Individual Project Writeup

Introduction:

The project that the paper focuses on is project one: Microarray Based Tumor Classification and specifically the analyst and biologist parts. In the analyst section there are two main parts: noise filtering and dimensionality reduction and then hierarchical clustering and subtype discovery. The purpose of the analysis is to reproduce the results from the comparison of the C3 and C4 tumor subtypes. Noise filtering was used in order to reduce the data dimensionality and then hierarchical clustering was used and important to the study in order to discover novel relationships among samples in a dataset. In the biologist section the focus was performing a more in depth analysis. The authors of Marisa et al wanted to understand the biological significance of different gene expression profiles for each tumor subtype by using gene set enrichment analysis. The primary focus is on the Hallmark, GO, and KEGG gene sets being compared to the top 1000 up and down regulated genes from the analysis results compared with the others using the Fisher's Exact Test.

Methods:

The first part of the analysis focused on noise filtering and dimensionality reduction on the data. Using the RMA normalized, ComBat adjusted expression matrix that was created by the programmer three filters were applied to the data. The first filter was for each gene at least 20% of the gene expression values must be > log2(15). The first filter function calculated the threshold which was log2(15), saving it as a variable and then also calculated what 20% of the data was by calculating .20*the number of rows in the data frame. The data was filtered by these thresholds and from 54,675 genes only 39904 genes passed the filter. The next filter was if the gene had a variance significantly different from the median variance of all probe sets using a threshold of $p < 0.01$. The dataset for the second filter was the genes that had passed the first filter. The variance of the data set for each row was calculated and then from there the median variance of the entire dataset was calculated. A chi-square test was run and a total of 24,806 genes passed the filter. The data set for the third and final filter was the genes that had passed both the first and second filter. The third filter was looking at if the data had a coefficient variation of > 0.186. The mean and standard deviation for each row by using the apply() function. The coefficient of variation was calculated by dividing the standard deviation by the mean and then the data was subsetted based on if the coefficient was > 0.186. All together only 2520 genes passed all three of the filters. CSV files were written containing the genes that passed each filter respectively.

The next part of the analysis focused on hierarchical clustering and subtype discovery. The dataset containing only the genes that passed all three filters was used in order to perform hierarchical clustering. The data was transposed and then created into a distance matrix by using the dist() in R and then clustering was performed by the hclust() function. The dendrogram was cut into two clusters using the cutree() function, specifying two clusters. A heatmap was created to show the gene-expression of each gene across all samples using the heatmap() function. The heatmap was colored by cit.coloncancermolecularsubtype, which contained the annotation matrix from the metadata.csv downloaded from SCC. If the subtype was C3 then it was colored red on the heatmap colorbar and if it was any other subtype it was colored blue. The next part of the analysis section was to take the dataset that contained the genes that passed all three filters and the clusters created earlier and find the differentially expressed genes using a welch t-test. After the Welch t-test was performed the p-values and t stat were pulled from the results so that it could be saved into a results dataframe. The padj values were calculated using the p.adjust() function and using the "fdr" method. A data frame was created with the probeids, t-statistics, p-value and padj values. The data frame was filtered by p value < 0.05 and there were 2012 genes that were differentially expressed of the 2520 genes in the original dataset (amount of genes that passed all three filters). The dataframe was also filtered by padj < 0.05 to be able to compare which is a better distinguisher of the most differentially expressed genes. The dataframe filtered by padj < 0.05 contained 1989 genes. The same process was repeated to find differentially expressed genes using the Welch t-test but this time using the dataset of genes that passed filter 2 instead of passing all three filters. The only difference was that in this process log2FC was also calculated when creating the data frame after the t-test to be used later on in the biologist role. Log2FC was calculated by taking the log2 of each cluster and subtracting by the mean for each row. When the data was subsetted to find the differential expressed genes when a p-value < 0.05 was used there were 13213 differentially expressed genes of 24,806 (amount of genes that passed the second filter). When p-adj < 0.05 was used on the data 13156 differentially expressed genes were found. It was decided that the padj < 0.05 provided more stringent thresholds for the differentially expressed genes so the dataframe that was filtered by padj was written out into a CSV to be used in the biologist section.

For the biologist section the following libraries were loaded in AnnotationDbi, hgu1333plus2.db(Carvalho), GSEABase(Morgan), affy(Rafael), and tidyverse(Wickham). The data that is used in the biologist section is the results of the differentially expressed genes from the last part of the analyst which used the dataset from filter 2. After the CSV was read into R, the select function from the hgu133plus package (Carvalho) was used to map the probeset ids to the gene symbols. In order to ensure that the correct gene symbols were matched, the probe id column of the data frame was used and all the key and columns were set to symbols. The results of the select were then merged into the dataframe using the probe ids. In order to make sure that there were no duplicate symbols a code was run to remove duplicates by using the duplicated() function. The top 1000 up-regulated genes were selected by first filtering the dataset to only contain the rows that had positive log2FC change. Once filtered the data set was ordered by ascending log2FC and the top 1000 values were saved. The same process was done to get the top 1000 down regulated genes except the data was filtered to only contain rows with negative log2FC. The KEGG, GO, and Hallmark gene sets were downloaded from MSigDB with gmt extensions and read into R using the getGMT function from the GSEABase library (Morgan). The fisher exact test was used to compute hypergeometric statistics and p-values comparing overlap for each gene set and each top 1000 up-regulated 1000 down-regulated genes. In order to do this a contingency table was created by creating a function that calculated the differentially expressed genes are in the gene set, differentially expressed genes that are not in the gene set, not differentially expressed genes but in the gene set, and not differentially expressed genes not in the gene set and then it returns a dataframe containing all these values. A function was created because we are using the three different gene sets Hallmark, KEGG, and GO. The function takes in the differentially expressed genes found before, the gene set, and then the non differentially expressed genes. For each of the three gene sets a for loop is created to go through each of the gene ids in the gene sets where the contingency tables were made and the fisher t-test was run and a resulting data frame containing the gene ids, p value and test statistic and where it was up or down regulated was created. In order to obtain the top 3 enriched gene set for each gene set, each result table was ordered by p-adj and then the top 3 values were returned and saved into three separate csv respectively.

Results:

Table 1: Table of Genes that Pass each Threshold. Three filters were applied to the expression matrix, at the end of the third filter is the number that passed each threshold. There were 54675 genes in the expression matrix before filtering.

Filter	Number Passing	Number Failing
Filter 1: Expressed in at least 20% of samples	**39904 genes**	**14771 genes**
Filter 2: Chi-Squared Test	**24806 genes**	**15098 genes**
Filter 3: Coefficient of variation > 0.186	**2520 genes**	**22286 genes**

In table 1, it shows how with each filter that is applied on the expression matrix how many pass and how many fail each one. The dataset originally contained 54,675 genes and after the three filters were applied only 2,520 genes passed. The first filter that was applied to the data was that for each gene there is at least 20% of the gene expression values must be > log2(15). Of the 54,675 genes 39,904 of the passed this filter when 14,771 genes failed. In the second filter it was having variance significantly different from the median variance of all probe sets using a threshold of $p < 0.01$. A chi-squared test was implemented in order to create this filter. From the 39,904 genes that passed the first filter 24,806 of them passed the second filter when 15,098 genes failed. In the last filter it was looking at the genes having a coefficient of variance that is greater than 0.186. Of the 24,806 genes that passed the first and second filter only 2,520 genes passed and 22,286 failed leaving a total of only 2,520 genes passed all three of the filters. Out of all the filters the filter that the most genes failed was filter 3 which leads to speculation that it was the most stringent filter of the three.

Table 2: Table of the number of samples in each cluster. There are two clusters: cluster 1 and cluster 2 and it shows how many samples are in each cluster

Cluster	Number of Samples
Cluster 1	60
Cluster 2	74

In table 2, it shows how the samples were divided into the two clusters from the hierarchical clustering dendrogram. There were 60 genes in cluster 1 and 74 genes in cluster 2. In figure 1 it shows a heat map of the genes and the samples, which was created in order to observe the gene expression differences of the genes across all samples. The samples are on the bottom of the heat map on the x-axis and the genes are on the right side on the y-axis. There is a color bar on the top of the heat map which shows red if its subtype is C3 and blue if its subtype is other. Overall, the distribution of the color bar since one side is mostly red and the other mostly blue shows that the gene expression does distinguish cancer molecular subtype and hierarchical clusters to some degree. The samples show grouping by molecular subtype based on gene expression, the order of the sidebar most of the C3 subtypes are on the left while the others are on the right side.

Figure 1: Heatmap of the genes and samples. There is a color bar that indicates which subtype each sample belongs to.

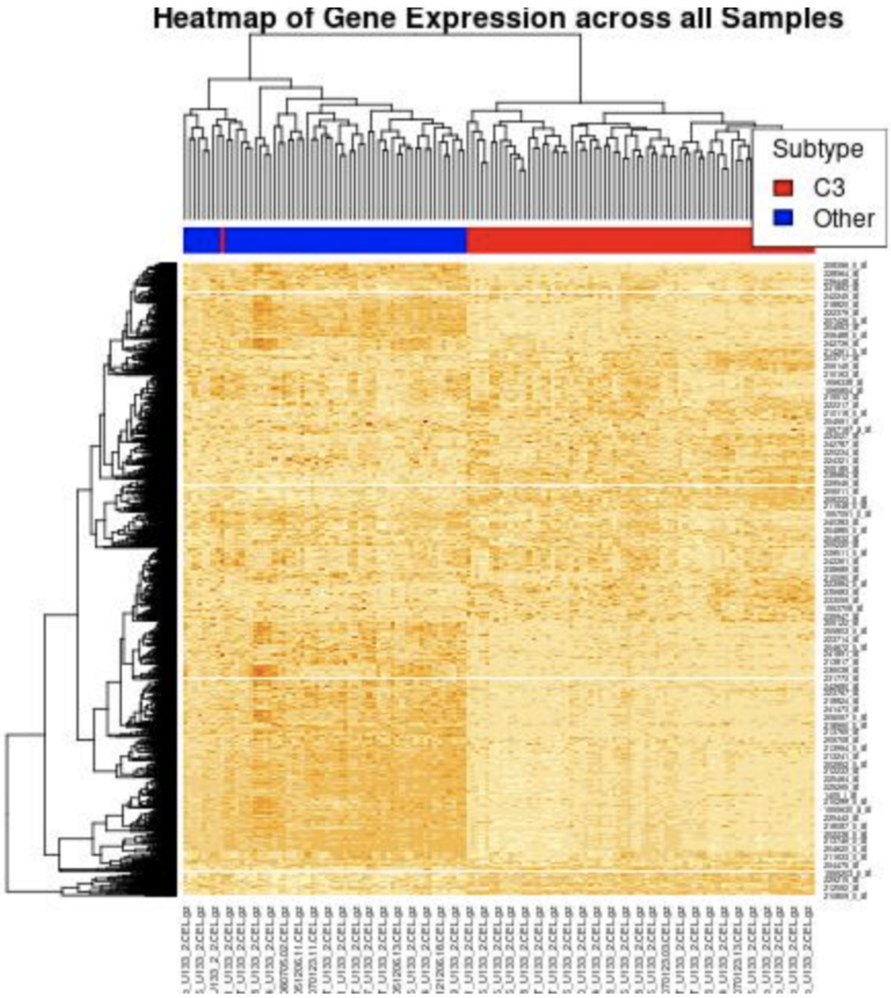

Table 3: Number of Differentially Expressed Genes at P < 0.05 between the two clusters

Expression Matrix	Number of DE Genes at $p < 0.05$	Number of DE Genes at padj < 0.05
Filter 2 (Chi-Squared)	**13213 genes**	**13156 genes**
Filter 3 (passed all three filters)	**2012 genes**	**1989 genes**

In table 3, it shows results of calculating the number of differentially expressed genes. Originally, the number of differentially expressed genes was calculated to show the genes that have a $p < 0.05$ and then again with a padj < 0.05. Both $p < 0.05$ and padj < 0.05 were calculated in order to compare and see if there is a difference in the number of genes. There is a difference in the number as shown in table 3 because there are less genes that pass the padj < 0.05 and ultimately using padj < 0.05 was used in teh rest of the analysis. The number of differentially expressed genes in the dataset that contained genes that passed filter two is 13,156 genes which is the data that is used in table 4 to determine the top up and down regulated probe sets.

Table 4: Top 10 up and down regulated probe sets. For each probeset the gene symbol, t-statistic, nominal p-value, and adjusted p-value are reported.

Probeset	Gene Symbol	P-Value	P-adj	T-statistic	Up/Down
242601_at	HEPACAM2	-13.16714	-2.5124e+04	-13.1671	UP
210107_at	CLCA1	-12.85128	-1.77105e+04	-12.8512	UP
203240_at	FCGBP	-15.98127	-3.96431e+05	-15.9812	UP
1554436_a_at	REG4	-8.883682	-6.92983e+02	-8.88368	UP
223597_at	ITLN1	-10.72716	-3.28516e+03	-10.7271	UP
207214_at	SPINK4	-12.40569	-1.2309e+04	-12.4056	UP
223969_s_at	RETNLB	-10.6470	-3.1071e+03	-10.6470	UP
219955_at	L1TD1	-10.47765	-2.7073e+03	-10.4776	UP
205815_at	REG3A	-6.347705	-1.2234e+02	-6.34770	UP
219727_at	DUOX2	-10.01371	-1.800e+03	-10.0137	UP
212904_at	LRRC47	-0.087555	-1.6845e-01	-0.0875	DOWN
201433_s_at	PTDSS1	0.021664	4.0892e-02	0.02166	DOWN
1553978_at	BORCS8	-0.038332	-7.3095e-02	-0.0383	DOWN
224288_x_at	FKSG49	0.0121790	2.3025e-02	0.01217	DOWN
201055_s_at	HNRNPA0	-0.005902	-1.1192e-02	-0.00590	DOWN
222650_s_at	SLC2A4RG	-0.010029	-1.9039e-02	-0.0100	DOWN
204031_s_at	PCBP2	-0.040686	-7.7618e-02	-0.0406	DOWN
203403_s_at	RNF6	-0.040562	-7.7375e-02	-0.0405	DOWN
208984_x_at	RBM10	-0.063778	-1.2213e-01	-0.0637	DOWN
222364_at	SLC44A1	0.009780	1.8500e-02	0.00978	DOWN

Table 4 contains 20 genes, the top ten up regulated and the top ten down regulated genes merged together into one dataframe. Determining if the gene was up and down regulated was determined by the log2FC values being positive or negative. Table 5 shows the three gene seets that are used in the last part of the analysis: Hallmark, KEGG, and GO and how many gene sets are present in each database.

Table 5: Gene Set Databases. Gene sets databases used and the number of gene set in each

Gene Set Databases	Number of Gene Sets
Hallmark (h.all.v7.5.1.symbols.gmt.txt)	50
KEGG (c2.cp.kegg.v7.5.1.symbols.gmt.txt)	186
GO (c5.go.v7.5.1.symbols.gmt.txt)	10402

Table 6: Number of Most Significantly Enriched Gene Sets. Significance was based on the padj < 0.05

Gene Set Database	Number of Significantly Enriched Gene Sets
Hallmark	4
KEGG	15
GO	125

Table 7: Top 3 Enriched Gene Sets for each Gene Set Type.

Gene Set Name	P-Value	P-adj	Tstat	Gene Set Database
HALLMARK_UNFOLDED_PROTEIN_RESPONSE	**1.634480e-03**	**0.04086**	**0.100490**	**Hallmark**
HALLMARK_OXIDATIVE_PHOSPHORYLATION	**6.263133e-05**	**0.00287**	**0.195242**	**Hallmark**
HALLMARK_MYC_TARGETS_V1	**1.063725e-0**	**0.000531**	**0.145235**	**Hallmark**
KEGG_NEUROTROPHIN_SIGNALING_PATHWAY	**0.001368270**	**0.0339**	**0.0**	**KEGG**

KEGG_O_GLYCAN_BIOSYNTHESIS	**0.0004991739**	**0.01326**	**6.257376**	**KEGG**
KEGG_LINOLEIC_ACID_METABOLISM	**0.0004057113**	**0.01160**	**8.207357**	**KEGG**
GOBP_INORGANIC_ANION_TRANSPORT	**0.0002954294**	**0.04936**	**2.892413**	**GO**
GOCC_MICROVILLUS_MEMBRANE	**0.0002966075**	**0.04936**	**7.040284**	**GO**
GOMF_PROTEIN_CONTAINING_COMPLEX_ BINDING	**0.0002869536**	**0.04853**	**0.550941**	**GO**

Table 6, shows the number of most significantly enriched genes for each database. This was calculated by putting a padj threshold of 0.05. The hallmark geneset database originally had 50 gene sets, shown in table 4 and of those 50 gene sets only 4 of them are significantly enriched. For the KEGG pathway out of the 186 gene sets, shown in table 4, 15 of them are significantly enriched. For the GO pathway out of the 1042 gene sets only 125 of them are significantly enriched. Table 7, shows the top three enriched geneset for each gene type: Hallmark, KEGG, and GO. When the gene set name was compared to the results in Figure 2 in Maria et al, there were no similarities found between them, which could be due to the differences in the way analysis were performed.

Discussion:

The goal of the analysis section was to employ noise filtering techniques to reduce data dimensionality on the expression matrix created by the programmer and then to perform hierarchical clustering in order to discover novel relationships between the samples in the dataset. The goal of the biologist section was to perform a more in depth analysis to understand the biological significance of the gene expression profiles by using gene set enrichment. The expression matrix originally contained 54675 genes and 134 samples. After going through three rounds of filtering: each gene at least 20% of the gene expression values must be > log2(15), a variance significantly different from the median variance of all probe sets using a threshold of $p < 0.01$, and having a coefficient variation of > 0.186 only 2520 genes remained. To determine the differential expressed genes, on the genes that passed all three filters the 2520 genes were used only with the clusters obtained from heriaraticla clustering. It was discovered that there were 1989 genes that were differentially expressed that passed all three criteria. In the paper the authors found that 1459 games passed the authors filters that were implemented, which is around 1000 less genes then what was discovered from filtering in this analysis. The difference in the results in filtering could be due to coding errors or differences in how the filters were created. In addition the authors described a discovery set of 443 samples and validation set of 1029 samples, when this analysis only contained 135 samples which also could play a role in the differences of the results. Differences in the filtering could have arisen based on sample size because a lot of the filters were dependent on sample size so it could fluctuate if one sample is smaller than the other thus leading to different results. The differential expressed genes for the genes that only passed filter one and two was also looked at. It was discovered that there were 13156 differentially expressed genes obtained from filtering at padj < 0.05. The amount of differentially expressed genes is slightly low since in the instructions it stated that there should be around 20k genes. It could be considered that finding 13156 differentially expressed genes is not that far from the 20k but the differences could be again due to the way the filter was being implemented and human error. From there the top 1000 up and down regulated genes were subsetted based on log2FC. After performing a fisher test to compute hypergeometric statistics and p-values on the up and down regulated genes and enriched gene sets were looked into. From the hallmark gene set database there were 50 genes with only 4 enriched, in KEGG there were 186 with 15 enriched, and in GO there were 10402 genes with 125 enriched. The top 3 enriched gene sets for each gene type were: HALLMARK_UNFOLDED_PROTEIN_RESPONSE, HALLMARK_OXIDATIVE_PHOSPHORYLATION, HALLMARK_MYC_TARGETS_V1, KEGG_NEUROTROPHIN_SIGNALING_PATHWAY, KEGG_O_GLYCAN_BIOSYNTHESIS, KEGG_LINOLEIC_ACID_METABOLISM, GOBP_INORGANIC_ANION_TRANSPORT, GOCC_MICROVILLUS_MEMBRANE, GOMF_PROTEIN_CONTAINING_COMPLEX_BINDING. Most of the pathways that were discovered have to do with some metabolism, cell communication, and protein folding after doing a quick search. When the gene sets were compared to the Maria et al Figure 2 results there were no similarities with the pathways found in this analysis. The pathways found in this analysis could fall in the same categories as the ones in the paper but there is none that share the same names. It would be interesting if time permitted to look more into and do a comparative analysis of these pathways to see if they share any molecular signatures. This difference in pathway results could be due to differences earlier in the filtering and determination of differential expressed genes which caused different pathways to appear enriched but also could be due to the differences in sample size as stated above.

References:

1. Marisa L, de Reyniès A, Duval A, Selves J, Gaub MP, Vescovo L, Etienne-Grimaldi MC, Schiappa R, Guenot D, Ayadi M, Kirzin S, Chazal M, Fléjou JF, Benchimol D, Berger A, Lagarde A, Pencreach E, Piard F, Elias D, Parc Y, Olschwang S, Milano G, Laurent-Puig P, Boige V. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. PLoS Med. 2013;10(5):e1001453. doi: 10.1371/journal.pmed.1001453. Epub 2013 May 21. PMID: 23700391; PMCID: PMC3660251.
2. Carvalho B (2015). *pd.ht.hg.u133.plus.pm: Platform Design Info for The Manufacturer's Name HT_HG-U133_Plus_PM*. R package version 3.12.0.
3. Morgan M, Falcon S, Gentleman R (2022). *GSEABase: Gene set enrichment data structures and methods*. R package version 1.58.0.
4. Gautier L, Cope L, Bolstad BM, Irizarry RA (2004). "affy—analysis of Affymetrix GeneChip data at the probe level." *Bioinformatics*, **20**(3), 307–315. ISSN 1367-4803, doi: 10.1093/bioinformatics/btg405.
5. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686.