Individual Project: Microarray Based Tumor Classification

Sooyoun Lee

Group: Saxophone

TA: Jing Zhang

GitHub repository:
https://github.com/leesu21/Sooyoun_Lee_Saxophone_BF528_Individual_Project

INTRODUCTION

Colon cancer (CC) is the most common type of cancer that is diagnosed for both males and females each year in the United States. CC is hard to accurately predict about the recurrence rate and there is no gene expression signature proven to be reliable for prognosis stratification in any clinical practice [1].

Doctors and many other research institutes are working to learn more about colon cancer, ways to prevent it, how to best treat it, and how to provide the best care to people diagnosed with this disease. Researchers are developing tests to analyze the stool samples to find the genetic changes associated with colon cancer. This has done by finding and removing the polyps or identifying cancer at the early stage, doctors have a better chance of curing the disease [5].

Marisa *et al* focuses on the molecular classification of the CC on the mRNA expression profile analyses and gathered clinical and pathological data of 750 patients with stages I to IV CC between 1987 to 2007. Overall, the microsatellite instability (MSI) found to have the most defective function of colon cancer however technologies need to improve and refine [1].

In this project, we have focused on comparing and identifying the C3 and C4 tumor subtypes. By using the 134 tumor samples, we have performed noise filtering, dimensionality reduction, hierarchical clustering, subtype clustering, and gene set analysis of the samples.

METHOD

Noise Filtering and Dimensionality Reduction

The method of noise filtering and dimensionality reduction was selected to analyze the data effectively. The sample expression data contains a total of 54,675 probes. Genes expressed under 20% of samples were removed which could be interpreted as that at least 20% of the gene-expression values must be higher than $log2(15)$. After the first filtering, the result returned to 22,028 probes.

In the second filtration process, the chi-square test was involved where the variance of a population was equal to the specified value [2]. The qchiq() function was used to allow the specific desired area in the tail and the number of degrees of freedom [3]. By using this function, it will compute the required x-value to get the specified area in the specified tail with a given number of degrees of freedom. The degrees of freedom are calculated by N-1 where N represents the number of columns in the original data, which will be 134-1=133. After the second filtering, genes that do not pass the threshold of $p < 0.01$ will be removed from the gene set.

For the last step of the filtration, the coefficient of variation was calculated across the samples. We filtered the genes with a coefficient of variation that is less than 0.186. As a result, a total of 1,534 probes have remained after the third filtration.

Hierarchical Clustering & Subtype Discovery

Cluster analysis is an important data mining method to discover in multidimensional data [4]. It identifies the patterns and the groups of similar objects within the data set of the interest. In this project, the tumor classification of C3 and C4 subtypes will be focused on and analyzed further. After filtering the data, the hierarchical clustering and the subtype discovery was performed. The hierarchical clustering was run by using the hclust() function to analyze the set of dissimilarities and methods for analyzing it. As a result, the dendrogram was created which contains two clusters: one with 77 genes and another with 57 genes.

Then the heatmap was obtained to graphically represent the data where the color intensity represented the respective gene expression level.

In order to identify the differentially expressed genes in two different clusters, we determined how many genes are differentially expressed at adjusted p<0.05 between the clusters for both lists. A Welch t-test was used to examine whether both groups are sampled from a normal distribution with equal variances. As a result, the p values were adjusted by using the t.test(), and the p.adjust() functions.

Gene Set Enrichment

By using the Bioconductor package of hgu133plus2.db package, we mapped gene symbols to the probeset IDs. We have analyzed the dataset of the probes corresponding to differentially expressed results by matching each probe corresponding gene symbol. By comparing the GO, KEGG, and Hallmark gene sets and two sets of 1000 probes, we determined how many up- and down-regulated genes would appear in each gene sets. Then by using the fisher.test function, we computed the hypergeometric statistics and p values comparing overlap for each gene set and each of the top 1000 increased and 1000 decreased genes. These gene sets contain 10182, 49, and 85 gene sets respectively.

RESULTS

In this project, our goal was to analyze and focus on reproducing the results from the comparison of the C3 and C4 tumor subtypes. The project was conducted into a two-phase design, first, an initial set of discovery samples were used to identify the patterns among the samples, and second, separate validation samples were used to test whether the results from the discovered sets were robust.

The noise filtering was used to delete the genes that do not express differentially across all of the samples. We have obtained 22,027 probes with expression intensity in at least 20% samples from 54,675 genes, and 15,508 genes were obtained with significant variation across all samples. Finally, after the third filter, we kept 1,533 genes that use to classify C3 and C4 cancer subtypes. After the noise filtering and the dimensionality reduction, the dendrogram for the hierarchically clustered samples was produced (Figure 1). The 134 samples represented in the dendrogram are separated into two clusters, one with 57 genes and the other with 77 genes. After conducting a Welch t-test, a set of 1249 significantly differentially expressed genes was obtained with an adjusted p-value below 0.05.

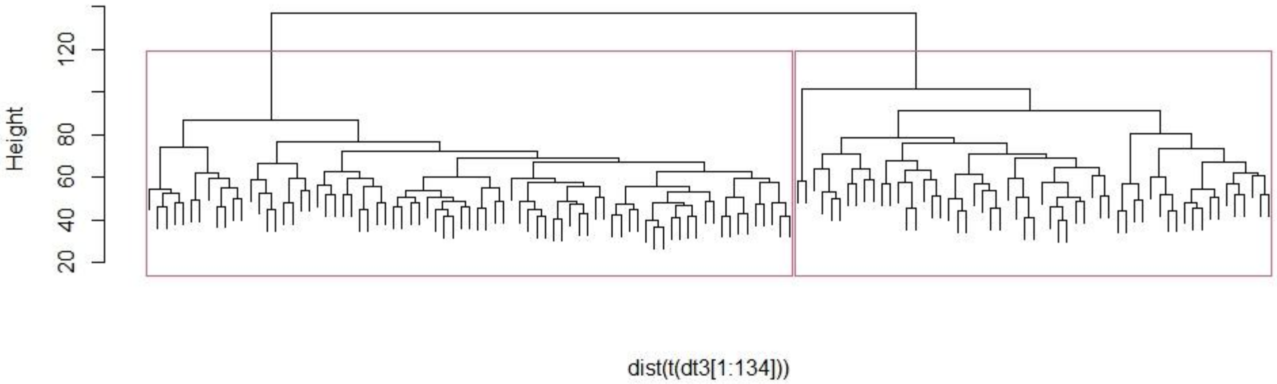

Figure 1. Dendrogram for the hierarchically clustered samples. All 134 sample GEPs are represented in the figure and the matrix is divided into two clusters: 77 genes as one cluster on the left side of the red box and another 57 genes as second clusters on the right.

To visualize the difference in gene expression profile between the C3 and C4 subtypes, a heatmap was created with 134 samples (Figure 2). The samples with the C3 subtype were colored in blue and the C4 subtype was colored in blue. The heatmap was divided into two clusters and then the Welch t-test was used on the two clusters on each probe.

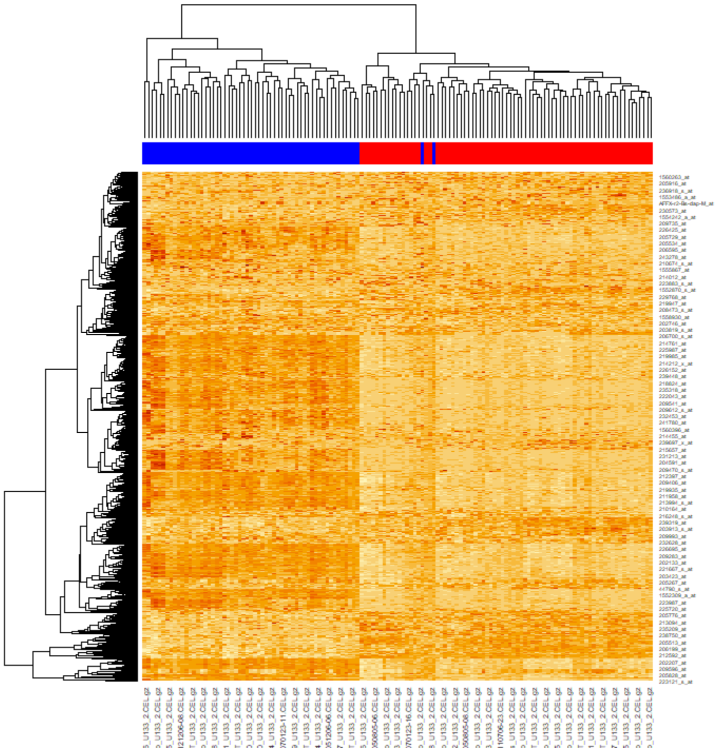

Figure 2. Heatmap of all filtered 134 samples. The red-colored represents subtype C4 and the blue-colored represents subtype C3 at the top of the figure.

The gene set enrichment analysis was performed after we have obtained a list of the most significantly differentially expressed genes. For the top 10 up-regulated genes (Table 1), we found that gene symbols such as STON1, SPOCK1, and MSRB3 were mentioned in the paper as the discriminant probe sets [4]. For the top 10 down-regulated genes (Table 2), a list of three gene symbols FCGBP, ST6GALNAC1, and LRRC31 was mentioned in the reference. The reference also shows that the epithelial-mesenchymal transition is upregulated in C3, and downregulated in C4 which is why we found both up-and down-regulated genes.

Table 1. Top 10 Up-Regulated Genes based on the t-test statistic value. The * represents the overlap with the results reported in the reference paper.

Probeset ID	T-Test Statistic	P value	Adjusted p value	Gene Symbol
207266_x_at	20.742112325	3.03E-43	6.69E-39	RBMS1
225651_at	20.34622276	1.76E-42	1.29E-38	UBE2E2
213413_at *	20.22100649	1.34E-38	1.84E-35	STON1
202363_at *	19.84792460	1.36E-40	4.99E-37	SPOCK1
225782_at *	19.66125008	6.83E-41	3.00E-37	MSRB3
212607_at	19.61426824	1.51E-38	1.96E-35	AKT3
225021_at	19.58786095	4.67E-39	7.91E-36	ZNF532
227561_at	19.36832050	2.64E-39	5.29E-36	DDR2
209210_s_at	19.26208339	3.84E-40	1.05E-36	FERMT2
227059_at	19.16760712	4.33E-39	7.91E-36	GPC6

Table 2. Top 10 Down-Regulated Genes based on the t-test statistic value. The * represents the overlap with the results reported in the reference paper.

Probeset ID	T-Test Statistic	P value	Adjusted p value	Gene Symbol
235350_at	-12.01387256	3.31E-21	9.02E-20	C4orf19
205489_at	-12.27246385	2.20E-23	7.85E-22	CRYM
211715_s_at	12.60580971	2.25E-23	8.01E-22	BDH1
214106_s_at	-12.75664924	9.69E-22	2.82E-20	GMDS
1568598_at	-12.76623720	1.88E-24	7.61E-23	KAZALD1
203240_at *	-12.90118227	1.35E-23	4.99E-22	FCGBP
218189_s_at	-13.09737298	5.78E-25	2.49E-23	NANS
222764_at	-13.11062327	1.48E-24	6.11E-23	ASRGL1

227725_at *	-13.66044649	9.50E-24	3.59E-22	ST6GALNAC1
220622_at *	-14.09766031	7.06E-28	4.60E-26	LRRC31

By using the GSEABase Bioconductor package, we checked significant gene enrichment. The gene enrichment analysis was performed on each genesets using the fisher test. As a result, GO genesets contains 10,182 gene sets, KEGG contains 49, and Hallmark contains 85 gene sets respectively. The top three enriched termf of each gene sets are selected (Table 3).

Table 3. Top3 enriched gene sets for GO, Hallmark, and KEGG genesets. The top 3 biological pathways that were enriched in genesets are represented as the BH values.

Geneset	P value	Estimate	Adjusted p value (BH)
GO			
GO_MITOCHONDRIAL_GENOME_MAINTENANCE	1	0.61215652	1
GO_REPRODUCTION	0.01946937	1.33828058	0.18669532
GO_SINGLE_STRAND_BREAK_REPAIR	1	0	1
HALLMARK			
HALLMARK_TNFA_SIGNALING_VIA_NFKB	0.00417823	1.92080893	0.01671295
HALLMARK_HYPOXIA	0.00001918	2.53330015	0.00011990
HALLMARK_CHOLESTEROL_HOMEOSTASIS	0.65672112	0.72182723	0.81076681
KEGG			
KEGG_N_GLYCAN_BIOSYNTHESIS	1	0.73446074	1
KEGG_OTHER_GLYCAN_DEGRADATION	1	0	1
KEGG_O_GLYCAN_BIOSYNTHESIS	1	0.52456508	1

DISCUSSION

Our study largely reproduced the results from Marisa *et al*. Comparing the genetic signatures of the C3 and C4 tumor subtypes. After conducting the three different filtration processes, we were able to classify these samples into two different clusters correspondingly whether it is the C3 or the C4 subtypes. The number of probe sets that passed the three filtration processes was 1,533, close to the number, 1,459, reported by Marisa *et al*. This shows that the clustering results in our project show high consistency with the results in the reference. Also, the number of genes from each cluster group; 77 genes in one cluster and another 57 genes in the second cluster, was exactly matched with the figure from Marisa *et al*. However, when performing hierarchical clustering of the 134 samples, we noticed that two samples of C4 subtype exhibit genetic signatures close to those of C3 subtype, indicating a possible samples mix-up during library preparation.

The gene set enrichment process shed light on the biological relevance of our results. From Marisa *et al*'s paper, we found 6 out of 20 gene symbols that overlap with our results, including STON1, SPOCK1, and MSRB3 from the top 10-upregulated genes, and FCGBP, ST6GNLNAC1, and LRRC31 from the top 10 down-regulated genes. We have also determined the gene set enrichment analysis for specific gene sets and as a result, the GO gene contains 10,182 gene sets, KEGG contains 49, and Hallmark contains 85 gene sets. The C3 subtypes were associated with the down-regulated ECM process and also the C4 subtype tumors were prone to metastasis [1]. Through this process, we have found that the biological process is related to cell death and its growth which were up-regulated in the C3 subtypes but this was not statistically shown in Marisa *et al*'s paper. However, we can assume that genes that are differentially expressed in the C3 are less significant when compared to the normal cells.

REFERENCES

[1] Marisa et al. Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value. PLoS Medicine, May 2013.

[2] 1.3.5.8. Chi-Square Test for the Variance. (n.d.). https://www.itl.nist.gov/div898/handbook/eda/section3/eda358.htm.

[3] Probability: Chi-Squared Distribution. (n.d.). http://courses.wccnet.edu/~palay/math160r/prob_chisq.htm.

[4] Alboukadel, & Khongfak, S. (2019, December 25). *The Ultimate Guide to Cluster Analysis in R*. Datanovia. https://www.datanovia.com/en/blog/cluster-analysis-in-r-practical-guide/.

[5] *Colorectal Cancer - Latest Research*. Cancer.Net. (2021, May 5). https://www.cancer.net/cancer-types/colorectal-cancer/latest-research.