChip 3000 7G was utilized for microarray scanning [3]. Quality measurements were assessed at these steps to ensure that the data was of high quality.

Methods

In this project, we read in CEL files for probe level data, did data normalization and the visualized distribution of the dataset with RLE and NUSE, corrected the batches effect and ran Principal component analysis on the normalized dataset. For the further statistical analysis, we filtered noise, did clustering analysis, subtype discovery and gene set functional enrichment. Data processing and statistical analysis were performed in R using package affy, affyPLM, AnnotationDbi, hgu133plus2.db, sva, factoextra, ggplot2.

Data read in:

To deal with CEL files, affy and affyPLM packages were used to read in data. We read in the 134 CEL files by ReadAffy() function in R. By this function, 134 CEL files can be read in simultaneously and saved in Ori_affyData as the original read in data.

Data Normalization:

The Robust Multi-Array Average expression measure (rma()) function in package affyPLM was used to normalize the original affy read in data, after data normalization. This function uses background correction and quantile normalization to normalize the original probe level data.

Data distribution with RLE and NUSE:

RLE is a powerful tool to visualize unwanted variation in microarray data. By RLE we assume the expression of a probe is not affected, so the median of each sample should close to 0. The NUSE is calculated based on an linear model, a good quality data should have a score around 1.[4]

fitPLM() function was used to fit affydata into PLM dataset which can be used to compute Relative Log Expression (RLE) and Normalized Unscaled Standard Error (NUSE) scores of the microarray samples.

Batch effects Correcting:

Combat function in the sva package was used to correct batch effects in this project. Follow-up data analysis was based on the data after batch effects correcting.

Principal Component Analysis:

Principal Component Analysis (PCA) was performed on the normalized data to analyse each principal component contributing how much to the dataset and reduce dimensions for data. First center and scale the data by the scale() function and then run PCA by prcomp() function. We got a rotation matrix of the original data by PCA$rotation.

Filtration of ComBat Adjusted Data:

After the data was ComBat adjusted, a filter was applied where only normalized intensity values that were greater than log2(15) and were expressed in at least 20% of the samples were retained as described by Marisa et. al.[1] This was done to select the genes that were expressed commonly enough to have some consensus and to ensure that the genes being selected weren't merely isolated occurrences that were rarely expressed in some samples. Thus, this would also filter out the probe sets that had a weak signal (intensity) due to the minimum expression criteria of log2(15) being used.

After that, the variance in expression values for every single probe-set across the samples was calculated and a median variance was obtained from those variances as that would come in handy for the next part of the analysis. To perform a two-tailed chi-square test, the variance test statistic was calculated as defined by the following formula : $T = (N-1)(s/\sigma_0)^2$ where N stands for the degrees of freedom (number of samples), s is the standard deviation (of expression values across samples) and σ_0 is the median variance.

Then, the chi-square distribution ranges were determined using a two tailed estimate. The R function *qchisq(*α */ 2, dof)* was used to obtain the lower tail of the chi-square distribution and *qchisq(1-*α */ 2, dof,lower.tail=FALSE)* was used to obtain the upper tail of the distribution where dof is the degrees of freedom (number of samples) and α is the confidence interval 0.01.[5] The variance test statistic was then compared to the chi-square distribution range for every gene. Probe sets were filtered where the test statistic fell in the range of the chi-square distribution as deviation from the chi-square range was crucial to obtain probe sets deviating from the distribution and to obtain probe sets with a higher, much significant degree of variance. The rationale behind this was to omit genes that were relatively average in expression and didn't vary much across samples.

For the final step of filtration, coefficient of variation was calculated for every probe set across the samples. The square root of variance for each individual gene was divided by the mean of the expression value for every single gene to obtain the coefficient of variation. This is better illustrated using the formula : $CV = \sigma / \mu$ where σ is the standard deviation and μ is the mean. Finally, after this, the third filter was applied where only the probe sets with a coefficient of variation greater than 0.186 were retained. This ensured that only highly varying probe sets were selected as they are important in determining significant expression and for further clustering.

Hierarchical Clustering and Exploratory Analysis

The filtered dataset was then used for clustering. Hierarchical clustering was performed for samples using the hclust() function in R. This resulted in 2 clusters of samples being formed. This cluster was then cut using the cutree() function in R. Molecular subtype annotation was loaded from the metadata file that provided the molecular subtype information for every sample and that information was leveraged to color code the samples falling into the C3 clusters as red and the rest as blue.

A heatmap was plotted using the heatmap.2() function from the gplots package in R.[6] The heatmap displays the cluster of samples along with the gene expression levels. The clusters are color coded and the intensity of the color as displayed in the color key represents the gene expression levels with yellow being highly upregulated and red being highly downregulated. The expression values follow the gradient described in the color key.

After the heatmap was plotted, a Welch t-test was performed on the probe sets between the two clusters. This resulted in T-test statistic values along with the p-values that detailed how significant the difference between the two clusters was. One of the significant characteristics of the Welch t-test is that it is used to determine differences in samples as the null hypothesis for the Welch t-test assumes that the means of the samples being compared are not equal. The Welch t-test was performed using the t.test() function in R and p.adjust() function was used to adjust the p-values that were obtained after the t-test using FDR correction.[5]

Mapping Probeset IDs to Gene Symbols Using Bioconductor Package: hgu133plus2.db

The t-test results of the differential expression matrix consisting of the probeset IDs were then mapped to gene symbols using the bioconductor package, hgu133plus2.db. Specifically, this package is the Affymetrix Human Genome U133 Plus 2.0 array. This process was implemented using the select() function with the keys being the probeset IDs and the columns being the databases symbols. Adding the mapped symbols to the probeset IDs resulted in “NA” values and duplicated mapped probeset IDs to gene symbols, which were both removed before continuing the analysis. The researchers in Marisa et al. (2013) chose the probeset with greatest variance out of the ones mapped to a single gene symbol, which is what was also implemented in this analysis by filtering results by the adjusted p-value and grouping the gene symbols together.

Differential Expression Analysis

Using the differential expression results and Chi-squared filtered results, the top 1,000 up- and down-regulated genes were selected based on the t-statistic. This was implemented using the slice_max() and slice_min() functions in R. Additionally, the top 10 up- and down-regulated genes were also sliced out.

Load in MSigDB Gene Sets using GSEABase

The KEGG, GO, and Hallmark gene sets were downloaded from MSigDB. MSigDB contains a collection of annotated gene sets. The KEGG gene set collection contains canonical pathways derived from the Kyoto Encyclopedia of Genes and Genomes. The GO gene set contains the gene ontology terms for biological processes, cellular components, and molecular functions. Lastly, the hallmark gene set represents well-defined biological states and processes. These gene sets will be useful when comparing overlaps between gene sets and top up/down-regulated genes. To load GMT files into R, the Bioconductor package GSEABase was used. Specifically, the get_Gmt() function loaded in my MSigDB files into R.

Building Contingency Tables

The results from the differential expression matrix with removed duplicated probeset IDs was used to build contingency tables, which were needed to implement Fisher's Exact Test. The contingency table for this analysis requires four numbers: number of differential expressed genes in the gene set, number of differentially expressed genes not in the gene set, number of not differentially expressed genes in the gene set, and number of not differentially expressed gene not in the gene set. A function that accepts a gene set from MSigDB, a gene set of differentially expressed genes, and a gene set of not differentially expressed genes was generated and returns a contingency table with the four required inputs. The gene sets of differentially expressed and not differentially expressed were generated using the GeneSet() function and the actual differentially and not differentially expressed genes symbols from the matrix with removed duplicated probeset IDs. Additionally, the KEGG, GO, and hallmark gene sets were each used in the contingency function to compare overlap between the annotated gene sets and differentially expressed genes extracted from this analysis. This was done by iterating through every element in the MSigDB gene sets to find overlap in the differentially expressed genes.

Fisher's Exact Test

The Fisher's Exact Test was performed on the KEGG, GO, and hallmark contingency tables using the fisher.test() function in R and the results from each test were held in a dataframe. The dataframe included the gene set name used, statistic estimate, and p-value. This was done by iterating through every element in the contingency tables and computed the Fisher's Test.

Statistical Analysis of Significantly Enriched Gene Sets

Using the results from the Fisher's Test for each gene set (KEGG, GO, and Hallmark), the p-values were adjusted using the Benjamin-Hochberg (FDR) procedure and the results were amended to the dataframes. This was implemented using the p.adjust() function and the 'BH' parameter in R. In order to find the statistically enriched gene sets, the nominal p-values were filtered using a 5% significance level. The top three gene sets from each gene set type were sliced out using slice_min(). These results were then compared to the results found in Marisa et al. (2013).

Results

We calculated the median of RLE and NUSE across the samples and used the histogram to visualize the distribution. Based on RLE function and NUSE function in affyPLM package, the data's distribution (RLE score and NUSE score) shown in Figure 1.For RLE, since we assume gene expression are not changing across samples, the median RLE value should be close to 0. For NUSE, standard error from the probe-level should be close to 1 across the samples. The frequency of median RLE and NUSE value for each sample shown in Figure 2 and Figure 3 with histogram.

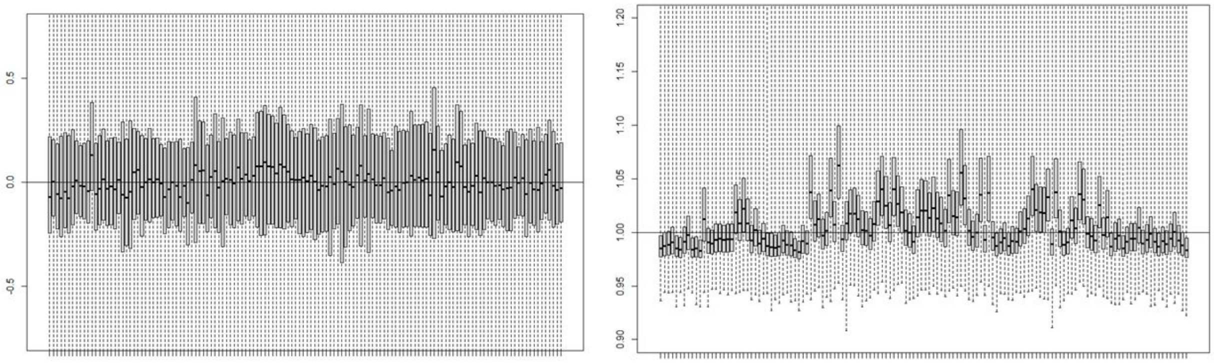

Figure 1 Left: RLE score for samples. Right:NUSE score for samples.

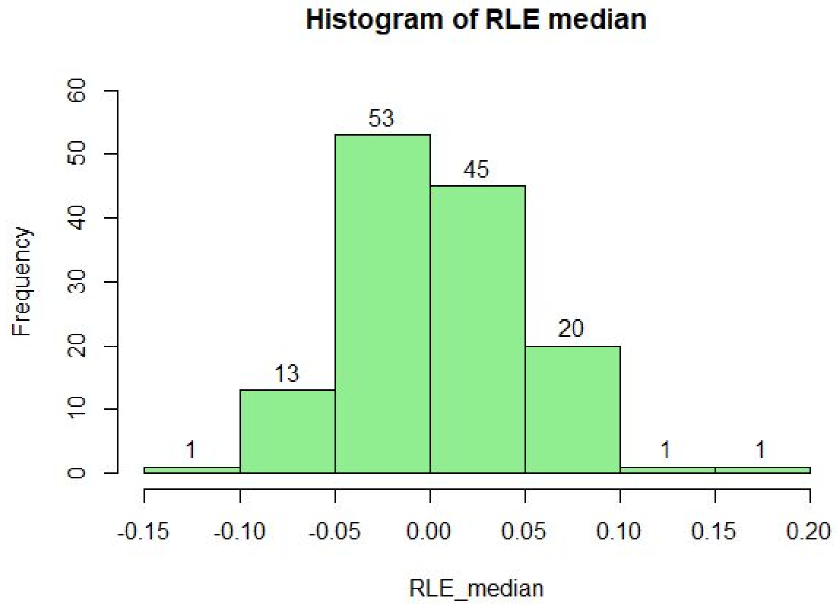

Figure 2 Distribution of median RLE score across the samples. 98/134(73.1%) samples are in the range of 0.05 close to 0.00. 131/134(97.8%) samples are in the range of 0.1 close to 0.00.

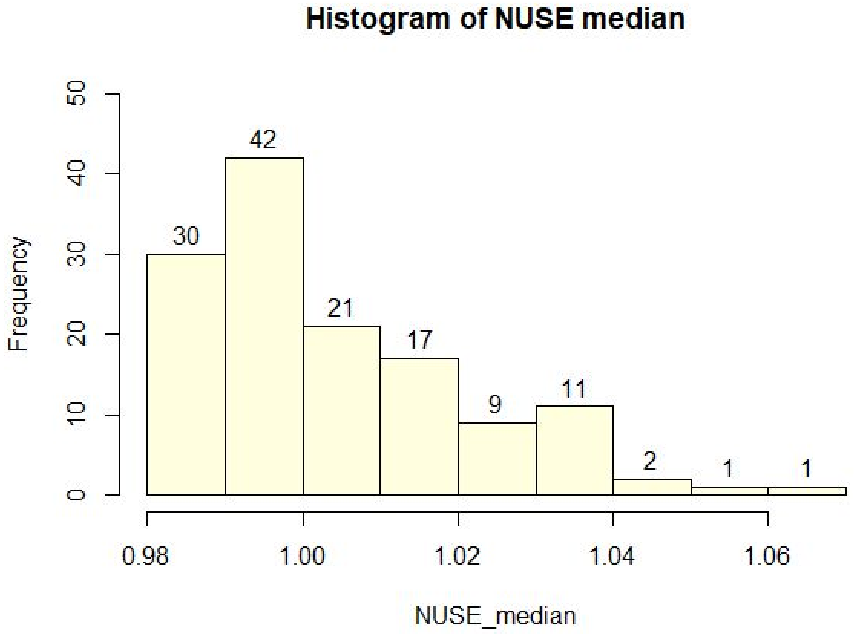

Figure 3 Distribution of median RLE score across the samples. 110/134(82.1%) samples have the median RLE score from 0.98 to 1.02 (0.02 close to 1). The higher RLE score means poor quality of data from 1.02 to 1.06. (24/134 = 17.9%)

A PCA analysis was conducted to analyse each principal component contributing how much to the dataset. PC1 vs PC2 graph plotted using the ggplot function shown in Figure 4. As shown in this PC1 vs PC2 plot, we could see that the points are distributed evenly. Here, we used the function fviz_eig() from package factoextra to visualize the contribution of each part from PC1 to PC10. From the result shown in Figure 5, PC1 took 11.3% and PC2 took 8.5% variance respectively, the other principal components took less than 5%.

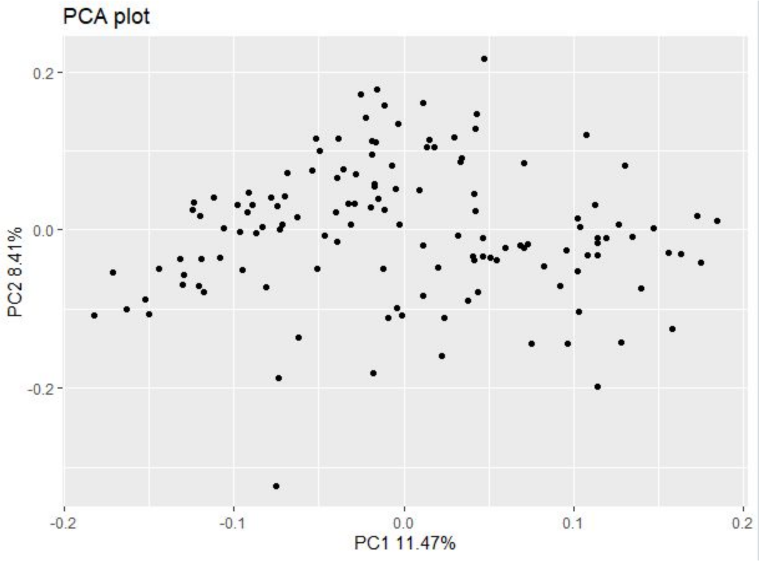

Figure 4 PCA plot PC1 vs PC2. The points scattered in the plot instead of gathering in one dimension because PC1 and PC2 only took about 10% of the variation.

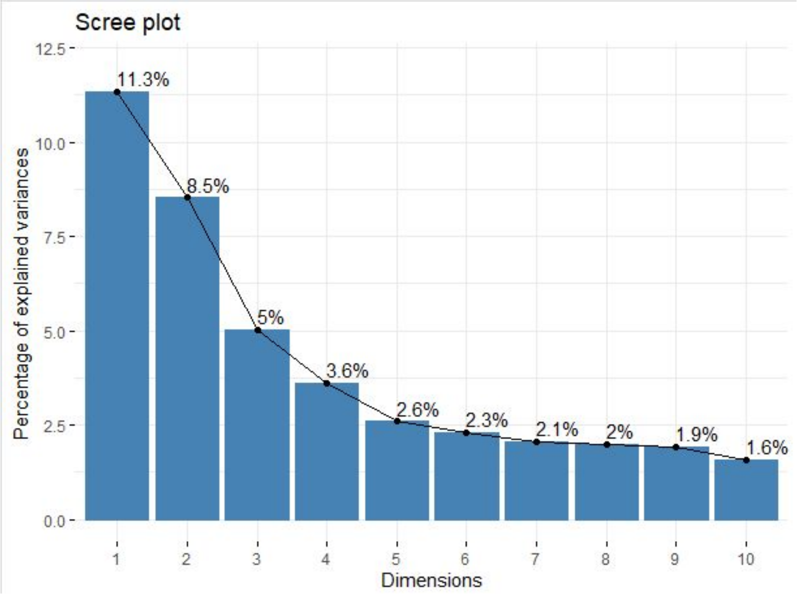

Figure 5 Percentage of explained variances. PC1 only took 11.3% and PC2 only took 8.5% of the variation. The other principal components took variation less than 5%.

The first filter (with the log2(15) and 20% samples) on the ComBat adjusted data resulted in 39678 probe-sets left in the dataset from an initial total of 54675 probe-sets. The second filter (with the chi-square distribution) resulted in 34537 probe-sets left. The third filter (with the coefficient of variation filter) was the most stringent one that resulted in just 1519 probe sets left from the previous pool of filtered genes. This is different from the reference paper by Marisa et. al. as that paper had a total of 1459 probe-sets remaining at the end of the filtration steps.[1] This could be due to our analysis primarily focusing on just two molecular subtypes : C3 and C4 while the original paper by Marisa et. al. focused on six subtypes.[1] The samples that were used in this analysis were also a mixture of the discovery and validation set samples from the original paper. This could affect the variances and other statistical aspects of the analysis due to the difference in samples and might be a reason why a different number of significant probe sets were observed in the reference paper and this analysis after the third filter was applied. Differences in statistical characteristics might also depend upon the range of chi-square distribution the variance test statistic was compared to as a distinctive method was not detailed in the original paper, however we discovered that choosing between a single-tailed and two-tailed chi-square distribution did not make a significant difference.

After the hierarchical clustering was performed and two clusters were generated, one cluster consisted of 53 samples while the other one contained 81 samples. The samples that were incorrectly clustered were: GSM972019, GSM972210, GSM972239, GSM972291, GSM972385 and GSM972412. This could be a result of the hierarchical clustering being used for this analysis instead of consensus clustering that used in the original paper. It could be also due to the combination of samples failing to properly cluster without a reference validation sample set and a difference in the expression values.

Finally, after the Welch t-test was performed and the obtained p-values were adjusted, a significance threshold of 0.05 was applied to get an estimate of the truly significant probe sets. This resulted in just 1236 probe-sets remaining that were significant. Some of the most differentially expressed significant genes were “204457_s_at”, “223121_s_at”, “223122_s_at”, “227059_a”, “213413_at”. This selection was based on the criteria of genes having the highest t-test statistic value and the lowest adjusted p-value. The genes that best represented the C4 cluster were “204457_s_at”, “223121_s_at”, “223122_s_at” as they were significantly upregulated compared to the C3 cluster and had a significant p-value. “209955_s_at”, “214261_s_at”, “224588_at” were some of the genes that represented the C3 cluster the best as they had a higher t-test statistic value compared to the other cluster. This is evident from the heatmap as well as there is a clear demarcation in the expression patterns between two clusters and the t-test statistic values and the low adjusted p-values further bolster this theory.

The t-test results of the differential expression matrix consisted of 34,537 probeset IDs and their corresponding calculated t-test statistic, p-value, and adjusted p-value. After mapping the probeset IDs to gene symbols from the hgu133plus2.db, we had a total of 37,147 mapped probeset IDs to gene symbols. Subsequently, after removing all the "NA" values from the matrix, we ended up with 32,521. Consistent with the findings in the paper, multiple probe sets mapped to the same gene symbol. In order to correct this, we removed the probes with the greatest variance, which is what was done in the paper. In the supporting information of Marisa et al. (2013), it states, "For all signatures used, genes were matched to our probe sets by the Gene Symbol annotation and only the most variant probe set (maximal rCV) was selected." After selecting the most variant probes, we had a total of 17,287 unique probe to gene symbol matches. Specifically, we had a total of 8,663 differentially expressed genes and 8,623 not differentially expressed genes. Both of which were statistically significant based on the adjusted p-value.

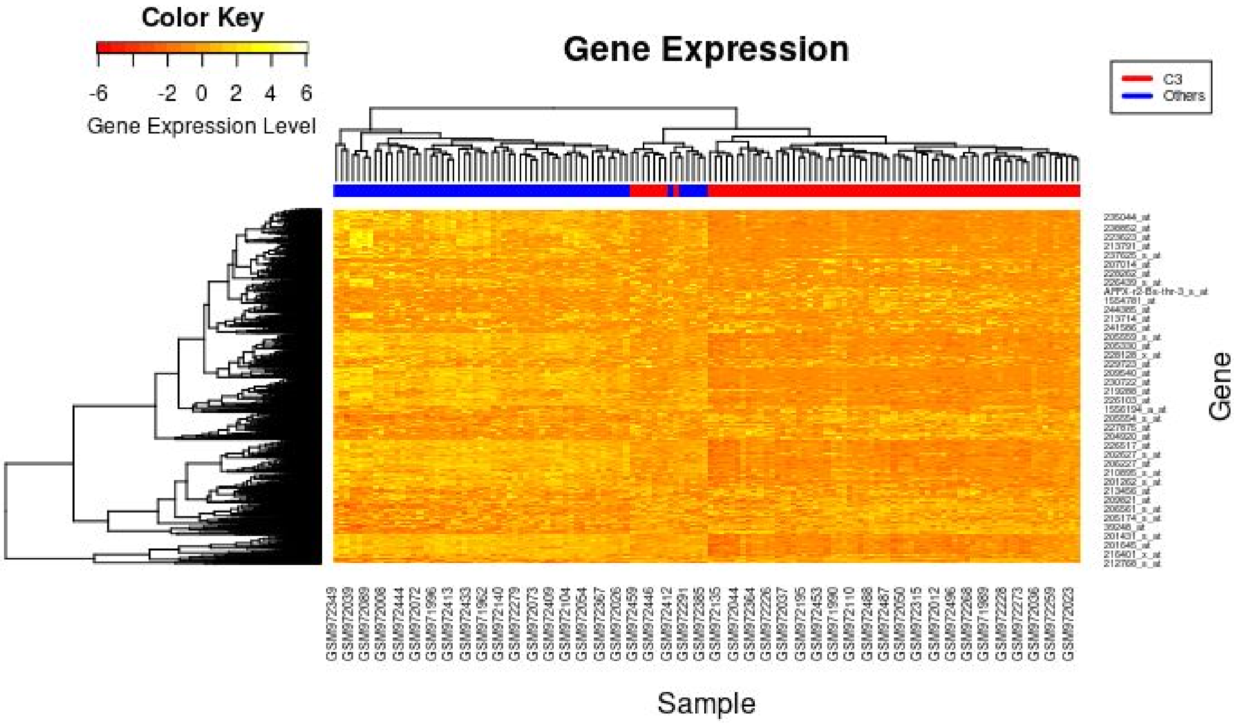

Figure 6: Heat map of the gene-expression of the 1519 probesetes (y-axis) across 134 samples (x-axis). The heatmap details the clustering of samples into 2 main categories and the gene expression levels for every probe-set. The column bar is color coded in red to indicate C3 samples and blue for others. The intensity of the color portrays gene expression level and corresponds to the Color Key legend provided indicating red for downregulation and yellow for upregulation.

Using the uniquely mapped probe-gene symbols, we selected the top 10 up- and down-regulated genes by using the t-statistic as the filtering parameter (Table 1). The top up-regulated and down-regulated genes were consistent with the most positive and negative t-statistic, respectively.

Table 1. Top and Bottom 10 Up- and Down-Regulated Genes. Up-regulated genes are denoted in red and down-regulated genes in blue.

Gene Symbol	**T-Test Statistic**	**P-Value**	**Adjusted P-Value**
GSDME	20.3033367554875	4.97911244260508E-38	5.21101837667429E-35
ARMCX1	20.2113758874161	7.29050504099791E-41	3.14740215751181E-37
COLEC12	20.1891710257631	1.18247922672344E-34	3.96497913139297E-32
NXN	20.0929427893491	1.16264339012998E-40	3.6503831604472E-37
SFRP2	19.9172836839369	1.45175561863521E-39	2.63890967372653E-36
FRMD6	19.7193601053122	1.43915542443423E-40	3.7149661777204E-37
SERPING1	19.3975922978825	1.27106020001626E-39	2.43881145155343E-36
SPOCK1	19.385541907646	5.81723912737999E-40	1.25568742338952E-36
ZFPM2	19.0610824935453	2.42320814127569E-33	5.81182913716934E-31
NDN	18.8478255016116	2.13518218679968E-38	2.63367097091074E-35
LRRC31	-12.7933431069625	1.74986213484781E-24	1.05287436500416E-22
FCGBP	-12.540963858046	1.95455036384631E-22	9.28807856397099E-21
ST6GALNAC1	-12.5226267031137	3.3598122940777E-21	1.3587568758848E-19
C4orf19	-11.8297141157842	1.89136593562142E-19	6.19166875057414E-18
CRYM	-11.6406924997473	1.83802884317493E-21	7.76989010486322E-20
NRARP	-11.5933936787504	2.52382053663349E-19	8.16918368076014E-18
MRAP2	-11.4166839788648	1.59155217720811E-20	5.90412862988576E-19
NXPE1	-11.3794411176717	5.86760736044167E-21	2.30283585690425E-19
NR3C2	-11.3085271746739	6.03854279877877E-19	1.85380580125709E-17
FOXA3	-11.2030480353915	2.90276634815245E-19	9.33452899126083E-18

We used the gene sets databases downloaded from MSigDB in the next part of our analysis. The KEGG, GO, and hallmark gene set databases contained 186, 14,765, and 50 gene sets, respectively. We generated contingency tables for each gene set within each gene set type and implemented Fisher's Exact Test. From these results, we found the total number of significantly enriched gene sets within each gene set type. The total number of significantly enriched gene sets in KEGG, GO, and hallmark gene set types were 42, 1397, and 30, respectively. This added up to be 1,469 significantly enriched gene sets. From each gene set type of significantly enriched gene sets, we selected the top three statistically significant (<0.05) enriched (Table 2).

Table 2. Top Three Significantly Enriched Gene Sets for each Gene Set Type. KEGG, GO, and Hallmark gene sets are colored green, yellow, and purple, respectively.

Gene Set	**Statistic Estimate**	**P-Value**	**Adjusted P-Value**
KEGG_DRUG_METABOLISM_CYTOCHROME_P450	2.80890872008474	0.000258506632614382	0.00801370561104584
KEGG_TRYPTOPHAN_METABOLISM	4.4895356921107	0.00032042273000624	0.00851408968302295
KEGG_PROTEASOME	0.273960910400999	0.000452911092991541	0.0105301829120533
GO_PEPTIDYL_LYSINE_ACETYLATION	0.524022570171075	0.000102655424838837	0.0103109343384043
GO_TRANSCRIPTION_COACTIVATOR_ACTIVITY	0.594429432256152	0.000106200204921836	0.0105949055788575
GO_CELL_CELL_ADHESION	1.33675849083288	0.000107953284514137	0.0106975184285318
HALLMARK_INFLAMMATORY_RESPONSE	1.62719087296858	0.00182339291574123	0.0151949409645103
HALLMARK_ESTROGEN_RESPONSE_LATE	1.57566076660704	0.00256329712557216	0.0174662991894616
HALLMARK_ALLOGRAFT_REJECTION	1.58818384672155	0.0030346976953996	0.0174662991894616

Discussions

In this project, the data consisted of 134 samples which were sourced from human tissue of primary colorectal adenocarcinoma and referenced in the Gene Expression Omnibus (GEO) database under accession number GSE39582. These 134 samples satisfied the RNA quality control measurements and were taken from the combined discovery and validation sets discussed in Marisa et al, 2013. The protocol for extracting total RNA samples ensured that the type was of high-quality. The total RNA underwent the sample preparation protocol implementing cRNA target generation and hybridization to the Affymetrix Human Genome U133 Plus 2.0 Arrays. Then, we read in the 134 CEL files for probe level data, did data normalization and saw the distribution of the dataset with RLE and NUSE. After correcting the batch effects, Principal component analysis was run on the normalized dataset. We then filtered noise, did clustering analysis, subtype discovery and gene set functional enrichment.

In the step of data preprocessing, we used methods like RLE and NUSE to visualize the data distribution across the samples of micarrays and to evaluate the quality of the expression array. For RLE, we assume gene expression is not changing across samples, so the median RLE value should be close to 0. For NUSE, standard error from the probe-level should be close to 1 across the samples. By the result shown in results the median value for RLE value around 0 and the median value of NUSE around 1, which indicates a good gene expression of the microarray data. With the Principal component analysis, the first two principal components PC1 and PC2 contributed about 20% variance. For microarray probes' level data, PCA can help us to show the clustering and the relationship between the genes and reduces the dimension of the dataset if we need. The noise filtering and dimensionality reduction process included filtration of probe-sets with low expression values and those that weren't expressed enough to have a consensus. After that, probe-sets with significantly low variance based on different criteria like chi-square test and coefficient of variation were filtered out. To perform exploratory analysis on the processed dataset, hierarchical clustering was used to cluster the samples into distinctive groups. A heatmap was plotted to visualize the clustering and the expression values of genes across samples. Furthermore, Welch's t-test was performed to further identify genes that were significantly different between two clusters and an adjusted p-value was calculated using FDR. 1044 genes in our dataset were identical to the 1459 genes found in the reference paper dataset after the third filter was applied. The clustering was also able to cluster samples accurately for the most part with some exceptions (6 samples). This signifies that the filtering steps applied were successful in retaining the genes that gave the clusters its unique signature.

In order to reveal statistically significant enriched gene sets within our tumor classification data, we needed to first map our differential expression matrix probeset IDs to gene symbols in annotated gene sets. Specifically, we used KEGG, GO, and Hallmark gene sets. After selecting the most variant probes and unique gene symbols pairs, we found the top up/down-regulated genes in the dataset. From here, we generated contingency tables for all the differentially and not differentially expressed genes. We then performed Fisher's Exact Test to find overlap between our gene sets and the annotated gene sets. This produced statistically significant enriched gene sets, which revealed important biological functions of our samples.

Our main biological findings include the statistically significant up- and down-regulated genes and the enriched gene set overlap between annotated gene sets and the ones we analyzed. Significantly high and low expression of genes implies relevant cellular activity that can reveal a multitude of biological processes. Specifically looking at CRC tumors, these findings can help answer fundamental questions regarding the pathological staging of CRC. In the paper, researchers found 57 probeset IDs corresponding to unique gene symbols compared to the 17,287 in our analysis. Their 57 unique genes can be found in Table S4 in the paper. Just looking at our top 10 up- and down- regulated genes, two up-regulated and two down-regulated are consistent with those found in the papers 57. The up-regulated genes are COLEC12 and FRMD6. The down-regulated genes are FCGBP and C4orf19. A possible explanation as to why our data differs from the paper is that the dataset we used is merged; however, in the paper, researchers split their data into discovery and validation sets.

Our results from the gene set enrichment analysis revealed significantly KEGG and GO pathways related to cancer hallmarks. Some pathways include: cell communication, growth/death, immune system, motility, replication and repair, angiogenesis, metabolism and main cancer signal transduction pathways. Our top three enriched KEGG pathways represent drug metabolism of cytochrome P450, tryptophan metabolism, and proteasome. Our top three Gene Ontology pathways reveal that we had significantly differentially expressed genes that acetylates peptidyl-lysine, activate transcription of specific genes, and generate cell adhesion. Our top three hallmark biological processes reveal we had genes in our data set that define an inflammatory response, a late response to estrogen, and gene up-regulated during transplant rejection. Our enriched gene sets are consistent with some of the pathways of cancer hallmarks. Compared to Figure 2 in the paper, none of the top three enriched gene sets in our data matches what was found in the paper; however, if we look at all of our KEGG and GO enriched gene sets, there are multiple overlaps. This discrepancy could be due to the different statistical methods used in the paper or the merged dataset we used for this project. Through this analysis, we were able to match biological pathways and processes to genes in the microarray data extracted from tumor samples.

Conclusion

According to the gene expression data, we were successfully able to replicate the sample clustering that was detailed in the paper with a 95.53 % accuracy. We were also able to successfully implement the filters described in the paper accurately as the number of genes that remained after filtering were very close to the genes described in the paper. Over 71.5% of the genes that remained after the third filter was applied were also found in the list of probe sets after the third filter in the reference paper. Some of the shortcomings of this analysis could be the result of ambiguity in some of the statistical methods used by Marisa et al. (2013) like the chi-square test. We overcame this difficulty by extensively testing using different statistical approaches and reading different literature. Another reason for the differences in the results between our analysis and the reference paper would be the clustering method used as hierarchical clustering was used for this analysis unlike the consensus clustering method that was used in the reference paper. The data being processed in this study could also be the cause of the differences in results between our analysis and the referenced paper. However, we were still able to replicate the results from the paper with a great accuracy despite the differences in the dataset and methods. Discrepancies between our analysis and the paper's analysis could also be due to the difference in data. Furthermore, the original study from the referenced paper focused on patients in France. However, if there were to be a more heterogeneous population being included in the analysis, we would have a more robust model that had a lower chance of bias and overfit. We would also need more covariates like lifestyle, eating habits and physical activity to better profile the signature of the molecular subtypes as that could also be a reason for overfit in a sample set.

Overall, utilizing microarray data sourced from primary tumor samples from patients with stage I to IV colon cancer and referenced in the Gene Expression Omnibus (GEO) database, we were able to reveal significantly enriched pathways of gene sets. This functional enrichment analysis discovers important biological functions that can help better improve the clinical treatment of colon cancer. Specifically, we found biological pathways and processes that are consistent with related cancer hallmarks. Our functional enrichment analysis was somewhat consistent with the findings in the paper. We had overlap with some of the biological pathways and processes found. Some challenges we encountered while performing the analyses needed to reach our functional enrichment goal included implementing unfamiliar packages and working with certain concepts in R that we had not used before. We were able to overcome these challenges by reading documentation on the statistical and bioconductor packages. Specifically, the biologist was required to use the hgu133plus2.db database to map probeset IDs to gene symbols. There were many useful online tutorials and documentation manuals available to help perform the mapping correctly.

References

1. Marisa, L., Reyniès, A. de, Duval, A., Selves, J., Gaub, M. P., Vescovo, L., Etienne-Grimaldi, M.-C., Schiappa, R., Guenot, D., Ayadi, M., Kirzin, S., Chazal, M., Fléjou, J.-F., Benchimol, D., Berger, A., Lagarde, A., Pencreach, E., Piard, F., Elias, D., … Boige, V. (2013). Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value. *PLOS Medicine*, *10*(5), e1001453. https://doi.org/10.1371/journal.pmed.1001453
2. Greenlee, R. T., Murray, T., Bolden, S., & Wingo, P. A. (2000). Cancer statistics, 2000. *CA: a cancer journal for clinicians*, *50*(1), 7–33. https://doi.org/10.3322/canjclin.50.1.7
3. *GEO Accession viewer*. (n.d.). https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM971958
4. Gandolfo, L. C., & Speed, T. P. (2018). RLE plots: Visualizing unwanted variation in high dimensional data. *PloS one*, *13*(2), e0191629. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0191629
5. The R Project for Statistical Computing https://www.r-project.org/
6. gplots: Various R Programming Tools for Plotting Data https://cran.r-project.org/web/packages/gplots/index.html
7. Cancer staging—National cancer institute. (2015, March 9). [CgvArticle]. https://www.cancer.gov/about-cancer/diagnosis-staging/staging