American Journal of Current Microbiology
http://ivyunion.org/index.php/ajcmicrob/

Research Article

# Identification of COVID-19 Drug Candidates through Computational Drug Repurposing

Henrik Zhang

Saratoga High School, 20300 Herriman Ave,
Saratoga, CA 95070, USA

**Abstract:** The recent outbreak of the COVID-19 pandemic has spread from China to the rest of the world in part because the virus has a high infection rate. The lack of a vaccine and ineffective antiviral treatments also contribute to the growing number of those who die from COVID-19. Though vaccines are on the verge of being delivered, drug treatments are still much needed because of the challenge of distributing the vaccine and the month long immunity development period where people remain vulnerable makes drug treatments essential. Drug repurposing reuses previously approved drugs to be used for another disease. Case in point are Remdesiver and Chloroquine, drugs previously used to treat ebola and malaria, which are now being repurposed to treat COVID-19. In this study, we used a computational approach to identify potential drugs to be repurposed in COVID-19 treatment based on patient genomic data and a public gene drug database. We retrieved datasets containing COVID-19 patient RNA-seq and proteomics data and used it to determine differentially expressed genes in each dataset and their interrupted functional pathways. Through the Drug Gene Interaction database, we found the specific drugs that target each differentially expressed gene and used the evidence scores from the database to create our final network of drug gene interactions. Our network contains a total of 22 unique drugs, 10 of which have been tested in a clinical setting, while the remaining drugs need to be validated in clinical studies in the future.

**Keywords**: Drug repurposing; COVID-19; Network; Gene expression; Proteomics; Computational
**Abbreviations**: COVID-19: coronavirus disease 2019, CQ: chloroquine, HCQ: hydroxychloroquine, BALF: Bronchoalveolar Lavage Fluid, PBMC: Peripheral blood mononuclear cells, LC–MS: Liquid Chromatography with tandem mass spectrometry, FDR: false discovery rate
**Received**: December 20, 2020; **Accepted**: January 27, 2021; **Published**: February 3, 2021
**Competing Interests**: The authors have declared that no competing interests exist.

***Correspondence to**: Saratoga High School, 20300 Herriman Ave, Saratoga, CA 95070, USA
**Email**: henrik.zhang04@gmail.com

# Introduction

There have been many pandemics during the 21st century, including SARS (2002), MERS (2012), Ebola (2013), the annual Influenza, and the most recent COVID-19 [1]. The COVID-19 pandemic, which originated in China, has spread across the whole world. COVID-19 is caused by the novel virus SARS-nCoV2, which researchers are searching for vaccines and treatments for. As of December 20th, 2020, both Pfizer and Moderna have had their vaccines approved by the FDA and both have shown high efficacy [2]. Despite the high potential of vaccines, there are still many problems yet to be addressed. Firstly, during the period of time when vaccines are being distributed and people are developing immunity, COVID-19 infections will not slow down. Medical workers and seniors are expected to receive the vaccine in December and essential workers will receive it early next year. By April, the vaccine should be available to the public. Second, the reported efficacy of the vaccine is around 90%, meaning that a significant percent of people that receive the vaccine will not derive any immunity. In the context of the whole world, a small percentage still indicates a vast amount of lives that are at stake.

Many viral infections have a series of responses by the host body, including apoptosis, stress response, and innate immunity. The body's response is dependent on a series of factors that include genetics, epigenetics, and lifestyle. This results in individuals who have good body capacity for protective responses to activate the antiviral defense mechanisms. These immune responses, however, severely decrease in the elderly and those with underlying disease, increasing the viral load and leading to severe respiratory problems and in some cases death. The nature of COVID-19's short replication times and high viral yields, makes the replication of positive-sense viral RNA prone to high error rates. This results in a virus that is elastic and difficult to design treatment for. Despite the lack of specific medications for the virus, there have been drugs used in clinics like the Regeneron antibody cocktail, which has shown that they work on individuals with severe symptoms. The FDA reported that there is very little difference between placebo groups and those that received the antibody cocktail in mild and moderate cases. However, in severe cases, the FDA noted a significant difference where the antibody cocktail is more effective, thus emergently approving the use of the Regeneron antibody cocktail in COVID-19 treatment [3]. Besides the antibody cocktail, there have also been drugs used in clinics such as Remdesivir, Chloroquine, Tocilizuman, Hydroxychloroquine, Umifenovir, Lopinavir, Oseltamivir, and Favipiravir to fight off COVID-19 infection. There have also been adjunctive agents such as zinc, vitamin D, Azithromycin, Ascorbic acid, Nitric oxide, Corticosteroids, IL-6 antagonists to treat patients as well [4].

Researchers have also started to focus their attention on drug repurposing because of its time and cost efficiency. On average, developing a brand new drug takes between 10 and 17 years. The cost to research and develop each successful drug is around 2.6 billion. Drug repurposing circumvents the need for research and development by reusing already established safe candidate compounds. As of now there have been many drug repurposing candidates for the COVID-19 treatment. Remdesivir has been shown in trials to display potent in vitro activity against COVID-19 infection [5]. Some other examples are chloroquine (CQ) and hydroxychloroquine (HCQ), which have shown beneficial impact in COVID-19 treatment. CQ was originally approved for the treatment against Malaria, and HCQ is a modified version of CQ with an

additional Hydrogen atom [6]. When it comes to identifying drugs to be repurposed there are two main methods: wet lab screening and dry lab in silico studies. Wet lab experiments like that of Riva et al developed a high-throughput assay to enable large-scale screening of known drugs, which one hundred molecules inhibit viral replications of SARS-nCoV-2, including 21 that have dose-response relationships. From this, thirteen molecules were found to have effective concentrations commensurate with probable achievable therapeutic dosages in patients [7]. The dry lab experiment of Stolfi et al demonstrates how network based computational methods can be used to identify possible candidates for drug repurposing. The team used the BioGRID database to retrieve 424,076 interactions and 500 COVID-19 genes. They then used DIAMOND to identify 1,500 COVID-19 proximal targets and then separate these targets into 5 specific tissue groups. The genes were then read into DrugBank to check for the drug-gene interactions and side effects in order to finalize the list of 18 proposed COVID-19 drug repurposing candidates [8].

In this study, we used computational methods to create a list of proposed COVID-19 drug repurposing candidates through patient and cell line genomic data. We used search engines like Google Scholar and PubMed to identify studies with publicly available data on COVID-19 patient's RNA-seq and proteomics data. In each dataset, we determined the differentially expressed genes that are potential COVID-19 drug targets. We then used said differentially expressed genes in the Drug Gene Interaction database to find specific drugs that target each differentially expressed gene. The Drug Gene Interaction database also provides a score for each drug gene interaction, and so when finalizing our list, we used a filter on the score to increase the accuracy of the drugs that we proposed for drug repurposing. We proposed a total of 22 drug repurposing candidates, 10 of which have already been used in clinical studies for the treatment of COVID-19. The remaining are predictions that could help treat patients diagnosed with COVID-19.

# Methods

### 1. Workflow

We used keywords such as "COVID-19" or "SARS-nCoV-2" in addition to "protein expression, or "gene expression" in Google Scholar to compile a list of 6 potential datasets. After checking the data availability, only 4 contained publicly available data to create our final list of datasets. Gene expression data was taken from four different studies regarding SARS-nCoV-2 infection. There were two datasets taken from cell lines, one from patient RNA expression data, and one from patient proteomics expression data. The cell lines consisted of the A549, NHBE, and Caco-2 cell lines. The patient expression data was taken from the BALF and PBMC of patients. The patient proteomics data was taken from the patients' blood. Each study had provided gene or protein expression data, which was then filtered through a fold change cutoff and a false discovery rate adjusted for each dataset to identify the differentially expressed genes in each dataset. We then created a network and function analysis for each dataset and compared the overlaps of the differentially expressed genes in each dataset. In the Drug Gene Interaction Database we found drugs that correlate with each differentially expressed gene. The Drug Gene Interaction Database provides a score for

each correlation between drug and gene based on the among of PMID's referencing such connections. We used the drugs that had scores greater than or equal to 5 to compile a final list of drugs that We recommend for the repurposing to fight off SARS-nCoV-2 infection.

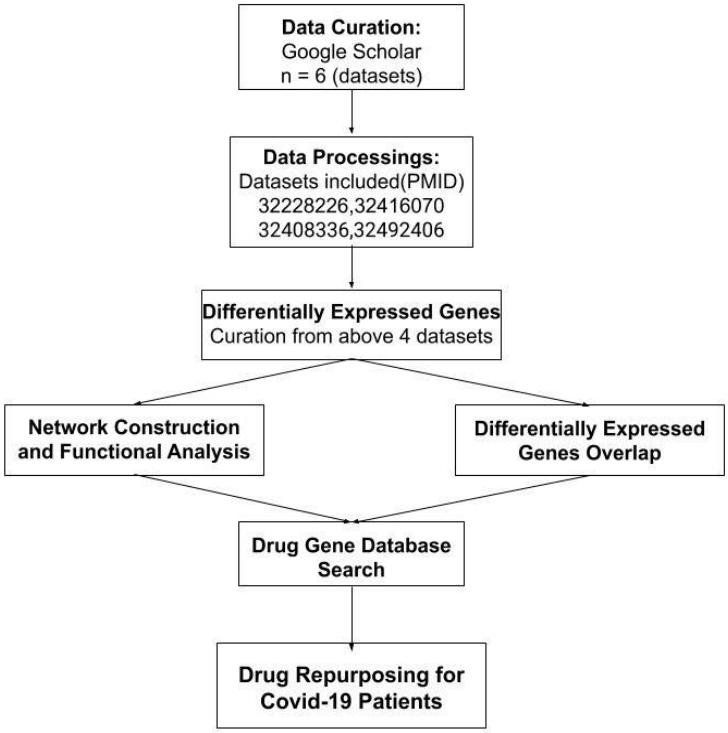


**Figure 1** Flow chart to detect host genes and change upon viral infection and drug repurposing for COVID-19 patients.

## 2. Data source

The first dataset that uses BALF and PBMC samples were obtained from 3 SARS-nCoV-2 patients and 3 healthy donors. The study used RNA-seq to detect transcriptome changes in both PBMC and BALF samples in both COVID-19 patients and healthy donors. The study used the DESeq2 package to determine their differentially expressed genes. We used the Supplementary File 1 provided sheets "BALF DEGs" and "PBMC DEGs". The second dataset uses the A549 and NHBE cell lines when infected with SARS-nCoV-2. They then used RNA-seq to detect the transcriptome changes in cells. The study used DESeq2 to determine the differentially expressed genes in each sample. We used the "media-2" and "media-6" supplementary file. The third dataset used the Caco-2 cell line and observed the viral infection over a 24 hour period and quantified the translatome and proteome changes detected by LC–MS/MS. We used the Supplementary File 01. The fourth dataset uses proteomics data from 46 COVID-19 patients and 53 control individuals. They split COVID-19 patients into non-severe (25 patients) and severe (28 patients). The study compared the comparison between the different severities of COVID-19 with controls. We took the "mmc6"

supplementary file and sheet “Prot severe vs healthy” and “Prot non-severe vs healthy”.

## 3. Identifying differentially expressed genes

To identify the differentially expressed genes in each dataset, we used a variety of filters on p-values, FDR and fold change. The *p*-value is the probability of obtaining similar test results as the results actually observed [9]. The FDR is the adjusted *p*-value, which represents the portion of false positives among all features equally or more extreme than the observed one [10]. Fold change is a measure describing the amount a quantity changes between two groups [11]. The fold change is calculated by the mean of the two groups as a multiple of one another. Differentially expressed genes are determined through a combination of FDR and fold change to reflect both the magnitude of the change and the statistical difference between two groups [12] . In the BALF and PBMC datasets, we used a fold change cutoff greater than 2 and false discovery rate (FDR) less than 0.01 to determine our differentially expressed genes. In the A549 and NHBE cell line datasets, we used fold change cutoff of 1.5 and a FDR cutoff at 0.05. In the Caco-2 cell line proteomics dataset, we used a fold change cutoff of 1.5 and a FDR cutoff at 0.05. In the patient proteomics dataset, we used a fold change cutoff of 1.5 and a FDR cutoff at 0.05. The differentially expressed genes from each dataset were then put into Venny, an online tool, to determine and present the overlap between different datasets [13].

## 4. Functional analysis

Functional analysis is important because it determines the targets of intervention that can be changed. Differentially expressed genes recognize individual genes as responsible for changes, whereas functional analysis detects a group of genes or pathway as responsible for changes. There are many tools that can be used to perform functional analysis. For example clusterProfiler, Ingenuity Pathway Analysis (IPA), AmiGO, and DAVID. However in this study, we used clusterProfiler, which is a free R package that implements methods to analyze and visualize functional profiles of gene clusters [14]. IPA is a commercial software that is very expensive [15]. In Rstudio, we used the enrichGO function inside of the clusterProfiler package, allowing us to read the differentially expressed genes for each dataset into clusterProfiler, which returned pathway analyses in the form of tables and figures. FDR less than 0.05 is used to consider a pathway as significant. The figures and bar plot present the top 20 enriched pathways for each dataset.

## 5. Drug gene association

The Drug Gene Interaction Database (DGIdb) is a database developed at Washington University that efficiently determines drug-gene interactions to help with clinical sequencing and personalized medicine [16]. The database uses information on drug-gene interactions from sources such as DrugBank, PharmGKB, Chembl, and Therapeutic Target Database. Genes were also categorized as potentially druggable based on their membership in selected pathways, molecular functions, and gene families from the Gene Ontology, the Human Protein Atlas, and numerous other studies. DGIdb contains over 40,000 genes and 10,000 drugs involved in over 100,000 drug-gene interactions or belonging to one of 42 potential druggable gene categories. DGIdb presents drug gene interactions and gene druggability information based off of papers, other databases and web resources. The database is a web based application that for example, reads in a list of genes, and will return with a list of drug interactions for each gene. Each suggested drug has a score

calculated based on the amount of publications that support such an interaction. In this study, we inputted the differentially expressed genes expressed in each dataset directly into DGIdb. Originally we were going to download the source code for DGIdb and use R to filter through each dataset and create a list of differentially expressed genes efficiently. However, the source code for the score calculation was unclear and so to ensure the accuracy of the study, we manually inputted each dataset of differentially expressed genes into the DGIdb website and copied the results into an excel spreadsheet. All the spreadsheets listed the genes, drugs, interaction type, sources, pmids, and score. For the BALF dataset, the total drug-gene interactions was 1138. The PBMC dataset had a total drug-gene interaction of 2284. The A549 cell line had a total drug-gene interaction of 262. The NHBE cell line had a total drug-gene interaction of 739. The Caco-2 cell line had a total drug-gene interaction of 246. The Proteomics non-severe vs healthy had a total drug-gene interaction of 25 and the proteomics severe vs healthy had a total drug-gene interaction of 111. The total drug-gene interaction after being filtered with a score greater than or equal to 5 across all datasets was 66. By filtering the drug-gene interactions that had a score greater than or equal to 5, we were able to create a final list of predicted drugs that can be used for repurposing to fight COVID-19.

## 6. Drug repurposing for COVID-19

Drug repurposing is a strategy for identifying new uses for approved or investigation drugs that are outside the scope of the original medical indication [17]. It is slowly becoming a more and more popular option for treating common and rare diseases as it uses de-risked compounds, with potential lower costs and shorter development timelines [18,19]. In terms of the COVID-19 pandemic, speed is lifesaving. Developing new drugs to treat COVID-19 takes too long, which is why drug repurposing is a viable option. The drug repurposing process begins with the identification of a candidate drug for a given indication, followed by mechanistic assessment of the drug effect in preclinical models and lastly the evaluation of efficacy in clinical trials [20,21]. When identifying candidate drugs, a computational approach is used. The computational approach uses data of gene expression. An example would be gene signature matching, which allows for drug-disease comparisons. Researchers use the gene expression profile of biological material such as cells or tissues before and after drug treatment. The differential gene expression signature is then compared with disease associated expression profile data from a healthy sample. Since the gene expression signature is upregulated and downregulated in the disease, the drug that regulates such genes could be a potential candidate for drug repurposing.

For our project, we took the same approach. We used the list of differentially expressed gene, drug interactions with evidence scores greater than or equal to 5. In order to create more accuracy, we required overlap between two datasets: BALF and PBMC, A549 and NHBE cell line, proteomics non-severe vs healthy, and severe vs healthy. A stricter overlap was placed across multiple datasets: BALF, PBMC, A549, and NHBE; BALF, PBMC, and Caco-2 cell line; A549, NHBE, and Caco-2; and Caco-2 and proteomics severe vs healthy. We then created a gene-drug network that matches each of the potential drugs for repurposing to their specific gene targets. The drugs displayed in Figure 5 have different prioritizations. The higher the evidence score given in DGI, which is shown in the figure by the size of the bubbles, the more potential there is for a drug to be a viable option in treating COVID-19. In Figure 5, the drugs that we recommend for COVID-19 treatment consist of tested and untested drugs that fight COVID-19. In order to verify our predictions, we searched individually for literature that supports usage of the drug in COVID-19

treatment. Many of the drugs that we have recommended in the figure match the current researcher's drug predictions; those that have no literature are the drugs that we predict to be viable options for COVID-19 treatment and should be tested in clinical trials.

## 7. Statistical analysis and data visualization

In this study, we used many applications to perform statistical analysis and data visualization. Namely, Venny, ClusterProfiler, RStudio, Microsoft Excel, and Powerpoint [13,14]. We used Microsoft Excel to filter the raw data of each dataset, into a list of differentially expressed genes. We then used Rstudio to prepare the differentially expressed genes into the required format needed to use ClusterProfiler. ClusterProfiler was used to perform pathway analysis on each dataset, and represented the data in bar plots and pathways. We also used Venny to display the overlap between the differentially expressed genes and their drug targets. Finally we used Microsoft Powerpoint to create a figure that displayed all differentially expressed genes with a score greater than or equal to 5 and their drug targets.

# Results

## 1. Detection of differentially expressed host genes in patients with COVID-19

The number of significant host genes generated in Table 1 comes from filtering through different criteria. In the patient BALF and PBMC datasets, which had 5 (3 healthy control and 2 COVID-19 patients) and 6 (3 healthy control and 3 COVID-19 patients) samples respectively, we used the criteria of the fold change cutoff greater than 2 and false discovery rate (FDR) less than 0.01 to identify the total amount of differentially expressed host genes with 1004 and 1024 (679, 423 up and 325, 601 down-regulated respectively) [22]. However, in the cell lines and both proteomics datasets, we used a fold change cutoff of 1.5 and a FDR cutoff at 0.05. The two lung cell lines, A549 and NHBE, are used to be affected by the COVID-19 virus, and a change of 102 (88 up and 14 down-regulated genes) and 204 (145 up and 59 down-regulated genes) differentially expressed genes were observed [23]. In another study, the Caco-2 cell line was used in a time series infection (2, 6, 10, and 24 hours) [24]. The 24-hour data was used because the proteome underwent extensive modulation in comparison to minor host proteome changes, resulting in 94 (29 up and 65 down-regulated genes) differentially expressed genes. The final proteomics dataset was created by comparing two different sets of patients with severe and non-severe infections of SARS-nCoV-2 [25]. The severe vs healthy had 41 DEG (28 up and 13 down-regulated), whereas the non-severe vs healthy had 11 DEG (4 up and 7 down-regulated). All the gene names of each dataset are included in Supplementary File 1.

The genes upregulated in COVID-19 infections include those that are involved in cytokine storms, which impacts the patients' severity [26,27]. Cytokines such as CCL2, CCL4, CCL8, CXCL1, CXCL2, CXCL6, CXCL8, CXCL10, and CXCL17 are up-regulated within COVID-19 patients. Another example is interferon-inducible genes such as IL-6, IL-10, IL8. The IL-6 gene has been shown in a meta-analysis comprising nine studies that showed patients with COVID-19 had three times higher IL-6 levels compared to those without, indicating high levels of IL-6 are associated with mortality risk [28]. These differentially expressed genes only contribute to increasing the mortality rate of patients with COVID-19. The increase in cytokines presents the risk of a cytokine storm, where rapid proliferation and hyperactivate of T cells,

macrophage, overproduction of inflammatory cytokines, and chemical mediators released by immune or non-immune cells [26]. This potentially leads to multiple organ failure due to the immune system's over detection of organs.

**Table 1** Significant Host Genes in Covid-19 Infection. Number of up and down-regulated significant host genes for each dataset

| Dataset | PMID | Up-regulated | Down-regulated | Total |
|---|---|---|---|---|
| Patient BALF | 32228226 | 679 | 325 | 1004 |
| Patient PBMC | 32228226 | 423 | 601 | 1024 |
| Cell line A549 | 32416070 | 88 | 14 | 102 |
| Cell line NHBE | 32416070 | 145 | 59 | 204 |
| Proteomics Cell line Caco-2(24 hour) | 32408336 | 29 | 65 | 94 |
| Proteomics severe vs healthy | 32492406 | 28 | 13 | 41 |
| Proteomics non-severe vs healthy | 32492406 | 4 | 7 | 11 |

The significant host genes vary from each dataset for a variety of reasons. The patient data had the most significant host genes because the immune system adapts in order to fight off infections. The T cells and B cells activate various genes to help fight off the infection, which would explain the high amount of significant host genes. Since cell lines are a simple system and have no immune system, gene expression will not fluctuate as much, explaining why there are fewer significant host genes. The select number of significant host genes in the proteomics dataset is explained simply because fewer proteins can be detected in comparison to RNA.

## 2. Functional annotation and pathway analysis of differentially expressed host genes in patients with COVID-19

Pathway analysis was carried out for each of the lists of differentially expressed genes in Table 1. The top 20 enriched pathways are shown in Figure 2. The BALF sample of COVID-19 patients indicates that the top enriched pathways include SRP-dependent cotranslational protein targeting to membrane (P = 2.7E-35) and viral transcription (P = 2.2E-23). The PBMC sample obtained by COVID-19 patients show that the top enriched pathways are complement activation, classical pathway (P = 5.7E-61), humoral immune response mediated by circulating immunoglobulin (P = 2.3E-57), regulation of acute inflammatory response (P = 1.1E-47). The A549 cell line demonstrates that some of the top enriched pathways are response to virus (P = 2.3E-31), type I interferon signaling pathway (P = 8.8E-28), and regulation of viral genome replication (P = 1.1E-18). The NHBE cell line indicates that the top enriched pathways are a response to virus (P = 3.3E-13),

humoral immune response (P = 9.2E-12), response to interferon-gamma (P = 3.2E-11), and regulation of inflammatory response (P = 9.3E-10). The Caco-2 cell line indicates that the top enriched pathways are protein-containing complex remodeling (P = 3.1E-9), protein-lipid complex remodeling (P = 3.1E-9), and plasma lipoprotein particle remodeling (P = 3.1E-9). The patient proteomics severe vs healthy data indicates that the top enriched pathways are platelet degranulation (P = 1.7E-17), acute inflammatory response (P = 2.8E-16), acute-phase response (P = 4.4E-14), protein activation cascade (P = 6.6E-10), negative regulation of hemostasis (P = 7.6E-10), and blood coagulation (P = 9.0E-10).

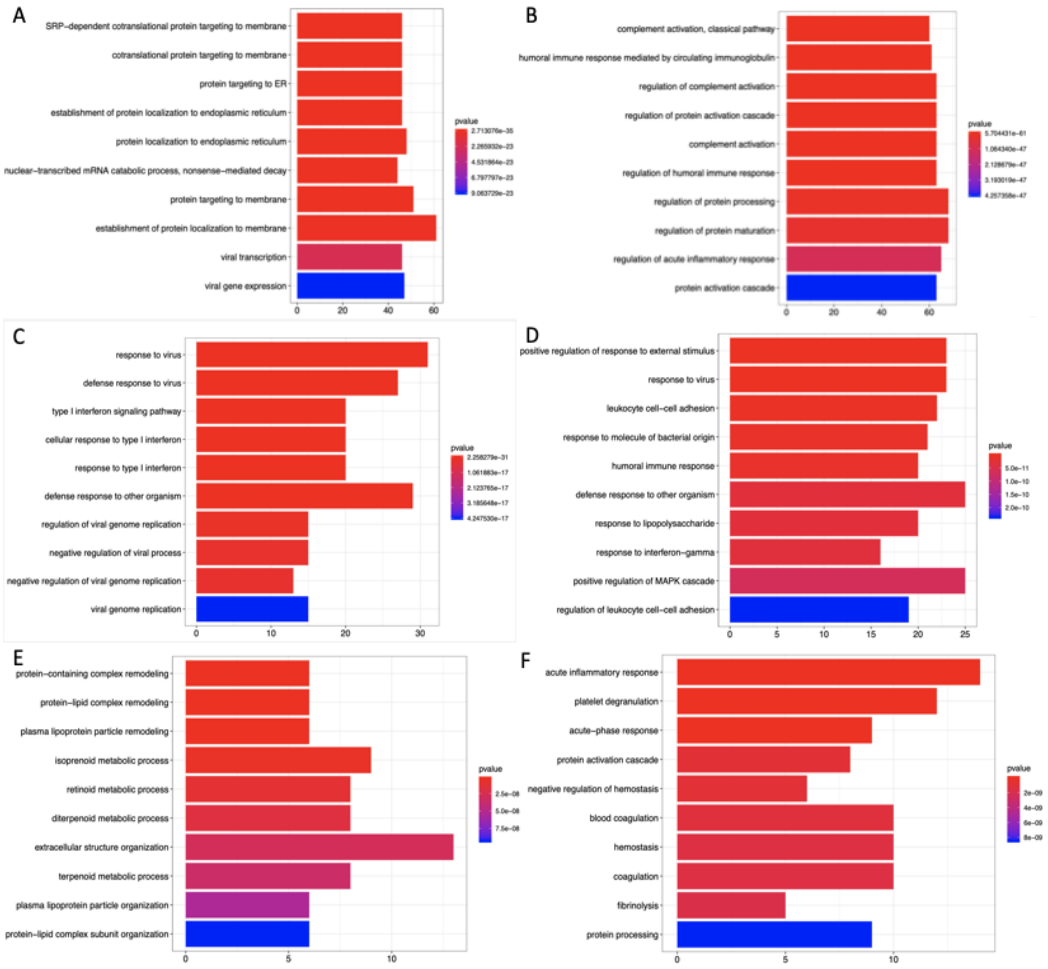


**Figure 2** Significantly enriched biological functions (by ClusterProfiler) associated with differentially expressed genes in COVID-19 patients and cell lines infected with SARS-nCoV-2. A shows the pathway analysis of the BALF patient dataset. B shows the pathway analysis of the PBMC patient dataset. C shows the pathway analysis of the A549 lung cell line. D shows the pathway analysis of the NHBE lung cell line. E shows the pathway analysis of the Caco-2 colon cell line. F shows the pathway analysis of the proteomics severe patient vs healthy control dataset.

Compared with cell lines, patient data offer different data points because patients have immune systems and a circulating system. This would explain why the patient pathway analysis shows T cell and B

cell mediated responses, blood coagulation, and platelet degranulation. However, both vitro and vivo systems can be infected by the virus and have immediate responses to the infection. Pathways in the proteomics severe vs healthy data, such as platelet degranulation and acute inflammatory response, involve the genes SERPINA3, ORM1, LGALS3BP, SERPING1, ITIH3, VWF, FGA, TAGLN2, FLNA, THBS1, HRG, APOH, SAA1, SAA2, HP, LBP, C9, SERPINA3, ORM1, APCS, SERPING1, C2, SAA4, CPN2, and CFHR5. The serum amyloid A genes serve as serum biomarkers for inflammatory response for patients infectected with the COVID-19 virus [29]. Additionally, the SERPINA3 gene seems to play a major factor in slowing down clot formations and inhibiting proteases released by inflammatory cells [30]. All the detailed genes with adjusted p values less than 0.01 involved in the pathway analysis are included in Supplementary File 2.

## 3. Overlap of differentially expressed genes across various studies

The comparison between various datasets is necessary because most datasets come from the same patients or have similar locations. The BALF and PBMC patient data come from the same patients, but are gathered from different areas of the body. The BALF data specimen comes from the bronchus, while the PBMC comes from the blood. The A549 and NHBE cell lines come from different people. The A549 is a carcinoma cell line in the lung, and the NBHE is a normal cell line in the lung. Some of the data we have is RNA-seq and some are from proteomics. By comparing the two types of data, we can determine whether the results will be consistent throughout the body.

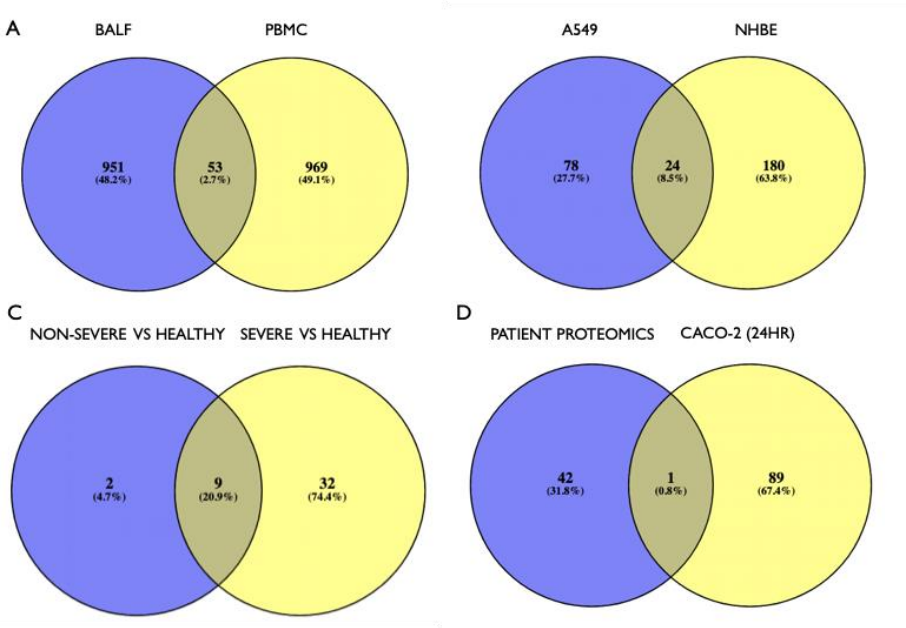


**Figure 3** Overlap of the differentially expressed genes across various datasets. A shows the overlaps between the BALF and PBMC datasets, B shows the overlaps between the A549 cell line and the NHBE cell line, C shows the overlaps between the non-severe vs healthy and severe vs healthy proteomics data, and D shows the overlaps between the patient proteomics data as a whole and Caco-2 (24hr) datasets. The proteomics data is combined because there is a

20% overlap between the two proteomics dataset, indicating a high correlation in differentially expressed genes.

The comparison in Figure 3 shows that there is a very low correlation between the BALF and PBMC datasets that are from the same patients. The venn diagram indicates that only 53 (2.7%) out of the 1973 total genes are present between the two datasets (1004 total genes for BALF and has 1022 total genes for PBMC). Additionally, there is also a low correlation between the A549 and NHBE cell lines. The venn diagram indicates that only 8.5% out of the 282 total genes are present between the two datasets (102 total genes for A549 cell line and 204 total genes for the NHBE cell line). The two proteomics datasets have a high correlation, which on the nonsevere vs healthy dataset, 9 out of the 11 overlapped (81.8%), whereas the proteomics severe vs healthy dataset had a total of 41 proteins. The two proteomics datasets displayed a high correlation between the differentially expressed proteins in the dataset. This indicates that most of the proteins are similar, indicating a stark similarity in the two proteomics datasets that allows us to combine them with each other for further comparisons. The combined proteomics datasets and the Caco-2 (24hr) cell line have an extremely low correlation (0.8%). Out of the 132 genes present, only 1 overlaps between the two datasets. The proteomics data had 43 total proteins, while the Caco-2 (24hr) had 90 total genes.

Since both the BALF and PBMC datasets originate from the same patients, but are extracted from different parts of the body, the low correlation makes sense. The BALF tissue comes from the bronchus, which would only have genes local to the respiratory system. COVID-19 infections trigger local immune responses that trigger macrophages and monocytes to respond to the infection, release cytokines and prime adaptive T and B cell immune responses. This process resolves infections most of the time, however in the case of failure, a dysfunctional immune response occurs, which can result in severe lung and even systemic pathology [31]. The PBMC sample comes from the bloods and since the circulatory system travels throughout the body, the genes present in the PBMC sample are global. The PBMC sample reflecting gene expression changes throughout the body, while the BALF tissue is strictly limited to the lung, explaining the low correlation between differentially expressed genes. Conversely, the A549 and NHBE cell lines are both from the lung. However, the A549 cell line is a carcinoma cell line explanted from a 58-year old caucasian male, whereas the NHBE cell line is a naturally occurring cell line in the lung and is explanted from a 79-year old caucasian female. Since the cell lines result from different humans, the genetic background differs extremely. This could be a potential explanation for the biological differences in this experiment. Additionally, the A549 cell line derives from a male, while the NHBE cell line derives from a female. This difference in gender can also contribute to biological differences in the experiment. It is commonly known that COVID-19 infections present a variety of symptoms throughout the severity of infection [32,33]. The proteomics data suggests that the non-severe infections are similar to that of healthy people because only 11 differentially expressed proteins change from healthy. However, the severe vs healthy proteomics data indicates that there is a change on the molecular level from the progression of the COVID-19 infection from non-severe to severe. Additionally the severe COVID-19 infection has 41 differentially expressed proteins that change from a healthy patient. When the patient proteomics data is compared with the Caco-2(24hr) cell line a star difference is shown. This indicates that there is a distinct difference between patient data and cell line data.

## 4. Drug and gene association network construction

The Gene Drug Interaction Database is a database that detects the relationship between genes and drugs. In our case, we filtered the differentially expressed genes and proteins from our datasets into the database. The database then compiles and returns each gene with a list of drugs that target that specific gene and a score indicating the amount of PubMed citations. By constantly increasing the score limit, the potential drugs decrease. For example, as shown in Table 2 in the Patient BALF data, the number of drugs started with 779 when the filter was set at a score greater than or equal to 1, but when the filter increased to greater than or equal to 5, the total number of drugs decreased to 117. In the PBMC, there were a total of 1,602 drugs that had a score greater than or equal to 1, while when the score was greater than or equal to 5, the total number of drugs decreased to 175. Based on the 3 cell lines, A549, NHBE, and Caco-2, we can predict that each gene had at least two unique drugs targeting it. The A549 cell line had a total gene count of 102 and had an initial drug count of 228, the NHBE cell line had a total gene count of 204 and had an initial drug count of 581, and the Caco-2 cell line had a total gene count of 94 and an initial drug count of 229. Then when the score increased to greater than or equal to 5, the three cell lines decreased to 41, 60, 20 total drugs, respectively. This trend applies to the patient proteomics data as well, which is shown by the severe vs healthy dataset having only 4 drugs and the non-severe vs healthy dataset having only 2 drugs when the score was set as greater than or equal to 5.

**Table 2** Number of drugs targeting differentially expressed genes in each dataset at different evidence scores between 1 and 5.

| Dataset | Total Genes | #Drug (score 1) | #Drug (score 2) | #Drug (score 3) | #Drug (score 4) | #Drug (score 5) |
|---|---|---|---|---|---|---|
| Patient BALF | 1004 | 779 | 437 | 246 | 162 | 117 |
| Patient PBMC | 1024 | 1602 | 998 | 471 | 305 | 175 |
| Cell line A549 | 102 | 228 | 163 | 81 | 50 | 41 |
| Cell line NHBE | 204 | 581 | 384 | 150 | 86 | 60 |
| Proteomics Cell line Caco-2 24 hour | 94 | 229 | 164 | 50 | 32 | 20 |
| Proteomics severe vs healthy | 41 | 108 | 47 | 14 | 6 | 4 |
| Proteomics non-severe vs healthy | 11 | 25 | 19 | 4 | 2 | 2 |

The results shown in Figure 4 are similar to those found in Figure 3. Figure 3 indicated the amount of similar genes across multiple datasets, while Figure 4 shows the amount of similar drugs across multiple datasets. It is not surprising that the overlap was very minimal at both the gene and drug level. Figure 4A shows a 7% (19 similar drugs) drug overlap between the BALF and PBMC patient datasets(BALF has 117

total drugs, PBMC has 175 total drugs). Figure 4B shows an even lower overlap as it only had 1%(1 similar drug) drug overlap between the A549 and NHBE cell line datasets(A549 has 41 total drugs, NHBE has 60 total drugs). Figure 4C shows a total overlap of the non-severe vs healthy proteomics dataset in comparison to the severe vs healthy proteomics dataset. Figure 4D shows a low similarity between the Caco-2 cell line and proteomics severe vs healthy dataset indicating a 9.1% (2 similar drugs) drug overlap(Caco-2 has 20 total drugs, Proteomics severe vs healthy has 4 total drugs).

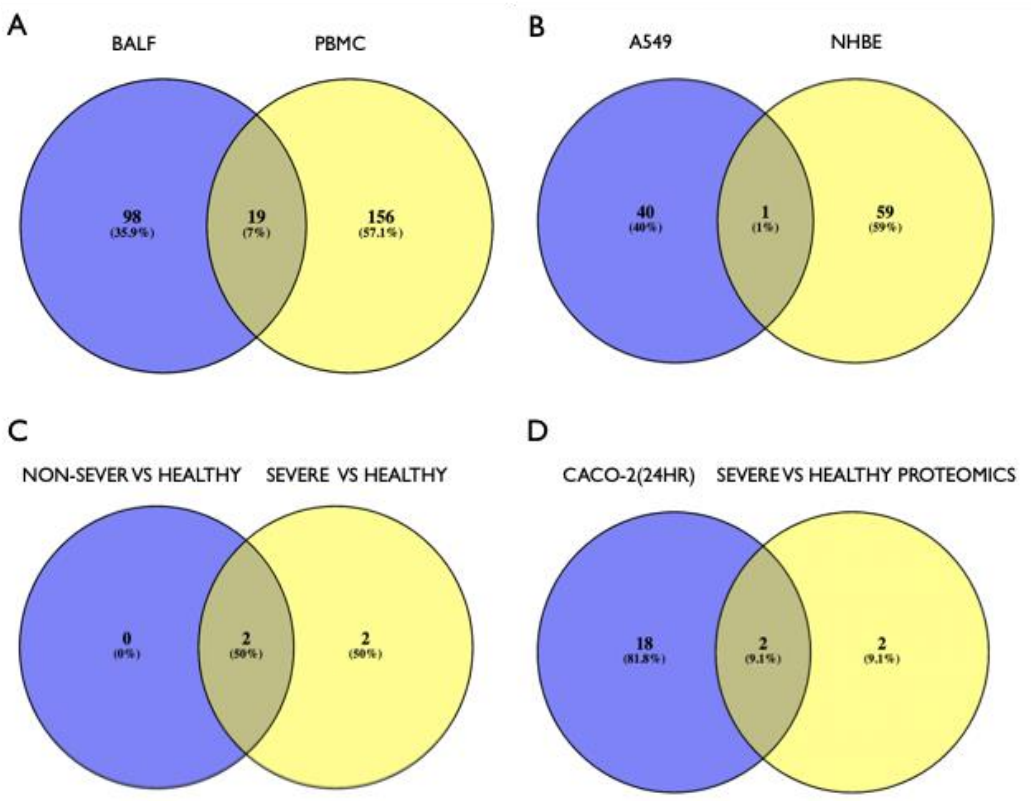


**Figure 4** Overlap between similar potential drugs when the score is set as greater than or equal to 5. Part A shows the drug overlap between the BALF and PBMC datasets, B shows the overlap between the A549 cell line and NHBE cell line, C shows the overlap between the proteomics non-severe vs healthy and severe vs healthy, and D shows the overlap between the Caco-2 cell line and proteomics severe vs healthy.

As the score for each increases, the total number of gene drug correlations decrease. we found that using a score of 5 was best suited to recommended enough drugs, while also being accurate. Even after setting the filter at an evidence score of 5, there is still a very low correlation between the drugs. This suggests that Covid-19 treatment might require personalized medicine or a cocktail of drugs. All the gene drug association with an evidence score greater than 5 is provided in Supplementary File 3.

## 5. Drug repurposing for COVID-19 treatment

A gene drug network shown in Figure 5 was constructed based on genes in Figure 4 which found the

overlapped drugs throughout different datasets. By finding the overlapped drugs from different datasets, it adds to the reliability of the drug predictions. After finding the overlapped drugs, we then used the previously found differentially expressed genes to find the differentially expressed genes that have a correlation with the overlapped drugs. The results from Figure 5 indicate that the drugs can target multiple genes.

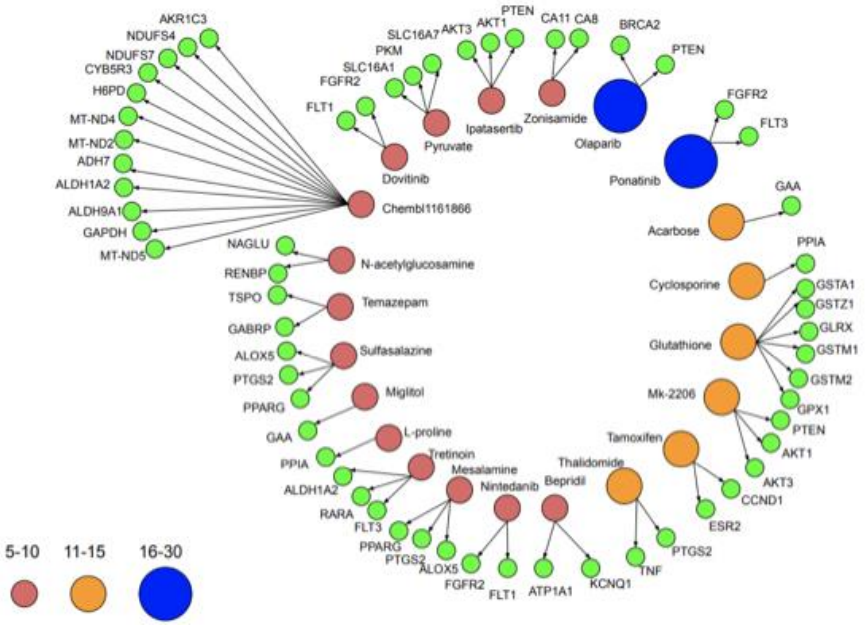


**Figure 5** Correlation between the suggested drugs to be repurposed for Covid-19 treatment and their gene targets. Drugs are split based on their scores in The Drug Gene Interaction Database. Multiple genes targets were found for each drug, each having its own score. We took the highest score for each drug to classify them into different groups. A blue circle indicates a high evidence score between 16 and 30. An orange circle indicates a medium evidence score between 11 and 15. A red circle indicates a low evidence score between 5 and 10. A green circle indicates a gene target of the drugs that come from the list of differentially expressed genes.

I split the finalized drugs into three different categories: high, medium and low evidence score. These scores were assigned based on the scores given in the DGIdb database such that a high evidence score was between a score of 16 and 30, a medium evidence score was between a score of 11 and 15, and a low evidence score was between 5 and 10. There are two high evidence score drugs, 6 medium evidence score drugs, and 14 low evidence score drugs. We found that the drugs that had a higher evidence score tended to be used more in the actual treatment of Covid-19 patients. One of the high evidence score drugs, Ponatinib, is already in use to treat patients diagnosed with Covid-19 [34]. Five out of the 6 medium evidence score drugs have had effectiveness in Covid-19 treatment. The drugs Cyclosporine, Glutathione, Mk-2206, Tamoxifen, and Thalidomide have been tested, meaning that the evidence scores are an accurate prediction for potential drugs that can be reused to treat Covid-19 patients [35–39]. Additionally, 4 out of the 14 low evidence score drugs have also seen usage in Covid-19 treatment. The drugs in use were Bepridil, Dovitinib, Chembl1161866, and Miglitol [40–43].

This leads us to suggest that drugs such as Olaparib and Acarbose can be used in Covid-19 treatment. Olaparib is one of the high evidence drugs, whereas Acarbose is a medium evidence drug. Olaparib targets both the BRCA-2 and PTEN genes. The BRCA-2 gene is known for being the tumor suppressor protein that regulates downstream genes in response to numerous cellular stress, and is frequently mutated in human cancer. SARS-nCoV-2 contains a spike protein that is divided into two subunits: S1 and S2. S1 infects human cells by binding to the human angiotensin-converting enzyme 2, while S2 mediates the membrane fusion process. After infecting the host receptor, a six-helix bundle fusion core is formed through the interaction of the HR-1 and HR02 domain. This brings viral and cellular membranes into close proximity for fusion and infection, which will ultimately unravel possible mechanisms of COVID-19 infection and its severity in humans who already are diagnosed with other diseases. The BRCA-2 gene, in addition to others, was found to have a strong interaction with the S2 subunit. BRCA-2 was also found to interact with the hepatic repeat-2 region through the C- terminal domain of the S2 subunit [44,45]. Emameh et al. conducted a study that applied single-cell transcriptomics data of human bronchial epithelial cells (2B4 cell line) infected with SARS-nCoV-2 [46]. Their reactome pathway analysis based on the differentially expressed genes revealed that the PTEN gene played a crucial role in COVID-19 infections through the activation of dendritic cells, the production of hyperactive B-cells and uncontrolled T-cells, and secretion of proinflammatory cytokines including interferons, TNF-alpha, IL-10, IL-4, and granulocyte monocyte colony stimulating factor. This all cooperates with other downregulated genes in the promotion of cytokine storms in COVID-19 patients [46]. Contrastly, Acarbose is a drug that indirectly treats the COVID-19 virus High blood sugar (Diabetes) is a known risk factor to COVID-19 [47]. A study in England found that patients diagnosed with both Diabetes and COVID-19 had a higher mortality rate. The odds ratios for in-hospital COVID-19 related deaths were 3.51 in people with type 1 diabetes and 2.03 in people with type 2 diabetes. Acarbose is a commonly used drug that lowers blood sugar and it is suggested that patients with lower blood sugar respond better to treatments for COVID-19, ultimately increasing the mortality rate.

Besides the drugs validated or predicted above, there are also 14 drugs with lower evidence scores. Among which, 4 have already been predicted to be used in Covid-19 treatment. For example, Bepridil was used in treatments for hypertension and chronic stable angina [48]. However, Bepridil comes in high dosages and can raise endosomal pH for slowing down SARS-nCoV-2 entry into human cell hosts and inhibiting $M^{Pro}$ in infected cells [41]. Despite not yet being in use for COVID-19 treatment, drugs such as sulfasalazine and mesalamine still have potential. Both drugs have a gene target, PPARG, which has been shown to have a reducing effect on cytokine storms in COVID-19, and in some cases, limits pulmonary inflammation [49,50]. Therefore, the 8 remaining drugs – zonisamide, ipatasertib, pyruvate, N-acetylglucosamine, temazepam,L-proline, tretinoin, and nintedanib – could have great potential in COVID-19 treatment.

# Discussion

The COVID-19 pandemic has had an effect on everyone in the world. The virus is highly contagious, and in severe cases it can cause death. So far, 76.55 million have been infected and 1.69 million have died globally. In the United States alone, 18.04 million have been infected and 322,936 have died [51]. These statistics will only increase as the trend for new cases continues to surge in the United States [52]. Many

await for a COVID-19 vaccine, but the need for treatment of COVID-19 infection still remains. The overall timetable for a vaccine's distribution, effectiveness, and time period to build up immunity allows for many more people to contract the virus. A faster development of drug treatments for the COVID-19 virus will undoubtedly save many lives. Therefore, there is much need for direct COVID-19 treatment.

New drug development takes lots of time and money. Researches have to first discover and develop the drug, go through long clinical trials, and then receive FDA approval in order for a newly developed drug to reach the marketplace. This process takes over 10 years to complete. Drug repurposing, however, is the fastest approach to finding a treatment for COVID-19 by repurposing pre-existing drugs. Repurposing drugs offers many benefits because it saves time and money on top of previously approved drugs that already exist in safe dosages [6]. In this study, we used a computational method based on patient gene expression data. We then used a database in parallel with patient expression data to find the relationship between the differentially expressed gene targets and already existing drugs to determine the potential candidates of COVID-19 drug repurposing. By doing so, potential drug repurposing candidates are identified much faster. The gene drug relationships are also curated and accurate, as the database ranks each interaction based on the amount of literature supporting it. Thus, by extension, the predictions made in this study are also reliable. However, the results from computational methods are only predictions. Wet lab experiments take longer, but compared to the computational prediction, has a higher likelihood of accuracy.

Besides from the computational approach taken in this study, there are many other methods of determining potential candidates for drug repurposing. Approaches such as binding assay, phenotypic screening, pathway based or network mapping, drug centric, target based, and signature based are all used to help researchers find repurposing drugs [6]. The pathway based or network mapping approach involves constructing networks based on the gene expression patterns, disease pathology and protein interactions. A signature based approach compares the signatures of a drug with that of another drug, disease or clinical phenotype [6]. In this study, we only used individual genes to determine candidates for drug repurposing, but by using a network or pathway based approach it would allow for drugs to target the hub genes, which can have a higher potential for treating the disease, because hub genes are likely to be the cause for the disease.

## Acknowledgement

I wish to thank Dr. Xue Gong for her continuous guidance and support in this study.

## References

1. Platto S, Xue T, and Carafoli E. COVID19: an announced pandemic. ***Cell Death Dis***. 2020, 11(9):799
2. Mahase E. Covid-19: Pfizer vaccine efficacy was 52% after first dose and 95% after second dose, paper shows. ***BMJ***. 2020, 371:m4826
3. Baum A, Ajithdoss D, Copin R, et al. REGN-COV2 antibodies prevent and treat SARS-CoV-2 infection in rhesus macaques and hamsters. ***Science***. 2020, 370(6520):1110-1115
4. Esakandari H, Nabi-Afjadi M, Fakkari-Afjadi J, Farahmandian N, Miresmaeili S-M, Bahreini E. A comprehensive review of COVID-19 characteristics. ***Biol Proced Online***
5. Sanders JM, Monogue ML, Jodlowski TZ, Cutrell JB. Pharmacologic Treatments for Coronavirus

Disease 2019 (COVID-19): A Review. ***JAMA***. 2020, 323(18):1824-1836
6. Parvathaneni V, Gupta V. Utilizing drug repurposing against COVID-19 - Efficacy, limitations, and challenges. ***Life Sci***. 2020, 259:118275
7. Riva L, Yuan S, Yin X, et al. Discovery of SARS-CoV-2 antiviral drugs through large-scale compound repurposing. ***Nature***. 2020, 586(7827):113-119
8. Stolfi P, Manni L, Soligo M, Vergni D, Tieri P. Designing a Network Proximity-Based Drug Repurposing Strategy for COVID-19. ***Front Cell Dev Biol***. 2020, 8:545089
9. Beers B. P-Value Definition. *Investopedia*. https://www.investopedia.com/terms/p/p-value.asp
10. False Discovery Rate. *Columbia Public Health*. https://www.publichealth.columbia.edu/research/population-health-methods/false-discovery-rate
11. Liang L, He Z, Yu H, et al. Selection and Validation of Reference Genes for Gene Expression Studies in Codonopsis pilosula Based on Transcriptome Sequence Data. ***Sci Rep***
12. Chen JJ, Wang S-J, Tsai C-A, Lin C-J. Selection of differentially expressed genes in microarray data analysis. ***Pharmacogenomics J***. 2007, 7(3):212-220
13. Oliveros JC. *Venny2.1*. https://bioinfogp.cnb.csic.es/tools/venny/
14. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. ***Omics J Integr Biol***. 2012, 16(5):284-287
15. *QIAGEN Ingenuity Pathway Analysis (QIAGEN IPA)*. QIAGEN https://digitalinsights.qiagen.com/products-overview/discovery-insights-portfolio/analysis-and-visualization/qiagen-ipa/
16. Griffith M, Griffith OL, Coffman AC, et al. DGIdb: mining the druggable genome. ***Nat Methods***. 2013, 10(12):1209-1210
17. Baker NC, Ekins S, Williams AJ, Tropsha A. A bibliometric review of drug repurposing. ***Drug Discov Today***. 2018, 23(3):661-672
18. Mercorelli B, Palù G, Loregian A. Drug Repurposing for Viral Infectious Diseases: How Far Are We? ***Trends Microbiol***. 2018, 26(10):865-876
19. Parvathaneni V, Kulkarni NS, Muth A, Gupta V. Drug repurposing: a promising tool to accelerate the drug discovery process. ***Drug Discov Today***. 2019, 24(10):2076-2085
20. Ciliberto G, Mancini R, Paggi MG. Drug repurposing against COVID-19: focus on anticancer agents. ***J Exp Clin Cancer Res CR***. 2020, 39(1):86
21. Pushpakom S, Iorio F, Eyers PA, et al. Drug repurposing: progress, challenges and recommendations. ***Nat Rev Drug Discov***. 2019, 18(1):41-58
22. Xiong Y, Liu Y, Cao L, et al. Transcriptomic characteristics of bronchoalveolar lavage fluid and peripheral blood mononuclear cells in COVID-19 patients. ***Emerg Microbes Infect***. 2020, 9(1):761-770
23. Blanco-Melo D, Nilsson-Payant BE, Liu W-C, et al. Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19. ***Cell***. 2020, 181(5):1036-1045.e9
24. Bojkova D, Klann K, Koch B, et al. Proteomics of SARS-CoV-2-infected host cells reveals therapy targets. ***Nature***. 2020, 583(7816):469-472
25. Shen B, Yi X, Sun Y, et al. Proteomic and Metabolomic Characterization of COVID-19 Patient Sera. ***Cell***. 2020, 182(1):59-72.e15
26. Sun X, Wang T, Cai D, et al. Cytokine storm intervention in the early stages of COVID-19

pneumonia. ***Cytokine Growth Factor Rev***. 2020, 53:38-42
27. Castelli V, Cimini A, Ferri C. Cytokine Storm in COVID-19: "When You Come Out of the Storm, You Won't Be the Same Person Who Walked in." ***Front Immunol***. 2020, 11:2132
28. Grifoni E, Valoriani A, Cei F, et al. Interleukin-6 as prognosticator in patients with COVID-19. ***J Infect***. 2020, 81(3):452-482
29. Li H, Xiang X, Ren H, et al. Serum Amyloid A is a biomarker of severe Coronavirus Disease and poor prognosis. ***J Infect***. 2020, 80(6):646-655
30. D'Alessandro A, Thomas T, Dzieciatkowska M, et al. Serum Proteomics in COVID-19 Patients: Altered Coagulation and Complement Status as a Function of IL-6 Level. ***J Proteome Res***. Published online August 14, 2020
31. Tay MZ, Poh CM, Ré nia L, MacAry PA, Ng LFP. The trinity of COVID-19: immunity, inflammation and intervention. ***Nat Rev Immunol***. 2020, 20(6):363-374
32. Gao Z, Xu Y, Sun C, et al. A Systematic Review of Asymptomatic Infections with COVID-19. *J Microbiol Immunol Infect Wei Mian Yu Gan Ran Za Zhi*. Published online May 15, 2020. doi:10.1016/j.jmii.2020.05.001
33. Zimmer K. Why Some COVID-19 Cases Are Worse than Others. ***The Scientist***. https://www.the-scientist.com/news-opinion/why-some-covid-19-cases-are-worse-than-others-67160. Published February 24, 2020.
34. Nguyen DD, Gao K, Chen J, Wang R, Wei G-W. Potentially highly potent drugs for 2019-nCoV. ***BioRxiv Prepr Serv Biol***. Published online February 13, 2020. doi:10.1101/2020.02.05.936013
35. Rudnicka L, Glowacka P, Goldust M, et al. Cyclosporine therapy during the COVID-19 pandemic. ***J Am Acad Dermatol***. 2020, 83(2):e151-e152
36. Khalil A, Kamar A, Nemer G. Thalidomide-Revisited: Are C OVID-19 Patients Going to Be the Latest Victims of Yet Another Theoretical Drug-Repurposing? ***Front Immunol***. 2020, 11:1248
37. Vatansev H, Kadiyoran C, Cumhur Cure M, Cure E. COVID-19 infection can cause chemotherapy resistance development in patients with breast cancer and tamoxifen may cause susceptibility to COVID-19 infection. ***Med Hypotheses***. 2020, 143:110091
38. Polonikov A. Endogenous Deficiency of Glutathione as the Most Likely Cause of Serious Manifestations and Death in COVID-19 Patients. ***ACS Infect Dis***. 2020, 6(7):1558-1562
39. Gassen N. Analysis of SARS-CoV-2-controlled autophagy reveals spermidine, MK-2206, and niclosamide as putative antiviral therapeutics. ***BioRxiv Prepr Serv Biol***. Published online April 15, 2020. https://doi.org/10.1101/2020.04.15.997254
40. Nambou K, Anakpa M. Deciphering the co-adaptation of codon usage between respiratory coronaviruses and their human host uncovers candidate therapeutics for COVID-19. ***Infect Genet Evol J Mol Epidemiol Evol Genet Infect Dis***. 2020, 85:104471
41. Vatansever EC, Yang K, Kratch KC, et al. Targeting the SARS-CoV-2 Main Protease to Repurpose Drugs for COVID-19. ***BioRxiv Prepr Serv Biol***. Published online May 23, 2020. doi:10.1101/2020.05.23.112235
42. Alexpandi R, De Mesquita JF, Pandian SK, Ravi AV. Quinolines-Based SARS-CoV-2 3CLpro and RdRp Inhibitors and Spike-RBD-ACE2 Inhibitor for Drug-Repurposing Against COVID-19: An in silico Analysis. ***Front Microbiol***. 2020, 11:1796
43. Cava C, Bertoli G, Castiglioni I. A protein interaction map identifies existing drugs targeting

SARS-CoV-2. ***BMC Pharmacol Toxicol***. 2020, 21(1):65

44. Singh N, Bharara Singh A. S2 subunit of SARS-nCoV-2 interacts with tumor suppressor protein p53 and BRCA: an in silico study. ***Transl Oncol***. 2020, 13(10):100814
45. Maremanda KP, Sundar IK, Li D, Rahman I. Age-dependent assessment of genes involved in cellular senescence, telomere and mitochondrial pathways in human lung tissue of smokers, COPD and IPF: Associations with SARS-CoV-2 COVID-19 ACE2-TMPRSS2-Furin-DPP4 axis. ***MedRxiv Prepr Serv Health Sci***. Published online June 16, 2020. doi:10.1101/2020.06.14.20129957
46. Zolfaghari Emameh R, Nosrati H, Eftekhari M, Falak R, Khoshmirsafa M. Expansion of Single Cell Transcriptomics Data of SARS-CoV Infection in Human Bronchial Epithelial Cells to COVID-19. ***Biol Proced Online***. 2020, 22:16
47. Fang L, Karakiulakis G, Roth M. Are patients with hypertension and diabetes mellitus at increased risk for COVID-19 infection? ***Lancet Respir Med***. 2020, 8(4):e21
48. COMPOUND SUMMARY Bepridil. *PubChem*. https://pubchem.ncbi.nlm.nih.gov/compound/bepridil
49. Huang S, Zhu B, Cheon IS, et al. PPAR-γ in Macrophages Limits Pulmonary Inflammation and Promotes Host Recovery following Respiratory Viral Infection. ***J Virol***. 2019, 93(9)
50. Ciavarella C, Motta I, Valente S, Pasquinelli G. Pharmacological (or Synthetic) and Nutritional Agonists of PPAR-γ as Candidates for Cytokine Storm Modulation in COVID-19 Disease. *Mol Basel Switz*. 2020, 25(9)
51. COVID-19 CORONAVIRUS PANDEMIC. *worldometer*. https://www.worldometers.info/coronavirus/?utm_campaign=homeAdvegas1
52. Schwartz M. Fauci Warns Of "Surge Upon A Surge" As COVID-19 Hospitalizations Hit Yet Another High. *NPR*. https://www.npr.org/sections/coronavirus-live-updates/2020/11/29/939863068/fauci-warns-of-a-surge-upon-surge-as-covid-19-hospitalizations-hit-yet-another-h