**FINAL PROJECT REPORT**

Jiayu Huang, Matthew Hibbs, and Jared Neikirk

University of Florida

CGS4144: Introduction to Bioinformatics

Professor Kiley Graim

November 22, 2022

**TABLE OF CONTENTS**

ABSTRACT…………………………………………………………………………………………………2

INTRODUCTION...…………………………………………………………………………………………3

METHODS..…………………………………………………………………………………………………4

RESULTS..…………………………………………………………………………………………………..8

CONCLUSION..…………………………………………………………………………………………….9

REFERENCES...…………………………………………………………………………………………...10

**ABSTRACT**

Using RNA-sequencing, we can reveal the presence and quantity of RNA in a biological sample at a given moment, analyzing the continuously changing cellular transcriptome. Specifically, RNA-sequencing facilitates the ability to look at differences in gene expression in different groups or treatments. We propose that using RNA-sequenced data from blood from participants of the COVID-19 Health Action Response for Marines (CHARM) study, that no gene features can be shown to exist in asymptomatic participants significantly more than symptomatic participants. The data was insufficient for identifying any features in asymptomatic participants that were different from participants with symptoms. The results of our methods were largely inconclusive in regards to any consistent significant differences between symptomatic and asymptomatic participants of the study. Based on these findings, we accept our null hypothesis that no statistical evidence exists within the study that shows a significant relationship between certain distinct features and symptomatic patients. This does not prove that no features exist that have an impact on a positive COVID-19 patient being asymptomatic. Our dataset was not large enough to prove that a significant relationship exists. Patterns underlying the data were discovered but disproven by inconsistent results between methods. Further studies will be relevant to explore if certain patterns found in this study persist consistently across a wider dataset.

**INTRODUCTION**

Using RNA-seq data from blood from participants of the COVID-19 Health Action Response for Marines (CHARM) study, what features can we identify from asymptomatic participants that are different from participants with symptoms and how strong of a predictor is the feature? The data used was acquired while investigating SARS-CoV-2 infections among U.S. Marine Recruits that underwent a 2-week quarantine at home after an initial outbreak (Letizia, 2022). The approach taken to address the question of specific differences between genomic features of asymptomatic and symptomatic patients was first standardizing, and examining the dataset for certain clustering group anomalies or significant differences. Standardization was required since sample data was taken at different timestamps which could disrupt accurate data from an unknown confounding variable. The motivation for exploring these differences comes from the abrupt nature of a worldwide pandemic that affected billions of people and threatened many with their lives (Gao, 2021) and has been previously the subject of many research and clinical studies (Yang, 2020). All of the statistical methods used to analyze and explore this dataset is readily available at the following GitHub repository: <https://github.com/jaredneikirk/CGS4144-Genome-Study>

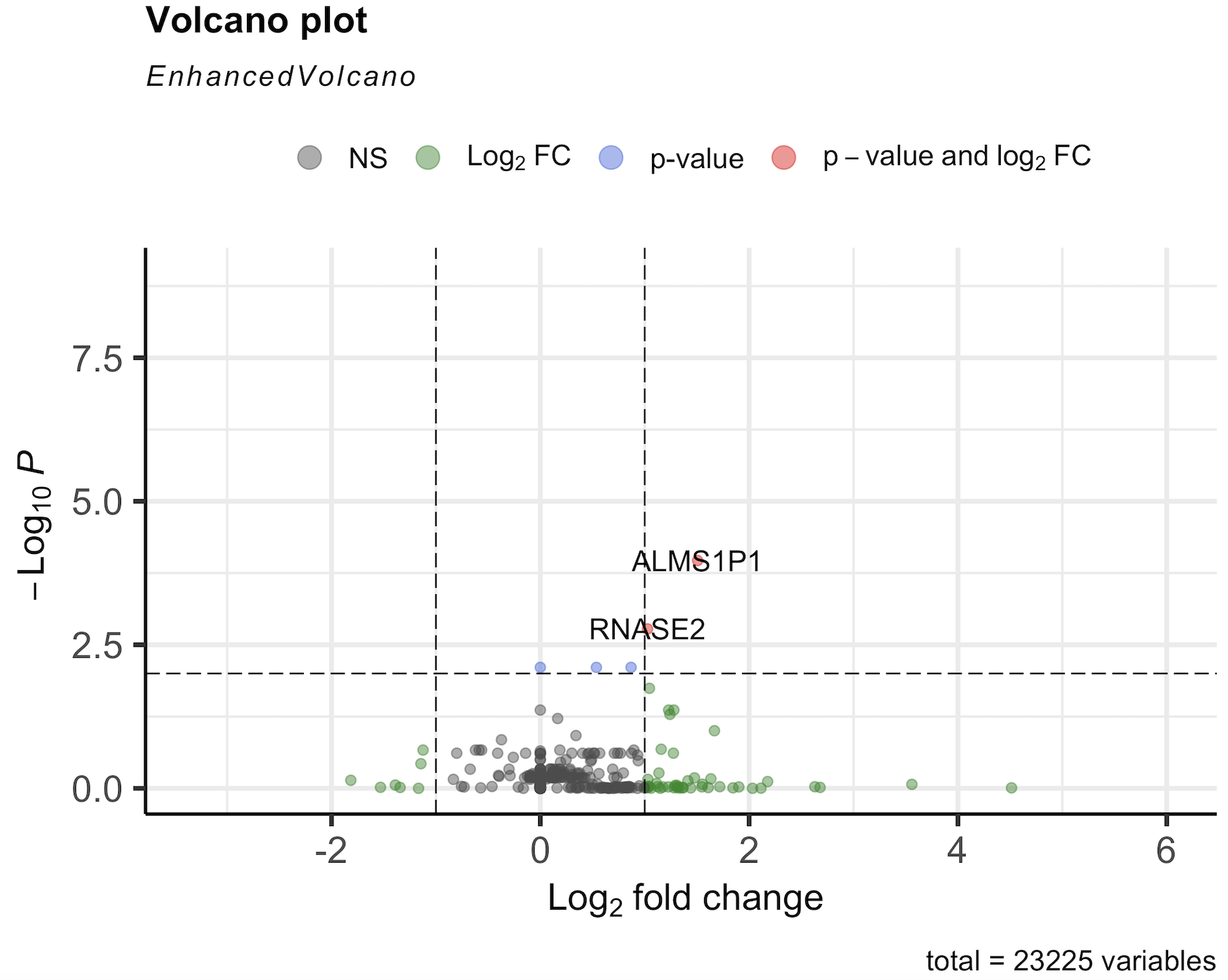
The dataset itself is available on GEO at the link below:

GEO Data: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE198449>

**METHODS**

**Differential Analysis**

To evaluate how genes differentially expressed among the dataset, we performed differential analysis and interpreted the result using volcano plot. Based on the volcano plot above, we observe that the ALMS1P1 and the RNASE2 gene are statistically significant and an outlier when attributing p-value and log2Fc. The log2 fold change of a couple of outlier genes is also observed 2-4 times greater than around the median log2 fold change. The gene data itself is also quite spread out amongst the x (log2 fold change) axis.

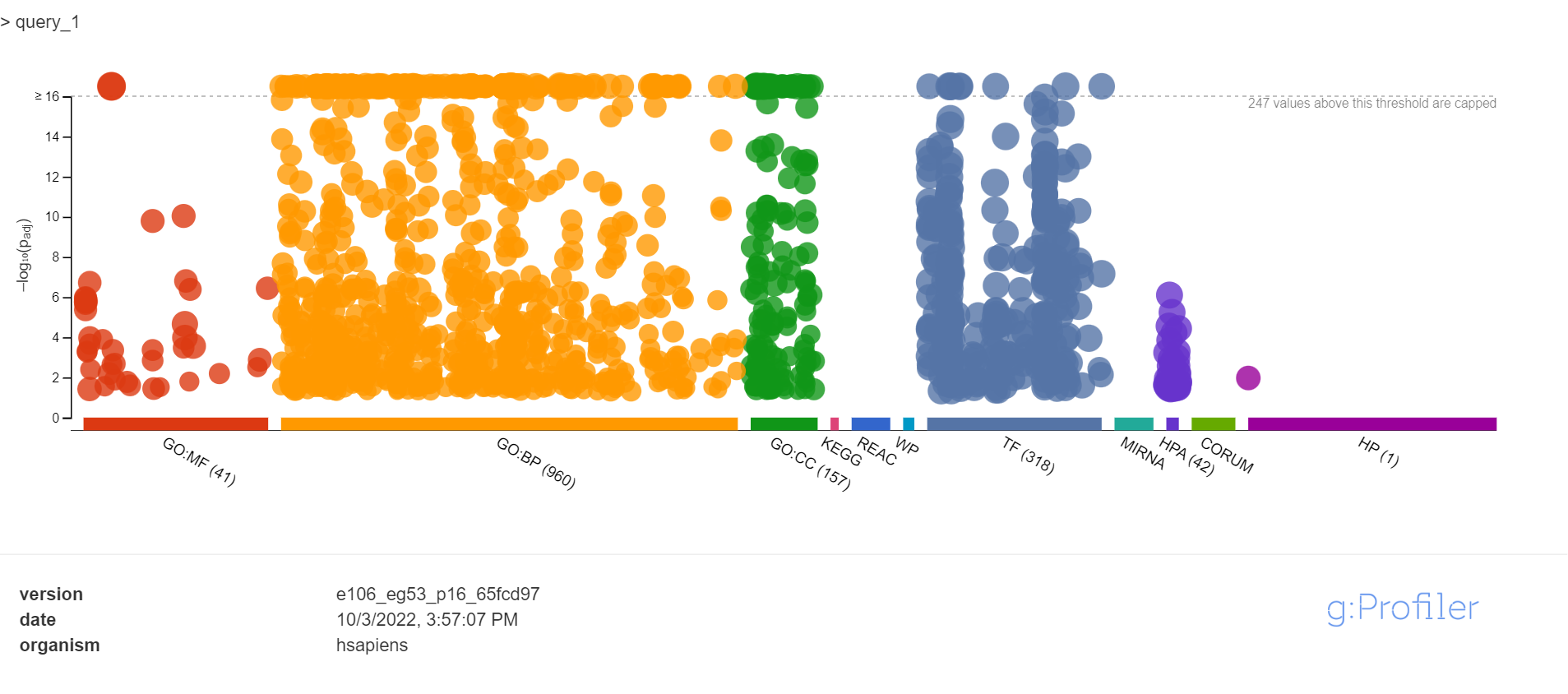


**Enrichment Analysis**

In order to categorize the genes we picked out, cluster analysis, and see if the functions of these genes have commonalities with our study, we implemented gene set enrichment analysis. In particular, we performed Gene Ontology (GO) term enrichment, a technique for interpreting sets of genes using the GO system of classification.

**gProfiler2**

gProfiler2 is a toolset for functional enrichment analysis and visualization, gene/protein/SNP identifier conversion and mapping orthologous genes across species. This tool allows us to better understand the biological functions of our significant genes.

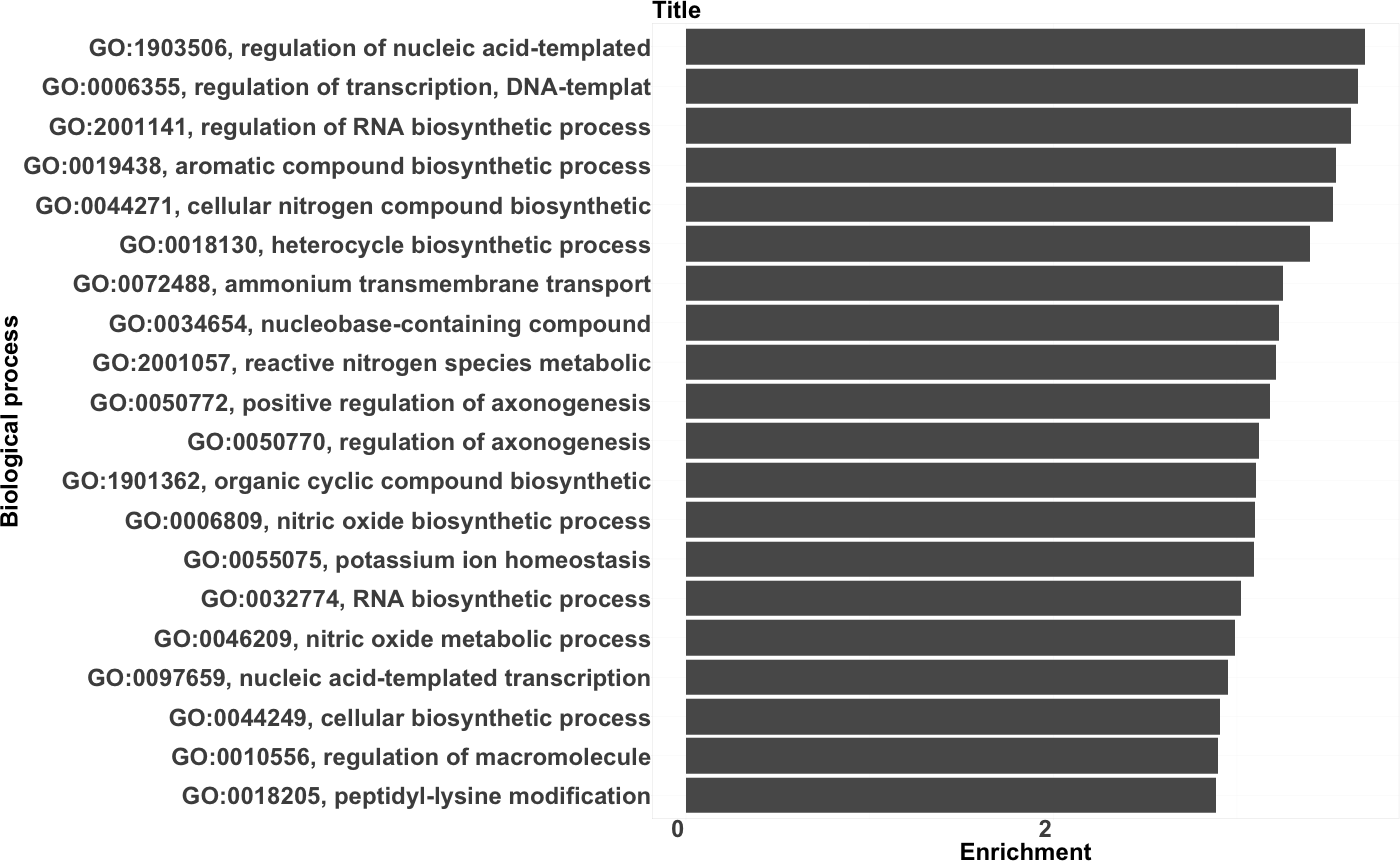


| **Source** | **Term ID** | **Term Name** | **P(adj)** |
| --- | --- | --- | --- |
| GO:BP | GO:0036293 | response to decreased oxygen levels | 3.053×10-2 |
| GO:BP | GO:0098542 | defense response to other organism | 4.288×10-2 |
| HP | HP:0000007 | Autosomal recessive inheritance | 1.071×10-2 |
| HPA | HPA:0210000 | fallopian tube | 1.228×10-3 |
| HPA | HPA:0060000 | bronchus | 1.150×10-2 |

R Script: <https://github.com/jaredneikirk/CGS4144-Genome-Study/blob/main/enrichment-scripts/gprofiler2.R>

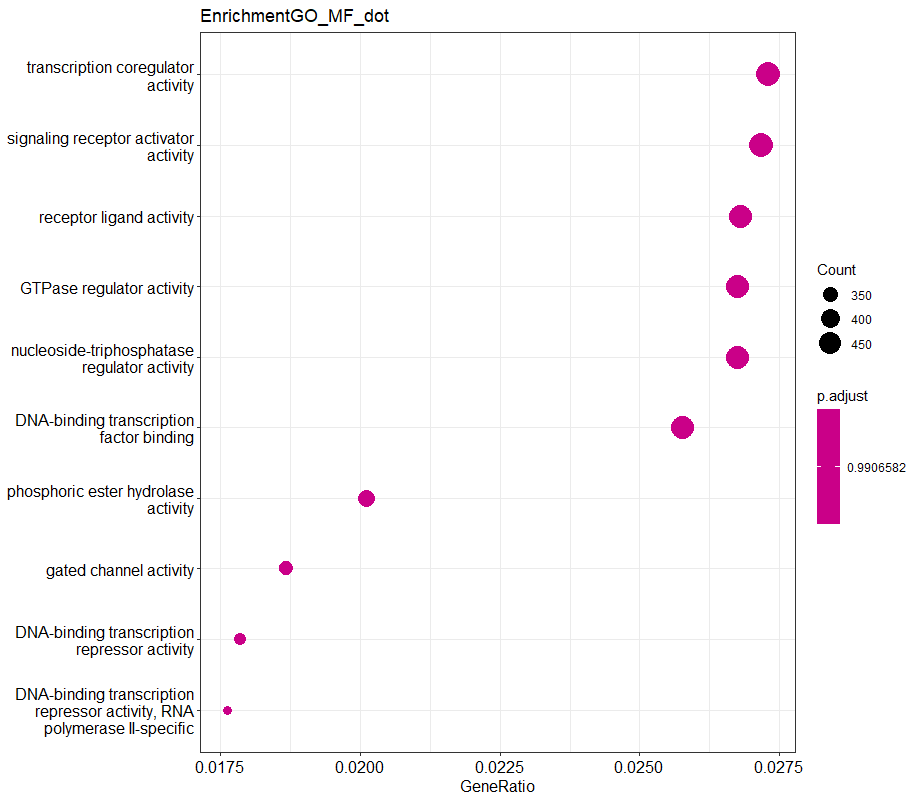
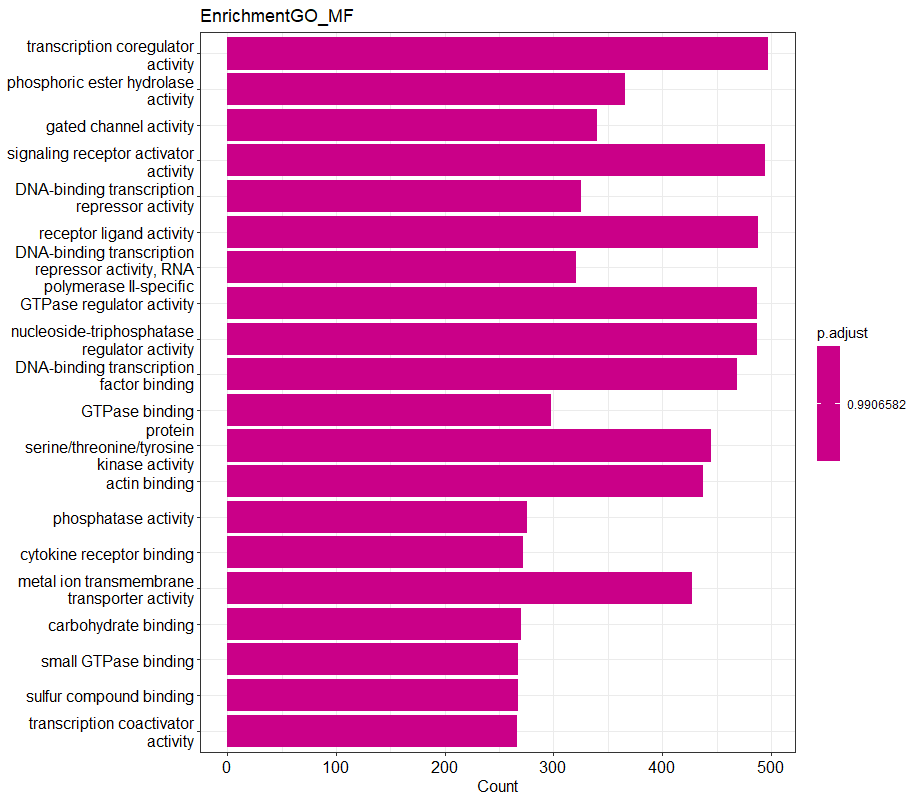
**topGo**

TopGo is a statistical package that provides tools for testing Gene Ontology terms that account for the Gene Ontology graph topography. In this case, a TopGo data object was generated using the significant genes from the dataset while using a Kolmogorov-Smirnov (K-S) test to visually rank the GO terms from the data object.

****

**clusterProfiler**

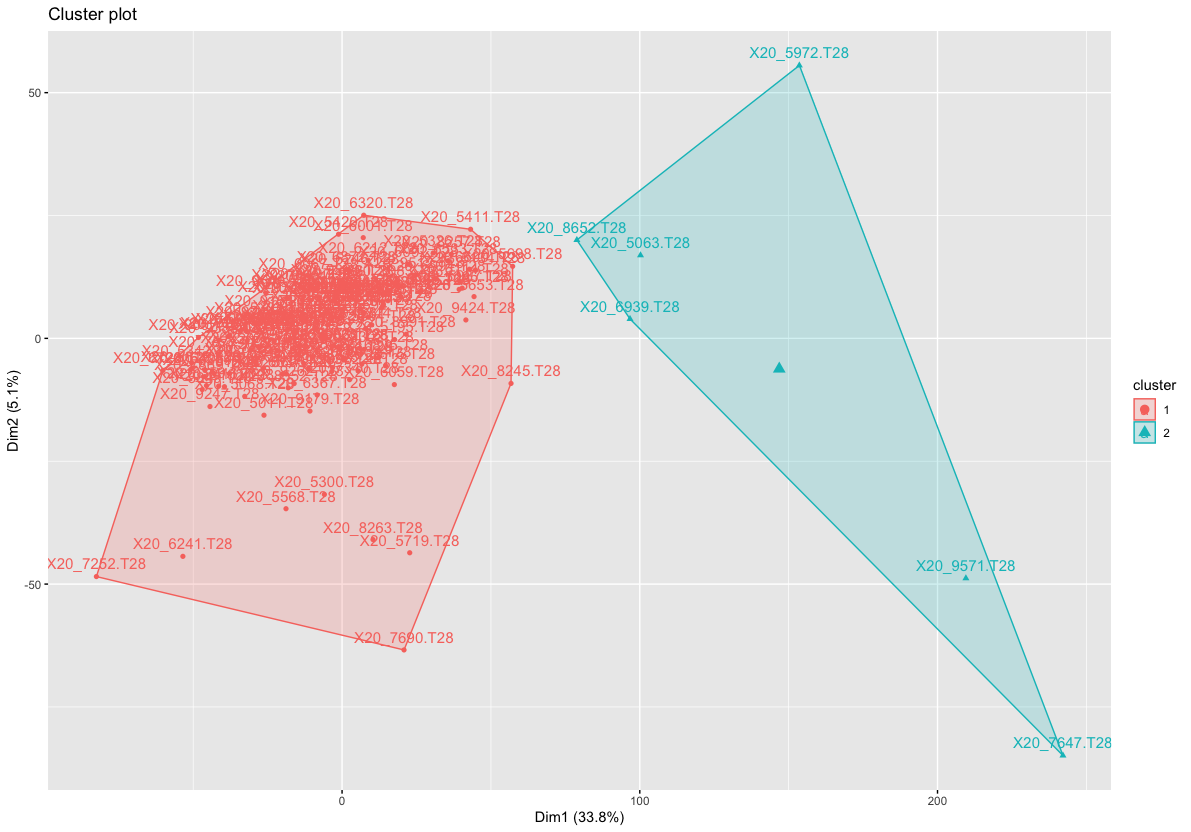
ClusterProfiler implements methods to analyze and visualize functional profiles of genomic coordinates, gene, and gene clusters.

****

**Clustering Algorithms**

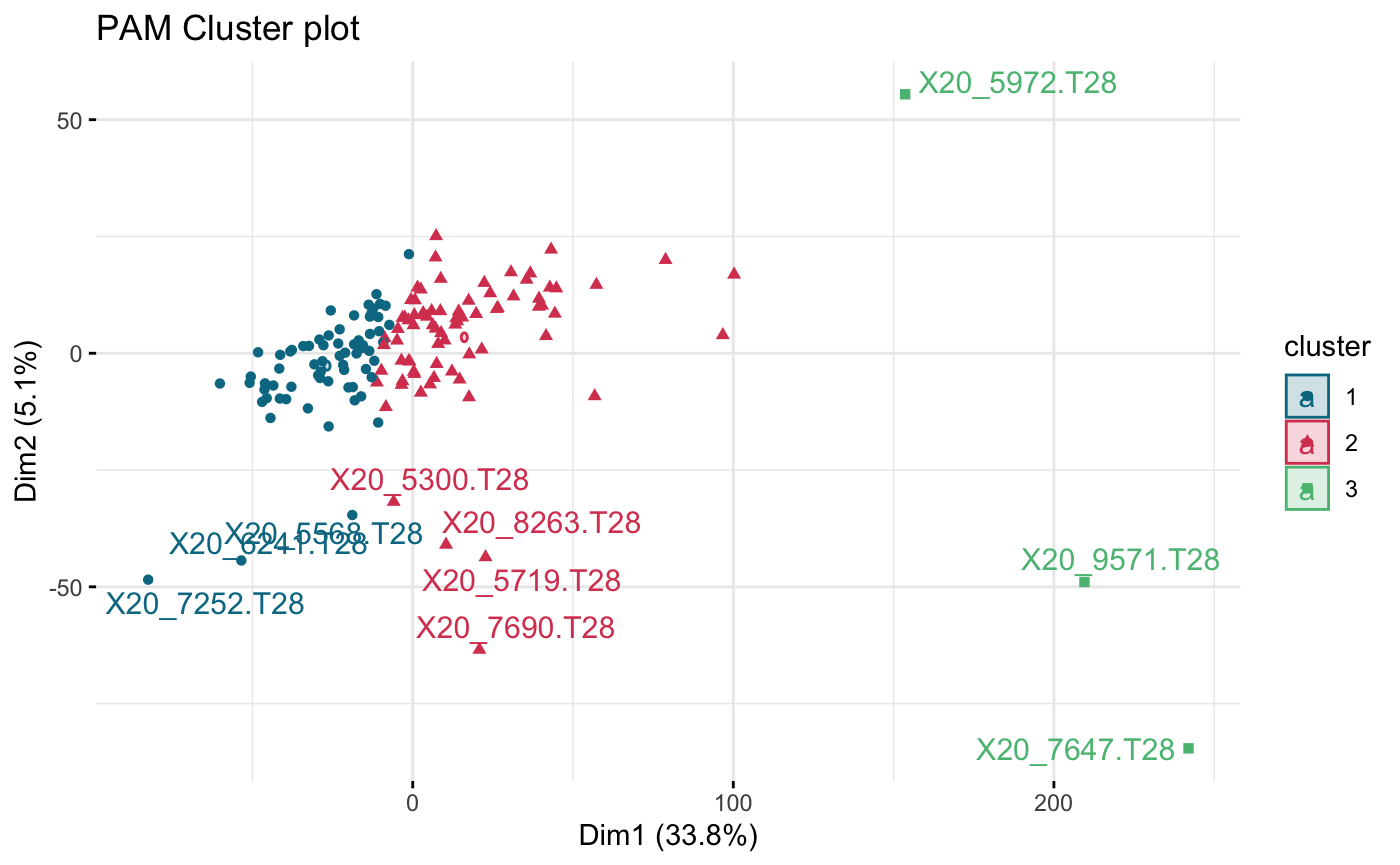
**K-means**

To further understand how differentially expressed genes varied between groups, we first implemented the K-means clustering algorithm with 5000 differentially expressed genes to analyze the expression data. K-means clustering is an unsupervised machine learning method that aims to group similar data points together and discover underlying patterns. To determine the best value for K, we compared the SSE for each K value from 1 to 10. When K equals 2, the SSE will not significantly decrease with each new addition of a cluster. In this case, 143 samples are clustered in group 1, and 6 samples are clustered in group 2. When K equals 3, there are 80 samples in group 1, 66 samples in group 2, and 3 samples in group 3.

****

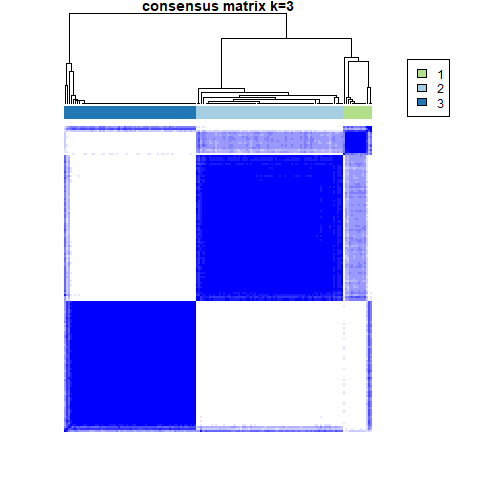
**PAM clustering**

After performing K-means clustering, a Silhouette analysis was performed on the 5000 most variable genes in the dataset, leading to an optimal *k* value of 3 for the PAM clustering algorithm. PAM is a clustering algorithm that partitions dimensional data around medoids, intending to find a sequence of medoids that are centrally located in the clusters.



**ConsensusClusterPlus**

In order to verify the results of K-means and PAM clustering, ConsensusClusterPlus was used to verify that the patterns underlying the data were not overly dependent on any one specific algorithm. ConsensusClusterPlus is also an unsupervised machine learning method that aims to group similar data points together and discover underlying patterns. When K equals 3, there are 75 samples in group 1, 70 samples in group 2, and 4 samples in group 3.

****

**Statistics**

|  | p (adj) | p-value |
| --- | --- | --- |
| k-means vs. PAM | 8.83 x 10-1 | 2.94 x 10-1 |
| k-means vs. ConsensusClusterPlus | 1.00 | 7.16 x 10-1 |
| PAM vs. ConsensusClusterPlus | 4.43 x 10-17 | 1.48 x 10-17 |

**RESULTS**

Upon in-depth analysis of the dataset, certain limitations were identified such as a small sample size of the symptomatic group when normalizing times for when participants were tested. Additionally, the dataset had inconsistent formatting where some genes were represented multiple times, so duplicates had to be removed and further decreased the total size of the dataset. Potentially because of these limitations, the results of our methods were largely inconclusive in regards to whether there were any consistent significant differences between the features of symptomatic and asymptomatic participants of the study. Certain methods may have revealed specific trends of clustering larger groups of genes of those that were symptomatic apart from asymptomatic individuals, but when comparing the results between the methods and trying to form a larger picture, the results lack consistency with the clustering of specific participants. Despite these results, other researchers have explored similar data and have had several interesting results, such as discovering significant differences in the allele frequency of the HLA genotype (*The influence of HLA*…, 2022)

**CONCLUSION**

The data collected using RNA-seq data from blood from participants of the COVID-19 Health Action Response for Marines study was insufficient for identifying what features from asymptomatic participants are different from participants with symptoms. The results of our methods were largely inconclusive in regards to any consistent significant differences between symptomatic and asymptomatic participants of the study. Based on these findings, we accept our null hypothesis that no statistical evidence exists within the CHARM study that shows a significant relationship between certain distinct features and symptomatic patients. Further research would be warranted to explore the genetic composition of patients who only exhibit extremely mild symptoms (Soares-Schanoski, 2022).

Taking ethics into consideration, all patient data was collected anonymously. Patient sex and age was not recorded directly in the dataset. The relationship between asymptomatic COVID patients and sex has been difficult to study due to the many confounding variables, including co-morbidities, differences in environment, fitness etc. (Sealfon, 2022). If we had to redo this project again, we would probably do more in depth research on a better dataset that would be more applicable to research on our hypothesis and offer a significantly larger sample size for both groups.

**REFERENCES**

*ASYMPTOMATIC SARS-COV-2 infection: A systematic review and Meta ...* (n.d.). Retrieved November 23, 2022, from <https://www.pnas.org/doi/10.1073/pnas.2109229118>

*The current status of gene expression profilings ... - wiley online library*. (n.d.). Retrieved November 23, 2022, from <https://onlinelibrary.wiley.com/doi/10.1002/ctd2.104>

Gao, W., Lv, J., Pang, Y., & Li, L.-M. (2021, December 2). *Role of asymptomatic and pre-symptomatic infections in covid-19 pandemic*. The BMJ. Retrieved November 22, 2022, from <https://www.bmj.com/content/375/bmj.n2342>

*The influence of HLA genotype on the severity of ... - Wiley Online Library*. (n.d.). Retrieved November 23, 2022, from <https://onlinelibrary.wiley.com/doi/full/10.1111/tan.14284>

Letizia, A. G., Al., E., Author AffiliationsFrom the Naval Medical Research Center, Michael, N. L., Others, M. R. K. and, Others, H. C. and, Others, M. L. O. D. and, Others, A. W. T. and, T. L. Cowger and Others, Y. Feng and F. Wang, & Mikuls, T. R. (2020, December 17). *SARS-COV-2 transmission among marine recruits during quarantine: Nejm*. New England Journal of Medicine. Retrieved November 22, 2022, from <https://www.nejm.org/doi/full/10.1056/nejmoa2029717>

Letizia, A. G., Arnold, C. E., Adhikari, B. N., Voegtly, L. J., Glang, L., Rice, G. K., Goforth, C. W., Schilling, M. A., Weir, D. L., Malagon, F., Ramos, I., Vangeti, S., Gonzalez-Reiche, A. S., Cer, R. Z., Sealfon, S. C., van Bakel, H., & Bishop-Lilly, K. A. (2021, December 8). *Immunological and genetic investigation of SARS-COV-2 reinfection in an otherwise healthy, Young Marine recruit*. MDPI. Retrieved November 22, 2022, from <https://www.mdpi.com/2076-0817/10/12/1589>

Sealfon, S. (2022, November 9). *Study explores sex differences in the effects of SARS-COV-2 in young adults*. Mount Sinai Health System. Retrieved November 22, 2022, from <https://www.mountsinai.org/about/newsroom/2022/study-explores-sex-differences-in-the-effects-of-covid-in-young-adults>

Sedegah, M., Porter, C., Hollingdale, M. R., Ganeshan, H., Huang, J., Goforth, C. W., Belmonte, M., Belmonte, A., Weir, D. L., Lizewski, R. A., Lizewski, S. E., Sealfon, S. C., Jani, V., Cheng, Y., Inoue, S., Velasco, R., Villasante, E., Sun, P., & Letizia, A. G. (n.d.). *Charm: Covid-19 health action response for marines–association of antigen-specific Interferon-gamma and IL2 responses with asymptomatic and symptomatic infections after a positive qpcr SARS-COV-2 test*. PLOS ONE. Retrieved November 22, 2022, from <https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0266691>

Soares-Schanoski, A., Sauerwald, N., Goforth, C. W., Periasamy, S., Weir, D. L., Lizewski, S., Lizewski, R., Ge, Y., Kuzmina, N. A., Nair, V. D., Vangeti, S., Marjanovic, N., Cappuccio, A., Cheng, W. S., Mofsowitz, S., Miller, C. M., Yu, X. B., George, M.-C., Zaslavsky, E., … Ramos, I. (1AD, January 1). *ASYMPTOMATIC SARS-COV-2 infection is associated with higher levels of serum IL-17C, matrix metalloproteinase 10 and fibroblast growth factors than mild symptomatic COVID-19*. Frontiers. Retrieved November 22, 2022, from <https://www.frontiersin.org/articles/10.3389/fimmu.2022.821730/full>

Yang, R., Gui, X., & Xiong, Y. (2020, May 1). *Comparison of clinical characteristics of patients with asymptomatic vs symptomatic coronavirus disease 2019 in Wuhan, China*. JAMA network open. Retrieved November 22, 2022, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7254178/>