Individual Project Write-Up

Contributor:

- Chillamcherla Dhanalakshmi Srija

Introduction

For the final project, the paper "An analysis of microarray dataset that generated and described in Marisa, et. al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value" has been used, and the roles of analyst and biologist have been explored further. Colorectal cancer (CRC) is one of the most common causes of death globally. In its final and advanced stages, the prognosis of the disease is detrimental. To classify colorectal cancer, the pathological stages are used in a clinical setting. This classification method has not been proven to yield the best results, so newer classification methods have been employed. This paper aims to classify CRC based on miRNA expression profile analysis. In this project, the processed data has been analyzed and interpreted. The noise reduction to arrive at a smaller feature containing dataset is essential for an accurate biological interpretation. The biologist's role is important to interpret the significance of the up and down-regulated genes biologically.

Methods

The processed data is analyzed for the project. The disproportionate number in the total samples and features causes the development of noise and this will impede the accuracy of the study. To reduce this, the data is normalized using a multitude of methods and filtered to reduce the discrepancy in the number of samples and features. This is done in the R version **4.2.0** (Vigorous Calisthenics). Many libraries in the R have been used for this. Mainly three filters have been employed and the data is passed through these. The first filter is to get the data that is expressed in at least 20% of the samples and has a value that is greater than log2(15). The second filter is to apply the variance throughout the dataset for each gene and then the median of the variance is calculated using a threshold of p less than 0.01. This is done by using the chi-squared test as qchisq() function in R. The third filter is the coefficient of the variation of the genes passing through this filter should a value that is greater than 0.186. This completes the first set of normalization and filters for the analyst role.

Table 1. Showing the number of genes through the various steps of filtering

Dataset	Number of genes
Initial (processed data)	54675
Filter 1	39750
Filter 2	15088
Filter 3	1658

For the second part of the analyst role, the filtered data is clustered based on a criterion. Clustering helps in identifying the new relationship between the data and helps discover the patterns. In the paper, consensus clustering was employed but due to it being very computationally intensive, hierarchical clustering is used here. hclust() function in R is utilized to perform the clustering. The number of clusters in the dendrogram is 2.

Table 2. Number of clusters in the hierarchical clustering

Cluster	Number of samples
Cluster 1	80
Cluster 2	54

Using heatmap() function, a heatmap of the gene across all the samples are produced and grouped based on the "C3" subtype. The heatmap was colored using cit-coloncancermolecularsubtype. The number of differentially expressed genes between the two clusters is identified using a Welch t-test. The resulting dataset would contain probeset ID, t-statistic, p-value, and adjusted p-value. p.adjust() function along with the method as "FDR" is used to create the last column to assess the differentially expressed genes.

Table 3. Number of differentially expressed genes

"Number of probes left with adjusted p-value < 0.05	1378

The same process was repeated to find differentially expressed genes using the welch t-test but this time the dataset is filtered with two filters instead of the three filters as we did at the beginning of the analyst role. Passing the dataset through these stringent filters helps in producing a cleaner dataset for the furtherance of the project methods.

The second chosen role for the project is the biologist role. This role utilizes the results from the analyst role to understand and decode the biological significance of the genes. A Bioconductor package is used to get the gene symbols for the probeset IDs from the analyst's role. To avoid duplicates in the results, the unique results along with the removal of the blank rows from this dataset. Gene sets from KEGG, GO and Hallmark pathways are downloaded from the MSigDB. These gene sets are downloaded in the form of GMT files. The top 1000 up- and down-regulated genes from the dataset are separated and written out into two files. The top 10 up- and down-regulated genes are written in tables. GSEABase package of R is used to load the gene sets that were downloaded from MSigDB.

Fisher.test() function is used to compute hypergeometric statistics and p-values. A contingency table was created to achieve this. The contingency table is made by calculating the differentially expressed genes from the downloaded gene sets, differentially expressed genes that have not passed the filters, the up-and down-regulated genes, and genes that contain all the values. Empty lists are initiated for the three pathway gene sets and the contingency table is modified along with a for loop to modify for all the three gene sets. The results from these are written out into data frames. The p-value is adjusted using the Benjamini Hochberg method and storing the data in separate files using the method "BH" and p.adjust() function. These files are written into comma-separated value(CSV) files for the project.

Results

The data is normalized and filtered based on different criteria. Three different filters were employed to achieve this. The filters are performed sequentially where the subsequent input was the output from the previous filter. The number of genes passing the filters reduced as the number of filters increased. This implies the presence of noise and contamination of the data. As we have seen in table 1, the number of genes that passed all the three filters is 1658 which significantly differs from the results in the reference study, 1459.

The stringent filters are important for proper and thorough data cleaning. After the filtering is done, the data is clustered using the hierarchical clustering method. Clustering is an unsupervised method. The cluster is cut into two main clusters, containing 80 and 54 in the first and second clusters respectively. This is different from the reference study as well it could be owed to the difference in the number of the genes passing all the filters since the paper used consensus clustering unlike the hierarchical clustering in this project. C3 and C4 subgroups of the clusters are used to color the heatmap.The threshold is 0.05 and the genes that passed through this filter were mapped into the heatmap. heatmap() function is used for this process. The number of probes left with an adjusted p-value < 0.05 is 1378.

Bioconductor package hgu133plus2.db is used to match the gene symbols with the probeset IDs. Unique and blank rows were removed through the filtering process. The top 1000 up-and down-regulated genes are written out into files. The top 10 up-and down-regulated genes from the two previously mentioned files are further studied to see the patterns. Tables 4 and 5 show the top 10 from up and down-regulated genes.

Figure 1. Heat map of the gene expression of the 1658 probesets (y-axis) across 134 samples (x-axis). The heatmap visualizes the clustering of the samples between the two cancer subtypes (C3 and C4). The color bar on the top of the heatmap identifies which of the 134 samples belong to either the C3 (red) or C4 (blue) cluster. The darker colors depict lower gene expression, yellow for downregulation, and red or dark orange for upregulated genes.

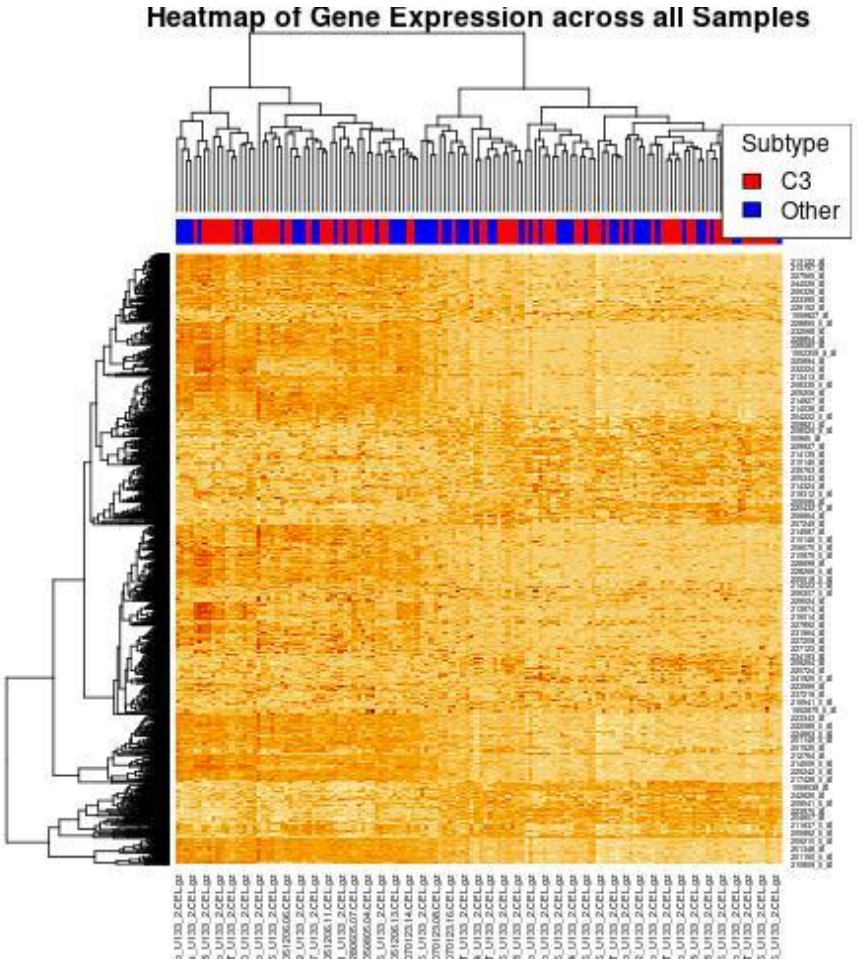

Table 4. Top 10 up-regulated genes. The results of matching gene symbols to their probeIDs, sorted by t-statistic and adjusted p-values to get the top 10 up- and down-regulated genes. Showing each probe ID to its corresponding t-statistic, p-value, adjusted p-value and gene symbol.

	PROBEID	T	P	P_adj	Symbol
2243	203240_at	13.70726	2.374168e-25	5.362492e-24	FCGBP
8949	220622_at	13.65436	1.259171e-26	3.198379e-25	LRRC31
13808	234008_s_at	12.86801	4.765254e-25	1.051143e-23	CES3
11883	227725_at	12.65593	1.015511e-21	1.560289e-20	ST6GALNAC1
5740	211715_s_at	12.63907	2.890720e-23	5.101190e-22	BDH1

6835	214106_s_at	12.45157	1.108356e-20	1.497124e-19	GMDS
971	1568598_at	12.44496	1.060051e-23	1.986839e-22	KAZALD1
3519	205489_at	12.41216	2.076247e-23	3.738236e-22	CRYM
849	1561387_a_at	12.35990	2.273805e-23	4.069652e-22	NXPE1
14083	235350_at	12.30246	1.104306e-21	1.691550e-20	C4orf19

Table 5. Top 10 down-regulated genes. The results of matching gene symbols to their probeIDs, sorted by t-statistic and adjusted p-values to get the top 10 up- and down-regulated genes. Showing each probe ID to its corresponding t-statistic, p-value, adjusted p-value and gene symbol.

	PROBEID	**T**	**P**	**P_adj**	**Symbol**
10547	225242_s_at	-21.74234	8.060678e-45	1.216195e-41	CCDC80
11533	227059_at	-22.01015	9.352116e-45	1.282770e-41	GPC6
2458	203695_s_at	-22.13179	6.930327e-42	3.485492e-39	GSDME
10918	225946_at	-22.17655	6.277484e-38	8.610425e-36	RASSF8
9018	221019_s_at	-22.18293	8.240801e-37	9.551792e-35	COLEC12
8178	218694_at	-22.37769	7.801419e-45	1.216195e-41	ARMCX1
9746	223122_s_at	-22.45322	2.208225e-46	9.829976e-43	SFRP2
4203	207266_x_at	-22.60903	6.179402e-47	4.661741e-43	RBMS1
6523	213413_at	-23.47071	1.122009e-45	3.385773e-42	STON1
2916	204457_s_at	-24.45588	2.354924e-50	3.553109e-46	GAS1

The three pathway gene sets GO, KEGG, and Hallmark that already have been downloaded are uploaded using Bioconductor package. Contingency tables for the three pathways are created and the results are written out. The top 3 from each pathway have been converted into tables for further analysis of the data. Benjamini Hochberg test is done on the data prior to this

Table 6. Top 3 Enriched GO Gene Sets. Fisher test performed on GO gene set collection.

	Setname	pvalue	estimate	exp	BH
1	GOBP_BIOLOGICAL_ADHESION	7.967431e-36	3.096417	Down	8.528339e-32
2	GOBP_VASCULATURE_DEVELOPMENT	5.483831e-26	3.316883	Down	2.934946e-22
3	GOBP_ANATOMICAL_STRUCTURE_FORMATION_INVOLVED_IN_MORPHOGENESIS	8.526327e-24	2.814047	Down	3.042194e-20

Table 7. Top 3 Enriched Hallmark Gene Sets. Fisher test performed on GO gene set collection.

	Gene set	pvalue	estimate	exp	BH
1	HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	2.655557e-57	12.329828	Down	2.655557e-55
2	HALLMARK_MYOGENESIS	1.772062e-12	4.323357	Down	8.860309e-11
3	HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	4.020172e-10	0.000000	UP	1.340057e-08

Table 8. Top 3 Enriched KEGG Gene Sets. Fisher test performed on GO gene set collection.

	Gene set	pvalue	estimate	exp	BH
1	KEGG_ECM_RECEPTOR_INTERACTION	4.827033e-15	8.969897	Down	1.795656e-12
2	KEGG_FOCAL_ADHESION	2.511755e-14	4.743836	Down	4.671865e-12
3	KEGG_DRUG_METABOLISM_CYTOCHROME_P450	5.537584e-10	7.322064	UP	5.791203e-08

Table 9. Gene set databases

Gene set database	Number of gene sets
Hallmark	50
Kegg	186
GO	10402

Discussion

The reference paper Marisa et. al originally used consensus clustering but since the method is computationally too intensive, hierarchical clustering was employed here. This resulted in the difference of results between the reference study and the project outputs. A dendrogram is cut into two clusters. The 1000 top up and down regulated genes show interesting results. “FCG3P”, “LRRC31”, “CES3” are the top three in the up regulated list. “CCDC80”, “GPC6”, “GSDME” are the top three from the down regulated genes. These genes are mainly related to protein folding, protein metabolism and it’s implication colorectal cancer.

No genes were noticed to be significantly expressed ($p < 0.05$). This could be attributed to the way the data is processed analyzed. The total of genes that passed all the three filters in the reference study is different from the project output. This could possibly be the reason for the discrepancies in the studies. The subgroups C3 and C4 show different genes in them, and this could be mainly accounted to characterize colorectal cancer for diagnostic purposes. Furthermore, distinct expression signatures between tumor subtypes further propels the theory of colorectal cancer subtypes, where each subtype can have varying prognostic classifiers.

The three gene sets that were used to create a contingency table are Kegg, Go, and Hallmark from MsigDB database. Most of the pathways seen here are a part of cell metabolism, protein modeling etc. The significant and top 3 results from the kegg gene set are

KEGG:

1. KEGG_ECM_RECEPTOR_INTERACTION (ECM -receptor interaction)
2. KEGG_FOCAL_ADHESION (focal adhesion)
3. KEGG_DRUG_METABOLISM_CYTOCHROME_P450 (drug metabolism).

Hallmark:

4. HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION (epithelial-mesenchymal interaction)
5. HALLMARK_MYOGENESIS (development of skeletal muscle)
6. HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION (epithelial-mesenchymal interaction)

GO:

7. GOBP_BIOLOGICAL_ADHESION (cell adhesion)
8. GOBP_VASCULATURE_DEVELOPMENT (development of blood vessels)
9. GOBP_ANATOMICAL_STRUCTURE_FORMATION_INVOLVED_IN_MORPHOGENESIS (involved in the development of anatomical structures)

References

1. Marisa L., de Reynies A., Duval A., Selves J., et al. (2013). Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value. PLoS Med 10(5): e1001453. doi:10.1371/journal.pmed/1001453 https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001453
2. Durinck, S., Spellman, P.T., Birney, E. and Huber, W., 2009. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nature protocols, 4(8), pp.1184-1191. https://www.nature.com/articles/nprot.2009.97
3. Database for hallmark pathway genes https://www.gseamsigdb.org/gsea/msigdb/cards/HALLMARK_MYOGENESIS.html
4. Subramanian, A., Kuehn, H., Gould, J., Tamayo, P., & Mesirov, J. P. (2007). GSEA-P: a desktop application for Gene Set Enrichment Analysis. Bioinformatics, 23(23), 3251-3253. https://academic.oup.com/bioinformatics/article/23/23/3251/289118?login=true
5. Tallarida, R. J., & Murray, R. B. (1987). Chi-square test. In Manual of pharmacologic calculations (pp. 140-142). Springer, New York, NY https://link.springer.com/chapter/10.1007/978-1-4612-4974-0_43