Deep Learning Analysis of Single-Cell Data Reveals Unique Genetic Features Of COVID-19 Severities

Elan Schonfeld (elan.schonfeld@columbia.edu)
Columbia Univeristy https://orcid.org/0000-0001-7368-1562

Research Article

Keywords:

Posted Date: June 12th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-1754568/v3

Abstract

COVID-19 was declared by the World Health Organization in 2020 to be a pandemic. Analysis of COVID-19 related genetic pathways allows for a better understanding of the possible effects and sequelae of the disease. Using 6178 scRNA sequenced human cells, having a status of control/mild/severe COVID-19 disease status, differential expression of genes and pathways was analyzed. Using Gene Set Enrichment Analysis (GSEA), mild COVID-19 was found to over-express the Influenza Pathway. In order to identify genes important in COVID-19 severity, a deep learning classifier was trained. Classifiers were repeatedly trained for this task using 10 randomly selected genes from the total number of 18,958 genes. The highest performing classifier (AUC=0.748) was trained using: *AC008626.1, SGO1, RHOBTB2, RBM41, NDUFAF4P1, COX5A, ZDHHC17, STX11, IPP, NUDT5* genes. These results further illustrate the other factors contributing to mild versus severe COVID-19, as well as evidence of potential misdiagnosis or overlapping pathway effects of Influenza and COVID-19.

Introduction

COVID-19 was declared by the World Health Organization in 2020 to be a pandemic.[1] The public health response to COVID-19 has consisted of three main factors: masks, isolation measures, and vaccination. In the case that someone is infected with COVID-19, their body begins an aggressive immune response. The body's first responses to infection are skin and inflammatory, and then move on to a T-cell and the lymphatic system response.[2]

To understand and characterize COVID-19 pathophysiology, single cell genomic expression has been used. Uses of this include differential expression and pathway analysis to analyze gene mutations in cancer, as well as to better understand the cardiovascular system and reasons for heart failure.[3,4] Pathway analysis, better known as Gene Set Enrichment Analysis (GSEA), is a method to identify groups of genes that are overexpressed or underexpressed between two groups of cells, and may have an association with disease phenotypes.[5] GSEA has been used in prior research on COVID-19 to determine which cell types are infected most and least.[6]

The analysis of pathways differentially expressed in mild and severe COVID-19 serves two main purposes. Firstly, it is used to lend insight into functional effects and mechanistic causes of COVID-19 and its symptoms. Secondly, it is used to gain an understanding of confounding variables that can impact COVID-19 diagnosis. Further exploring these subjects can potentially increase the ability of researchers to explore the pathophysiology of COVID-19, as well as inform clinical understanding. Information regarding the body's immune response can also be uncovered through differential expression and pathway analysis.

The specific genes implicated in COVID-19 are not fully known, but through deep learning, more information on this topic can be discovered. An increased understanding of the genetic changes undergone during mild and severe COVID-19 infections can be invaluable for those treating this disease, as it can offer information about how COVID-19 affects the body, the body's immune response, and inform a research direction.

Deep learning methods have been applied to COVID–19 to predict diagnosis from CT scans as well as to design inhibitor drugs with structural prediction making support.[7,8] Using similar algorithms, it is the goal of this work to shed light on genetic differences between mild and severe COVID cases. One key focus of this study was to determine factors that continue to the severity of a COVID-19 infection, and if COVID-19 severity has a notable relationship with genetic expression. Lastly, we aimed to investigate if the differential genetic expression can be used to predict the severity of a COVID-19 infection.

Methods

The cells from the Single Cell Expression Atlas database were labeled according to the following definitions. Mild disease was defined as having no or limited clinical symptoms and not requiring computed tomography (CT) scanning or hospitalization. Moderate disease was defined as being symptomatic, with dyspnea and radiological findings of pneumonia upon thoracic CT scan, requiring hospitalization with a maximum of 9 L/min of oxygen. Severe disease was defined as respiratory distress requiring admission into the ICU.[9] Mild and moderate cases were then combined for data analysis. Raw data was downloaded from the Single Cell Expression Atlas website.[9] Next, the package "Seurat" was used to combine the raw data, gene names, and cell names into a Seurat object.[10] The data was then normalized, and features (important genes) were determined. UMAP reduction was performed on the Seurat object with the purpose of visualization of genetic differences.[10] PCA was then used to perform linear dimensionality reduction.[10] Metadata from the Single Cell Expression Atlas was integrated into the dataset, and cells were grouped into vectors by their COVID-19 status (control, mild COVID-19, severe COVID-19).[9] Cells were then clustered using hierarchical clustering. Differential expression of genes between COVID-19 status groups was calculated, and, using this information, the fGSEA package was used to calculate differentially expressed pathways between these groups. Next, mild and severe COVID-19 cell data were exported. All cells with a control/normal phenotype were not included in subsequent analysis. Preprocessed data was loaded into Python 3. 4527 cells were separated into a training set of 3527 training set cells and 1000 validation set cells. The validation set was held out and not used for the training of any models. A uniform random sample of 1000 genes was selected from the total population of 18,958 genes (data available upon request). A fully connected, deep learning model was built using TensorFlow and Keras.[11,12] (Figures 1 and 2)

The model was trained for 50 epochs. To determine the performance of the model, an Area Under Curve (AUC) metric was calculated using the validation set. The AUC is a metric to account for unequally weighted classes. It is the area under the Receiver Operating Characteristic Curve which is a plot of True Positive Rate versus False Positive Rate for varying thresholds to judge the predictive value of an algorithm. An AUC of 0.50 would represent no predictive value and an AUC of 1.0 would represent perfect predictive value.

Next, to determine predictive genes for Mild Versus Severe COVID-19 status, a random sample of 10 equally weighted genes is selected from the population of all 18,958 genes. A model, using the architecture and hyperparameters as that specified above, was built. Expression data of the 10 randomly selected genes was used as input for training the model. This process was repeated 1000 times. On every iteration, the validation AUC was compared to the current best performing model's validation AUC. If the current model scored higher, the current 10 random genes were recorded. The maximum validation AUC and its corresponding 10 genes are returned.

Results

GSEA returned notable results when comparing pathway expression between cells with mild COVID-19 and those with severe COVID-19 (Figure 3). Figure 3 details which pathways are most over and under-expressed in mild COVID-19 compared to severe COVID-19. As seen in figure 3, the Reactome Selenoamino Acid Metabolism is significantly overexpressed in mild COVID-19 compared to in severe COVID-19, with an NES of 5.44 and an adjusted p-value of . The Reactome Influenza Infection is also significantly overexpressed in mild COVID-19 compared to severe COVID-19, with an NES of 5.20 and an adjusted p-value of . The Zhou Inflammatory Response FMA Up and Zhou Inflammatory Response Live Up pathways were both less expressed in mild COVID-19 compared to severe COVD-19. The Zhou Inflammatory Response FMA Up pathway had an NES of -3.18 and an adjusted p-value of. The Zhou Inflammatory Response LIVE Up pathway had an NES of -3.26 and an adjusted p-value of .

It was further determined that the Influenza pathway is overexpressed due to upregulation of the majority of genes, rather than a select few outliers (Figure 4). Figure 4 details the Reactome Influenza Infection pathway, and each gene. The bars on the x-axis are in the order of the ranked genes (genes sorted from most overexpressed to most underexpressed in mild COVID-19 compared to severe COVID-19), with each bar representing a gene in the Reactome Influenza Infection pathway. Enrichment scores are listed on the y-axis, and the points are connected, showing that many genes that are part of the Reactome Influenza Infection pathway are overexpressed in mild COVID-19 compared to severe COVID-19, and that the significant NES obtained is not due to outliers, but rather to a general overexpression of genes contained in the pathway.

Upon repeated random selection of 10 genes to be used as input when training the deep learning models, the highest performing model (Validation AUC of 0.748) was trained using input of the following genes: *AC008626.1, SGO1, RHOBTB2, RBM41, NDUFAF4P1, COX5A, ZDHHC17, STX11, IPP, NUDT5* (Figure 5). Figure 5 details that certain genes differentiate control cells from those infected with COVID-19, ie. *SGO1* and *RHOBTB2*. Additionally, according to figure 5, genes COX5A is highest expressed in severe COVID-19. NUDT5, however, is highly expressed in both control cells and cells infected with mild COVID-19, but less so in those infected with severe COVID-19.

Discussion

As determined by Differential Expression and GSEA, pathway analysis suggests that severe cases of Influenza may have been improperly diagnosed as mild COVID-19. Not only is the Influenza pathway highly expressed in the mild COVID-19 group, but the selenoamino acid metabolism is also positively enriched. The selenoamino acid metabolism pathway is a well-known pathway that is implicated in successful immune responses to Influenza infection.[14] However, misdiagnosis may not be the cause of this result. Influenza may have similar mechanisms as does mild COVID-19, causing the overlapping pathway expression.

The highest achieving AUC (0.748) obtained by the deep learning training demonstrates that the neural network can successfully model to a reasonable extent a genetic component of COVID-19 severity. However, it is likely that there are non-genetic factors contributing towards COVID-19 severity (eg: socioeconomic circumstances, region, and demographics). Ten genes were discovered in this study to be sufficient to train a highly predictive deep learning model. SGO1 is involved in cell replication by promoting centromeric cohesion.[15] *RHOBTB2* is a GTPase also involved in mitotic cell division.[16] *COX5A*, which assists in establishing the proton gradient in the mitochondrial respiratory chain, possibly implicates this pathway as a defining characteristic of severe COVID-19.[17] *IPP* codes for an actin interacting protein; actin is involved in cell movement, cell structure, and stability which responds to GTPase activity and thus may be linked to *RHOBTB2*.[18] GTPase function has been implicated by some members of the 10 gene set; furthermore, an additional member of the gene set, *NUDT5* is responsible for hydrolysis of modified nucleosides.[19] *ZDHHC17*, a Zinc Finger Palmitoyltransferase has been implicated in initiating endocytosis at the plasma membrane; endocytosis is crucial for SARS–CoV-2 cellular uptake.[20-22] Related to the endocytosis events, *STX11* is involved in fusion of transport vesicles subcellularly.[23] Single cell expression data of these genes was sufficient to train a deep learning model with high predictive value, thereby implicating their expression patterns in the distinction between mild versus severe COVID–19.

A limitation of this study is that the database utilized cells with COVID-19 severity determined by severity of clinical symptoms, as well as whether a CT scan, oxygen supply, or hospitalization was necessary. Statistically, these metrics accurately label the cells, however, the diagnosis of these cells is subjective and future work should use more systematic approaches to label the COVID-19 cell severity. Possibility for future analysis can include a control group of normal cells as well as a control group of Influenza infected cells. This would allow for the comparison of Influenza infected cells to mild COVID-19 infected cells. Additionally, future experiments should include a larger number of cells to improve the statistical power of the analysis. scRNA and informatics will offer crucial insights into fighting COVID-19 on both the cellular and epidemiological level.

References

1. Lv M, Luo X, Estill J, et al. Coronavirus disease (covid-19): A scoping review. Eurosurveillance. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7175649/. Published April 2020. Accessed November 16, 2021.

2. Chowdhury MA, Hossain N, Kashem MA, Shahid MA, Alam A. Immune response in COVID-19: A Review. Journal of infection and public health. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7359800/. Published November 2020. Accessed November 16, 2021.
3. Zhang Y, Wang D, Peng M, et al. Single-cell RNA sequencing in cancer research. Journal of Experimental & Clinical Cancer Research. https://jeccr.biomedcentral.com/articles/10.1186/s13046-021-01874-1#:~:text=scRNA%2Dseq%20shows%20that%20one,tumor%20formation%2C%20and%20cancer%20relapse. Published March 1, 2021. Accessed November 16, 2021.
4. Ramirez Flores RO, Lanzer JD, Holland CH, et al. Consensus transcriptional landscape of human end-stage heart failure. Journal of the American Heart Association. https://www.ahajournals.org/doi/10.1161/JAHA.120.019667. Published March 31, 2021. Accessed November 18, 2021.
5. Higdon R. Gene set and protein set expression analysis | SpringerLink. Encyclopedia of Systems Biology. https://link.springer.com/referenceworkentry/10.1007%2F978-1-4419-9863-7_1209. Accessed November 16, 2021.
6. Li G, Ruan S, Zhao X, Liu Q, Dou Y, Mao F. Transcriptomic signatures and repurposing drugs for COVID-19 patients: Findings of bioinformatics analyses. Computational and structural biotechnology journal. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7719282. Published 2021. Accessed November 16, 2021.
7. Zhavoronkov A, Aladinskiy V, Zhebrak A, et al. Potential Covid-2019 3c-like protease inhibitors designed using generative deep learning approaches. ChemRxiv. https://chemrxiv.org/articles/preprint/Potential_2019-nCoV_3C-like_Protease_Inhibitors_Designed_Using_Generative_Deep_Learning_Approaches/11829102/2. Published February 19, 2020. Accessed November 16, 2021.
8. Babukarthik RG, Adiga V, Sambasivam G., et al. Prediction of COVID-19 Using Genetic Deep Learning Convolutional Neural Network (GDCNN). IEEE Xplore temporarily unavailable. https://ieeexplore.ieee.org/abstract/document/9201297. Published 2020. Accessed November 16, 2021.
9. Silvin A, Chapuis N, Dunsmore G et al. Elevated Calprotectin and Abnormal Myeloid Cell Subsets Discriminate Severe from Mild COVID-19. Cell. https://www.cell.com/cell/fulltext/S0092-8674(20)30993-4?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0092867420309934%3Fshowall%3Dtrue#article. Published 2021. Accessed November 16, 2021.
10. Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573-3587.e29. doi:10.1016/j.cell.2021.04.048
11. Martín Abadi, Ashish Agarwal, Paul Barham, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Published online 2015. https://www.tensorflow.org/
12. Chollet F, Others. Keras. Published 2015. https://keras.io
13. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv.org. https://arxiv.org/abs/1412.6980. Published January 30, 2017. Accessed November 16, 2021.
14. Sheridan PA, Zhong N, Carlson BA, Perella CM, Hatfield DL, Beck MA. Decreased selenoprotein expression alters the immune response during influenza virus infection in mice. J Nutr. 2007;137(6):1466-1471. doi:10.1093/jn/137.6.1466
15. Zhang Q, Liu H. Functioning mechanisms of Shugoshin-1 in centromeric cohesion during mitosis. Essays Biochem. 2020;64(2):289-297. doi:10.1042/EBC20190077
16. David M, Petit D, Bertoglio J. Cell cycle regulation of Rho signaling pathways. Cell Cycle. 2012;11(16):3003-3010. doi:10.4161/cc.21088
17. Vercellino, I., Sazanov, L.A. The assembly, regulation and function of the mitochondrial respiratory chain. Nat Rev Mol Cell Biol (2021). https://doi.org/10.1038/s41580-021-00415-0
18. Roca-Cusachs P, Iskratsch T, Sheetz MP. Finding the weakest link: exploring integrin-mediated mechanical molecular pathways. J Cell Sci. 2012;125(Pt 13):3025-3038. doi:10.1242/jcs.095794
19. Takao Arimori, Haruhiko Tamaoki, Teruya Nakamura, et al. Diverse substrate recognition and hydrolysis mechanisms of human NUDT5, Nucleic Acids Research, Volume 39, Issue 20, 1 November 2011, Pages 8972–8983, https://doi.org/10.1093/nar/gkr575

20. ZDHHC17 zinc finger DHHC-type palmitoyltransferase 17 [homo sapiens (human)] - gene - NCBI. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/gene/23390. Accessed November 16, 2021.
21. Roshni R. Singaraja, Shinji Hadano, Martina Metzler, et al. HIP14, a novel ankyrin domain-containing protein, links huntingtin to intracellular trafficking and endocytosis, Human Molecular Genetics, Volume 11, Issue 23, 1 November 2002, Pages 2815–2828, https://doi.org/10.1093/hmg/11.23.2815
22. Ole H Petersen, Oleg V Gerasimenko, Julia V Gerasimenko, Endocytic uptake of SARS-CoV-2: the critical roles of pH, Ca2+, and NAADP, Function, Volume 1, Issue 1, 2020, zqaa003, https://doi.org/10.1093/function/zqaa003
23. Offenhäuser, C., Lei, N., Roy, S., et al. (2011), Syntaxin 11 Binds Vti1b and Regulates Late Endosome to Lysosome Fusion in Macrophages. Traffic, 12: 762-773. https://doi.org/10.1111/j.1600-0854.2011.01189.x

Declarations

Ethics approval and consent to participate:

This research did not contain any studies involving animal or human participants, nor did it take place in any private or protected areas. No specific permissions were required for corresponding locations.

Consent for publication:

I give my full permission for the publication of this manuscript.

Availability of data and materials:

The data that support the findings of this study are available in the Single Cell Expression Atlas at https://www.ebi.ac.uk/gxa/sc/experiments/E-MTAB-9221/results/tsne.

Competing interests:

The author declares no conflict of interest.

Funding:

The author declares no sources of funding.

Authors' contributions:

Elan Schonfeld: conceptualization, data curation, formal analysis, software, investigation, visualization, methodology, and writing—original draft, review, and editing.

Acknowledgments:

Not applicable.

Figures

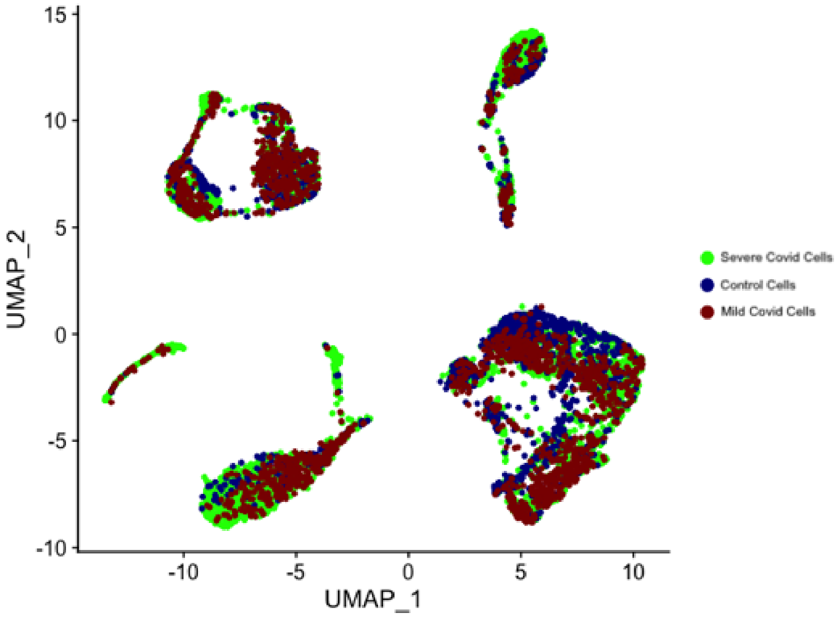
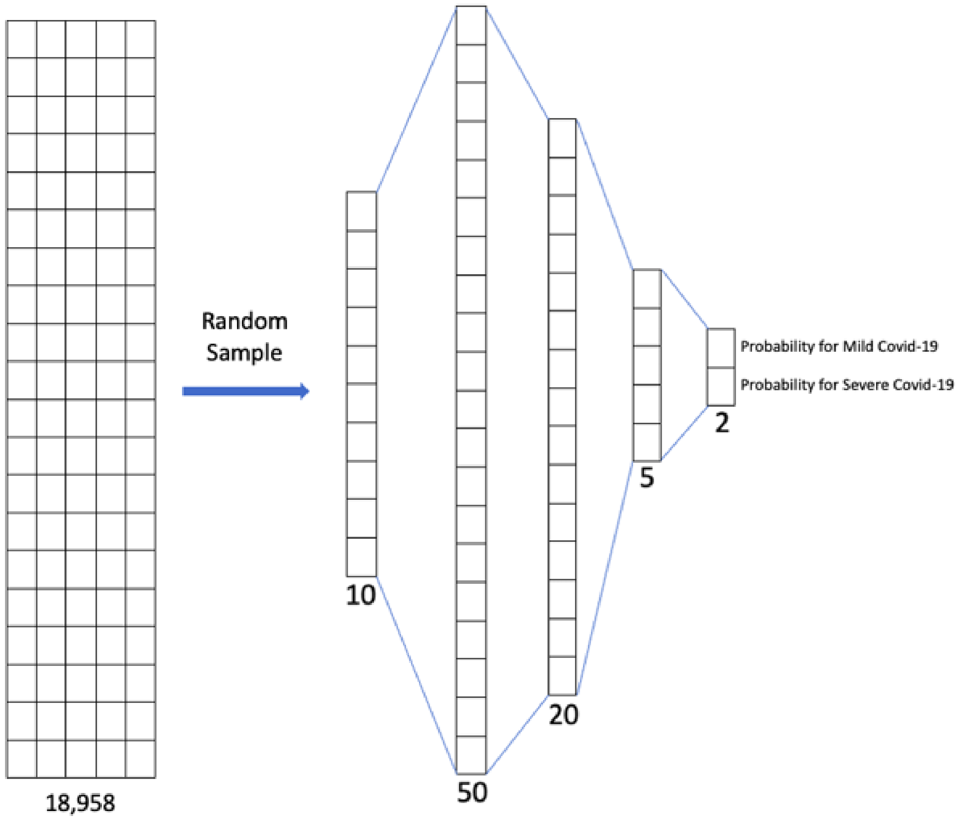


Figure 1

Dimensional Reduction Plot of All Cells, Labeled by Control/COVID-19 Severity

The UMAP dimensional reduction plot provides evidence that there does not exist any linear boundary that would have great predictive power in classifying COVID cell severity. Thus, deep learning was necessary rather than a simple logistic regression that is confined to a linear decision boundary prediction.

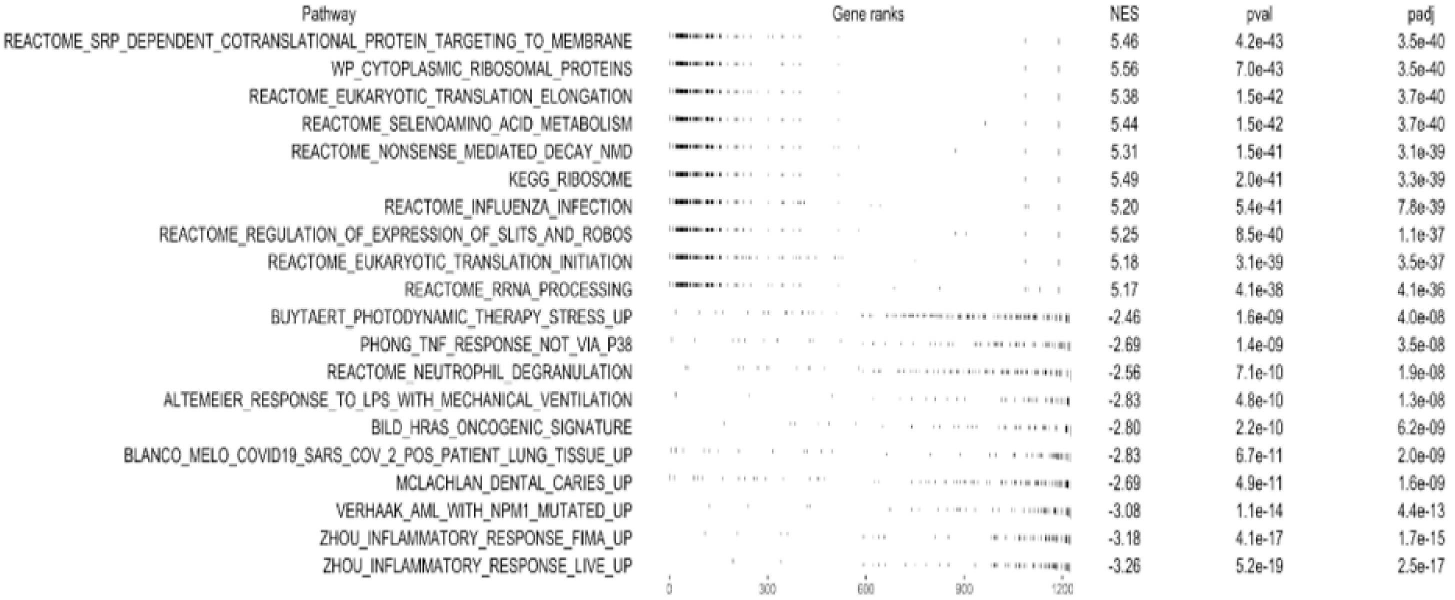


18,958
Random Sample
10
50
20
5
2
Probability for Mild Covid-19
Probability for Severe Covid-19

Figure 2

Deep Learning Model Construction

Layer sizes were: 50, 20, 5, 2; the input layer was of size 10. Activations of ReLU were used except a sigmoid activation for the last layer. ADAM optimizer and a Sparse Categorical Cross Entropy loss function were used.[13]



Pathway	NES	pval	padj
REACTOME_SRP_DEPENDENT_COTRANSLATIONAL_PROTEIN_TARGETING_TO_MEMBRANE	5.46	4.2e-43	3.5e-40
WP_CYTOPLASMIC_RIBOSOMAL_PROTEINS	5.56	7.0e-43	3.5e-40
REACTOME_EUKARYOTIC_TRANSLATION_ELONGATION	5.38	1.5e-42	3.7e-40
REACTOME_SELENOAMINO_ACID_METABOLISM	5.44	1.5e-42	3.7e-40
REACTOME_NONSENSE_MEDIATED_DECAY_NMD	5.31	1.5e-41	3.1e-39
KEGG_RIBOSOME	5.49	2.0e-41	3.3e-39
REACTOME_INFLUENZA_INFECTION	5.20	5.4e-41	7.8e-39
REACTOME_REGULATION_OF_EXPRESSION_OF_SLITS_AND_ROBOS	5.25	8.5e-40	1.1e-37
REACTOME_EUKARYOTIC_TRANSLATION_INITIATION	5.18	3.1e-39	3.5e-37
REACTOME_RRNA_PROCESSING	5.17	4.1e-38	4.1e-36
BUYTAERT_PHOTODYNAMIC_THERAPY_STRESS_UP	-2.46	1.6e-09	4.0e-08
PHONG_TNF_RESPONSE_NOT_VIA_P38	-2.69	1.4e-09	3.5e-08
REACTOME_NEUTROPHIL_DEGRANULATION	-2.56	7.1e-10	1.9e-08
ALTEMEIER_RESPONSE_TO_LPS_WITH_MECHANICAL_VENTILATION	-2.83	4.8e-10	1.3e-08
BILD_HRAS_ONCOGENIC_SIGNATURE	-2.80	2.2e-10	6.2e-09
BLANCO_MELO_COVID19_SARS_COV_2_POS_PATIENT_LUNG_TISSUE_UP	-2.83	6.7e-11	2.0e-09
MCLACHLAN_DENTAL_CARIES_UP	-2.69	4.9e-11	1.6e-09
VERHAAK_AML_WITH_NPM1_MUTATED_UP	-3.08	1.1e-14	4.4e-13
ZHOU_INFLAMMATORY_RESPONSE_FIMA_UP	-3.18	4.1e-17	1.7e-15
ZHOU_INFLAMMATORY_RESPONSE_LIVE_UP	-3.26	5.2e-19	2.5e-17

Figure 3

Top Up and Down Expressed Pathways Between Mild and Severe COVID-19

Figure 3 gives a column of gene ranks, where each vertical line represents a gene in the pathway and shows how over or under-expressed that particular gene is. The NES column contains normalized enrichment scores, a quantitative measurement of how much a pathway is over/under-expressed. P-values and adjusted p-values (adjusted as the dataset is very large) are also listed to determine if the normalized enrichment scores are significant.The Reactome Selenoamino Acid Metabolism (a well-known pathway in Influenza) and the Reactome Influenza Infection pathway are both significantly overexpressed in mild COVID-19.[14]

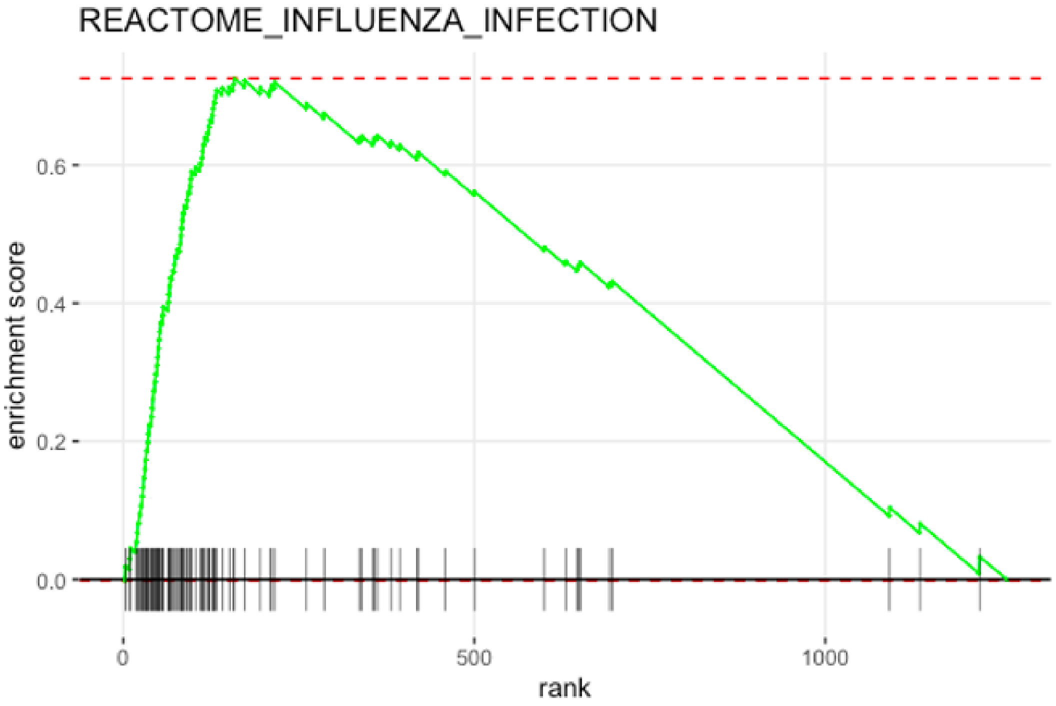


Figure 4

Enrichment Plot for Influenza Infection Pathway

The enrichment score reflects the degree to which the genes in a gene set are overrepresented at the top or bottom of the entire ranked list of genes. This demonstrates how many genes are in support of this pathway being higher expressed in mild COVID-19 than in Severe COVID-19.

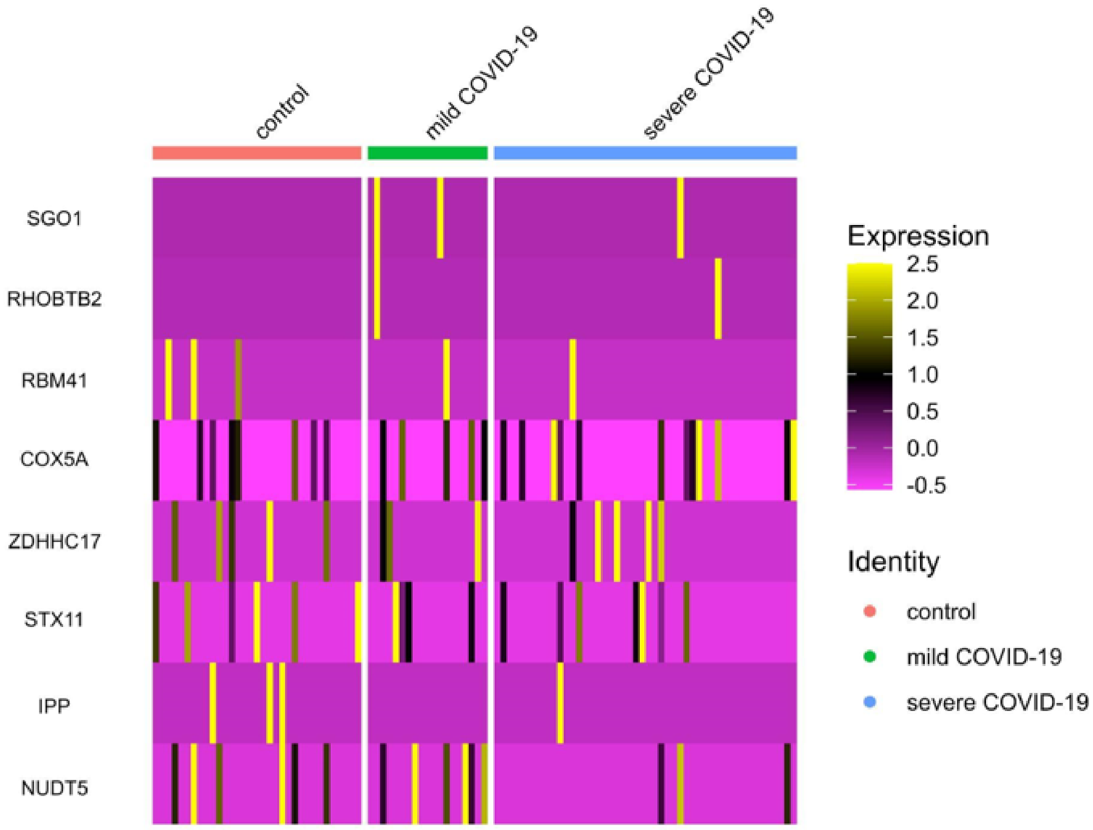


Figure 5

Heat Map of DL Model Input Genes Differentially Expressed by COVID-19 Severity

Out of the 10 genes used to train the deep learning model, 8 of them were found to be differentially expressed between COVID-19 status (control, mild, severe). 2 genes (*AC008626.1* and *NDUFAF4P1*) were not differentially expressed, and therefore are not included in Figure 5. Figure 5 is a heat map that depicts expression levels of genes between COVID-19 status, where yellow is most expressed, and pink is least expressed. This offers a visual interpretation of how these genes are used to differentiate COVID-19 status.