Final Project Report

by Team 4: Adassa Coimin, Daniela Villalva, Vince Duarte

Abstract

In liquid-based sequence analysis, blood has become the most widely used source for sampling. Different techniques and methods of minimally invasive derivations of samples are being pursued, including tumor-educated platelets in blood. Our data was produced through this technique of converting RNA-sequenced platelet data into images, which was then analyzed to determine the accuracy of which it would identify ovarian cancer (OC). From this data, we wanted to answer the question of how we can use this RNA-sequence data ourselves, to identify the presence of ovarian cancer in blood platelets based on gene expression. We used R to analyze the expression of the genes, and generate various plots to convey the relationship that might exist between certain genes, and malignancy of ovarian cancer. From our analysis, we found that the data was pretty consistent with what we would expect. When looking at the relationship between gene expression and the presence of ovarian cancer, we found that a large number of genes had a high expression and enrichment score for ovarian cancer. This enrichment score was based on the gene ontology of our data, and suggests that a lot of genes in a transcriptome are correlated with the diagnosis of ovarian cancer. The ovaries are made of three layers, of which the middle layer is made of connective tissue, while the innermost layer contains blood and lymphatic vessels. Tissues are determined by gene expression patterns, suggesting that there are specific genetic pathways and regulations that influence tissue specificity. Connective tissues are the most differentiated and abundant tissues in the human body, while the blood and lymphatic system are two of the main transport systems that are closely related to each other and essential in various biological pathways. From our findings, it would also suggest that a lot of the genes that show a high upregulation perform an essential role in other pathway interactions in other areas of the body. We anticipate that our approach and the analysis that we conducted can be used to understand a bit more about the connection between the genes that are responsible for encoding certain molecules, such as proteins, in the ovaries, with the disruption presented by ovarian cancer, on the important biological processes throughout the rest of the body. This may allow cancer researchers to have a more targeted approach with countering the rate of spread, deterioration and stress on the body due to OC, and aid in the development of new medicine.

Introduction

The data we used was RNA-sequenced data of tumor-educated platelets in ovarian cancer. There were a total of 69 samples that represented females who were of different ages and if they had cancerous or non-cancerous tumors. Liquid biopsies using blood were implemented to obtain tumor-educated platelets from each person sampled. Liquid biopsies are less invasive and can be used to detect tumor-specific mutations (8, 9, 10). Other studies have used bone marrow as a source to determine prognostic significance of cancers, but this method was more invasive and less preferred by patients (7). Our group wanted to determine if it was possible to identify the presence of ovarian cancer in blood platelets based on gene expression. To answer our question we used enrichment analysis and unsupervised analysis to

process the information of the samples. Gene expression studies have previously shown that breast cancers are distinct at the transcriptomic level, and that prognosis can be determined by the expression of proliferation-related genes (1, 2). Additionally, lung cancer and thyroid cancer can be detected using gene expression profiles (3, 4, 5). One study used hierarchical cluster analysis to separate tumors based on their estrogen receptor status and determined different survival marker genes. In our Assignment 3, we used unsupervised analysis and different clustering methods to separate the samples based on the type of tumor. If breast cancers can be diagnosed using gene expression, then we wanted to determine if gene expression can also determine the presence of ovarian cancer.

Methods

Our data set had 69 samples and over 50,000 genes per sample, which were labeled with Ensembl IDs instead of HUGO gene names. We first converted the Ensembl IDs to HUGO genes using the mapIds method. To see the variation in the data, the data was log scaled and the gene expression ranges were plotted on a density plot. To analyze the RNA-seq data and to test for differential expression, we used DESeq2. We used both t-SNE and UMAP to visualize the variance in lower dimensions. To perform additional differential expression analysis we used a volcano map and extracted significantly differentially expressed genes plotting them to a heatmap. Gene ontology was used to run enrichment analysis using different methods: topGO, clustProfiler, gProfiler2, and GenomicSuperSignature. Aside from expression analysis, we also performed unsupervised analysis on the dataset. K-means, Hierarchical clustering, ConsensusClusterPlus, and PAM clustering were used to cluster the data based on different k-values, either manually-selected or auto-selected. Further analysis was done using different numbers of genes to see how the amount of genes affected clustering. Once the results of clustering with different numbers of genes were generated, statistical tests were performed on the data. The p-value of the chi-squared test was adjusted, as well.

GeneID Conversion

- mapIds- used to map Ensembl IDs to their associated HUGO names

Data analysis and cleansing

- Log scaling
- Variation analysis

Differential Analysis

- DEseq2 - identifies differentially expressed genes between two experimental groups

Enrichment Analysis using GO (gene ontology)

- topGO- used for semi-automated enrichment analysis for Gene Ontology (GO). (A2)
- clustProfiler- used to analyze and visualize functional profiles of genomic coordinates, gene and gene clusters. (A2)

- gProfiler2 - performs functional enrichment analysis and visualization of gene lists, converts gene/protein/SNP identifiers to numerous namespaces, and maps orthologous genes across species. (A2)
- GenomicSuperSignature- method for interpreting the RNA-seq dataset in the context of a large-scale database of previously published and annotated results (A2)

Clustering/ Unsupervised Analysis (Algorithms)

- K-Means - unsupervised non-linear algorithm that clusters data based on similarity or similar groups. (A3)
- Hierarchical clustering - method of cluster analysis that seeks to build a hierarchy of clusters. (A3)
- ConsensusClusterPlus - used to determine cluster count and membership by stability evidence in unsupervised analysis (A3)
- PAM clustering - searches for k representative objects in a data set and then assigns each object to the closest medoid/ group in order to create k number of clusters. (A3)

Statistical Analysis

- Chi-squared test for independence - nonparametric hypothesis test used to evaluate whether two categorical variables are related to each other

GitHub Link

All the code is available in our GitHub

- https://github.com/AdassaC/CGS4144-Group-Project

Results

Part 1: Differential and Enrichment Analysis

To process our data, we began by performing the crucial step of mapping the Ensembl IDs names of the genes to be converted into their associated HUGO names. This step was particularly important because it streamlined the identification of the genes into the more common classification system derived from the HUGO Gene Nomenclature Committee (HGNC).

From there we were able to continue prepping the data by performing variation analysis. This entailed assessing the gene expression counts in the gene expression matrix for major discrepancies and differences in the counts across the samples. From there we normalized the counts by log-scaling the data. The normalized data provided a basis for us to then display the distribution of the per-gene expression ranges in a density plot. As you can see in the (Figure 1) below, we are able to discern that our data exhibits some slight skewness on the right tail, suggesting that the average of the ranges is greater than the median.

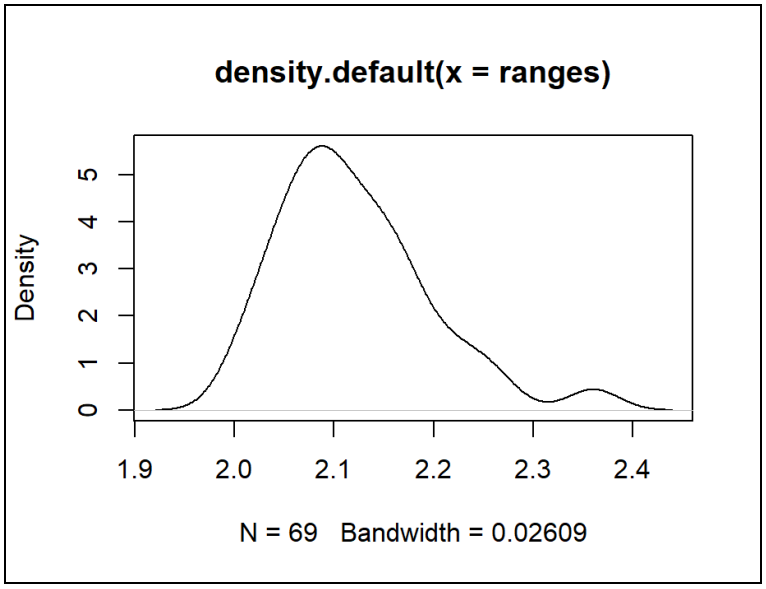


Figure 1: Density Plot

After thoroughly analyzing and normalizing the data, the next step was to perform a differential analysis. Differential gene expression analysis refers to the comparison that is made in the gene counts between the sample group types and leads to the interpretation and further study of the differences in the abundance of certain genes within the transcriptome. Our data was organized in a table of counts, as displayed in Figure 2, of the differentially expressed genes that we would use to assess what genes were upregulated in correlation with the presence of ovarian cancer.

	SRR12705190	SRR12705191	SRR12705193	SRR12705194	SRR12705195	SRR12705196
ENSG00000000003	0.8731599	0.8731599	0.8731599	0.8731599	0.8731599	0.8731599
ENSG00000000005	0.8731599	0.8731599	0.8731599	0.8731599	0.8731599	0.8731599
ENSG00000000419	4.8230719	3.3905432	0.8731599	3.4312689	0.8731599	3.0021309
ENSG00000000457	0.8731599	2.7521476	0.8731599	0.8731599	2.3631764	2.5174945
ENSG00000000460	4.1393557	3.6925095	2.8007129	0.8731599	0.8731599	4.1518858
ENSG00000000938	4.4054539	6.3186872	2.2830342	2.1326700	4.7554527	4.5959007
ENSG00000000971	0.8731599	3.2100534	0.8731599	0.8731599	0.8731599	0.8731599
ENSG00000001036	2.2392325	3.2100534	4.3601538	2.6055159	2.9035195	4.1946814
ENSG00000001084	2.2392325	3.5498103	0.8731599	2.6055159	2.3631764	0.8731599
ENSG00000001167	0.8731599	0.8731599	0.8731599	0.8731599	0.8731599	0.8731599
ENSG00000001460	0.8731599	0.8731599	2.2830342	0.8731599	0.8731599	0.8731599
ENSG00000001461	0.8731599	3.3905432	0.8731599	2.1326700	0.8731599	3.9159348
ENSG00000001497	3.3807386	0.8731599	0.8731599	3.9392905	0.8731599	0.8731599
ENSG00000001561	0.8731599	0.8731599	0.8731599	0.8731599	0.8731599	0.8731599
ENSG00000001617	0.8731599	0.8731599	0.8731599	0.8731599	0.8731599	0.8731599
ENSG00000001626	0.8731599	0.8731599	0.8731599	0.8731599	0.8731599	0.8731599
ENSG00000001629	4.7295561	3.2100534	4.2193429	3.4312689	2.9035195	3.2739356

Showing 1 to 17 of 57,736 entries, 60 total columns

Figure 2: Differential expression counts table

To determine differential gene expression, we used the commonly used algorithm known as DESeq2. DESeq2 is a preferred method for conducting this type of analysis due to the way it provides a relatively good estimation of variance across a negative binomial distribution. This distribution is particularly good for RNA-seq data which can incur variance from both experimental and biological sources. Another advantage that comes from using this algorithm arises from the fact that it uses a generalized linear model to reduce the effects that are determined by confounding variables. These confounding effects can be due to these variables such as age, sex, or biological origin of the sample. As such, DESeq2 tends to have conservative false discovery rates (FDRs) compared to other programs.

Our samples were categorized into these groups based on the cancer status of the sample, which were cancerous samples, where ovarian cancer was observed, non-cancerous, where there was no cancer observed and the tissue type was normal, and neither, a categorical identifier for samples that were undeterministically cancerous or non-cancerous.

From our analysis, we were able to use a PCA plot and a t-SNE plot to portray the variance-mean in our data across the samples, as can be seen in Figure 3. In the PCA plot, there is a higher variance in PC1 than PC2. There are 2 main clusters: one on the -10 line and one around the 15 line. The first cluster is heavily influenced by PC2 and the second cluster is influenced by PC1. In the t-SNE plot, The pink dots represent the normal samples and they tend to be closer to other normal samples. The blue dots (cancerous samples) closest neighbors are blue dots. The NA samples were undetermined samples. These plots were generated using the three groups we initially identified in our data.

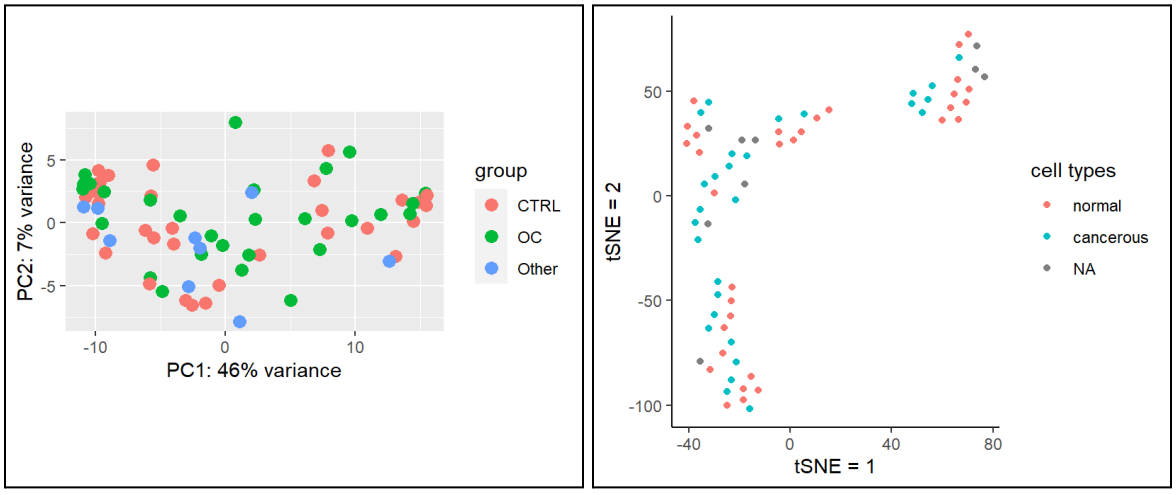


Figure 3: PCA plot and tSNE plot

The UMAP in Figure 4, plots differently from the t-SNE and the PCA Plot in the fact that there seems to be more randomness and scatterness to the data points in it, relating to the cancer status of the samples in the dataset. This UMAP procedure was performed on normalized counts and took the data from thousands of genes and reduced to two variables X1 and X2.

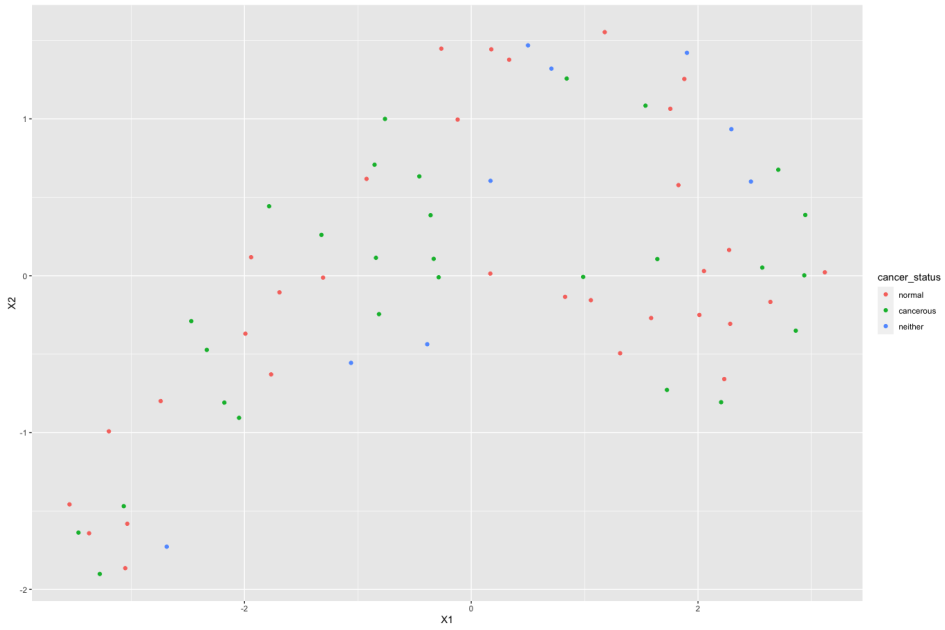


Figure 4: UMAP plot

From our DESeq2 analysis findings, we were also able to gather information about the distribution of our data through the table of differentially expressed genes of high throughput values to convey the strength of the expression levels across the samples. The volcano plot, as can be seen in Figure 4, was derived using the results from DESeq2 with a p-value cuttoff of about 1, and plots the Log2 fold change against the - Log10 P values for all the genes that were positively expressed. The heatmap in Figure 6, is generated by extracting the list of significantly expressed genes.

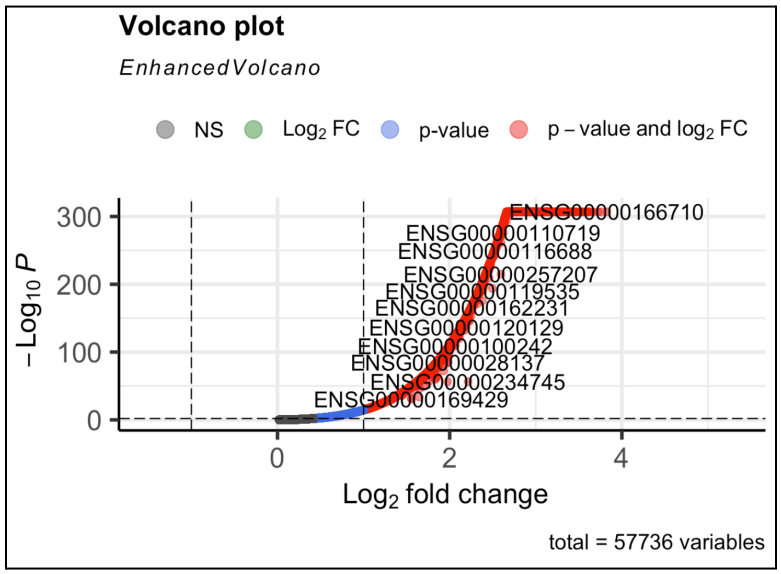


Figure 5: Volcano plot using Enchanced Volcano

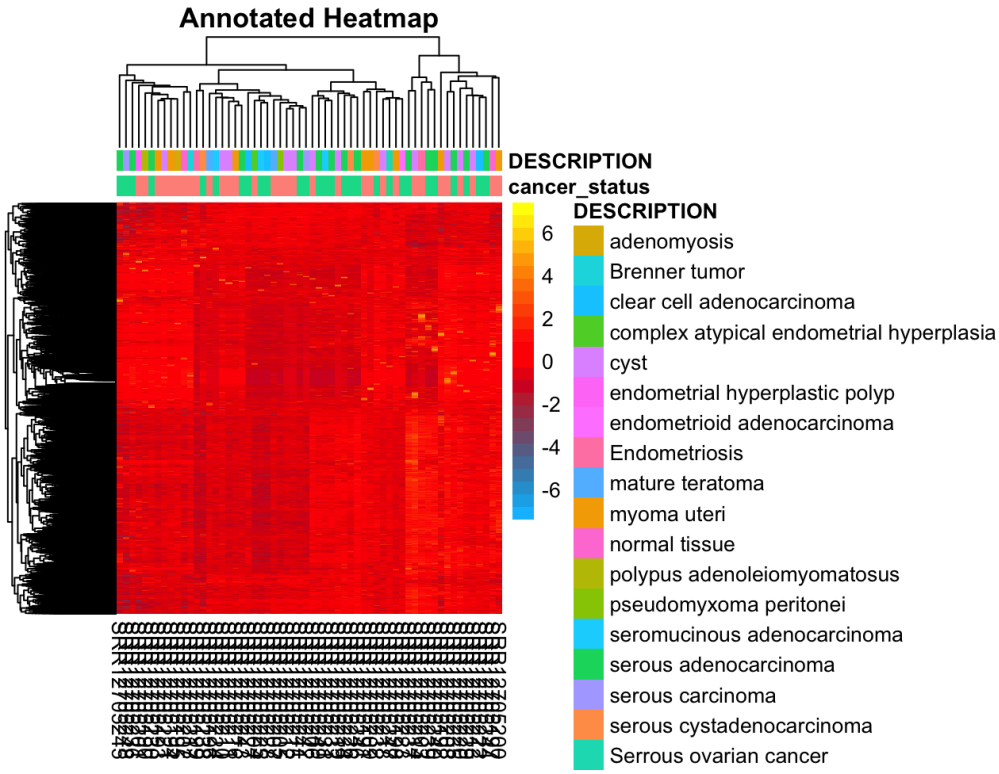


Figure 6: Heatmap colored by sample groupings

Furthermore, we were then able to conduct enrichment analysis across gene ontology, using four different methods. Gene set enrichment analysis is a process that allows researchers to identify subsets of genes and proteins, along with the manner by which they are represented as way of classifying the associations between phenotypes for certain diseases. The first method we used was gProfiler2 algorithm that conducts functional enrichment analysis and visualizations of gene lists, converting gene identifiers while mapping the genes across the species to assesse. These results are displayed in the Figure 7, below:

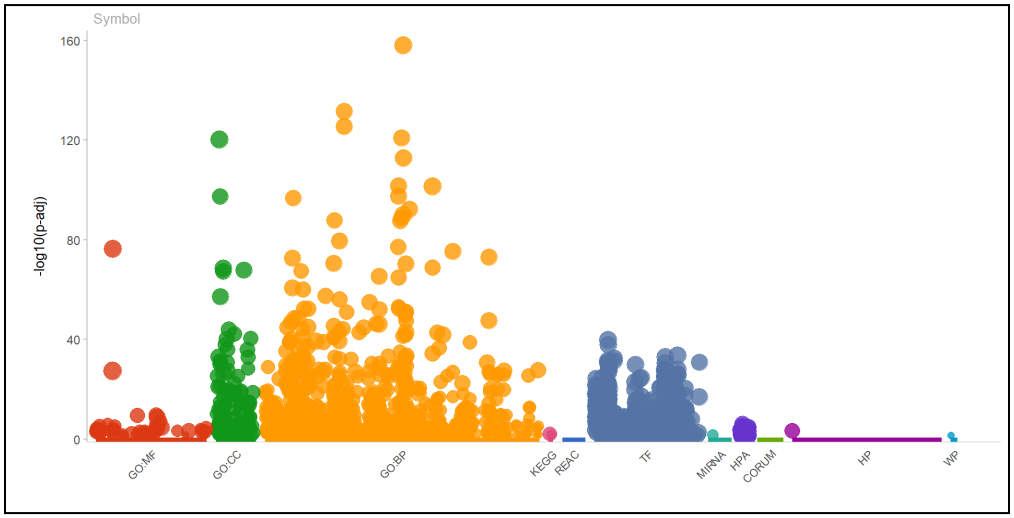


Figure 7: GSEA using gProfiler2

The second method we used was TopGO, which facilitates semi-automated enrichment analysis for Gene Ontology terms, which plots the p-value classic versus the p-value elim in Figure 8.

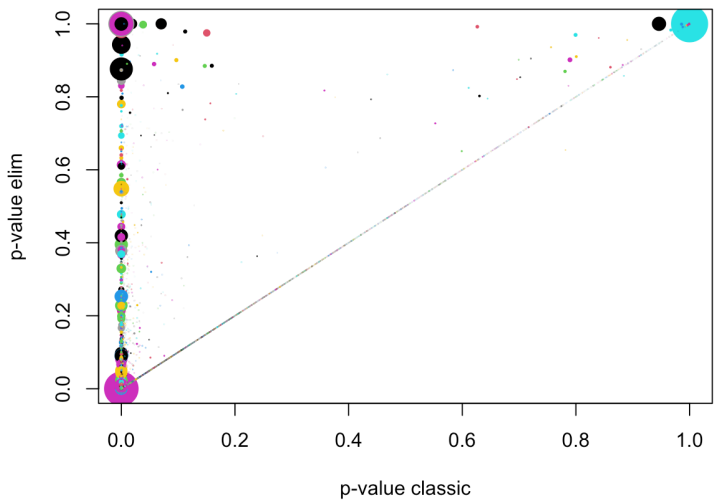


Figure 8: GSEA using TopGo

The third method to conduct the enrichment analysis for gene ontology was clustProfiler. From this algorithm, we were able to generate an enrichment map (Figure 9) of the expression of the gene ontology terms, as a directed acyclic graph. It organized the terms into a network with the edges being the connections between the group of genes that overlap. It determines the strength of relationships between various pathways and cell functions/structures within the genes. From this algorithm, we were also able to gather a GSEA plot shows the manner in which the gene set from our dataset is distributed along with the enrichment score, and ranked position of the genes.

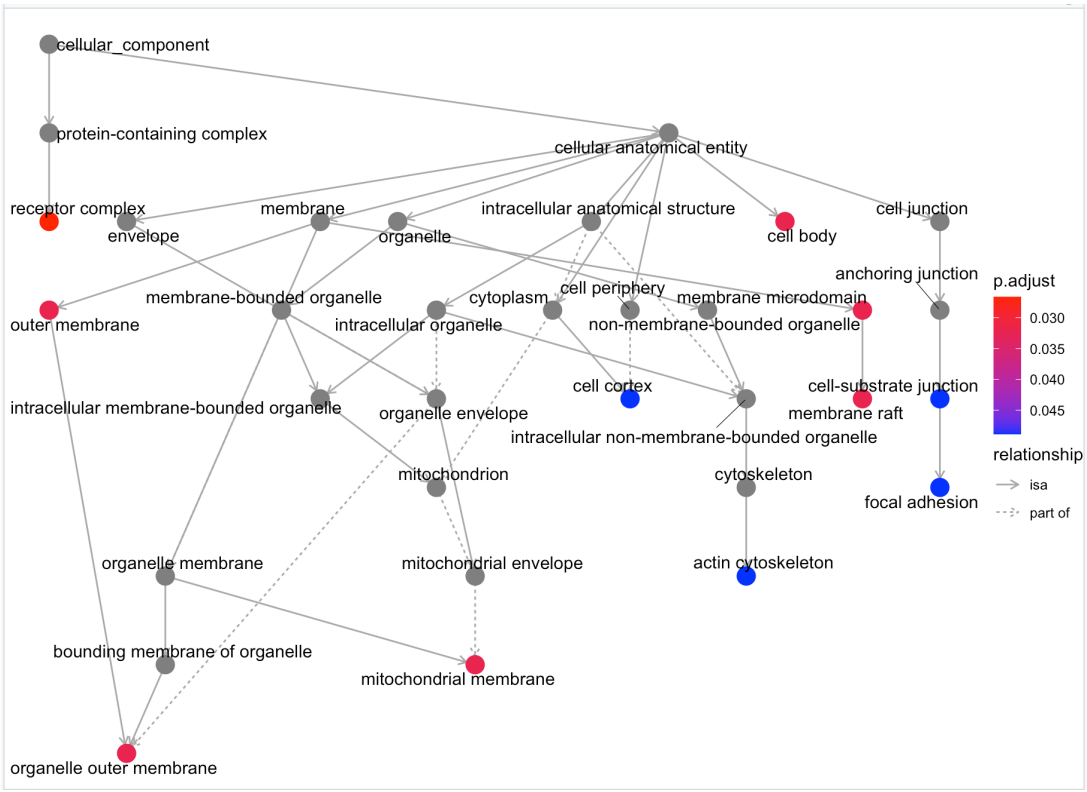


Figure 9: Enrichment map of GO terms using clustProfiler

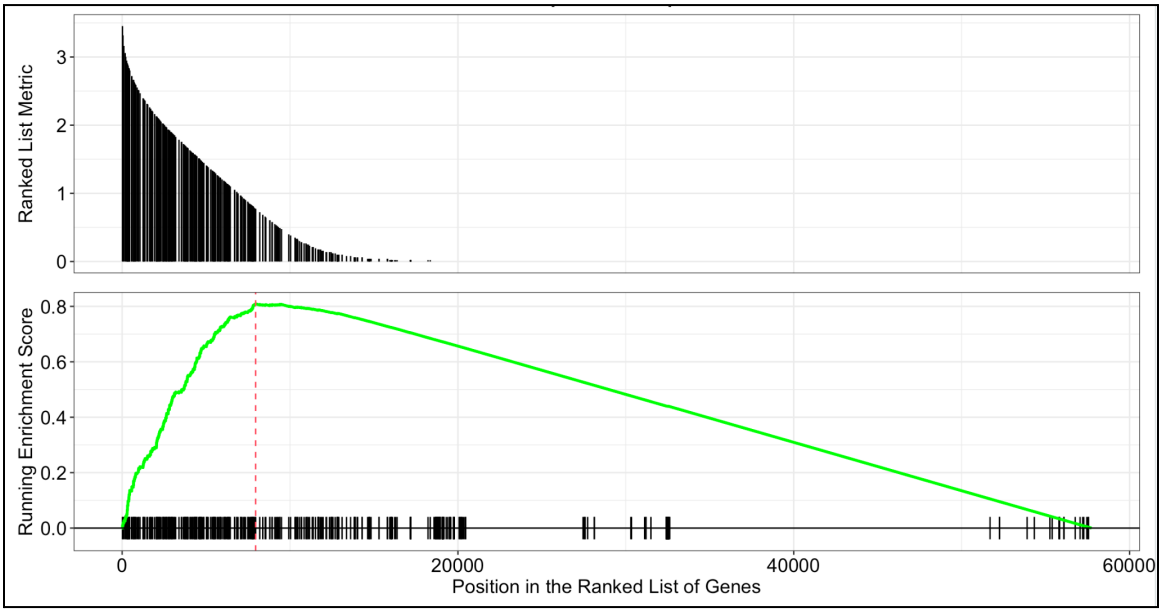


Figure 10: GSEA plot

Our final method for conducting GSEA was through the Genomic Super Signature algorithm which leverages a pretrained RAV model in an attempt to interpret new transcriptomic datasets. The

differentially expressed genes are grouped into clusters and then assigned a validation score after being passed into the model. As shown below in Figure 11, the width of each cluster forms a relationship with the validation score.

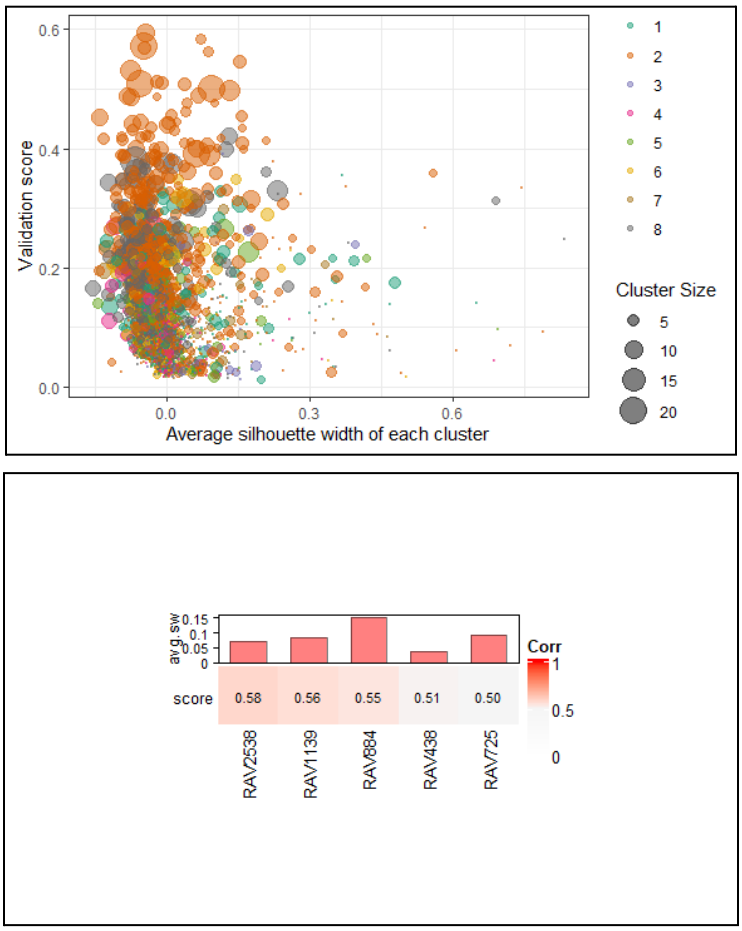


Figure 11: GSEA results using Genomic Super Signature

Part 2: Unsupervised Cluster and Statistical Analysis

When performing analyzing our data by performing cluster analysis, we used 4 different types of clustering algorithms, at various number of genes such as 10, 100, 1000, 5000, and 10,000. Each algorithm used a method of clustering our data around certain number of clusters by either returning a pre-determined K-value, or by allowing the researcher to input a specific value to evaluate the manner by which our data formed clusters. We found that our samples were most optimally clustered around k-values of 2, as the number of genes in our dataset increased, and k-values of 3 at lower amounts of genes present in dataset. This is likely due to the fact that we encountered three groups in the classification of our data, on cancer status

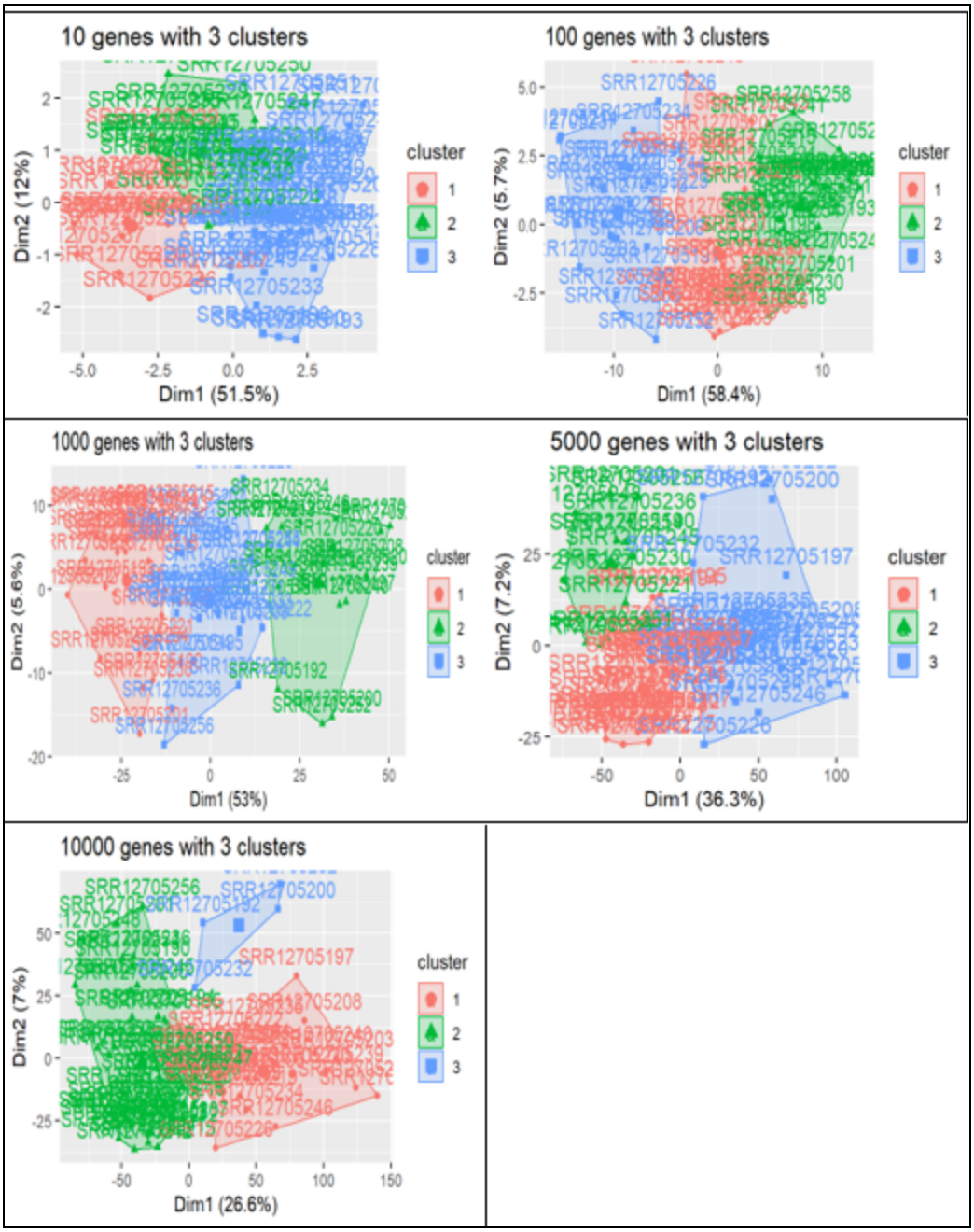


Figure 12: K-means with different number of genes

In the K-means algorithm, the K-value can be picked and changing the k creates different cluster groups. If a smaller K is picked then the clusters were seen more distinctly. A larger K mixed all the clusters together and it was hard to distinguish groups. A different number of genes ranging from 10 to 10000 were grouped among 5 clusters. The next algorithm we used was the Consensus Cluster Plus, which provides a way in which the user can detect pairwise consensus values– the proportion that two items occupied the same cluster out of the number of times they occurred in the same subsample, are calculated and stored in a symmetrical consensus matrix for each k. Moreover, we used hierarchical clustering to

build a hierarchy of clusters by identifying the distance between the samples across gene expression. As can be seen in the figure below (Figure 13), as we increased the number of genes in the dataset the formation of clusters around the K-value of 2, became more agglomerative, compared to lower amounts at k-value of 3.

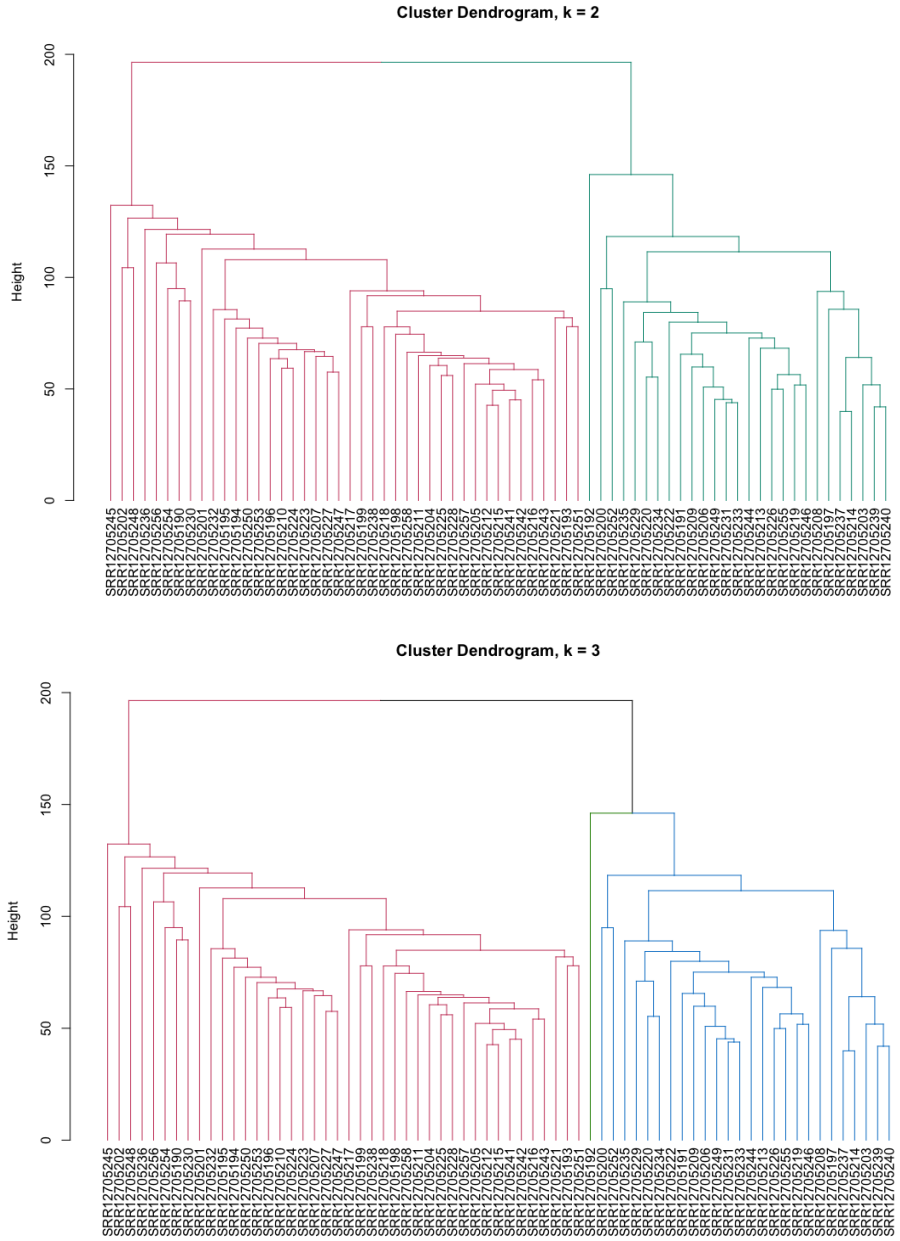


Figure 13: H-clust of 5000 genes

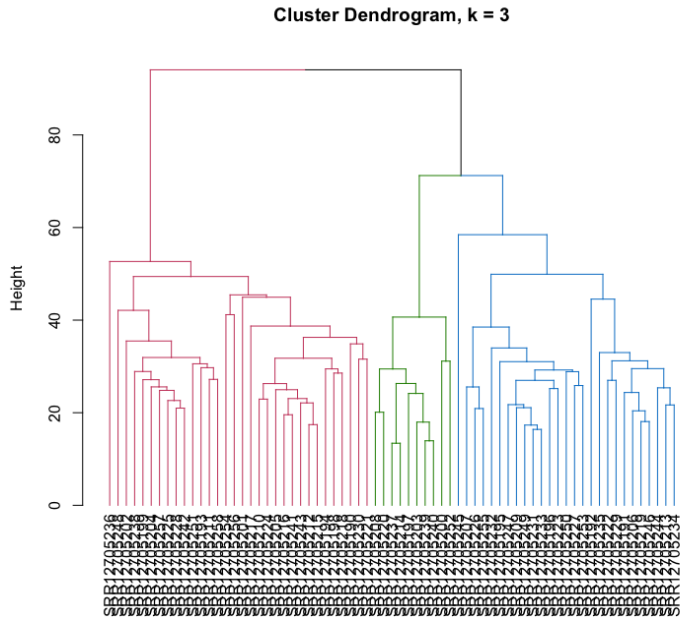
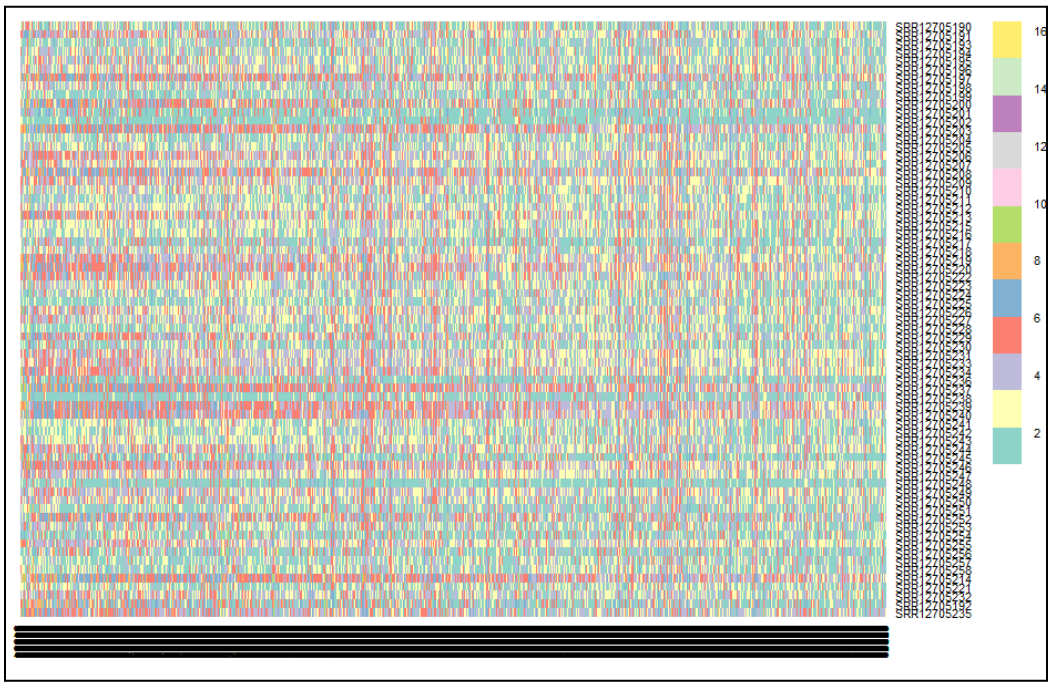


Figure 14: Clustering of 10 Genes at k = 3

Below, are our generated heatmaps derived form our clustering methods.



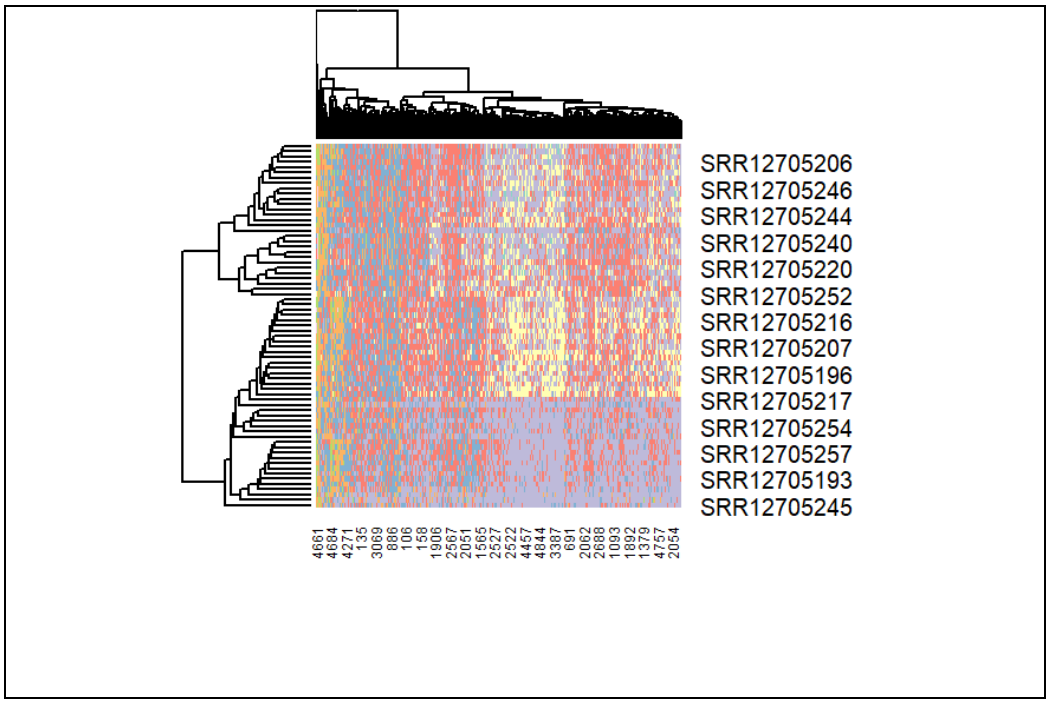


Figure 14: Generated heatmaps of the clusters identified

	regular	adjusted
1	1.000000e+00	1.000000e+00
2	1.243473e-28	3.730419e-28
3	3.964025e-06	4.756830e-06
4	7.602992e-08	1.140449e-07
5	1.540431e-15	3.080863e-15
6	1.130228e-32	6.781369e-32

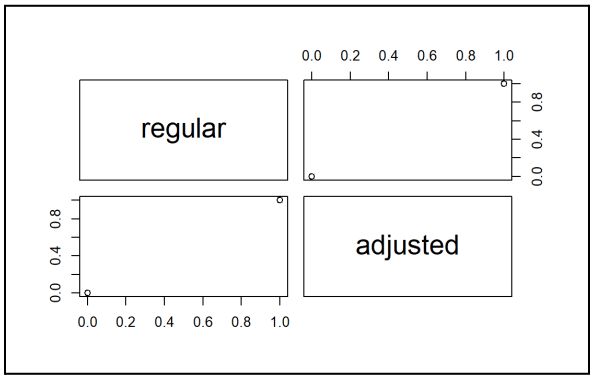


Figure 15: P-values of clusters derived from K-means using X^2 Test for Independence

Figure 15 shows the comparison of clusters from samples of 5000 genes, and attempts to convey a relationship between the groups we identified from part 1, and the 5 clusters identified in this particular run of K-means.

The “Partition Around Medoids” (PAM) clustering algorithm is a technique used to locate the sequence of objects located in clusters. This algorithm intends to form Medoids, where the representative objects are minimized centrally.
K=3

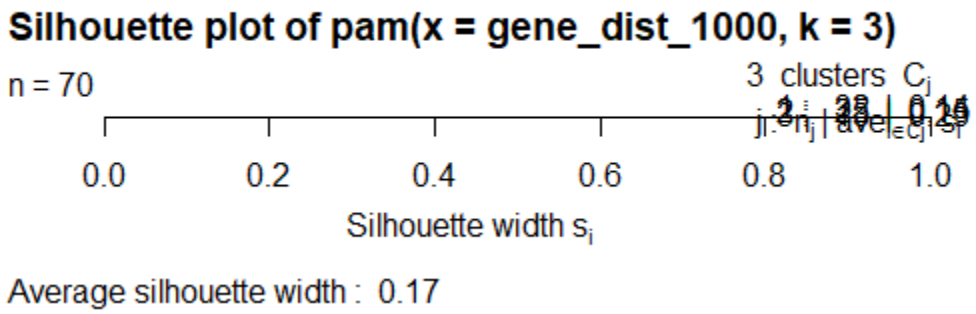


Figure 16: PAM clustering on a 1000 size gene distribution on 3 folds

The algorithm is intended to find a sequence of objects called medoids that are centrally located in clusters. For a cluster with k-means 3, the average silhouette width was 0.17.

K=5

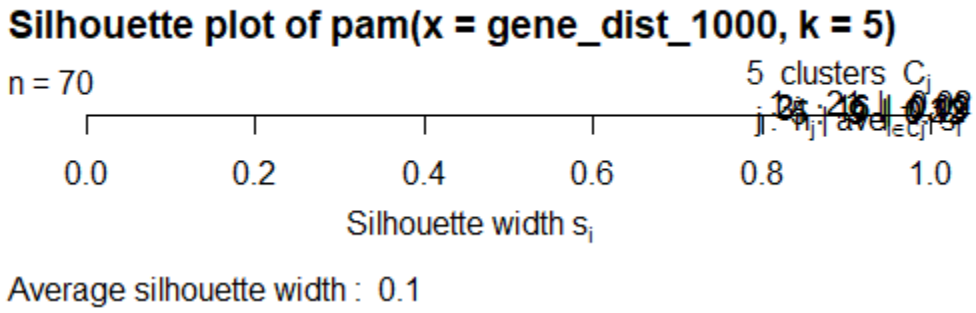


Figure 17: PAM clustering on a 1000 size gene distribution on 5 folds

Partitioning (clustering) of the data into k clusters “around medoids”, a more robust version of K-means. *K*-Means clustering aims at minimizing the intra-cluster distance (often referred to as the total squared error). In contrast, K-Medoid minimizes dissimilarities between points in a cluster and points considered as centers of that cluster. For an increase in the amount of folds to 5 clusters, the average silhouette width increased to 0.1.

Figure 18: PAM clustering Heatmap on a 1000 size gene distribution

This heatmap shows the Gene values mapped against the PAM clustering method at K-means 3 clusters. The x-axis shows cluster output values at regular intervals.

Chi Squared Test

```
        Chi-squared test for given probabilities

data:  pam_5$clustering
X-squared = 51.335, df = 69, p-value = 0.9449
```

Figure 19: Pearson's chi-squared test on the PAM clustering model results

As a result of the chi-squared test, for a high p-value we fail to reject the null hypothesis for the given probabilities.

Part 3: Project Outlook and Significance

To provide an overview of the overall progress of the project, throughout the completion of these various methods of analyzing the gene expression data using the R language, we accomplished a lot in understanding how our data was organized through our endeavor of attempting to answer our scientific question. What our findings show is that there is an overall trend with the way in which some genes are expressed and show an upregulation with the presence of cancer. This can be identified in the way that our data forms clusters of 3 consistent with the way our dataset was group across cancer status, as the number of genes present in our dataset increases, and by this way the accuracy of these clusters increase as they have more genes to form association or disassociate amongst.

Despite these revelations, we had some trouble answering our original question due to some errors in the implementation of the differential analysis, and statistical analysis portion of our project. Due to this initial error, a couple of our plots derived from the DESeq2 algorithm seem to be off and displaying inaccurate information (in how they cluster). This thankfully seems to be overwritten or displayed more accurately in our cluster analysis. Moreover, the curve of the volcano plot seems to be incredibly unequal and shows upstreams of essentially all of our genes, as positive, which should not be the case, since

statistically not all genes show positive expression with ovarian cancer. Normally a volcano plot shows differentiation at some median point on either side.

Our PCA plot and t-SNE plot also had inconsistencies in being able to split our data into groups within the sample. In the same way, the GSEA plot shows an overexpression of all of our genes. Our plot is particularly interesting due to the way in which we have a high enrichment score, across the board, while all of our genes seem to be positively expressed. This ideally probably shouldn't be the case.

As we continued to progress through the project, we realized that the samples in our data had three groups, a control group, a cancer group, and one that was neither. We figured that this was probably due to the fact that the researchers who gathered this data could not positively identify whether or not the sample was cancerous or non-cancerous, and might have exhibited some pre- or post-cancerous attributes. This led us to believe perhaps we should have taken out the samples in our data that were classified as neither. This may have contributed to the errors that we experienced from part 1, and the statistical analysis portion of part 2.

We conducted this investigation by using a dataset from NCBI Gene Expression Omnibus. The results were collected from a sample of 401 non-small cell lung cancer patients and 62 sarcoma patients. We have avoided bioethical issues by maintaining the independence of the genome data headers. In the future, to further improve ethical guidance, we could reach out to the researchers `Tomasz Stokowy` at the Department of Clinical Science at the University of Bergen to receive consent for our analysis.

There are some ways our group could improve the project if we were to do it again. We could reevaluate the technology used in the project. Instead of using R, there is increasing support for python modules in the machine learning space. For example, scikit-learn provides a module for K-Means clustering and other unsupervised learning techniques. By porting the data into a pandas dataframe and extracting the KMeans model, we would have more fine control over hyperparameter tuning.

> *class* `sklearn.cluster.`**`KMeans`**(*n_clusters=8, *, init='k-means++', n_init=10, max_iter=300, tol=0.0001, verbose=0, random_state=None, copy_x=True, algorithm='lloyd'*) [source]

Figure 20: KMeans parameter tuning via scikit-learn documentation

Conclusion

Our group wanted to determine if it was possible to identify the presence of ovarian cancer in blood platelets based on gene expression. Our results from performing the differential analysis and gene set enrichment analysis, suggest that a very large number of genes are directly correlated with the presence and diagnosis of ovarian cancer. This was determined using the enrichment score, and taking a close look at the way in which our data formed clusters around the deterministic and categorical groups prescribed by our data, such as the predetermined cancerous samples, non-cancerous samples, and those determined to be neither. From the methods of unsupervised analysis, using clustering algorithms we determined the formation of clusters across our samples, was influenced a lot by the number of genes present in our matrix. This suggests that the more genes present in our matrix the more upregulated they were, and the more accurate our classifications of clustering for cancerous and noncancerous groups. Overall, we

learned a great deal about how important genes and their expression patterns are in the diagnosis of malignant tissue, and how they can be helpful in deriving medicines in understanding how these same genes interact in biological processes that take place in these tissues and those similar across the body.

References

1. Reis‐Filho JS & Pusztai L (2011) Gene expression profiling in breast cancer: classification, prognostication, and prediction. Lancet 378, 1812–1823. https://pubmed.ncbi.nlm.nih.gov/22098854/
2. Sotiriou C, Neo S‐Y, McShane LM, Korn EL, Long PM, Jazaeri A, Martiat P, Fox SB, Harris AL & Liu ET (2003) Breast cancer classification and prognosis based on gene expression profiles from a population‐based study. Proc Natl Acad Sci USA 100, 10393–10398. https://pubmed.ncbi.nlm.nih.gov/12917485/
3. Alexander EK, Schorr M, Klopper J, Kim C, Sipos J, Nabhan F, Parker C, Steward DL, Mandel SJ & Haugen BR (2014) Multicenter clinical experience with the Afirma gene expression classifier. J Clin Endocrinol Metab 99, 119–125. https://pubmed.ncbi.nlm.nih.gov/24152684/
4. Jarzab B, Wiench M, Fujarewicz K, Simek K, Jarząb M, Oczko‐Wojciechowska M, Włoch J, Czarniecka A, Chmielik E, Lange Det al. (2005) Gene expression profile of papillary thyroid cancer: sources of variability and diagnostic implications. Cancer Res 65, 1587–1597. https://pubmed.ncbi.nlm.nih.gov/15735049/
5. Matsubara T, Ochiai T, Hayashida M, Akutsu T & Nacher JC (2019) Convolutional neural network approach to lung cancer classification integrating protein interaction network and gene expression profiles. J Bioinform Comput Biol 17, 1940007. https://pubmed.ncbi.nlm.nih.gov/31288636/
6. Gimeno L, Serrano‐López EM, Campillo JA, Cánovas‐Zapata MA, Acuña OS, García‐Cózar F, Martínez‐Sánchez MV, Martínez‐Hernández MD, Soto‐Ramírez MF, López‐Cubillana Pet al. (2020) KIR+ CD8+ T lymphocytes in cancer immunosurveillance and patient survival: gene expression profiling. Cancers (Basel) 12, 2991. https://pubmed.ncbi.nlm.nih.gov/33076479/
7. Aktas B, Kasimir‐Bauer S, Heubner M, Kimmig R & Wimberger P (2011) Molecular profiling and prognostic relevance of circulating tumor cells in the blood of ovarian cancer patients at primary diagnosis and after platinum‐based chemotherapy. Int J Gynecol Cancer 21, 822–830. https://pubmed.ncbi.nlm.nih.gov/21613958/
8. Iwahashi N, Sakai K, Noguchi T, Yahata T, Matsukawa H, Toujima S, Nishio K & Ino K (2019) Liquid biopsy‐based comprehensive gene mutation profiling for gynecological cancer using CAncer personalized profiling by deep sequencing. Sci Rep 9, 10426. 10.1038/s41598-019-47030-w - DOI https://pubmed.ncbi.nlm.nih.gov/31320709/
9. Asante D‐B, Calapre L, Ziman M, Meniawy TM & Gray ES (2020) Liquid biopsy in ovarian cancer using circulating tumor DNA and cells: ready for prime time? Cancer Lett 468, 59–71. https://pubmed.ncbi.nlm.nih.gov/31610267/
10. Rajeev Krishnan S, De Rubis G, Suen H, Joshua D, Lam Kwan Y & Bebawy M (2020) A liquid biopsy to detect multidrug resistance and disease burden in multiple myeloma. Blood Cancer J 10, 37. 10.1038/s41408-020-0304-7 - DOI https://pubmed.ncbi.nlm.nih.gov/32170169/
11. https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/differential-g
12. ene-expression

13. https://blog.bioturing.com/2022/06/02/the-basics-of-deseq2-a-powerful-tool-in-differential-expression-analysis-for-single-cell-rna-seq/
14. https://www.biostars.org/p/367191/
15. https://my.clevelandclinic.org/health/body/22999-ovaries
16. https://gtexportal.org/home/tissue/Ovary?tissueSelect=Ovary
17. https://pubmed.ncbi.nlm.nih.gov/29069589/