Prediction of Human Transcriptional Biomarkers for Severe Infection with SARS-CoV-2

Jeffrey Clancy, Curtis S. Hoffmann, Brett E. Pickett

Introduction

Human infections with severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) have resulted in hundreds of millions of confirmed cases and millions of deaths globally. The large diversity in the human response to infection, combined with large numbers of infections, contributed to strained hospital capacity and, the demand for biomarkers capable of predicting severity based on one of these factors has continually grown. Consequently, the aim of this study is to perform a meta-analysis of existing human transcriptomics data from collected blood samples to predict transcriptional prognostic markers. Such markers of disease severity could then contribute to making informed decisions concerning the care of infected patients who seek treatment at the hospital.

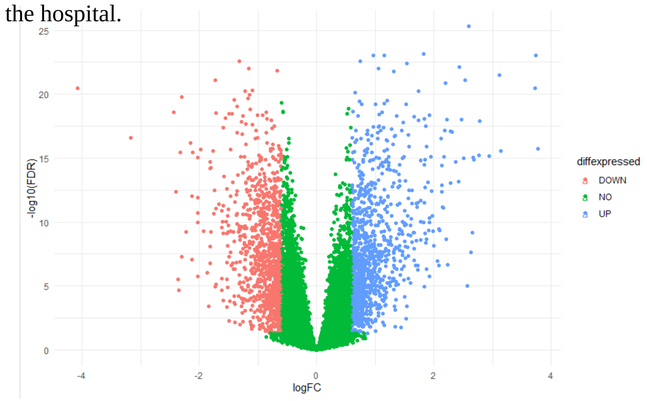

Figure 1. Volcano plot of all differentially expressed genes in severe vs. mild human infection with SARS-CoV-2. Genes that are up or down regulated from blood samples collected from patients having severe symptoms or mild symptoms during infection with SARS-CoV-2. Genes showing statistically significant up-regulation (blue), down-regulated (red), or no significant change (green). X-axis shows the $\log_2$ fold-change values while the y-axis displays false-discovery rate-adjusted p-values to account for multiple hypothesis testing.

Methods

356 fastq RNA-sequencing files from patients having either “mild” or “asymptomatic” were obtained from the Sequence Read Archive (SRA) at NCBI using the sratools software. A python-based snakemake workflow was used to perform RNA processing and calculate differential gene expression. A random forest classification method (using the R randomforest package) was then trained using a randomly-selected 70% of the dataset and then applied to the remaining 30% to identify the transcripts that were most useful in predicting severity. This algorithm calculated the Gini impurity values for each feature, which were then sorted to rank the importance of genes.

Results

Overall, we identified 8435 significant differentially expressed genes after applying a multiple hypothesis correction with log2 fold-change values ranging from -4.2 to 3.78 (Figure 1). We constructed a table with the read quantification data for all transcripts from each gene represented as columns and the samples represented as rows. We then used this table to generate a receiver-operator characteristic (ROC) curve (Figure 2) we generated a ROC curve for the six expressed genes having the highest Gini Impurity values and calculated its area under the curve (AUC) to be 94.3%. We repeated this process for only the top two expressed genes (GIMAP7 and S1PR2) and only the top expressed gene (GIMAP7). This analysis quantified the AUC to be 89.8% for the two combined genes and 84.4% for only the top gene

Table 1: List of expressed genes that best predicted SARS-CoV-2 disease severity in human blood cells.

Gene Symbol	Gini Impurity	Severity Level	Mean Reads	Standard Deviation	Median Reads
GIMAP7	0.44	High	1150.79	840.69	932
		Low	3208.53	1999.66	2796
S1PR2	0.41	High	44.81	67.55	36
		Low	137.56	163.18	70
PRR5L	0.4	High	148.06	111.09	112
		Low	387.78	201.51	351
RABGAP1L	0.38	High	834.23	247.90	816.5
		Low	1800.47	1233.70	1368
TRERF1	0.35	High	246.87	127.21	222.5
		Low	477.12	201.68	451
GPR174	0.33	High	109.62	93.16	81
		Low	329.52	238.31	286
CRTAM	0.32	High	52.33	38.58	43.5
		Low	133.43	78.61	119
GPR68	0.31	High	30.68	29.67	22.5
		Low	98.79	71.99	84
CD2	0.3	High	567.39	564.07	449.5
		Low	1732.99	1329.63	1386
GPR18	0.3	High	44.76	37.26	36
		Low	123.45	74.27	115

Conclusion

We ultimately produced a list of biomarkers capable of predicting disease severity that is up to 96% accurate. These results present a potential prognostic for the severity of disease in patients infected with SARS-CoV-2. Incorporating these biomarkers as additional data points can enable assessment of severe disease when resources may be limited, and priority must be given to those with the greatest risk of severe infection. These biomarkers include both up and down regulation association, rather than just statistically significant correlation, allowing us to detect previously undetected biomarkers such as GIMAP7 and S1PR2.

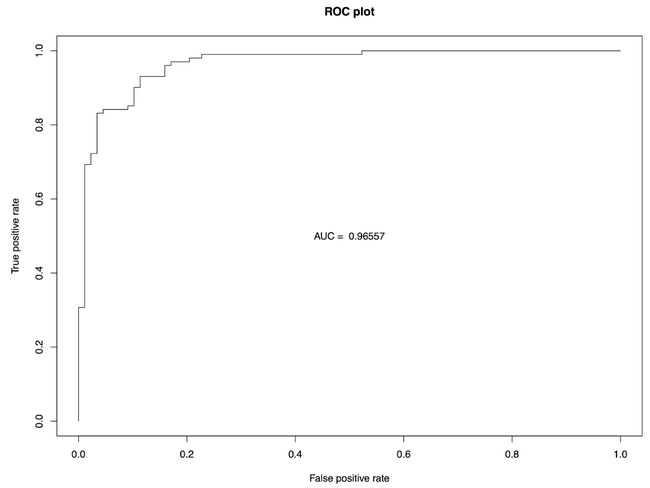

Figure 2. Receiver-operator characteristic (ROC) curve constructed from all expressed genes in severe vs. mild human infection with SARS-CoV-2. Constructing a ROC curve from all RNA-sequencing read quantification values showed that this technique achieved an area-under-the-curve (AUC) value of greater than 96%.

Acknowledgments

We are grateful to the BYU Office of Research Computing for providing the computational resources needed to complete this study. We also thank the original clinicals, patients, and others who provided the RNA-sequencing data.

BYU

References:
Arunachalam PS, Wimmers F, Mok CKP, Perera RAPM et al. Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans. *Science* 2020 Sep 4;369(6508):1210-1220. PMID: 32788292
Overmyer KA, Shishkova E, Miller IJ, Balnis J et al. Large-Scale Multi-omic Analysis of COVID-19 Severity. Cell Syst 2021 Jan 20;12(1):23-40.e7. PMID: 33096026