OXFORD

BIOLOGY
Methods & Protocols

Biology Methods and Protocols, 2022, bpac013
https://doi.org/10.1093/biomethods/bpac013
Advance Access Publication Date: 30 May 2022
METHODS ARTICLE

MLBioIGE: integration and interplay of machine learning and bioinformatics approach to identify the genetic effect of SARS-COV-2 on idiopathic pulmonary fibrosis patients

Sk. Tanzir Mehedi [1,3,†], Kawsar Ahmed [2,3,*,†], Francis M. Bui [2], Musfikur Rahaman[1], Imran Hossain[1], Tareq Mahmud Tonmoy[1], Rakibul Alam Limon[1], Sobhy M. Ibrahim[4] and Mohammad Ali Moni [5,*]

[1]Department of Information Technology, University of Information Technology and Sciences, Baridhara, Dhaka-1212, Bangladesh,
[2]Department of Electrical and Computer Engineering, University of Saskatchewan, Saskatoon, SK S7N 5A9, Canada,
[3]Group of Bio-PhotomatiX, Department of Information and Communication Technology, Mawlana Bhashani Science and Technology University, Santosh, Tangail 1902, Bangladesh,
[4]Department of Biochemistry, College of Science, King Saud University, Riyadh 11451, Saudi Arabia and
[5]Faculty of Health and Behavioural Sciences, School of Health and Rehabilitation Sciences, The University of Queensland, St Lucia, QLD 4072, Australia

*Correspondence address. (K.A.) Department of Information and Communication Technology, Group of Bio-Photomatix, Mawlana Bhashani Science and Technology University, Santosh, Tangail-1902, Bangladesh. E-mail: kawsar.ict@mbstu.ac.bd and (M.A.M.) Faculty of Health and Behavioural Sciences, School of Health and Rehabilitation Sciences, The University of Queensland, St Lucia, QLD, 4072, Australia. E-mail: k.ahmed@usask.ca, m.moni@uq.edu.au
†These authors contributed equally to this work.

Abstract

SARS-CoV-2, the virus that causes COVID-19, is a current concern for people worldwide. The virus has recently spread worldwide and is out of control in several countries, putting the outbreak into a terrifying phase. Machine learning with transcriptome analysis has advanced in recent years. Its outstanding performance in several fields has emerged as a potential option to find out how SARS-CoV-2 is related to other diseases. Idiopathic pulmonary fibrosis (IPF) disease is caused by long-term lung injury, a risk factor for SARS-CoV-2. In this article, we used a variety of combinatorial statistical approaches, machine learning, and bioinformatics tools to investigate how the SARS-CoV-2 affects IPF patients' complexity. For this study, we employed two RNA-seq datasets. The unique contributions include common genes identification to identify shared pathways and drug targets, PPI network to identify hub-genes and basic modules, and the interaction of transcription factors (TFs) genes and TFs–miRNAs with common differentially expressed genes also placed on the datasets. Furthermore, we used gene ontology and molecular pathway analysis to do functional analysis and discovered that IPF patients have certain standard connections with the SARS-CoV-2 virus. A detailed investigation was carried out to recommend therapeutic compounds for IPF patients affected by the SARS-CoV-2 virus.

Keywords: SARS-CoV-2; COVID-19; machine learning; idiopathic pulmonary fibrosis; gene ontology; differentially expressed genes

Introduction

Coronaviruses have various variants that can infect humans and animals [1]. The variants of this virus are responsible for various diseases, ranging from common fever and cold cough to more serious illnesses such as Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome (MERS) [2]. Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) is a new type of coronavirus, which got a lot of attention at the end of 2019 because it was a new variant of coronavirus that had never been observed in humans previously. Coronavirus Disease 2019 (COVID-19) is the name of the new coronavirus, which was first discovered in Wuhan, China, in December 2019 [3]. Chinese officials reported 44 instances of pneumonia with unknown causes to the World Health Organization (WHO) between 31 December 2019, and 3 January 2020 [4]. The first fatality of COVID-19 occurred in Wuhan on 9 January 2020, while the first death outside of China occurred in the Philippines on 1 February 2020. Within a few days, the disease had spread worldwide and was out of control in many nations [4]. On 30 January 2020, the WHO designated the virus as a Public Health Emergency (PHE) of worldwide concern [5]. This virus was declared a pandemic by the same organization on 11 March 2020, after a total of 4500 deaths were reported in 30 countries and territories throughout the world [5]. Italy surpassed China, with the highest reported death cases of this virus reported on 19 March 2020 [4]. The USA has surpassed both China and Italy as the country with the highest confirmed virus cases on 26 March 2020 [4]. On a global basis, the bloodiest week was 13–19 April 2020, when nearly 7460 deaths were officially reported each day by this virus. The pandemic's epicenter migrated to Latin America and the Caribbean in June 2020. Between 15 July 2020 and 15 August 2020, the region had an average of almost 2500 deaths per day. With over 78 000 cases on 30 August 2020, India surpassed the US record for the highest cases in a single day, and a second wave hit India on 9 April 2021.

Received: March 13, 2022. **Revised:** April 18, 2022. **Accepted:** May 25, 2022

There were 281 808 270 confirmed cases from December 2020 to December 2021, with 5 411 75 deaths by this virus [6]. On 26 November 2021, WHO designated a new variant (B.1.1.529) of SARS-CoV-2 named Omicron in South Africa. On 26 November, WHO designated Omicron as a variant of concern. The first COVID-19 case associated with the Omicron variant was reported in the USA on 1 December 2021, and at least one Omicron variant had been detected in 22 states as of 8 December 2021. Recently, this new variation of this virus has spread worldwide and is out of control in several countries, putting the outbreak into a terrifying phase.

SARS-CoV-2 is a single-stranded RNA virus that is positive in a sense. The Spike (S), envelope (E), membrane (M), and nucleocapsid (N) proteins are the four proteins found in SARS-CoV-2. Spike proteins are responsible for attaching to a host cell's membrane. Idiopathic pulmonary fibrosis (IPF) disease is a long illness marked by the thickening and stiffening of lung tissue associated with scar tissue formation [7]. In this condition, the sponge or meaty section of the lung becomes scarred or fibrotic. It is a slow-progressing, highly fatal disease that affects roughly 80% of people within 3–5 years of diagnosis [7]. Pulmonary fibrosis affects people in different ways. Various common, easily curable diseases might cause similar symptoms. Shortness of breath and a persistent dry, hacking cough are the most common indications and symptoms of IPF. Many impacted people also notice a decrease in appetite and weight loss over time. Due to a lack of oxygen, some people with IPF acquire enlarged, rounded tips on their fingers and toes (clubbing) [8]. IPF's cause is not understood. The following are some of the most common risk factors for IPF: Almost all patients with IPF are over 50 years. Genetics, up to 20% of patients with IPF have another family member who suffers from the condition. Approximately 75% of people with IPF smoke now or have in the past. Gastroesophageal reflux or heartburn affects about 75% of people with IPF. Male patients account for roughly 65% of IPF patients [9]. Radiation treatments to the chest or the use of certain chemotherapy medications have been shown to enhance the risk of pulmonary fibrosis [10, 11]. SARS-CoV-2 contains spike protein, which has a greater interaction with ACE2, and IPF patients have a lot of this enzyme, confirming IPF as a risk factor for this disease [12, 13]. These investigations have revealed several linkages between IPF and COVID-19, which raises concerns.

Contributions

In this article, we used a variety of combinatorial statistical approaches, machine learning algorithms, and bioinformatics tools to investigate how the SARS-CoV-2 virus affects IPF patients' complexity. The following are the main contributions of this article:

- the experiments have been conducted using a real-time dataset. We have observed common gene identification by machine learning algorithms and various bioinformatics analyses to identify shared pathways and drug targets;
- the Protein-Protein Interaction (PPI) network was examined to discover hub-genes and modules. The interactions of transcription factors (TFs) genes and TFs-miRNAs with common differentially expressed genes (DEGs) were also discovered. Furthermore, we used gene ontology (GO) analyses and molecular pathway analyses to do functional analysis and discovered that IPF patients have certain common connections with SARS-CoV-2 infection;
- a comprehensive analysis has been conducted to suggest drug molecules for IPF patients with SARS-CoV-2 infections. In the context of molecular-based knowledge and several pathway-based analyses, which illustrate the utility of the biological system for both SARS-CoV-2 and IPF; and
- finally, the current challenges and future research directions of integration and interplay between machine learning and bioinformatics have been discussed.

The remainder of this study is organized in the following manner. The 'Materials and methods' section begins with a full description of the dataset with preprocessing and an overview of selected methodology. The 'Result analysis' section discusses the evaluation and interpretation of experimental outcomes for these methodologies. In addition, 'Discussion' section contains a lengthy explanation and discusses some application areas for scientific society. Finally, 'Conclusions' section contains an overview of the findings and possible future directions.

Materials and methods

In this section, we have thoroughly detailed the overview of the analysis, including the dataset transformation process and various transcriptome analyses.

Overview of approach

We applied machine learning and transcriptomic analysis to identify shared associations between SARS-CoV-2 and IPF by employing selected datasets shown in the block diagram in Fig. 1. The machine learning approaches have been used to identify common DEGs of the selected datasets. Furthermore, these shared or common DEGs were used to construct gene–disease association networks, identify GO, pathways, PPI network, hub-genes, transcription factor (TF)–gene, TF–miRNA, and identify candidate drugs.

Dataset analysis

This section has performed a series of operations on the dataset without changing its properties. Also, we have thoroughly explained the overview of the selected dataset.

Dataset description

We have identified common genetic interrelationships between SARS-CoV-2 and IPF using Ribonucleic Acid Sequencing (RNA-Seq) datasets from the Gene Expression Omnibus (GEO) collection of the National Center for Biotechnology Information (NCBI) directory [14, 15]. The transcriptional responses to SARS-CoV-2 infection are contained in the SARS-CoV-2 dataset with GEO accession ID GSE147507 and GEO platform ID GPL18573. In contrast, the transcriptome analysis reveals differential splicing events in IPF lung tissue that are contained in the IPF dataset GEO accession GSE52463 and GEO platform ID GPL11154 [16]. SARS-CoV-2-affected Lung Epithelial Cell (LECs) are found in the GSE147507, while IPF-affected lung tissues are found in the GSE52463 dataset. The GSE147507 dataset contains two types of samples (control and SARS-CoV-2-affected cells) taken from SARS-CoV-2-affected LECs, while the GSE52463 dataset has two types of samples (control and IPF-affected cells). Metadata and count data are also included in both databases. The RNA sequence was extracted from the GSE147507 dataset using high-throughput sequencing technologies on the Illumina NextSeq 500 (*Homo sapiens*) platform [17]. The IPF dataset, on the other hand, comprises mRNA sequencing of eight IPF-affected lung tissues and seven control lung tissue samples, all of which were sequenced on the Illumina Hi-Seq 2000 (*H.*

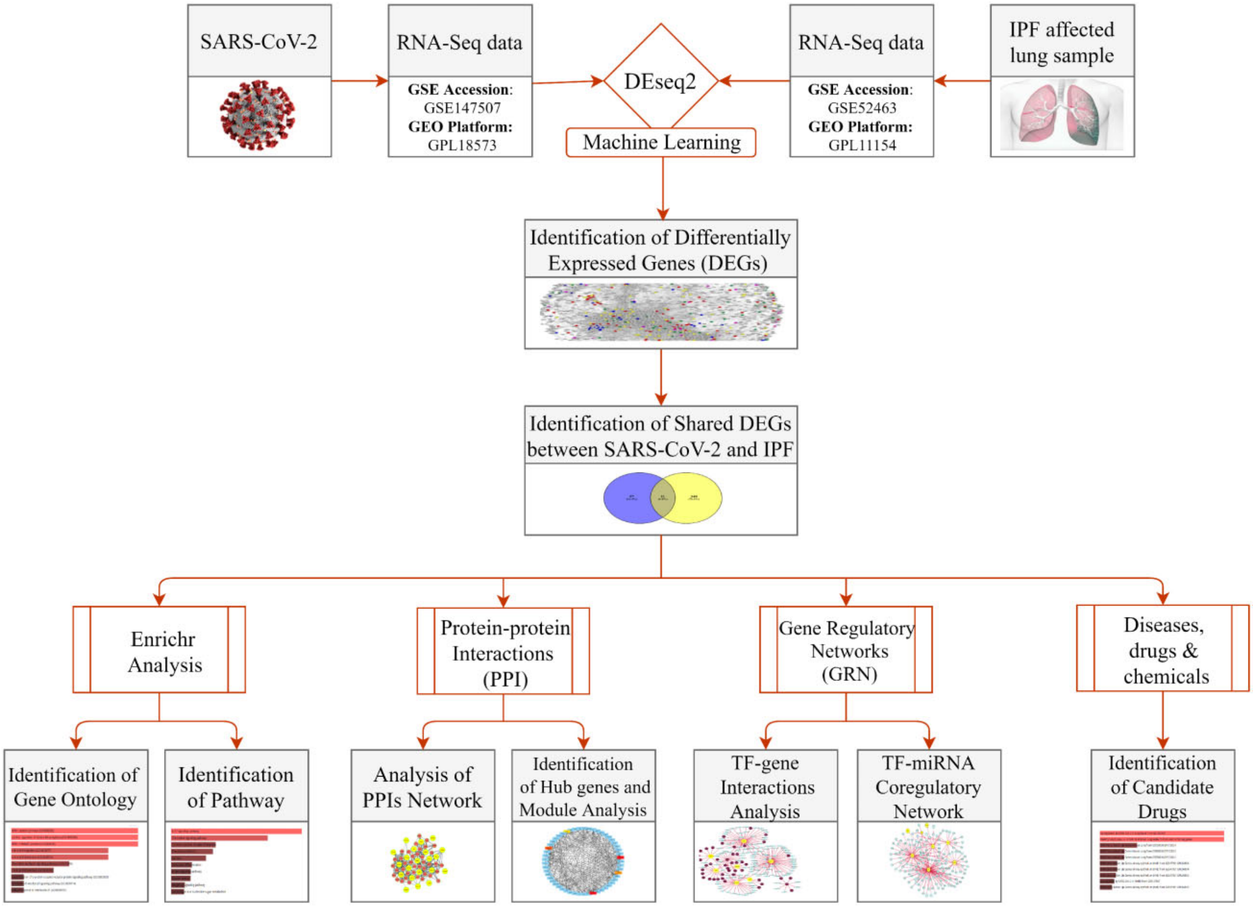
SARS-CoV-2
RNA-Seq data
GSE Accession: GSE147507
GEO Platform: GPL18573
DEseq2
Machine Learning
RNA-Seq data
GSE Accession: GSE52463
GEO Platform: GPL11154
IPF affected lung sample
Identification of Differentially Expressed Genes (DEGs)
Identification of Shared DEGs between SARS-CoV-2 and IPF
Enrichr Analysis
Protein-protein Interactions (PPI)
Gene Regulatory Networks (GRN)
Diseases, drugs & chemicals
Identification of Gene Ontology
Identification of Pathway
Analysis of PPIs Network
Identification of Hub genes and Module Analysis
TF-gene Interactions Analysis
TF-miRNA Coregulatory Network
Identification of Candidate Drugs

Figure 1: The complete workflow for the current investigation. Two types of samples (control cells, affected cells) were collected from SARS-CoV-2-infected lung epithelial cells and both are included in the GSE147507 dataset. The GSE52463 dataset contains IPF-affected lung samples. Common DEGs were identified from both the datasets using machine learning technique. From the common DEGs, GO identification, pathway analysis, PPIs network, TF–gene analysis, TF–miRNA analysis, and hub-gene identification were designed and based on those analysis drug molecule identification was performed.

Table 1: Contents of the datasets

Properties	SARS-CoV-2	IPF
GEO Accession	GSE147507	GSE52463
GEO Platform	GPL18573	GPL11154
Organisms	*Homo sapiens*	*Homo sapiens*
Assay type	RNA-Seq	RNA-Seq
Type of the datasets	Transcriptional response to SARS-CoV-2 infection	In IPF lung tissue, transcriptome analysis indicates distinct splicing events.
Instrument	Illumina NextSeq 500	Illumina HiSeq 2000
Total GEO samples	110	15
Experiment type	High-throughput sequencing for expression profiling	High throughput sequencing for expression profiling

sapiens) platform utilizing high-throughput sequencing technology [18]. Table 1 lists the datasets used in this study and their geo-features and sequencing methods.

Data preparation

To achieve optimal performance, it is necessary to clean and prepare the dataset before applying machine learning methods. Data preparation is generally done by removing unnecessary features, checking the variation of independent features, converting non-numerical features, removing outliers, and replacing missing values if they exist. The two fundamental steps apply during the data preparation process. The first is data preprocessing, and the second is the data transformation step.

Data preprocessing

This dataset originates from multiple heterogeneous sources. Due to its vast size, this dataset is highly susceptible to missing and noisy data. This section discusses the essential steps in data preprocessing: data cleaning and data integration.

- Data Cleaning: First, we applied various techniques to remove noise and clean inconsistencies in the metadata and count-data from both datasets. For example, Rosner's test for outliers checking and the predictive mean matching method for imputing missing values. Then, to apply machine learning techniques, we converted the qualitative values into quantitative values by applying various techniques (e.g. Biobase (version

2.30.0), GEOquery (version 2.40.0), limma (version 3.26.8), and Bioconductor) packages of the R programming language, which is a free, open-source, and open-development software project for the analysis and comprehension of genomic data.

- Data Integration: To improve the accuracy, the data integration technique helped us reduce and avoid redundancies in the resulting dataset. This dataset originates from multiple heterogeneous sources. So, it is essential to check both datasets for redundancy and correlation analysis. This analysis has measured how strongly one feature implies the other. Figure 2a and b shows the correlation between different features for the two datasets, GSE147507 and GSE52463, respectively. For our analysis, we have evaluated the correlation between all the features using the following Pearson's product-moment coefficient equation.

$$r = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}$$

where $\bar{a}$ is the meaning of x variable and $\bar{y}$ is the meaning of y variable, x_i and y_i are values in tuple i.

Data transformation

We applied this processing step to achieve more efficient resulting processes and easily understand the patterns. Some selected features have larger values than others, which leads to incorrect performance. We have implemented these strategies to scale the selected feature values within a range between [0.0] and [1.0] without changing the characteristics of the data.

$$N=(X-Xmin)/(Xmax-Xmin)$$

where N is the output normalized values, X is an original value and Xmax and Xmin is the maximum and minimum values of the feature, respectively.

As shown in the following equation, a technique called minimum–maximum normalization has been used to scale the selected feature values within the range. We have also evaluated the density plot for both datasets. The density plot shows the

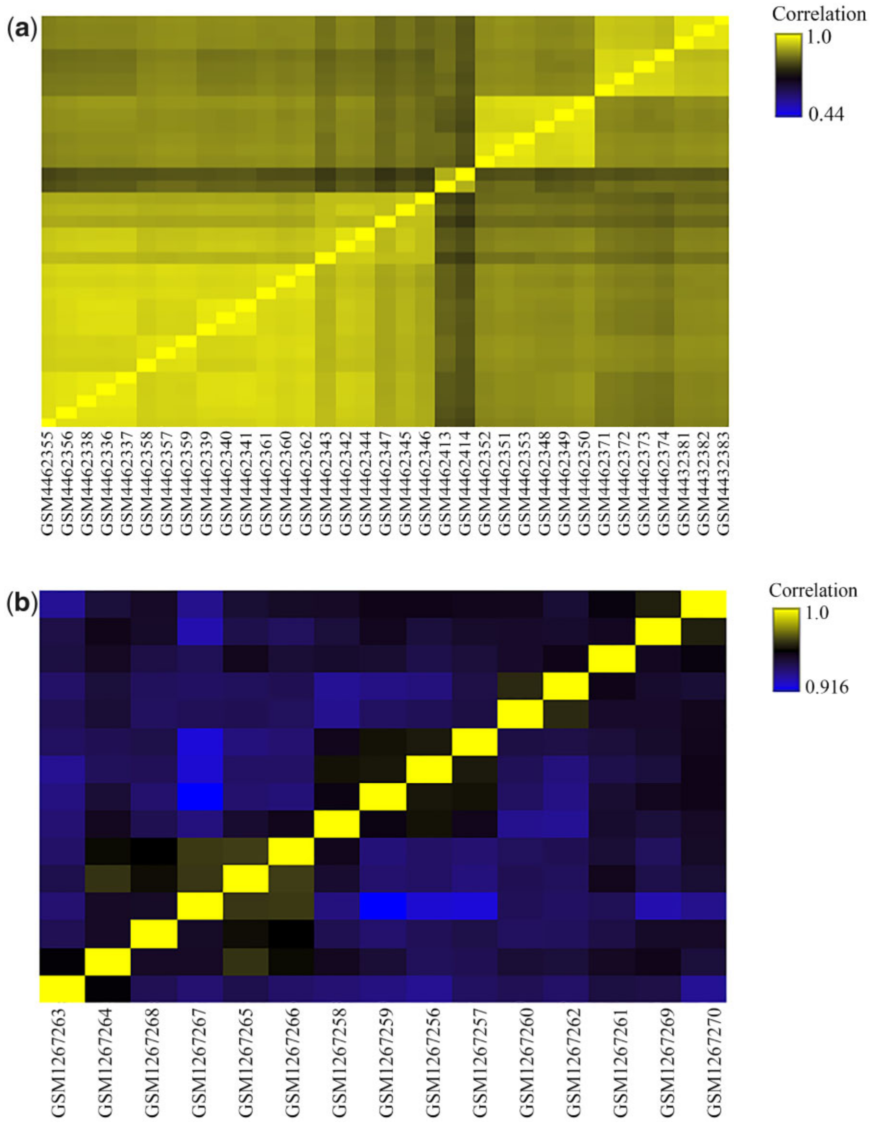

Figure 2: The correlation analysis between different features for the two datasets (a) GSE147507 and (b) GSE52463. This analysis has measured how strongly one feature implies the other.

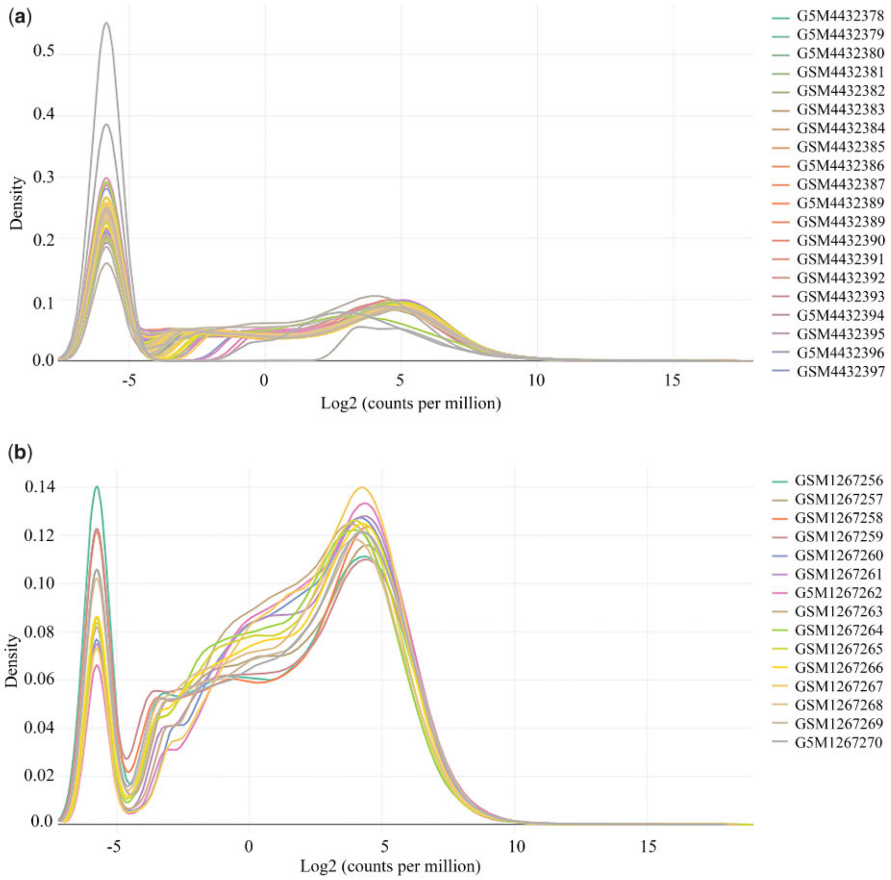

Figure 3: The density plots of the two datasets (a) GSE147507 and (b) GSE52463. The density plot shows the smooth distribution of the points along the numeric axis. The peaks of the density plot are at the locations where there is the highest concentration of points.

smooth distribution of the points along the numeric axis. The peaks of the density plot are at the locations where there is the highest concentration of points. Figure 3a and b shows density plots for the two datasets, GSE147507 and GSE52463, respectively.

Spotting DEGs and shared DEGs between SARS-CoV-2 and IPF

A gene is differentially expressed when a statistical discrepancy exists between several test settings during the transcription phase [19]. The major purpose of this study is to identify DEGs that are shared between the GSE147507 and GSE52463 datasets. The DESeq2 and lima packages of the R programming language were used to access data generated by microarray analysis. DEGs from both datasets were identified using a machine learning method. Listing 1 shows the applied procedure of the machine learning algorithm to identify DEGs from both datasets. Across all datasets, significant DEGs were identified using cutoff criteria (P-value < 0.05 and $|\log\ \mathrm{Fc}| \geq 1.00$). The shared DEGs of the GSE147507 and GSE52463 datasets were found using the online VENN analysis platform JVENN tool [20].

Identifying of GO and molecular pathway

Enrichment analysis of gene set is a technique for identifying DEGs linked to a biological process or molecular function [21]. GO is a classification system that divides genes into biological mechanisms, molecular functions, and cellular components [22]. The purpose of analyzing GO concepts is to understand the molecular activity, cellular structure, and the position in the cell where genes fulfill their functions [23]. We used four databases to find common molecular pathways in IPF and COVID-19: Kyoto Encyclopedia of Genes and Genomes (KEGG) [24], Wiki Pathways [24], Reactome [25], and BioCarta [25]. Various gene annotations may be found in the KEGG, which is commonly used to characterize metabolic pathways. A web-based platform Enrichr has been used to obtain GO, and molecular pathways for the common genes mentioned earlier in this research [26, 27]. To derive GO and molecular pathways, we utilized 20 sorted genes.

Analysis of PPI network

The role of PPIs in cellular biology is projected to be a major focus of research, and it serves as a requirement for system biology [28]. Proteins finish their journey within a cell with a comparable protein affiliation established by a PPI network, indicating the protein processes. Proteins interact with other proteins to carry out their activities inside cells, and the information created by a PPI network informs individuals about the protein's function [29]. We built the PPI network of DEGs proteins using the STRING resource to exchange activity and physical linkages between IPF and COVID-19 [30]. The STRING generates experimental and predicted outcomes based on the data and the interaction generated by the online tool, which is determined by 3D structures, accessory data, and confidence scores [31]. The confidence score was set using the STRING platform that was different categorized confidence scores (low, medium, and high). We have been worked on the PPI network with a medium confidence score (0.400). We get the exact information,

Listing 1. The procedure of the machine learning algorithm to identify DEGs from both datasets.

```
Input: RNA-Seq dataset
Output: Identification of differentially expressed genes (DEGs)

outputFileName=open(“result.csv”,”a”)

datasetName=os.listdir(folderPath)
for i in range(len(datasetName)):
    fileName=open(“GSE”+str(name[i])+”.csv”)
    fileName.next()
    for dataset in fileName:
#Extract countdata and store in a matrix
datasetName=os.listdir(folderPath)
countData = read.csv((“GSE”+str(name[i])+”filtered_countdata.csv”):
        countDataFrame = data.frame(countData)
        countDataFrameRound=mutate(across(where(is.numeric),round , 3))
#Extract metadata and store in a matrix
metaData = read.csv((“GSE”+str(name[i])+”filtered_metadata.csv”):
    countDataFrame = data.frame(metaData)
#Analyze count data using DESEQ2
applyDESeq = DESeqDataSetFromMatrix(countData=countDataFrameRound,
              colData=metaDataFrame, design=~treatment, tidy=TRUE)
applyDESeq = DESeq(applyDESeq)
result = results(applyDESeq)

#Result Analysis
#Check and Omit the null value
checkNull = is.na(result)
resultsOmitNa = na.omit(result)

#Count the up regulated gene
resultOmitNaFilterUp = filter(resultOmitNa, log2FoldChange
                         > 1 & Padj < 0.05)
#Count the down regulated gene
resultOmitNaFilterDown = filter(resultOmitNa, log2FoldChange
                           <-1 & Padj < 0.05)
#ABS logFC value and setup cuttoff criteria for P adj value
resultFinal = filter(resultOmitNa, abs(log2FoldChange)
               > 1 & Padj < 0.05)
outputFileName.write(resultFinal)
outputFileName.close()
```

using the network type “full string network” (the edges indicate both functional and physical protein resources) and a selected number of 10 interactors. Then, we consume our PPI network into Cytoscape (version 3.7.1) for visual representation and further PPI network experimental studies. And with that the purpose of identifying hub-genes, the obtained PPIs are analyzed through Cytoscape. Cytoscape is an open-source network visualization framework that serves as a versatile method for combining several datasets to optimize efficiency for various interactions such as protein–protein interactions, genetic interactions, and protein–DNA interactions, among others [32, 33].

Identifying of hub-genes and module analysis

The PPI networks are nodes, edges, and connections, with hub-genes being the most entangled nodes. The PPI networks are used to identify hub-genes. Hub-genes provide dense areas identified as important parts of the PPIs network. The hub-genes for the associated PPI networks are indicated by CytoHubba, a Cytoscape application plugin [34]. CytoHubba is the most popular Cytoscape hub-genes identification plugin for its user-friendly interface. CytoHubba has 20 different methods for topological analysis (e.g. MCC, Degree, DMNC, MNC, EPC, Bottleneck, etc.). The degree analysis method was employed to find the hub-genes for this study. Because the degree method facilitates analysis by suggesting large, closely compacted modules in the PPI network, it is employed instead of another approach [35]. The Molecular Complex Detection (MCODE) plugin in the Cytoscape software is utilized to locate the most profound modules in the PPIs network [36]. The MCODE method is based on a graph-theoretic clustering algorithm that detects densely connected regions in large protein–protein interaction networks that may represent molecular complexes [36]. The method has the advantage over other graph clustering methods of having a directed mode that allows fine-tuning of clusters of interest without considering the rest of the network and allows examination of cluster inter-connectivity, which is relevant for protein networks. Furthermore, the method is not affected by the known high rate of false positives in data from high-throughput interaction techniques [37]. Moreover, the method is relatively easy to implement and, since it is local density based, has the advantages of both a directed mode and a complex connectivity mode. The MCODE method has also been employed in the PPIs network to locate highly bound areas in the molecular complexes.

TF–gene analysis

TFs bind to individual genomes and regulate their levels of expression. As a result, it is required for molecular recognition [38]. In all species, TFs control gene expression and play a critical role in transcription. TFs play an important role in a variety of biological processes, including cell cycle regulation and development. TF–gene linkage with the newly discovered top 12 common DEGs among 90 DEGs was used to investigate the effects of TF–genes on functional pathways and genomic levels. By using the Network Analyst tool to find topologically relevant TFs from the ENCODE database, which was used in the TF–gene interaction network [39–41], we were able to exploit TF–gene interactions with previously established common genes. Network Analyst is a web-based tool for doing transcriptional research and meta-analysis on various species, including humans [42, 43]. The TF–gene interaction network has made up of 190 nodes and 301 edges. Moreover, the network has 12 DEGs and 178 TF–genes, where HSPB6 is regulated by 85 TF–genes, EPAS1 is regulated by 68 TF–genes, and FCGR2A is regulated by 37 TF–genes according to their degree value. These 178 TF–genes are regulated by more than one common DEG, which indicates high interaction of the TF–genes with common DEGs.

TF–miRNA interaction with the common DEGs

The miRNAs are short non-RNAs that are expressed by RNA polymerase II and then regulated by a shared biogenic pathway in a step-by-step method. Using a combination of experimental and computational techniques, miRNAs have been discovered in a variety of species. By binding to the 3′-untranslated, miRNA regulates gene expression at the post-transcriptional stage. The RegNetwork database was utilized to collect TF–miRNA coregulatory interactions, which helps to identify the miRNAs and regulatory TF–genes that regulate DEGs of interest at the transcriptional and post-transcriptional phases [43]. We found miRNAs that interact with common DEGs and then utilized the Network Analyst tool to analyze how they interact. With this platform, researchers can find complex datasets and determine biological traits and functions [44]. The network of miRNA–gene interactions was examined using Cytoscape software. By classifying top miRNAs to higher levels, this software aids researchers in determining biological roles and features. The TF–miRNA coregulatory network has 191 nodes and 216 edges. According to research, DEGs engage with 87 miRNAs and 93 TF–genes.

Candidate drugs identification

Predicting PDI or drug molecule recognition is important for this research. We identified a therapeutic molecule based on the common DEGs of SARS-CoV-2 and IPF using the Enrichr tool and DSigDB database. There are 22 527 gene sets in the drug signatures database. To acquire access to the DSigDB database, the Enrichr platform is employed [45, 46]. Enrichr is a well-known web portal with many gene-set libraries that may be used to look into gene-set enrichment on a genome-wide scale [26].

Result analysis

The overall performance of the analysis is discussed in this section. Beginning with a discussion of DEGs and mutual DEG identification, the article progresses to a description of the candidate drug identification procedure.

DEGs and mutual DEGs identification

We investigated the interrelationships and implications of disrupted genes that activate COVID-19 and IPF using the NCBI's human RNA-seq and microarray datasets. The GSE147507 dataset determines DEGs for SARS-CoV-2, and its GEO platform identifier is GPL18573. There are 926 upregulated and 799 downregulated genes in the GSE147507 dataset, resulting in 1725 DEGs. In the GSE52463 dataset, which has the GEO platform identifier GPL11154, we discovered a total of 1008 DEGs, with 669 upregulated and 339 downregulated genes. The quantitative measurement of the selected datasets is shown in Table 2. After cross-comparative analysis using JVENN, a trustworthy web platform for Venn analysis, we discovered 90 similar DEGs from the GSE147507 and GSE52463 datasets. Twenty common DEGs were chosen for further study from 90 common DEGs based on the *P*-value (MDK, HP, HSPB6, CHIT1, TNFAIP6, EPAS1, MMP1, CCL18, CXCL6, CCL11, IL1RN, LAMP3, CD207, ARRB1, RNASE2, LILRA1, FCGR2A, STAT4, CD69, and SAMSN1). Additional study has been conducted using these 20 frequent DEGs. Figure 4 depicts the common DEGs as a Venn diagram, with 90 genes discovered to be shared in the GSE147507 and GSE52463 datasets.

GO and molecular pathway analysis

Enrichment analysis of gene sets is a technique for identifying DEGs linked to a biological process or molecular function. For this study, we looked at the most prevalent DEGs. GO processes are divided into biological, cellular components, and molecular functions. Table 3 shows the biological process connected to GO keyword identification findings based on the combined score. Table 4 shows the results of the identification of molecular function-related GO keywords based on the combined score. Table 5 also shows the results of the cellular component-related GO keywords identification based on the combined score. The KEGG, Wiki Pathways, Reactome, and BioCarta have been used to

Table 2: Quantitative measurements of the datasets used in this analysis

Properties	GSE147507	GSE52463
Common gene analysis	DESeq2 and the lima package	DESeq2 and the lima package
Cutoff criteria	$P < 0.05$ and $\lvert\log Fc\rvert \geq 1.0$	$P < 0.05$ and $\lvert\log Fc\rvert \geq 1.0$
Total DEGs count	1725 genes	1008 genes
Upregulated DEGs count	926 genes	669 genes
Downregulated DEGs count	799 genes	339 genes

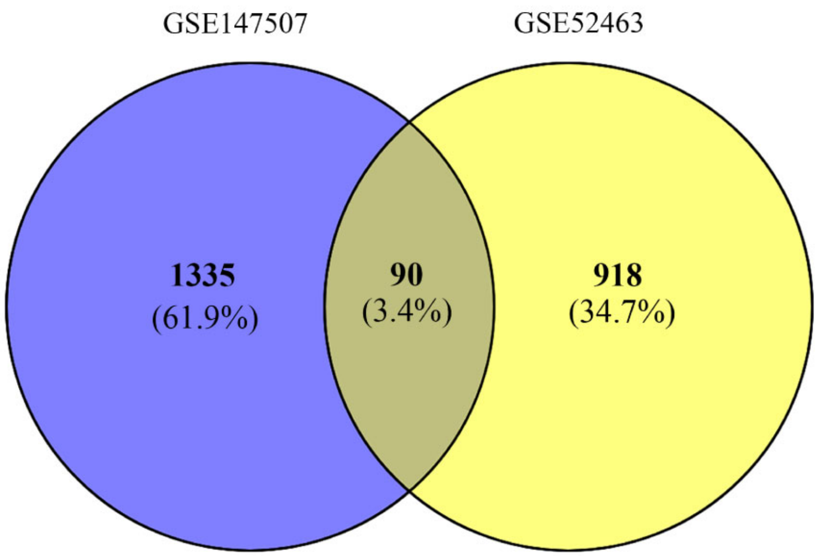

Figure 4: Common DEGs representation through a Venn diagram. There are 90 genes were found common from the 1635 DEGs of SARS-CoV-2 infection and 918 DEGs of IPF patients. The common DEGs were 3.4% among total 2553 DEGs.

Table 3: The combined score was used to identify biological process-related GO keywords

Group	GO ID	GO pathways	P-value	Genes
GO biological process	GO: 0006032	Chitin catabolic process	6.98E-03	CHIT1
	GO: 0090240	Positive regulation of histone H4 acetylation	6.98E-03	ARRB1
	GO: 0006030	Chitin metabolic process	6.98E-03	CHIT1
	GO: 0072677	Eosinophil migration	2.59E-04	CCL11; CCL18
	GO: 0048245	Eosinophil chemotaxis	2.59E-04	CCL11; CCL18
	GO: 0070098	Chemokine-mediated signaling pathway	1.83E-05	CXCL6; CCL11; CCL18
	GO: 0030593	Neutrophil chemotaxis	1.94E-05	CXCL6; CCL11; CCL18
	GO: 0002029	Desensitization of G-protein coupled receptor protein signal	7.97E-03	ARRB1
	GO: 0038114	Interleukin-21-mediated signaling pathway	7.97E-03	STAT4
	GO: 0098757	Cellular response to interleukin-21	7.97E-03	STAT4

Table 4: The combined score was used to identify GO keywords linked to molecular functions

Group	GO ID	GO pathways	P-value	Genes
GO molecular function	GO: 0019966	Interleukin-1 binding	5.98E-03	IL1RN
	GO: 0008009	Chemokine activity	1.26E-05	CXCL6; CCL11; CCL18
	GO: 0004568	Chitinase activity	6.98E-03	CHIT1
	GO: 0042379	Chemokine receptor binding	1.53E-05	CXCL6; CCL11; CCL18
	GO: 0005537	Mannose binding	1.09E-02	CD207
	GO: 0048020	CCR chemokine receptor bind	6.54E-04	CCL11; CCL18
	GO: 0005041	Low-density lipoprotein receptor	1.29E-02	TNFAIP6
	GO: 0005125	Cytokine activity	1.53E-05	CXCL6; IL1RN; CCL11;
	GO: 0005149	Interleukin-1 receptor binding	1.49E-02	IL1RN
	GO: 0005159	Binding of insulin-like growth factor receptors	1.49E-02	ARRB1

find the most impactful pathways of the shared DEGs between IPF and SARS-CoV-2. Tables 6, 7, 8, and 9 show the essential pathways discovered in the datasets. The graphical view of GO terms and pathways analysis are shown in Figs. 5 and 6.

Analysis of PPI network for the identification of hub-genes

The PPI network analysis is the most important element. This network has conducted hub-gene recognition, module analysis, and drug identification. In STRING, the specific DEGs have been provided as input. The analysis file was re-imported into the Cytoscape software for visualization. For the most frequent DEGs, a PPI network has been created. Finally, the PPIs network results connect to therapeutic compound suggestions, placing the PPIs analysis as the research's focus. Figure 7 shows the PPI network with 60 nodes and 308 edges. For SARS-CoV-2 and IPF, the PPI network was developed to discover hub-genes and medicinal compounds.

Identification of hub-genes for therapeutic solutions and module analysis

CytoHubba, a Cytoscape software plugin, was used to track the hub-genes from the PPIs network. The degree meaning of the

Table 5: The combined score was used to identify cellular component-related GO keywords

Group	GO ID	GO pathways	P-value	Genes
GO cellular component	GO: 1904724	Tertiary granule lumen	1.37E-03	CHIT1; TNFAIP6
	GO: 0030669	Clathrin-coated endocytic vesicle membrane	3.25E-02	CD207
	GO: 0045334	Clathrin-coated endocytic vesicle	4.88E-02	CD207
	GO: 0070820	Tertiary granule	1.15E-02	CHIT1; TNFAIP6
	GO: 0030659	Cytoplasmic vesicle membrane	5.27E-02	ARRB1
	GO: 0035580	Specific granule lumen	6.02E-02	CHIT1
	GO: 0031410	Cytoplasmic vesicle	1.92E-02	CD207; ARRB1
	GO: 0005769	Early endosome	2.04E-02	LAMP3; CD207
	GO: 0031901	Early endosome membrane	7.05E-02	CD207
	GO: 0030665	Clathrin-coated vesicle membrane	7.79E-02	CD207

Table 6: Pathway analysis results in identification through KEGG using the combined score

Database	Pathways	P-value	Gene
KEGG	IL-17 signaling pathway	1.05E-04	CXCL6; CCL11; MMP1
	Chemokine signaling pathway	3.39E-05	CXCL6; CCL11; ARRB1; CCL18
	Cytokine–cytokine receptor interaction	1.84E-04	CXCL6; IL1RN; CCL11; CCL18
	Rheumatoid arthritis	3.69E-03	CXCL6; MMP1
	Asthma	3.06E-02	CCL11
	Osteoclast differentiation	7.05E-03	FCGR2A; LILRA1
	Relaxin signaling pathway	7.37E-03	MMP1; ARRB1
	Bladder cancer	4.02E-02	MMP1
	Hedgehog signaling pathway	4.59E-02	ARRB1
	Amino sugar and nucleotide sugar metabolism	4.69E-02	CHIT1

Table 7: Pathway analysis results in identification through Wiki pathways using the combined score

Database	Pathways	P-value	Gene
Wiki Pathways	Thymic Stromal Lymphopoietin Signaling Pathway	1.00E-03	CCL11; STAT4
	Amplification and Expansion of Oncogenic Pathways as Metastatic Traits	1.69E-02	EPAS1
	Matrix Metalloproteinases	2.95E-02	MMP1
	Signal transduction through IL1R	3.25E-02	IL1RN
	Type 2 papillary renal cell carcinoma	3.34E-02	EPAS1
	Photodynamic therapy-induced NF-kB survival signaling	3.44E-02	MMP1
	Bladder Cancer	3.92E-02	MMP1
	Integrated Cancer Pathway	4.31E-02	MMP1
	Hedgehog Signaling Pathway	4.31E-02	ARRB1
	Hepatitis C and Hepatocellular Carcinoma	4.79E-02	MMP1

Table 8: Pathway analysis results in identification through Reactome using the combined score

Database	Pathways	P-value	Gene
Reactome	PTK6 Expression	4.99E-03	EPAS1
	Regulation of gene expression by Hypoxia-inducible Factor	9.96E-03	EPAS1
	Chemokine receptors bind chemokines	1.42E-03	CXCL6; CCL11
	Oxygen-dependent proline hydroxylation of Hypoxia-inducible Factor Alpha	1.78E-02	EPAS1
	Activation of SMO	1.78E-02	ARRB1
	Regulation of Insulin-like Growth Factor transport and uptake by Insulin-like Growth Factor Binding Proteins	2.08E-02	MMP1
	NOTCH2 Activation and Transmission of Signal to the Nucleus	2.08E-02	MDK
	Basigin interactions	2.47E-02	MMP1
	Regulation of hypoxia-inducible Factor by oxygen	2.56E-02	EPAS1
	Cellular response to hypoxia	2.57E-02	EPAS1

hub-genes, which represents the number of interactions between the genes in the PPI network, has been categorized. Hub-genes are the bulk of interconnected nodes in a PPI network. The topological analysis identified the top five genes (AKT1, IL1B, CCL5, MMP9, and ARRB1) classified as hub-genes based on their degree value. Table 10 shows the results of the topological analysis. These hub-genes could be exploited as biomarkers, leading to new therapeutic approaches for the studied diseases. The network has 50 nodes and 283 edges, and we utilized a degree-sorted circle structure to lay it out. The network of hub-genes is depicted in Fig. 8, with the top five hub-genes AKT1, IL1B, CCL5, MMP9, and ARRB1.

Table 9: Pathway analysis results in identification through BioCarta using the combined score

Database	Pathways	P-value	Gene
BioCarta	Beta-arrest ins in GPCR Desensitization Pathway	3.54E-04	CCL11; ARRB1
	NO2-dependent IL12 Pathway in NK cells Pathway	8.96E-03	STAT4
	Role of Beta-arrestins in the activation and targeting of MAP kinases Pathway	4.06E-04	CCL11; ARRB1
	G-Protein Signaling Through Tubby Proteins Pathway	9.95E-03	CCL11
	Roles of Beta-arrestins-dependent Recruitment of Src Kinases in GPCR Signaling Pathway	5.23E-04	CCL11; ARRB1
	Activation of PKC through G-protein coupled receptors Pathway	1.09E-02	CCL11
	Visual Signal Transduction Pathway	1.29E-02	ARRB1
	Attenuation of GPCR Signaling Pathway	1.29E-02	ARRB1
	IL12- and Stat4-dependent Signaling Pathway in Th1 Development	1.49E-02	STAT4
	Cystic fibrosis transmembrane conductance regulator (CFTR) and beta 2 adrenergic receptor (b2AR)	1.98E-02	CCL11

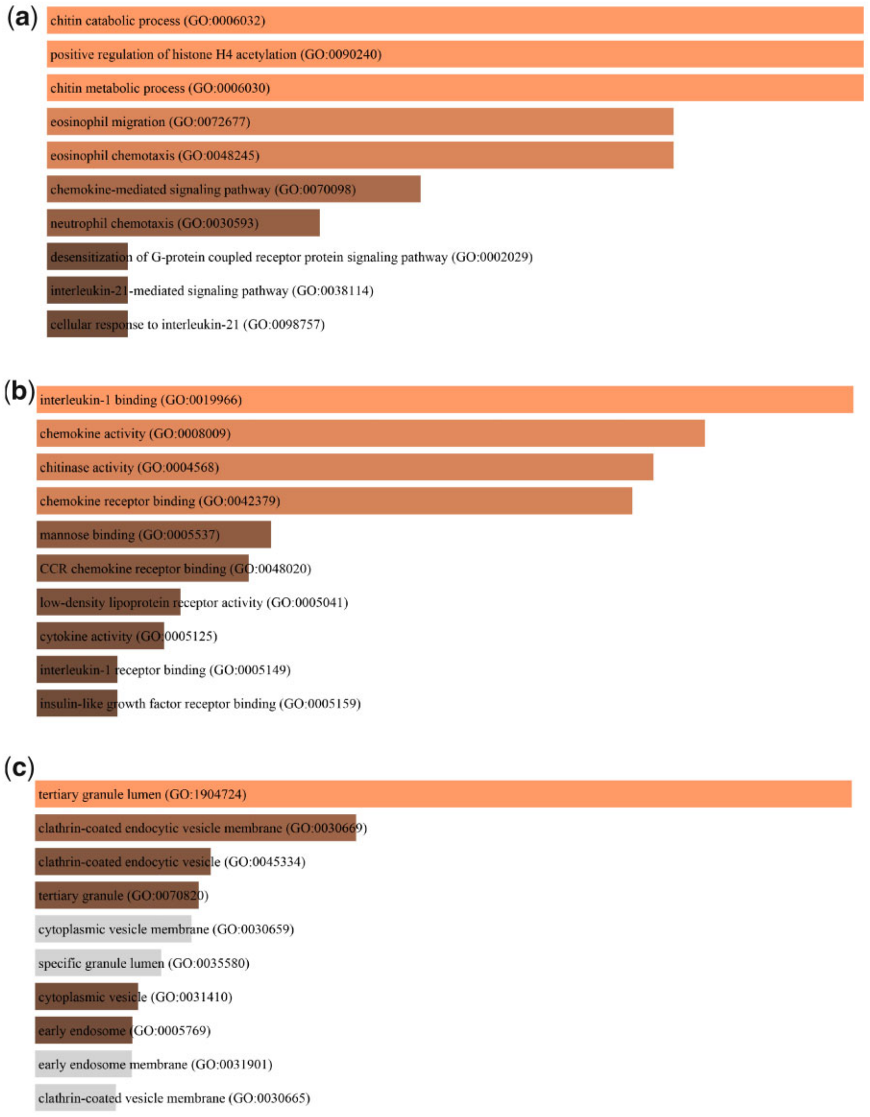

Figure 5: According to the combined score, (a) biological, (b) molecular function, and (c) cellular component relevant GO keywords were identified. The higher the enrichment score, the higher number of genes are involved in a certain ontology.

TF–gene analysis

The Network Analyst platform was used to investigate TF–gene interactions. The common DEGs were used to examine the TF–gene network. There are 190 nodes and 301 edges in the TF–gene network. Furthermore, the network contains 12 DEGs and 178 TF–genes, with 85 TF–genes regulating HSPB6, 68 TF–genes regulating EPAS1, and 37 TF–genes regulating FCGR2A according to their degree value. These 178 TF–genes are regulated by several

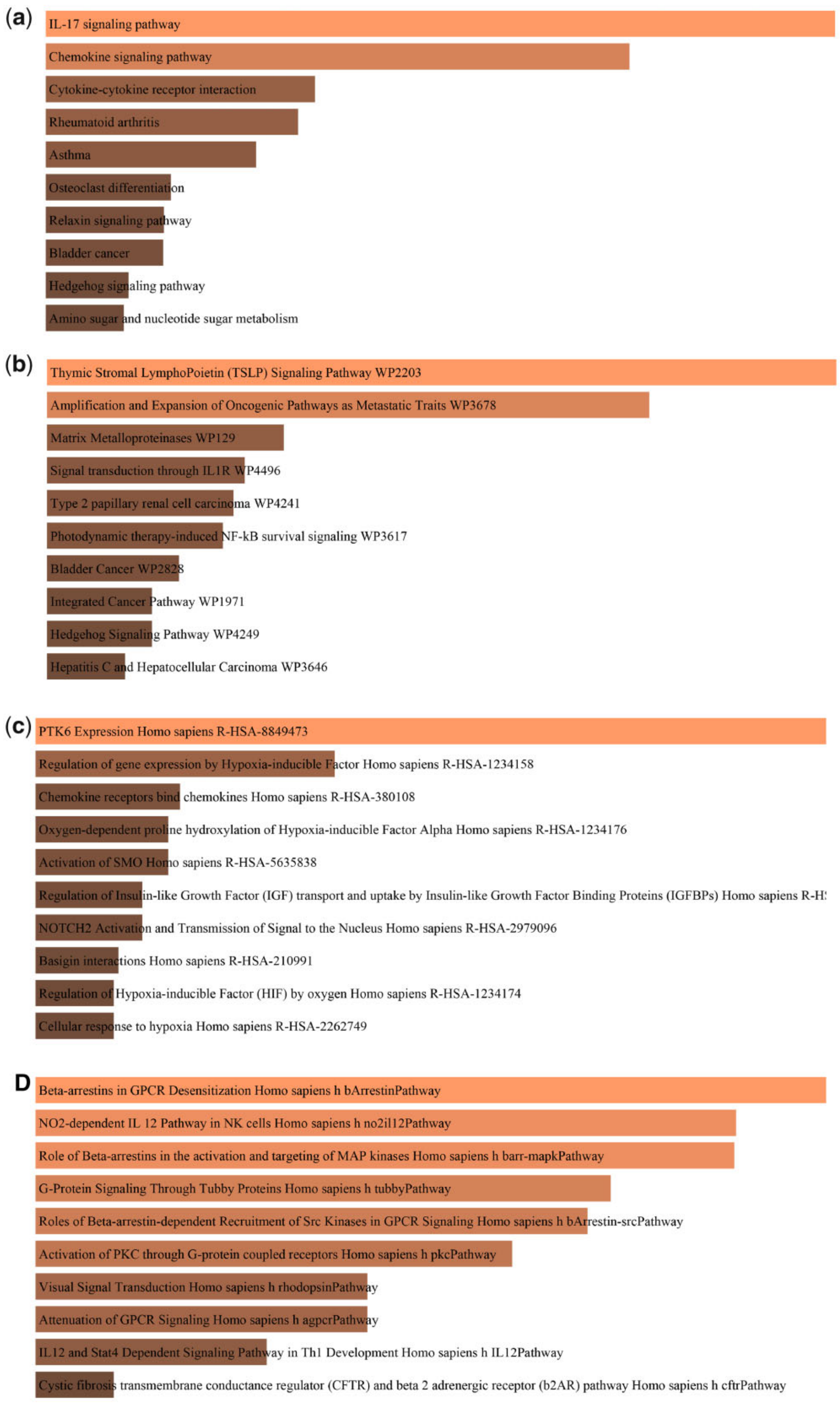

Figure 6: The pathway analysis results were identified using (a) KEGG, (b) Wiki Pathways, (c) Reactome, and (d) BioCarta. The results of the pathway terms were identified through the combined score.

common DEGs, indicating a high level of interaction between the TF–genes and common DEGs. The TF–gene network is shown in Fig. 9.

TF–miRNA analysis

The TF–miRNA coregulatory network was built using the Network Analyst tool. Analyzing this TF–miRNA coregulatory

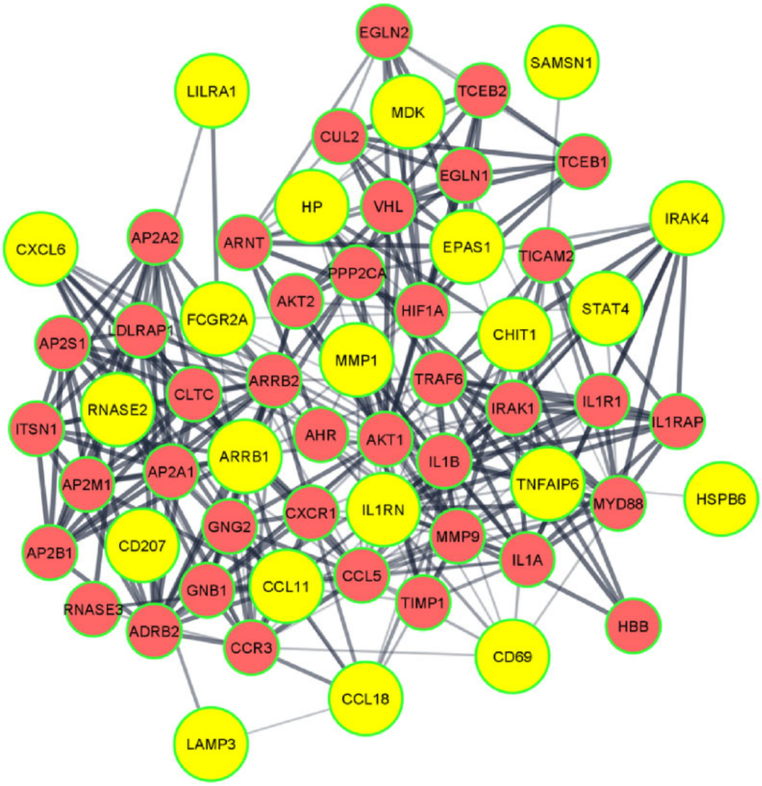

Figure 7: A network of PPIs discovered common DEGs in two illnesses (SARS-CoV-2 and IPF). The orange nodes denote common DEGs, whereas the edges denote the relationship between two genes. The network under investigation has 60 nodes and 308 edges.

Table 10: Exploration of topological results for the top five hub-genes

Hub gene	Degree	Stress	Close ness	Between ness	Bottle neck	Clustering coefficient	EcCentricity	Radiality
AKT1	27	3322	42.25000	637.30186	26	0.25356	0.25000	4.47458
IL1B	26	2172	42.33333	475.08574	03	0.34154	0.33333	4.52542
CCL5	22	1216	38.25000	238.70899	14	0.35931	0.25000	4.23729
MMP9	22	1808	39.16667	322.49125	07	0.35498	0.33333	4.33898
ARRB1	19	1630	37.55000	291.37776	06	0.43865	0.25000	4.25424

network revealed the connection of miRNAs and TFs with common DEGs. There are 191 nodes and 216 edges in this coregulatory network. DEGs interact with 87 miRNAs and 93 TF–genes, according to this study. Figure 10 shows the TF–miRNA coregulatory network.

Candidate drugs identification and validation

Drug compounds for common DEGs have been discovered using the Enrichr platform. Using the DSigDB database, we discovered 10 candidate medicinal compounds. The top 10 chemical compounds have been extracted based on the combined score of *P*-value and adjusted *P*-value. NICKEL SULFATE CTD 00001417, Clonidine HL60 UP, and THYMOLPHTHALEIN CTD 00006891 are the three-drug compounds most genes interact with, according to the data. These medicines are common pharmaceuticals for COVID-19 and IPF since these signature drugs have been discovered for common DEGs. Table 11 displays the most efficient medications for the most common DEGs from the DSigDB database.

Computationally predicted results usually need experimental verification, but it has more difficulty and limitations in practical implementation. Thus, similar to Zhang *et al.* [47], they found a novel validation process for suggested drug compounds based on the Receiver Operator Characteristic (ROC) curve. We tried to validate our suggested drug compounds using the ROC curve mechanism. Figure 11 shows the validation performance comparison between the top five suggested drug compounds using the ROC curve. We considered the top five suggested drug compounds, where Nickel Sulfate has a higher validation accuracy than the others, according to the ROC curve. Other suggested drug compounds, as shown in Fig. 11, were also validated using the same procedures, which is much more valuable to the medical community.

Discussion

COVID-19 is more common in people who have lung disease. This study contributes to the development of a bioinformatics and machine learning model to identify the Genetic Effect of SARS-CoV-2- and IPF-affected patients. Shortness of breath, cough, and chest pain are the most typical symptoms of these two diseases. About 1725 and 1008 DEGs were found in GSE147507 and GSE52463, respectively, using bioinformatics-related techniques. Common DEGs between the GSE147507 and GSE52463 datasets have been discovered for better coordination. There is a total of

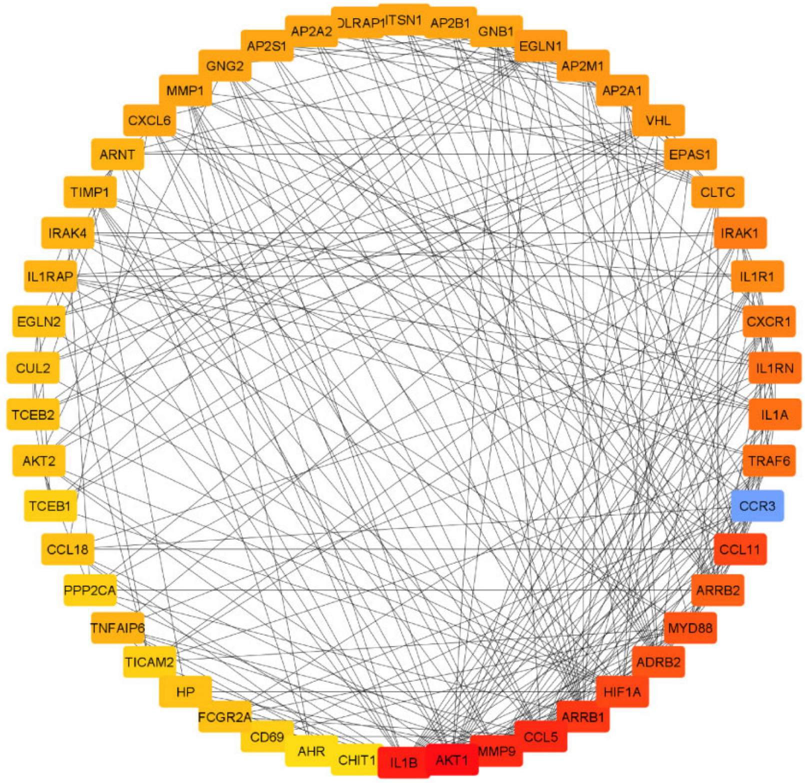

Figure 8: The PPIs network was used to find hub-genes. There are 50 nodes and 283 edges in the network. AKT1 and IL1B have degrees of 27 and 26, respectively, according to topological analysis. CCL5, MMP9, and ARRB1 had degrees of 22, 22, and 19, respectively.

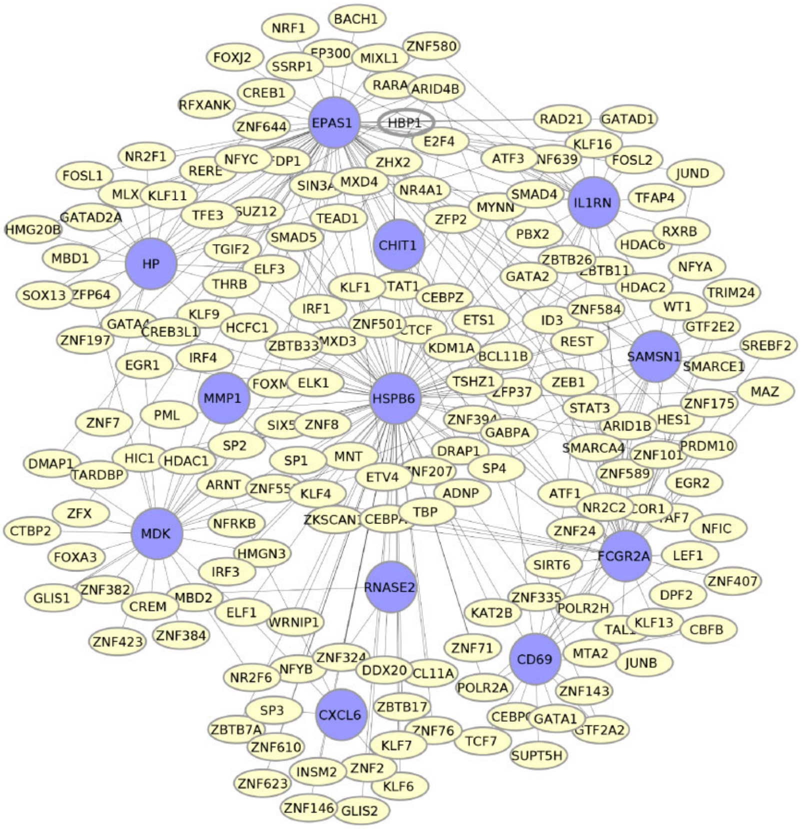

Figure 9: The interaction of TF–genes with common DEGs is represented via a network. The common genes are shown by the highlighted yellow color node, while TF–genes are represented by the other nodes. There are 190 nodes and 301 edges in the network.

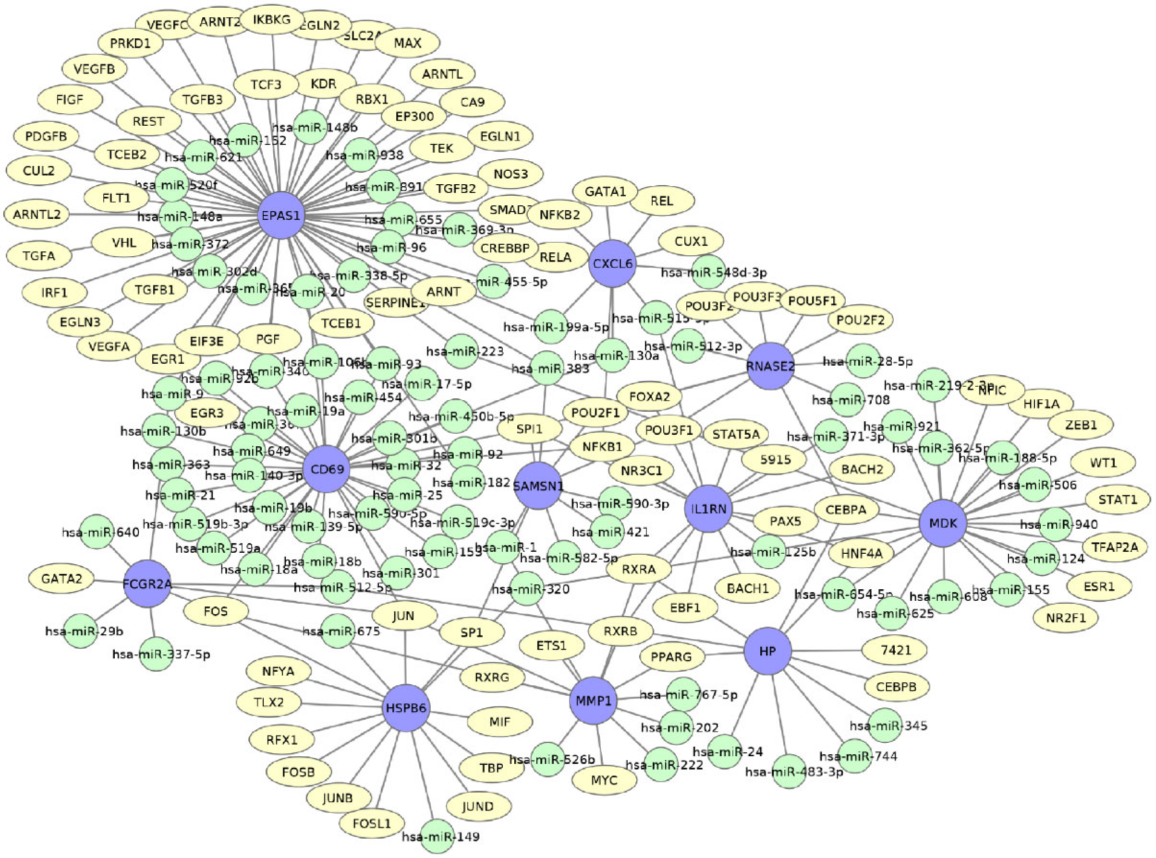

Figure 10: There are 93 TF–genes, 87 miRNAs, and 11 DEGs in the TF–miRNA network. There are 191 nodes and 216 edges in the network. DEGs are represented by blue nodes, while miRNA is represented by green nodes, and TF–genes are represented by other nodes.

Table 11: The top 10 drug compounds suggested for common DEGs

Name of the drugs	P-value	Adjusted P-value	Name of the genes
Nickel Sulfate CTD 00001417	1.37E-12	8.81E-10	CXCL6; IL1RN; CCL11; TNFAIP6; EPAS1; MMP1; LAMP3; CD207; STAT4; CD69; SAMSN1
Clonidine HL60 UP	1.04E-06	3.36E-04	IL1RN; FCGR2A; RNASE2; SAMSN1
Thymolphthalein CTD 00006891	3.80E-04	1.01E-02	EPAS1; ARRB1
Peptidoglycan CTD 00006490	4.34E-04	1.07E-02	TNFAIP6; MMP1
Lithocholic acid HL60 UP	4.63E-04	1.10E-02	CD69; SAMSN1
Beclomethasone CTD 00005468	3.93E-05	3.21E-03	IL1RN; CCL11; RNASE2
Salmeterol CTD 00002421	4.92E-04	1.13E-02	CCL11; RNASE2
Mephentermine HL60 UP	4.48E-05	3.21E-03	IL1RN; EPAS1; CD69
Colchicine HL60 UP	8.09E-06	1.04E-03	IL1RN; FCGR2A; EPAS1; SAMSN1
Bromocriptine HL60 UP	6.94E-05	4.07E-03	FCGR2A; TNFAIP6; SAMSN1

90 DEGs that have been identified. Twenty common DEGs were chosen for further study from 90 common DEGs based on the P-value (MDK, HP, HSPB6, CHIT1, TNFAIP6, EPAS1, MMP1, CCL18, CXCL6, CCL11, IL1RN, LAMP3, CD207, ARRB1, RNASE2, LILRA1, FCGR2A, STAT4, CD69, and SAMSN1). The analysis of GO, KEGG, Wiki Pathways, Reactome, BioCarta pathway analysis, PPIs, TF–gene, TF–miRNA coregulatory network, and candidate drug detection has been continued in the research project.

DEGs that have been identified as common have been used to find GO words. GO keywords were identified using the combined score. Biological process, molecular function analysis, and cellular component analysis are the three categories of GO analysis [48]. KEGG, Wiki Pathways, Reactome, and BioCarta were used to identify pathway analysis results. For the most prevalent DEGs, the KEGG pathway has been determined. KEGG is a database that aids researchers in understanding the high-level functions and utility of biological systems. Because hub-gene recognition, module analysis, and drug identification are all strongly dependent on the PPI network, it is the significant part of the research. Common DEGs were also subjected to PPI analysis. The identification of hub-genes in the PPI network was studied. The five genes that have been highlighted are AKT1, IL1B, CCL5,

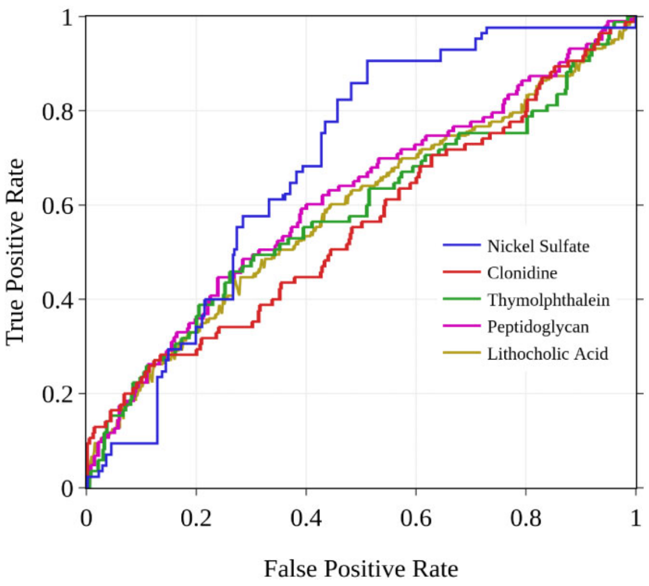

Figure 11: Performance comparison of the top five suggested drug compounds based on the ROC curve. We considered the top five suggested drug compounds, where Nickel Sulfate has a higher validation accuracy than the others, according to the ROC curve.

MMP9, and ARRB1. These five genes are classified as hub-genes based on their degree value. The aim of concentrating on a small area is to suggest a more effective medication component.

The interaction of TF–genes and miRNAs was investigated to identify transcriptional and post-transcriptional regulators of common DEGs. The specific DEGs have been used to investigate TF–gene interactions. TF–genes act as regulators of gene expression, which can contribute to cancer cell formation. About 85 TF–genes regulate HSPB6, 68 TF–genes regulate EPAS1, and 37 TF–genes regulate FCGR2A according to their degree value in the network, with 12 DEGs and 178 TF–genes. The TF–miRNA coregulatory network depicts the interactions between miRNAs and TF–genes tested for their ability to influence common DEGs. There were 87 miRNAs and 93 TF–genes discovered. Several studies have found evidence of altered miRNA expression in IPF samples, and members of the miR-200 family play a significant role in IPF sample management [49]. Taz *et al.* [50] investigated only 69 samples, whereas we analyzed 110 SARS-CoV-2 samples. As a result, this research will ideally integrate COVID-19 with IPF risk factor treatment. Chemical testing can be used to verify the drugs' efficacy.

In addition, we thoroughly discussed the application areas of our research for the scientific society. First of all, researchers can use the same approach to investigate the impact of SARS-CoV-2 on other diseases. Also, if a new virus appears, our research will serve as a useful starting point for further investigation. Furthermore, our research suggests several viable drugs, so scientists will be able to find a treatment for SARS-CoV-2 with more research. Finally, our research is an example of a virus's genetic relationship with a certain type of patient. So, researchers can use this methodology to figure out the genetic relationships between different viruses and patients.

Conclusions

COVID-19 infections have been associated with a high-risk factor for IPF patients. Shortness of breath, cough, and chest pain are the most typical symptoms of these two diseases. We used machine learning and bioinformatics analysis to summarize the relationships between these two disease genes as part of our research. We analyzed DEGs from two selected datasets, analyzed the results using shared gene identification, and discovered SARS-CoV-2- and IPF-affected lung-cell infection responses. As a consequence, we discovered 90 genes that are linked across these datasets. These interconnected genes built the PPI network, which identified the five most important hub-genes. In addition, we looked at SARS-CoV-2 and IPF to see if they might predict the outcomes of identifying infections of other diseases. The therapeutic goals are logically presented because they are executed from the discovery of hub-genes and could work as an effective precursor to meanwhile licensed medications. We believe that the biomarkers, pathways, and molecular markers we discovered will be valuable in developing pharmacological therapies.

Declarations

Ethical Approval

Not applicable (there is no human-related data. So, ethical approval is not taken from the external body of the committee).

Consent to Participate

Not applicable (there is no human-related data. So, consent is not necessary to take from the participant).

Consent to Publish

Not applicable (there is no human-related data. So, consent to publish is not necessary to take from the participant).

Funding

This work was supported by Researchers Supporting Project number (RSP-2021/100), King Saud University, Riyadh, Saudi Arabia. This work was supported in part by funding from the Natural Sciences and Engineering Research Council of Canada (NSERC).

Acknowledgement

The authors are grateful for support from Researchers Supporting Project number (RSP-2021/100), King Saud University, Riyadh, Saudi Arabia.

Author contribution

Conceptualization was done by K.A., M.A.M.; Data curation, Formal analysis, Investigation, and Methodology by Sk. T.M., M.R., I.H., T.M.T., R.A.L.; Funding acquisition by S.M.I., F.M.B.; Project administration by K.A.; F.M.B., M.A.M.; Resources, Software by K.A., M.A.M.; Supervision by Sk. T.M., K.A.; Validation by Sk. T.M., K.A.; F.M.B.; Visualization and Writing of the original draft by Sk. T.M., M.R., I.H., T.M.T., R.A.L., K.A., F.M.B., M.A.M.; Writing: review editing by Sk. T.M., M.R., I.H., T.M.T., R.A.L., S.M.I., K.A., F.M.B., M.A.M.

Conflict of interest statement. None declared.

References

1. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat Microbiol* 2020;**5**:536–44.
2. National Foundation for Infectious Diseases. *Coronaviruses*. https://www.nfid.org/infectious-diseases/coronaviruses (24 March 2021, date last accessed).
3. Zou L, Ruan F, Huang M *et al.* SARS-CoV-2 viral load in upper respiratory specimens of infected patients. *N Engl J Med* 2020;**382**: 1177–9.
4. Wikipedia. *COVID-19 Pandemic*. https://en.wikipedia.org/wiki/COVID-19_pandemic (19 May 2021, date last accessed).
5. World Health Organization. *Transmission of SARS-CoV-2: Implications for Infection Prevention Precautions*. https://bit.ly/3wpdk4p (9 July 2020, 25 February 2021, date last accessed).
6. World Health Organization. WHO Coronavirus (COVID-19) *Dashboard*. https://covid19.who.int (26 May 2021, date last accessed).
7. Raghu G, Collard HR, Egan JJ et al.; ATS/ERS/JRS/ALAT Committee on Idiopathic Pulmonary Fibrosis. An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. *Am J Respir Crit Care Med* 2011;**183**:788–824.
8. Medline Plus. *Idiopathic Pulmonary Fibrosis*. https://medlineplus.gov/genetics/condition/idiopathic-pulmonary-fibrosis (18 August 2020, 2 March 2021, date last accessed).
9. Breathe the Lung Association. *Idiopathic Pulmonary Fibrosis Causes*. https://www.lung.ca/lung-health/lung-disease/idiopathic-pulmonary-fibrosis/causes (1 August 2014, 5 March 2021, date last accessed).
10. Mayo Clinic. *Pulmonary Fibrosis*. https://www.mayoclinic.org/diseases-conditions/pulmonary-fibrosis/symptoms-causes/syc-20353690 (6 March 2018, 10 March 2021, date last accessed).
11. Lindell KO, Olshansky E, Song MK *et al.* Impact of a disease-management program on symptom burden and health-related quality of life in patients with idiopathic pulmonary fibrosis and their care partners. *Heart Lung* 2010;**39**:304–13.
12. Brake SJ, Barnsley K, Lu W *et al.* Smoking upregulates angiotensin-converting enzyme-2 receptor: a potential adhesion site for novel coronavirus SARS-CoV-2 (Covid-19). *J Clin Med* 2020;**9**:841.
13. Sohal SS, Hansbro PM, Shukla SD *et al.* Potential mechanisms of microbial pathogens in idiopathic interstitial lung disease. *Chest* 2017;**152**:899–900.
14. Barrett T, Wilhite SE, Ledoux P *et al.* NCBI GEO: archive for functional genomics data sets – update. *Nucleic Acids Res* 2013;**41**: D991–D995.
15. National Center for Biotechnology Information. *GEO Dataset*. https://www. ncbi.nlm.nih.gov/geo (15 March 2021, date last accessed).
16. Clough E, Barrett T. The gene expression omnibus database. *Methods Mol Biol* 2016;**1418**:93–110.
17. Blanco-Melo D, Nilsson-Payant BE, Liu WC *et al.* Imbalanced host response to SARS-CoV-2 drives development of COVID-19. *Cell* 2020;**181**:1036–45.
18. Nance T, Smith KS, Anaya V *et al.* Transcriptome analysis reveals differential splicing events in IPF lung tissue. *PLoS One* 2014;**9**:e92111.
19. Anjum A, Jaggi S, Varghese E *et al.* Identification of differentially expressed genes in RNA-seq data of Arabidopsis thaliana: a compound distribution approach. *J Comput Biol* 2016;**23**:239–47.
20. Bardou P, Mariette J, Escudie F *et al.* Jvenn: an interactive Venn diagram viewer. *BMC Bioinformatics* 2014;**15**:293.
21. Subramanian A, Tamayo P, Mootha VK *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005; **102**:15545–15550.
22. Doms A, Schroeder M. GoPubMed: exploring PubMed with the gene ontology. *Nucleic Acids Res* 2005;**33**:783–786.
23. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**:27–30.
24. Slenter DN, Kutmon M, Hanspers K *et al.* WikiPathways: a multi-faceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res* 2018;**46**:661–7.
25. Fabregat A, Jupe S, Matthews L *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res* 2018;**46**:649–55.
26. Kuleshov MV, Jones MR, Rouillard AD *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* 2016;**44**:W90–7.
27. Enrichr. *GO Terms and All Pathways Analysis*. https://maayanlab.cloud/Enrichr (20 March 2021, date last accessed).
28. Ewing RM, Chu P, Elisma F *et al.* Large-scale mapping of human protein–protein interactions by mass spectrometry. *Mol Syst Biol* 2007;**3**:89.
29. Ben-Hur A, Noble WS. Kernel methods for predicting protein–protein interactions. *Bioinformatics* 2005;**21**:i38–46.
30. STRING. *Multiple Proteins by Names/Identifiers*. https://string-db.org (4 April 2021, date last accessed).
31. Szklarczyk D, Franceschini A *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* 2010;**39**:561–8.
32. Cytoscape C. *Network Data Integration, Analysis, and Visualization in a Box*. https://cytoscape.org (10 April 2021, date last accessed).
33. Shannon P, Markiel A, Ozier O *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**:2498–504.
34. Cytoscape App Store. CytoHubba. https://apps.cytoscape.org/apps/cytohubba (11 April 2021, date last accessed).
35. Chin CH, Chen SH, Wu HH *et al.* CytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Syst Biol* 2014;**8**:S11.
36. Cytoscape. *Molecular Complex Detection (MCODE)*. https://apps.cytoscape. org/apps/mcode (11 April 2021, date last accessed).
37. Bader GD, Hogue CWV. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 2003;13.
38. Caramori G, Ito K, Adcock IM. Transcription factors in asthma and COPD. *IDrugs* 2004;**7**:764–70.
39. Ye Z, Wang F, Yan F *et al.* Bioinformatic identification of candidate biomarkers and related transcription factors in nasopharyngeal carcinoma. *World J Surg Oncol* 2019;**17**: 60.
40. Network Analyst. A Comprehensive *Network Visual Analytics Platform for Gene Expression Analysis*. https://www.networkanalyst.ca (25 April 2021, date last accessed).
41. Zhou G, Soufan O, Ewald J *et al.* Network Analyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. *Nucleic Acids Research* 2019;**47**: W234–241.
42. ENCODE. *Encyclopedia of DNA Elements*. https://www.encodeproject.org (26 April 2021, date last accessed).
43. Liu ZP, Wu C, Miao H *et al.* RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database* 2015;**2015**:bav095.

44. Xia J, Gill EE, Hancock REW. Network analyst for statistical, visual and network-based meta-analysis of gene expression data. *Nat Protoc* 2015;**10**:823–44.
45. Chen EY, Tan CM, Kou Y *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* 2013;**14**:128.
46. Enrichr. *Pathway's Analysis*. https://maayanlab.cloud/Enrichr (27 April 2021, date last accessed).
47. Zhang F, Wang M, Xi J *et al.* A novel heterogeneous network-based method for drug response prediction in cancer cell lines. *Sci Rep* 2018;**8**:3355.
48. Yang D, Guo X, Chen Y *et al.* Leukocyte aggregation in vitro as a cause of pseudoleukopenia. *Laboratory Medicine* 2008;**39**: 89–91.
49. Wang L, Huang W, Zhang L *et al.* Molecular pathogenesis involved in human idiopathic pulmonary fibrosis based on an integrated microRNA-mRNA interaction network. *Mol Med Rep* 2018;**18**:4365–73.
50. Taz TA, Ahmed K, Paul BK *et al.* Network-based identification genetic effect of SARS-CoV-2 infections to Idiopathic pulmonary fibrosis (IPF) patients. *Brief Bioinform* 2021;**22**: 1254–13.