Meta Analysis of COVID-19 Patient Data[1]

Amiel Hundley [1], Owen Gift [1], Kasey Bryan [1]

[1] School of Biomedical Engineering, Drexel University, USA

Course : Bmes483/543

Instructor: Ahmet Sacan

Date : 2022-06-07

Original Paper(s) : “Geo accession viewer,” National Center for Biotechnology Information. [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE183356&fbclid=IwAR0aBJrwZLemtXP1Laogrnc6XfWx7Gi3bDTnpWmFAOIEZeETSqTzgof-DJE. [Accessed: 07-Jun-2022].

Dataset(s) : https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE177477
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE164805
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE182279

[1] .

ABSTRACT

For this study, three datasets were used to analyze similar differentially expressed genes in covid-19 patients and healthy patients. If successful, this could greatly assist with treatment due to the pathophysiology information now being understood and cataloged. The three data sets were analyzed separately using MATlab, then integrated at the end to be compared. Since the data sets were so large, only the top 10% was used in this analysis to allow for MATlab to handle the large amounts of data. We found data sets 1 and 2 had a large quantity of shared differentially expressed genes.

1 INTRODUCTION

Need:

The first case of the SARS-CoV-2 virus was reported on December 1st, 2019, and since then over 6.5 million reported deaths have occured due to Covid-19 [1]. Symptoms of the virus vary from mild to moderate respiratory illness symptoms, such as (but not limited to) headache, cough, sore throat, fever or chills, and shortness of breath or difficulty breathing. On March 11, 2022 The World Health Organization declared Covid-19 a global pandemic due to the high-risk nature of the virus. Effective vaccines have been developed that reduce the risk of infection as a result of the virus; however, the vaccine does not fully prevent transmission of the virus, even when individuals are fully up-to-date on booster shots. Most people who are exposed to the virus are able to recover at home, but severe cases require antiviral treatments, IV infusions, and even ventilators for patients who have difficulty breathing. Far too many individuals and families have been impacted due to the Covid-19 virus, which exemplifies the need to conduct further research to find a cure.

Biology/Physiology:

In the 21st century, three coronaviruses have emerged in the population, SARS-CoV, MERS-CoV, and most recently SARS-CoV-2, which caused the current coronavirus disease COVID-19. The SARS-CoV-2 virus has been considered the most contagious of this century. To date, different variants have evolved which causes further implications regarding detection, transmission, vaccination, and severity.

To gain entry into the host cell, SARS-CoV-2 binds to the angiotensin-converting enzyme 2 (ACE2), which is located on the membrane surface of the host cells. Proteolytic activity allows for fusion of the S protein, which facilitates fusion of the virus to the host membrane. Next, the S protein is cleaved into S1 and S2 subunits. The S2 subunit is composed of fusion peptide (FP), and heptad repeat domains (HR) 1 and HR2. The FP is inserted into the host membrane, and the HR1 and HR2 interact to form a six-helix bundle which brings the viral envelope and cell membrane close enough for host penetration. Once in the cell, Coronavirus RNA transcription occurs in the host cytoplasm via the catalytic subunit RdRp. The discontinuous nature of the transcription process leads to extraordinary recombination rates, which can be as high as ~25%.

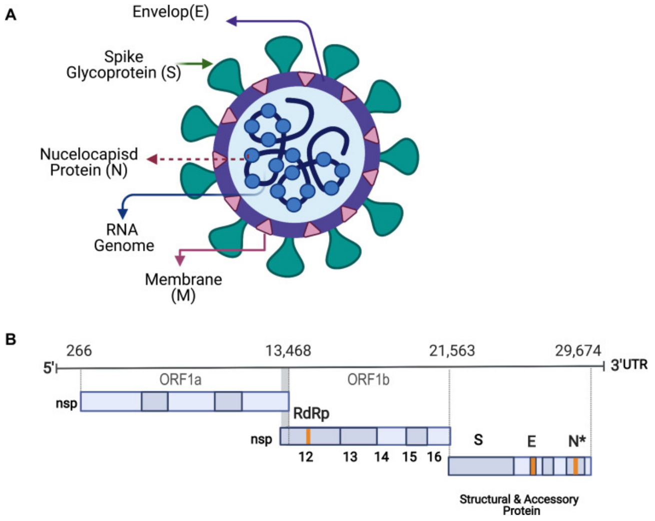


Figure 1: Schematic diagram of SARS-CoV-2

Three factors have been identified as possible contributors to the virulence of SARS-CoV-2, and they might impact pathogenicity. They include the following: 1) differences in the S protein RBD, 2) differences in properties of accessory proteins, and 3) the addition of a polybasic cleavage site in the S protein [2]. These distinguishers are important to note because they could be attributed to the high-risk factors of Covid-19, such as higher transmissibility and contagiousness.

Goals:

The goal of this study is to define a hypothesis identifying key mediators driving the pathophysiology of COVID-19 for diagnostic and therapeutic target testing. If successful, common differentially expressed genes between groups of patients will be identified which will help understand the principle host protective mechanisms against COVID-19 infection. As a result, this may enable improvement in treatment options in the future.

Related Work:

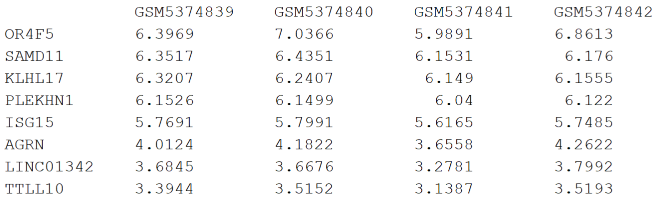
* Provide a short survey of related work. Put this study in the context of others. Check out the papers cited by your project paper and the papers that cite your project paper to find out what else is being done in this area. Use a citation manager (e.g., Endnote) to automate creation of references, e.g., [1].

2 Dataset

For this study, three initial datasets were used. The first study was published in 2021 and evaluated chilblain-like lesions during the COIV-19 pandemic compared to seasonal chilblain. This experiment was conducted through expression profiling by array, and the goal was to assess the gene expression of skin chilblain-like lesions and compare it with seasonal chilblain and healthy skin. For this study, 18 samples were included: 10 were chilblain-like lesions, 4 were seasonal chilblain, and 4 were healthy skin. The control data was "disease- 'healthy control'" and the experimental data was "disease: chilblain like lesion COVID-19" [3]. The second study was published in 2021 and evaluated the specific transcriptional signatures of severe COVID-19. This experiment was conducted through expression profiling by array, and the overall design was that microarray analysis of the whole genome transcriptome was applied to peripheral blood mononuclear cells (PBMCs) taken from severe and mild covid-19 patients as well as healthy controls. The control data was "PBMC, HC " and the experimental data was "PBMC, severe patient" [4]. The third and final study was published in 2021 and evaluated the Type I interferon pathways in SARS-CoV-2 infected individuals. The experiment was conducted through expression profiling by array, and the goal was to compare blood transcriptomes from COVID-19 patients and healthy patients using the Clariom S RNA microarray, Affymtrix assay. The patients were recruited from Pakistan/ The control data was "Disease state: healthy control" and the experimental data was "disease state: SARS-CoV-2 positive" [5].

3 METHODS

The goal was to find differentially expressed genes using MATlab. First, the initial datasets were read using bmes_downloadandparsegse() and bmes_downloadandparsegpl(), both provided by Dr. Ahmet Sacan and unaltered.. Once the GSE and GPL information was downloaded, the data was to extract the two columns each containing the ID and ORF information for each dataset. Then, for the genes column the gene names were extracted from the SPOT_ID description. Next, for each gseprobe,the gplprobes were searched and the corresponding gene was used. A map container was used to speed us through the process. If entries were not found, the probe name was kept in place.



```
            GSM5374839    GSM5374840    GSM5374841    GSM5374842
OR4F5       6.3969        7.0366        5.9891        6.8613
SAMD11      6.3517        6.4351        6.1531         6.176
KLHL17      6.3207        6.2407         6.149        6.1555
PLEKHN1     6.1526        6.1499          6.04         6.122
ISG15       5.7691        5.7991        5.6165        5.7485
AGRN        4.0124        4.1822        3.6558        4.2622
LINC01342   3.6845        3.6676        3.2781        3.7992
TTLL10      3.3944        3.5152        3.1387        3.5193
```

Figure 2: Example of data after formatted/cleaned

The data was then filtered to only include the top 10% of genes that varied most across samples using the genevarfilter function, and then compared using the pdist function to return the euclidean distance between pairs of observations. The data was displayed in hierarchical clustering graphs using the dendrogram function (as shown in figure 3 below) and heatmap clustergrams using the clustergram function (as shown in figure 4 below). Next, for each data set, the top 10 most different genes between the experimental and control groups were reported using the ttest2 function and displayed in a table with the gene name, P-value, and foldchange values that were calculated, as shown in figures 5, 6, and 7 below. Next, similar differentially expressed genes were found between the three data sets by comparing the top 10 for each of the three data sets and finding similarities. Finally, the similarly differentiated genes were displayed in a dendrogram graph, as shown in figure 8 below.

4 Experiments and Results

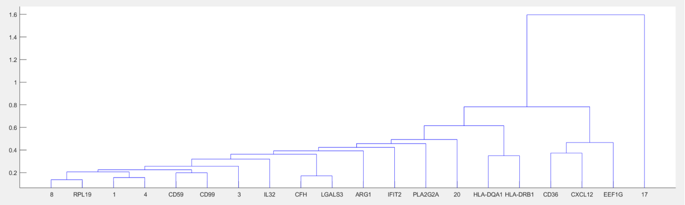


figure 3: Hierarchical clustering of samples

Figure 3 shows the result of hierarchical clustering for the first dataset. This process was done for each dataset. We chose to display the data in both a dendrogram and a heatmap in this stage in the data analysis. We chose to use hierarchical clustering because it is commonly used in exploratory data analysis, and it was a good way to help visually interpret the data. In this graph, genes that join together closer in the tree are more similar to each other, so we were able to see the similarity between the different genes.

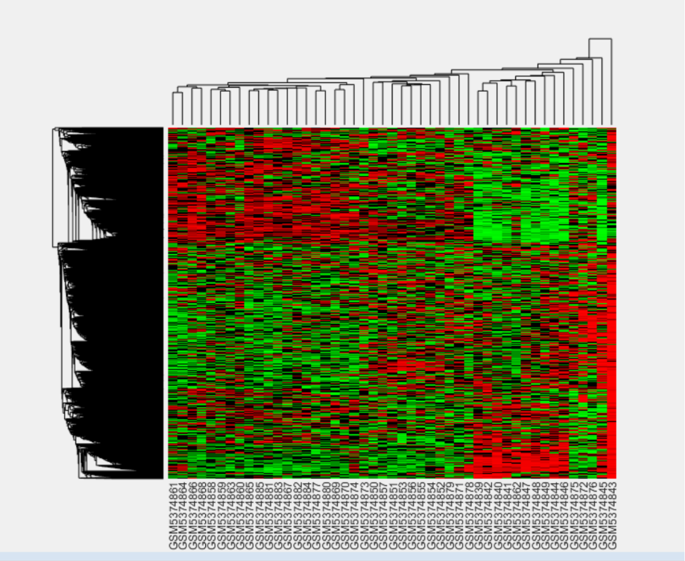
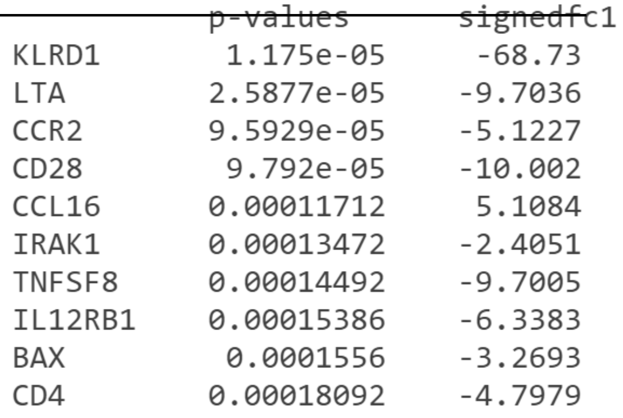


figure 4: Heatmap of gene expression

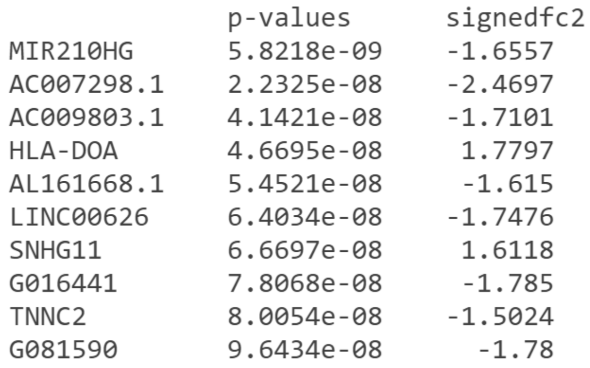
Figure 4 shows the result of heatmap gene expression for the first dataset. This process was done for each dataset. The heatmap shows how the genes are clustered, with red being more intense and green being less intense.



```
              p-values       signedfc1
KLRD1          1.175e-05      -68.73
LTA           2.5877e-05     -9.7036
CCR2          9.5929e-05     -5.1227
CD28           9.792e-05     -10.002
CCL16         0.00011712      5.1084
IRAK1         0.00013472     -2.4051
TNFSF8        0.00014492     -9.7005
IL12RB1       0.00015386     -6.3383
BAX            0.0001556     -3.2693
CD4           0.00018092     -4.7979
```

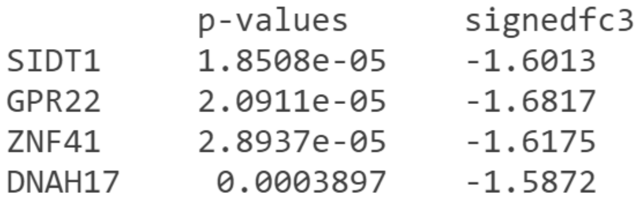
Figure 5: Report of the 10 most different genes between experimental and control groups in dataset 1

Figure 5 shows the top most different genes between the experimental and control groups in dataset 1. To determine these genes, a ttest was ran on each gene, and P-values<=0.01 were returned. Fold change is a measure describing how much a quantity changes between an original and a subsequent measurement.



```
             p-values      signedfc2
MIR210HG     5.8218e-09    -1.6557
AC007298.1   2.2325e-08    -2.4697
AC009803.1   4.1421e-08    -1.7101
HLA-DOA      4.6695e-08     1.7797
AL161668.1   5.4521e-08     -1.615
LINC00626    6.4034e-08    -1.7476
SNHG11       6.6697e-08     1.6118
G016441      7.8068e-08     -1.785
TNNC2        8.0054e-08    -1.5024
G081590      9.6434e-08      -1.78
```

Figure 6: Report of the 10 most different genes between experimental and control groups in dataset 2



```
           p-values       signedfc3
SIDT1      1.8508e-05     -1.6013
GPR22      2.0911e-05     -1.6817
ZNF41      2.8937e-05     -1.6175
DNAH17      0.0003897     -1.5872
```

Figure 7: Report of the 10 most different genes between experimental and control groups in dataset 3

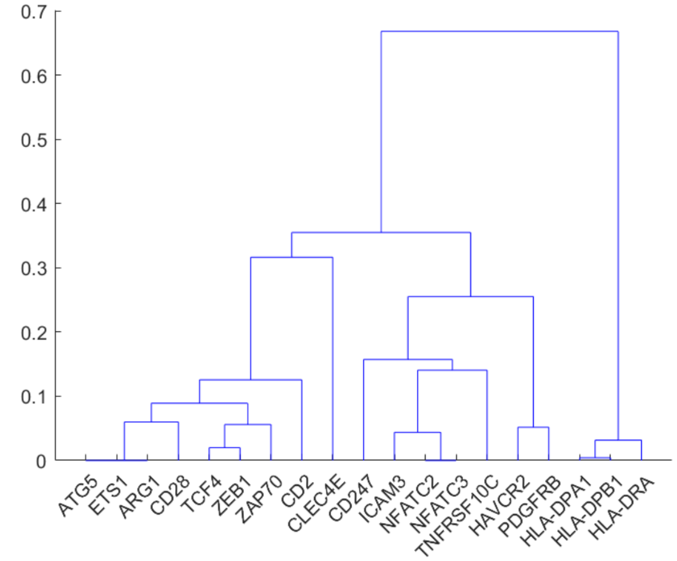


Figure 8: similar differentially expressed genes between all three data sets

Figure 8 shows a dendrogram of all similar differentially expressed genes between all three data sets. This graph was created using the top 10 most different genes from each of the three data sets

5 Discussion

Gene	Purpose
ATG5	responsible for the activation and the differentiation of various immune cells in innate and adaptive immunity
ETS1	associated with regulation of immune cell function and with an aggressive behavior in tumors that express it at high levels
ARG1	provides instructions for producing the enzyme arginase. This enzyme participates in the urea cycle, a series of reactions that occurs in liver cells. The urea cycle processes excess nitrogen, which is generated when proteins and their building blocks (amino acids) are used by the body
CD28	the CD28 pathway plays a central role in immune responses against pathogens, autoimmune diseases, and graft rejection
TCF4	plays a role in the maturation of cells to carry out specific functions (cell differentiation) and the self-destruction of cells (apoptosis)
ZEB1	Protein coding gene. Diseases associated with ZEB1 include Corneal Dystrophy, Posterior Polymorphous.
ZAP70	Protein coding gene: diseases associated with ZAP70 include immunodeficiency and autoimmune disease
CD2	Protein Coding gene. Diseases associated with CD2 include Mastocytosis, Cutaneous and Mastocytosis
CLEC4E	Protein Coding gene. Diseases associated with CLEC4E include Subcutaneous Mycosis and Fungal Keratitis
CD247	plays an essential role in adaptive immune response
ICAM3	functions not only as an adhesion molecule, but also as a potent signaling molecule. Among its related pathways are Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell
NFATC2	plays a central role in inducing gene transcription during the immune response.
TNFRSF10C	function as an antagonistic receptor that protects cells from TRAIL-induced apoptosis
HAVCR2	a Th1-specific cell surface protein that regulates macrophage activation, and inhibits Th1-mediated auto- and alloimmune responses, and promotes immunological tolerance
PDGFRB	essential for normal development of the cardiovascular system and aids in rearrangement of the actin cytoskeleton.
HLA-DPA1	plays a central role in the immune system by presenting peptides

	derived from extracellular proteins
HLA-DPB1	plays a central role in the immune system by presenting peptides derived from extracellular proteins
HLA-DRA	plays a central role in the immune system and response by presenting peptides derived from extracellular proteins

Listed above is a table listing all of the genes that were found to be differentially similar across the control and experiential groups. After researching the genes, they all had relevance to COVID-19 symptoms which is important because this validates the data. Although none explicitly mention COVID-19, genes relating to immune responses and cardiovascular pathways are notable for this study.

Summary:

This study was helpful in understanding differentially expressed genes in different studies related to COVID-19. This is important because COVID-19 has made a significant impact on almost every family in one way or another all over the world, and any research that is able to find relationships is important for the progression of finding a cure/treatment to the virus.

6 References

1. "Ihme: Covid-19 projections," *Institute for Health Metrics and Evaluation*. [Online]. Available: https://covid19.healthdata.org/global?view=cumulative-deaths&tab=trend. [Accessed: 07-Jun-2022].
2. S. Alsobaie, "Understanding the molecular biology of SARS-COV-2 and the COVID-19 pandemic: A Review," Infection and drug resistance, 16-Jun-2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8215902/. [Accessed: 07-Jun-2022].
3. "Geo accession viewer," *National Center for Biotechnology Information*, 19-Aug-2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE182279. [Accessed: 07-Jun-2022].
4. "Geo accession viewer," *National Center for Biotechnology Information*, 15-Jan-2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE164805. [Accessed: 07-Jun-2022].
5. "Geo accession viewer," *National Center for Biotechnology Information*, 01-Dec-2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE177477. [Accessed: 07-Jun-2022].
6. "Genecards - human genes | gene database | gene search," *GeneCards: the human gene database*. [Online]. Available: https://www.genecards.org/. [Accessed: 07-Jun-2022].