



Daffodil
International
University

IDENTIFICATION OF BIOMARKERS AND PATHWAYS
FOR THE SARS-COV-2 INFECTIONS THAT MAKE
COMPLEXITIES IN PULMONARY ARTERIAL
HYPERTENSION PATIENTS

By

TASNIMUL ALAM TAZ

171-35-241

A thesis submitted in partial fulfillment of the requirement for the degree
of Bachelor of Science in Software Engineering.

Department of Software Engineering

DAFFODIL INTERNATIONAL UNIVERSITY

Semester Summer – Year-2021

**IM: Dr. Imran Mahmud; SMR: S A M Matiur Rahman; BKP: Bikash Kumar Paul; RZ:
Raihana Zannat; NH: Nayeem Hasan; SI: Mr. Shariful Islam; SFR: SK. Fazlee Rabby; MA:
Marzia Ahmed; RM: Md. Rajib Mia.**

APPROVAL

This Thesis titled “**Identification of biomarkers and pathways for the SARS-CoV-2 infections that make complexities in pulmonary arterial hypertension patients**”, submitted by **Tasnimul Alam Taz, ID: 171-35-241** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc in Software Engineering and approved as to its style and contents.

3IK081A
12.06.2021

Bikash Kumar Paul

Lecturer (Senior Scale)

Department of Software Engineering

Faculty of Science and Information Technology

Daffodil International University

THESIS DECLARATION

I, Tasnimul Alam Taz do hereby declare that this report has been done by me under the supervision of Bikash Kumar Paul, Senior Lecturer, Dept. of Software Engineering, Daffodil International University. We also declare that this report or any portion of this report has not been submitted elsewhere for award of any degree.

Supervised By,
Bikash Kumar Paul
Lecturer (Senior Scale)
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Submitted By,
Tasnimul Alam Taz
ID: 171-35-241
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

ACKNOWLEDGEMENT

First, I am expressing my gratitude to the almighty Allah for giving me the ability to complete this thesis work. I would like to express our sincere gratitude to my honorable supervisor, Bikash Kumar Paul, Senior Lecturer, Department of Software Engineering. This thesis would not have been completed without his support and guidance. His constant encouragement gave me the confidence to carry out my work. I would also like to give special gratitude to one of my favorite teachers Kawsar Ahmed, Assistant Professor, Dept. of Information and Communication Technology, Mawlana Bhashani Science and Technology University. His proper direction and guidance help me to prepare this thesis work without any difficulty.

I express my heartiest gratitude towards the entire department of Software Engineering at Daffodil International University for providing good education and knowledge.

I also express my gratitude to all my teacher's SAM Matiur Rahman, Associate Professor; Dr. Imran Mahmud, Head, Dept. of Software Engineering. I would like to convey my sincere gratitude to Md. Shariful Islam, Lecturer, Dept. of Software Engineering. The knowledge that I have learned from the classes in my degree of bachelor's in software engineering level were essential for this thesis. In course of conducting the study necessary information were collected through books, journals, electronic media and other secondary sources. I also want to thank my friends for providing me support and encouragement. Their optimism and encouragement have allowed to overcome any obstacle at any phase.

ABSTRACT

This study aimed to identify significant gene expression profiles of the human lung epithelial cells caused by SARS-CoV-2 infections. We performed a comparative genomic analysis to show genomic observations between SARS-CoV and SARS-CoV-2. A phylogenetic tree has been carried for genomic analysis that confirmed the genomic variance between SARS-CoV and SARS-CoV-2. Transcriptomic analyses have been performed for SARS-CoV-2 infection responses and pulmonary arterial hypertension (PAH) patients' lungs as a number of patients have been identified who faced PAH after being diagnosed with COVID-19. Gene expression profiling showed significant expression levels for SARS-CoV-2 infection responses to human lung epithelial cells and PAH lungs as well. Differentially expressed genes (DEGs) identification and integration showed concordant genes (SAA2, S100A9, S100A8, SAA1, S100A12, and EDN1) for both SARS-CoV-2 and PAH samples including S100A9 and S100A8 genes that showed significant interaction in the Protein-protein interactions (PPIs) network. Extensive analyses of Gene ontology and signaling pathways identification provided evidence of inflammatory responses regarding SARS-CoV-2 infections. The altered signaling and ontology pathways that have emerged from this research may influence the development of effective drugs especially for the people with pre-existing conditions. Identification of regulatory bio-molecules revealed the presence of active promoter gene of SARS-CoV-2 in TF-miRNA coregulatory network. Predictive drug analyses provided concordant drug compounds that are associated with SARS-CoV-2 infection responses and PAH lung samples and these compounds showed significant immune response against the RNA viruses like SARS-CoV-2, which is beneficial in therapeutic development in the COVID-19 pandemic.

TABLE OF CONTENT

Contents	Page
APPROVAL	ii
THESIS DECLARATION	iii
ACKNOWLEDGEMENT	iv
ABSTRACT	v
TABLE OF CONTENT	vi
LIST OF TABLES	viii
LIST OF FIGURES	ix
LIST OF ABBREVIATION	x
CHAPTER 1: INTRODUCTION	1
1.1 Background	1
1.2 Motivation of the Research	2
1.3 Problem Statement:	3
1.4 Research Question	3
1.5 Research Objectives	3
1.6 Research Scope	3
1.7 Thesis Organization	3
CHAPTER 2: LITERATURE REVIEW	5
2.1 Definition of The Thesis Topic or Area:	5
2.2 The Direction of Research	5
2.3 Theoretical Framework	5
2.4 Resources	5
CHAPTER 3: METHODOLOGY	6
3.1 Comprehensive genomic-level phylogenetic study	6
3.2 Details information of the datasets	6
3.3 Data filtering and retrieval of DEGs, and identification of common DEGs between SARS-CoV-2 and PAH	7
3.4 Gene ontology and cell informative pathway analysis	7
3.5 Designing of Protein-protein Interactions (PPIs) network	8
3.6 Establishment of the topological algorithm on the PPIs network and detection of hub nodes	8
3.7 Analysis of TF-miRNA coregulatory network	8
3.8 Therapeutic drug compounds prediction	8
CHAPTER 4: RESULT AND DISCUSSION	9

4.1 Genomic and phylogram differences between SARS-CoV and SARS-CoV-2.....	10
4.2 Gene expression analysis of PAH patients and SARS-CoV-2 infected human lung epithelial and associative cells.....	10
4.3 Common DEGs identifications for further molecular analysis and ensure the efficiency of predictive drugs.....	12
4.4 Gene Ontology and pathway analysis based on the similar DEGs.....	13
4.5 PPIs network construction to perceive hub nodes.....	19
4.6 Hub nodes identification based on the topological analyses and module detection from the PPIs network.....	20
4.7 Analysis of TF-miRNA coregulatory network.....	22
4.8 Predictive drug compounds.....	23
CHAPTER 5: CONCLUSION.....	24
5.1 Findings.....	24
5.2 Recommendation for Future works.....	24
REFERENCES.....	25

LIST OF TABLES

TABLE 1: THE ASSOCIATION OF CONCORDANT GENES IN GO TERMS AND GO PATHWAYS AND THE PROPORTIONAL P-VALUES.....	13
TABLE 2: THE ASSOCIATION OF CONCORDANT GENES IN KEGG, WIKIPATHWAYS, REACTOME AND BIOCARTA DATABASES AND THE PROPORTIONAL P-VALUES.....	15
TABLE 3: EXPLORATION OF TOPOLOGICAL RESULTS FOR TOP 3 HUB GENES.....	22
TABLE 4: PREDICTIVE DRUG COMPOUNDS OF CONCORDANT GENES.....	23

LIST OF FIGURES

- FIGURE 1:** FTHE WORKFLOW OF CURRENT ANALYSIS. GENOMIC DIFFERENCES BETWEEN SARS-COV AND SARS-COV-2 ARE VISUALIZED THROUGH A PHYLOGENETIC ANALYSIS. TWO DATASETS GSE147507 AND GSE117261 ARE COLLECTED ACCORDING TO SARS-COV-2 INFECTION IN HUMAN LUNG EPITHELIAL CELLS AND PAH LUNG, RESPECTIVELY. DIFFERENTIALLY EXPRESSED GENES (DEGS) WERE IDENTIFIED USING R PROGRAMMING LANGUAGE AND SIMILAR DEGS WERE IDENTIFIED FROM TOTAL DEGS OF BOTH THE DATASETS. CORRESPONDING SIMILAR DEGS WERE USED TO PERFORM TRANSCRIPTOMIC ANALYSES. THE GENE EXPRESSION PROFILING WAS PERFORMED FOR BOTH THE DATASETS AND GENE ONTOLOGY (GO) TERMS, CELL INFORMATIVE PATHWAYS, PPIS NETWORK, HUB GENE IDENTIFICATION AND TF-MIRNA BASED ANALYSES WERE PERFORMED. ACCORDING TO THE CORRESPONDING SIMILAR DEGS, DRUG COMPOUNDS WERE PREDICTED 6
- FIGURE 2:** PHYLOGRAM OF SARS-COV AND SARS-COV-2 THAT PROVIDE GENOMIC DIFFERENCES BETWEEN HUMAN CORONAVIRUSES OF 2003-2018 (SARS-COV) AND 2019-2020 (SARS-COV-2). TWO COLORS ARE IMPLEMENTED TO DIFFERENTIATE SARS-COV (PURPLE) AND SARS-COV-2 (GREEN)..... 10
- FIGURE 3:** GENE EXPRESSION PROFILING OF SARS-COV-2 INFECTION IN HUMAN LUNG EPITHELIAL CELLS FOR THE TOP 20 GENES AND SELECTED 24 SAMPLES FROM THE GSE147507 DATASET. 11
- FIGURE 4:** (A) GENE EXPRESSION VISUALIZATION OF HEALTHY CONTROLS (GSM3290083, GSM3290086 AND GSM3290085) AND PAH SAMPLES. (B) VOLCANO PLOT SHOWS THE REGULATION OF GENES (UP-REGULATED AND DOWN-REGULATED) FOR GSE117261..... 11
- FIGURE 5:** (A) CONCORDANT GENE IDENTIFICATION BETWEEN GSE147507 AND GSE117261 DATASET THAT PROVIDE EVIDENCE OF 6 COMMON DIFFERENTIALLY EXPRESSED GENES IN BETWEEN 108 GENES OF GSE147507 (COVID-19) AND 59 GENES OF GSE117261 (PAH) DATASET. (B) HEAT MAP ACCORDING TO THE LOG FOLD CHANGES FOR THE SHARED COMMON DEGS BETWEEN COVID-19 DATASET AND PAH DATASET. 12
- FIGURE 6:** HEAT MAP FOR THE IDENTIFICATION OF HIGHLY RISK PRONE NATURE OF S100A9 AND S100A8 GENES. (B) RISK GROUP COMPARISONS BETWEEN THE SHARED COMMON GENES OF SARS-COV-2 AND PAH. 13
- FIGURE 7(A):** GO TERMS REGARDING BIOLOGICAL PROCESS, MOLECULAR FUNCTION AND CELLULAR COMPONENT ACCORDING TO THE ASSOCIATIVE P-VALUES. 19
- FIGURE 7(B):** CELL INFORMATIVE PATHWAYS (KEGG, BIOCARTA, REACTOME AND WIKIPATHWAYS) ANALYSIS RESULT REGARDING ASSOCIATIVE P-VALUES. 19
- FIGURE 8:** PPIS NETWORK FOR IDENTIFIED COMMON DEGS THAT REFERS TO SARS-COV-2 INFECTIONS IN HUMAN LUNG AND PAH LUNG. THE COMMON GENES ARE HIGHLIGHTED WITH PURPLE NODE (SAA2, S100A9, S100A8, SAA1 AND S100A12). THE NETWORK CONSISTS OF 125 NODES AND 136 EDGES. 20
- FIGURE 9:** HUB GENE DETECTION FROM THE SIMILAR DEGS BASED ON THE PPIS NETWORK. THE HIGHLIGHTED NODES S100A9 (RED), S100A8 (ORANGE) AND SAA1 (YELLOW) ARE REGARDED AS HIGHLY INTERCONNECTED NODES CONSIDERED AS HUB NODES. THE NETWORK IS MADE UP OF 124 NODES AND 135 EDGES. 21
- FIGURE 10:** HIGHLY INTERCONNECTED REGIONS (MODULE) IDENTIFICATION NETWORK THAT CONSISTS OF 13 NODES AND 13 EDGES. THE HUB GENES S100A9 (ORANGE) AND S100A8 (ORANGE) ARE VISUALIZED IN THE CORRESPONDING MODULE NETWORK 21
- FIGURE 11:** HIGHLY TF-MIRNA COREGULATORY NETWORK VISUALIZATION. THE NETWORK INCLUDES 69 NODES AND 77 EDGES. ACCORDING TO THE NETWORK, THERE EXIST 35 TF GENES (BLUE) AND 28 ARE MIRNAS (RED) AND THEY ARE INTERACTED WITH 6 COMMON DEGS (GREEN). 22

LIST OF ABBREVIATION

NCBI = National Center for Biotechnology Information;

COVID-19= Coronavirus disease (COVID-19);

SARS-CoV-2= severe acute respiratory syndrome coronavirus 2;

SARS-CoV= severe acute respiratory syndrome coronavirus;

PAH= Pulmonary Arterial Hypertension;

GEO= Gene Expression Omnibus;

SAA2= Serum Amyloid A2;

SAA1= Serum Amyloid A1;

S100A8= S100 Calcium Binding Protein A8;

S100A9= S100 Calcium Binding Protein A9;

S100A2= S100 Calcium Binding Protein A12;

EDN1= Endothelin 1;

PPIs= Protein-protein interactions.

1.1 Background:

COVID-19 is caused by a virus called SARS-CoV-2, which belongs to the Coronaviridae family (Gorbalenya, 2020). The widespread behavior of this virus has immensely influenced the death rate and proved it as the most internecine global epidemic of the 21st century. Angiotensin-converting enzyme 2 (ACE2), which is used by SARS-CoV-2 forms an entrance in host human cells and binds with human ACE2 that eventually leads to the intense spread of this lethal virus among human (Walls, 2020). Spike protein is considered to be a potential therapeutic target against SARS-CoV-2 (Chi X, 2020).

The first severe case of COVID-19 that led to death eventually was indicated on January 11, 2020 (Al-Awadhi, 2020). As of September 10, 2020, the number of confirmed COVID-19 cases all over the world is 27,688,740 including 899,315 deaths (<https://covid19.who.int/>). A large proportion of the total patients of COVID-19 are male (54.3%), where the mortality rate of the elderly patients is higher (15%), compare to younger patients (Nain, 2020). Due to the rapid spread of COVID-19, the pace of vaccine production has not been able to keep pace with demand. The transference of lethal SARS-CoV-2 from one person to another mostly occurs through respiratory droplet transmission (Mackenzie, 2020). The prevalence of SARS-CoV-2 is increasing because pre-symptomatic infectious diseases are difficult to detect (Ferretti, 2020).

Pulmonary arterial hypertension (PAH) is considered to be a progressive disorder and causes right heart affliction and the arteries of human lungs get affected by PAH as well (Schermuly, 2011). Dyspnea, fatigue, chest pain are among the major symptoms of PAH, which is significantly associated with lung vascular scheme and causes premature death (Lai, 2014). Although early diagnostic therapy can certainly reduce the death rate of PAH (Natarajan, 2011). COVID-19 has caused many people to suffer from cardiac, age-related, and pulmonary diseases, including pulmonary arterial hypertension (Horn, 2020). Meanwhile, researchers have produced results that demonstrate the activity of SARS-CoV-2 in promoting pulmonary microthrombi, vascular leak through different ways including inflammation, damage of DNA, and mitochondrial dysfunction (Chen, 2020). Based on these studies, PAH can be considered as a major risk factor of COVID-19. Due to the mentioned reasons, it is revealed that there may be a number of pathological compatibility between COVID-19 and PAH. To get an idea of this compatibility we have tried to identify altered pathways that are common for SARS-CoV-2 infections and PAH-affected samples. To accomplish these tasks, large-scale transcriptomic datasets have been used in this research.

Large-scale microarray datasets are important for uncovering gene expression based biological information (Wichert, 2004). High-throughput sequencing has immensely influenced the advancement of biomedical research by contributing to the rapidly growing genome sequencing field. High-throughput sequencing-based analysis has already been implemented on severe acute respiratory syndrome coronavirus (SARS-CoV) that has also produce remarkable gene expression results (Irigoyen, 2016).

The significance of the research is that we performed the largest comparative and transcriptomic study against SARS-CoV-2 infection responses to human lung epithelial cells. The potential biomarkers we have been able to figure out have proved the significance in terms of appropriate immune responses. The following analyses attempt to find cell informative pathways and drug compounds based on the transcriptomic analysis on SARS-CoV-2 and PAH. But initially, the genomic analysis was introduced to identify genomic differences of SARS-CoV and SARS-CoV-2 effect on Homo sapiens. This genomic level study eventually allows the research to put emphasis on SARS-CoV-2 and the major risk factors. As a result, two datasets (GSE147507 and GSE117261) were selected for the transcriptomic level study. Hence, the research went through the identification process of finding out differentially expressed genes (DEGs) from GSE147507 and GSE117261. However, similar DEGs were conducted as input data for a further molecular level study that includes Gene ontology (GO) terms identification and predictive analysis on cell informative pathways. The visualization of the Protein-protein interactions (PPIs) network is regarded as the focal point of the analysis as hub nodes and significant modules were identified from the PPIs. Herein, transcriptional regulators are also traced based on the similar DEGs of GSE147507 and GSE117261. Finally, potential drug compounds are suggested.

1.2 Motivation of the Research:

The study of COVID-19 is the most essential research in the world today. The outbreak of the SARS-CoV-2 virus appear to be exacerbated during the time period of last two years. The principal target of this catastrophic virus is the human respiratory system. This is why in our research we try to see its compatibility with Pulmonary Arterial Hypertension disease. A number of current studies have shown a similarity between the SARS-CoV-2 virus and PAH, which has made us more interested in revealing the link between the two diseases and the solution in terms of bioinformatics study. Spike protein of SARS-CoV-2 virus which enters the host body and helps in the transmission of the virus to the host body. The most frightening thing is that as soon as it enters the host body, it does not show any risk symptoms other than a few common symptoms. Infected patients with this disease show no symptoms other than mild problems like fever, chills, headache, sore throat, and cough are seen at an early stage. Symptoms are associated with final phases, but then there is no prior system to treat the patient as any substantial vaccine has not yet been developed. However, the vaccine that has been developed is still in a face of crisis as the demand is extremely high.

1.3 Problem Statement:

Since learning about previous similar works, I've noticed some weaknesses, which may take this to the next level of the study.

- ✓ Limited volumes of data is used for COVID-19 analysis.
- ✓ No work related to SARS-CoV-2 and PAH has been done yet.
- ✓ Unable to generate Genomic analysis.
- ✓ Did not work with Transcriptomic analysis.
- ✓ No computational biology approach has been done yet.

1.4 Research Question:

A research question is an answerable inquiry into a specific problem or topic. In a research, this is the initial step. Once, an understanding of what to explore, the 'initial step' means that the research topic is the first important step in the research. Some questions are explored. The list of questions is given below:

- ✓ How to collect the data set for selected diseases? and where?
- ✓ How to mine the data set?
- ✓ How to show difference in between SARS-CoV and SARS-CoV-2 viruses?
- ✓ How to identify Differentially Expressed genes?
- ✓ How to analyze the Topological Properties in finding significance of PPI network?
- ✓ How to add the references?

1.5 Research Objectives:

The key goals of this thesis are given below:

- ✓ Identify the Differences between SARS-CoV and SARS-CoV-2.
- ✓ Perform transcriptomic and genomic analyses.
- ✓ Generate PPI diagram.
- ✓ Apply topological algorithm on the PPI network.
- ✓ Predict drug compounds in terms of the hub nodes of PPI network.

1.6 Research Scope:

The study's scope applies to the areas covered by the research. The scopes are given below:

- ✓ Downloading larger data set.
- ✓ Differentially expressed genes identification using R.
- ✓ PPI network.
- ✓ Analyzing Topological Properties.
- ✓ Gene ontology and Pathway Identification.
- ✓ Prediction of potential therapeutic drugs.

1.7 Thesis Organization:

In a certain section, the whole paper is organized. Where the relevant analysis is discussed in Section 2 to extract the conceptual framework. The source of supporting

Proof of arguments to not only validate the statement or theory, but also serve as a basis for the results. The Proposed methodology is presented in the section 3. I have clarified my results and final outcome in section 4 with the support of some graphs and network. In that section, all the outcomes have been described. A small summery of the whole research is presented as conclusion in section 5.

2.1 Definition of the Thesis Topic or Area:

Computational biology and bioinformatics are interdisciplinary fields in which computational techniques are created and applied in order to analyze vast biological data sets, including genetic sequences, cell populations or protein samples, to allow new hypotheses or to explore new biology. In molecular biology and medicine, bioinformatics is the application and implementation of computer science techniques to solve problems. There is a reliable healthy workplace potential for the next several years with the success and faster growth in this area. A computational biologist has a very successful research profile in the research and innovation field of biotech, pharmaceutical, and analytical software firms.

2.2 The Direction of Research:

The present study attempted to examine the relationship between SARS-CoV-2 and PAH. Differentially expressed genes among the selected diseases are identified using the method of computational biology. The top 6 weighted genes are selected from common gene results. It establishes and analyzes PPI, Topological Properties, Hub nodes identification and module analysis.

2.3 Theoretical Framework:

In this study, PPI network is discussed. Protein-protein interactions are incredibly efficient biological connections formed between two or more protein molecules as a result of the interaction that involve electrostatic forces, hydrogen bonding and the hydrophobic effect guided by biological pathways. Based on PPI network, topological properties are produced through Cytoscape tool. TF-miRNA coregulatory network is a significant mechanism in deciding the activity, delivery, rate of excretion, and toxicity of drug molecules is the binding of drugs to proteins in the blood, serum or plasma. The TF-miRNA coregulatory network is also developed in Cytoscape.

2.4 Resources:

In this research, COVID-19 and PAH are selected. All genes are obtained from GEO database (NCBI), based on the diseases. NCBI is a database which is trusted. The genes that have been found to be substantially differentially expressed are obtained for analysis. All the genes have been identified as Homo sapiens. The tools that are used in the whole analysis are all downloaded from the internet.

The experimental workflow of the ongoing research is presented in Figure 1.

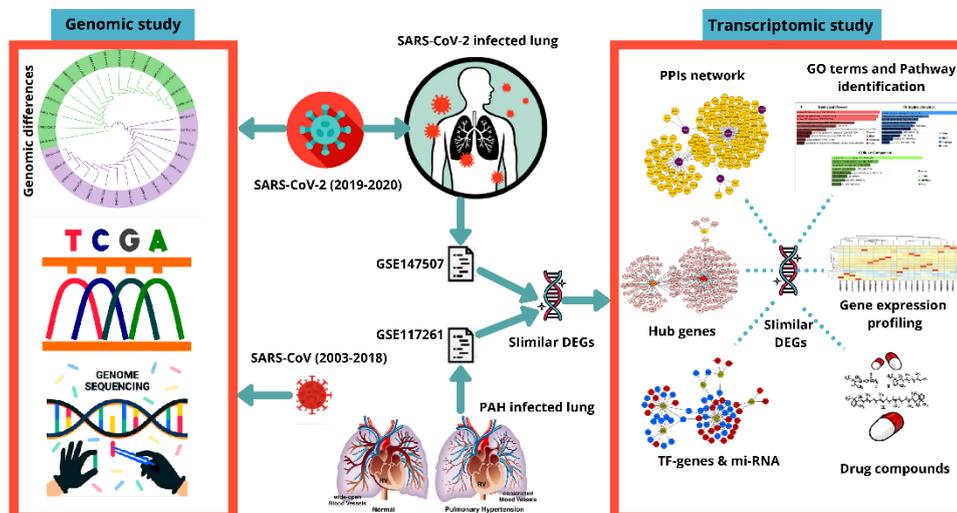


Figure 1: The workflow of current analysis. Genomic differences between SARS-CoV and SARS-CoV-2 are visualized through a phylogenetic analysis. Two datasets GSE147507 and GSE117261 are collected according to SARS-CoV-2 infection in human lung epithelial cells and PAH lung, respectively. Differentially expressed genes (DEGs) were identified using R programming language and similar DEGs were identified from total DEGs of both the datasets. Corresponding similar DEGs were used to perform transcriptomic analyses. The gene expression profiling was performed for both the datasets and Gene ontology (GO) terms, cell informative pathways, PPIs network, Hub gene identification and TF-miRNA based analyses were performed. According to the corresponding similar DEGs, drug compounds were predicted.

3.1 Comprehensive genomic level phylogenetic study

Comparison between SARS-CoV and SARS-CoV-2 at the viral genomic level is generated with the collection of a number of genome sequences. The sequences were gathered from the Virus Pathogen Database and Analysis Resource (ViPR). Total 32 sequences were analyzed where SARS-CoV and SARS-CoV-2 both contain 16 sequences respectively. The sequences for SARS-CoV: JN247391, JN247392, JN247393, JN247394, JN247395, JN247396, JN247397, GU553363, GU553364, AY274119, MK062179, MK062180, MK062181, MK062182, MK062183, and MK062184. Besides, sequences for SARS-CoV-2: MT008022, MT008023, MN988668, MN988669, LC521925, LC522972, LC522973, LC522974, LC522975, MN938385, MN938387, MN938384, MN938388, MN938386, MN938389 and MN938390. According to the sequences, a PHYLIP formatted comprehensive phylogenetic guided tree was designed using Clustal Omega. Clustal Omega contains significant features and exploits comprehensive information based on sequence alignments (Sievers, 2011]. The phylogenetic tree was re-designed using the interactive tree of life (iTOL). iTOL provides graphical representations of numerous phylogenetic trees and the representations can be customized (Letunic, 2007).

3.2 Details information of the datasets

Last GSE147507 and GSE117261 datasets were assembled from the Gene Expression Omnibus (GEO) database (Clough, 2016). GEO database provides gene expression-

based analysis, which is under the platform of National Center for Biotechnology Information (NCBI) (Edgar, 2002). GSE147507 dataset interprets host responses to SARS-CoV-2 and transcriptional responses in lung epithelium cells. GPL18573 Illumina NextSeq 500 (Homo sapiens) platform is utilized for GSE147507 to retrieve the analysis of RNA sequence. The contributor of the GSE147507 dataset was Blanco-Melo D et al. However, the GSE117261 dataset represents transcriptomic analysis and systems biology representation on PAH lung. GPL6244 platform was used for GSE117261 dataset, which is [HuGene-1_0-st] Affymetrix Human Gene 1.0 ST Array [transcript (gene) version]. GSE117261 consists of a total of 83 samples that include PAH lung: 58 samples and control lung: 25 samples.

3.3 Data filtering and retrieval of DEGs, and identification of common DEGs between SARS-CoV-2 and PAH

The Transcriptomic datasets GSE147507 for SARS-CoV-2 infection in human lung epithelial cells and GSE117261 for PAH lung is used for this research. The initial preprocessing phase of the research goes through the retrieval of DEGs for both datasets. Identification of DEGs for the dataset GSE147507 is achieved with the assist of the R programming language. Herein, limma (Smyth, 2005) and DESeq2 (Love, 2014) packages of R programming language are used for obtaining DEGs for the GSE147507 dataset. Absolute log₂ fold change >1.0 and an adjusted p-value <0.05 were considered as cutoff criteria to determine significant DEGs from the GSE147507 dataset. GEO2R, which is a web-based platform for the analysis of microarray datasets is used for the identification of DEGs for the GSE117261 dataset. GEO2R performs the analysis in a comparative manner by comparing infected samples vs controlled samples and the comparison is generated through limma and GEOquery (Davis, 2007) packages from Bioconductor (Gentleman, 2004) project in the platform of R programming language. Benjamini-Hochberg methodology was implemented for GSE147507 and GSE117261 datasets with the purpose of the false discovery rate (FDR) controlling (Benjamini, 1995). Similar DEGs were also acquired using the R programming language.

3.4 Gene ontology and cell informative pathways analysis

Gene Gene set enrichment analysis (GSEA) is generally a computational and statistical methodology that defines whether a set of determined genes show statistical significance in different biological conditions (Subramanian, 2007). The resources of Gene Ontology (GO) provides structural and computational information considering the gene product based functions (Podder, 2020). GO can be categorized into three subsections including molecular function, biological process, and cellular component for annotation of gene products (Doms, 2005). GO terms for the current study are obtained using Enrichr (<https://amp.pharm.mssm.edu/Enrichr/>) platform. Enrichr is a web-based program that contains large gene sets consisting of 102 libraries and performs experiments that are genome-based (Kuleshov, 2016). For cell informative pathway analysis KEGG (Kanehisa, 2000), Reactome (Fabregat, 2018), WikiPathways (Slenter, 2018), and BioCarta databases are employed. The results from the databases are also implemented using the Enrichr platform.

3.5 Designing of Protein-protein Interactions (PPIs) network

Prominent information about the functions of protein is achieved with the analysis of protein interactions, which is regarded as the primary step in drug discovery and systems biology (Sikic, 2009). The number of complex biological processes are determined with the advanced study of Protein-protein interactions (PPIs) networks (Pagel, 2005). Identified similar DEGs for SARS-CoV-2 and PAH lung, were provided as an input in InnateDB (Breuer, 2013) using the NetworkAnalyst (<https://www.networkanalyst.ca/>) web-based platform. Numerous omics data analysis is achieved through a visual representation of NetworkAnalyst platform including complex PPIs network (Xia, 2015). The network was further designed using Cytoscape (<https://cytoscape.org/>). Cytoscape software can be regarded as a prominent source in integrating protein-interactions and genetic-interactions (Shannon, 2003).

3.6 Establishment of the topological algorithm on the PPIs network and detection of hub nodes

Hub nodes generally defined by the highly-interconnected nodes in a large-scale complex PPIs network (Hsing, 2008). The hub nodes for the current research are determined by the degree topological algorithm. The degree algorithm is applied to the PPIs network using a plugin of Cytoscape software, which is cytoHubba (<http://apps.cytoscape.org/apps/cytohubba>). cytoHubba is a comprehensive plugin of Cytoscape software that consists of 11 topological algorithms to rank the nodes in a specific network (Chin, 2014). In the areas where the hub genes are highly-interconnected, these areas are regarded as prominent modules from the PPIs network. Distinguishing the modules from the PPIs network will provide better visualization of the hub nodes in separated modules. For specific module analyses for the corresponding PPIs network is generated by ClusterViz (<http://apps.cytoscape.org/apps/clusterviz>), which is also a Cytoscape plugin. Cluster identification and detection of functional modules from a number of networks including PPIs network, metabolic network, and gene network are determined by ClusterViz plugin (Wang, 2014).

3.7 Analysis of TF-miRNA co-regulatory network

RegNetwork repository was used to generate the analysis of the TF-miRNA co-regulatory network (Liu, 2015). The miRNAs and TFs are identified from the co-regulatory network, which is responsible for the regulation of DEGs at transcriptional and post-transcriptional levels. The visualization of the network was provided using NetworkAnalyst web-based platform. For system-level data understanding, NetworkAnalyst has been used as a leading bioinformatics tool as a demand of immensely growing gene expression-based datasets (Zhou, 2019).

3.8 Therapeutic drug compounds prediction

According to similar DEGs, a number of drug compounds are predicted from the Drug Signatures database (DSigDB) using the Enrichr platform. DSigDB consists of gene sets: 22,527, gene: 19,531 and unique compound: 17,389 (Yoo, 2015). DSigDB predominantly predicts drugs on gene expression based datasets and each set of the gene are regarded as targeted genes considering a compound. Performing genome-

based characterization including RNA, DNA, and protein-based biomedical, pharmacological, and biological information can be gathered with more accuracy and at an inexpensive cost using the Enrichr web-platform (Chen, 2013).

4.1 Genomic and phylogram differences between SARS-CoV and SARS-CoV-2

Genomic differences are observed through phylogenetic analysis of SARS-CoV and SARS-CoV-2. The 16 genome sequences for SARS-CoV are the sequences from the year 2003-2018 and the host responses were for humans. However, another 16 genome sequence sample for SARS-CoV-2 are the sequences from the year 2019-2020 and host responses were for humans as well. The result of the phylogenetic analysis shows that SARS-CoV and SARS-CoV-2 do not produce any clade between them but the samples share ancestral origin among themselves. This distinguishes SARS-CoV and SARS-CoV-2 at the genomic level. Phylogenetic visualization of SARS-CoV and SARS-CoV-2 genome sequences are displayed in Figure 2.

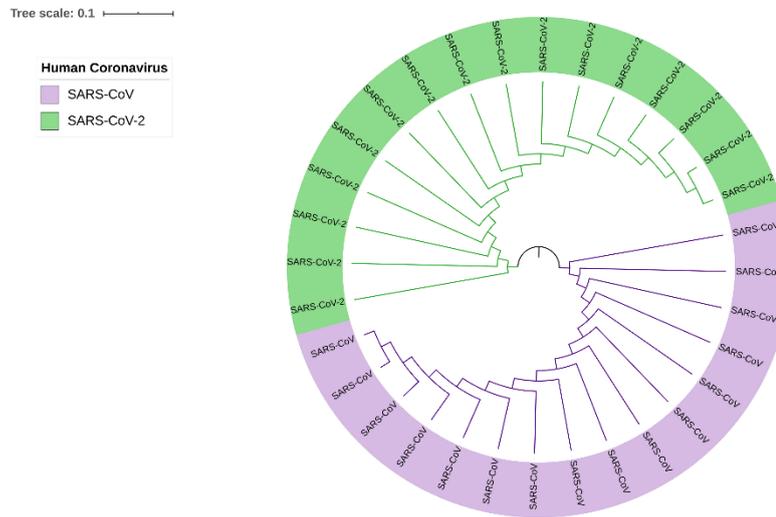


Figure 2: Phylogram of SARS-CoV and SARS-CoV-2 that provide genomic differences between human coronaviruses of 2003-2018 (SARS-CoV) and 2019-2020 (SARS-CoV-2). Two colors are implemented to differentiate SARS-CoV (purple) and SARS-CoV-2 (green).

4.2 Gene expression analysis of PAH patients and SARS-CoV-2 infected human lung epithelial and associative cells

The Form the GSE147507 dataset, 24 samples were filtered, and those samples were involved with SARS-CoV-2 infection to primary human bronchial epithelial cells, lung adenocarcinoma, and lung biopsy cells. The gene expression of the top 20 genes from the selected samples has been visualized in Figure 3, which provides the report of the high expression profile of S100A9 and KRT5 gene. Besides, among all 83 samples of PAH lung and healthy controls, characterization of gene expression is presented for 20 samples including 3 healthy controls (GSM3290083, GSM3290086, and GSM3290085), and the remaining of them are PAH samples. Differentiating PAH

samples and healthy controls provide evidence of distinct groups of PAH samples according to hierarchical clustering and comparing both samples at RNA level provides different infection response of PAH sample compared to healthy controls (Figure 4A). A volcano plot is visualized and considered the adjusted p-value <0.05 , which showed the up-regulated and down-regulated genes that have been identified through a comparative analysis between PAH samples and normal samples for the GSE117261 dataset (Figure 4B).

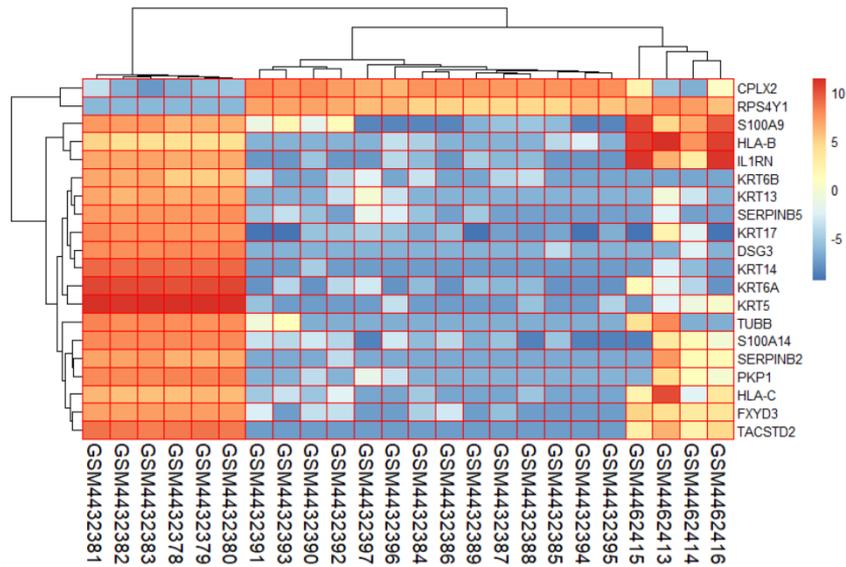


Figure 3: Gene expression profiling of SARS-CoV-2 infection in human lung epithelial cells for the top 20 genes and selected 24 samples from the GSE147507 dataset.

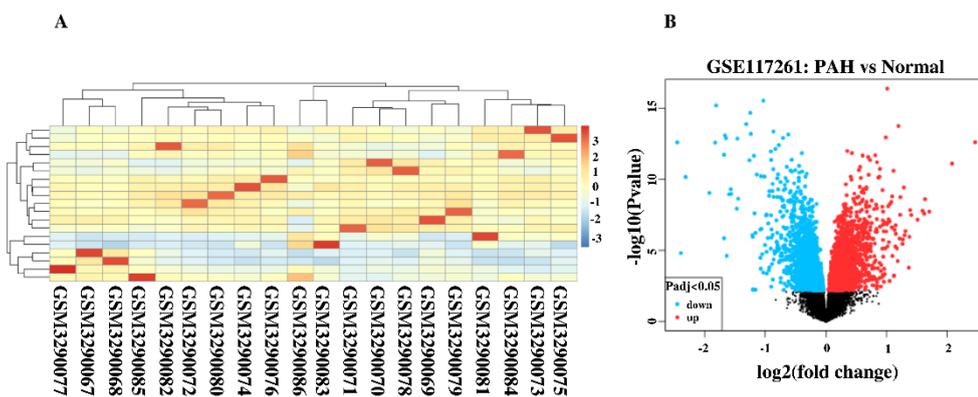


Figure 4: (A) Gene expression visualization of healthy controls (GSM3290083, GSM3290086 and GSM3290085) and PAH samples. (B) Volcano plot shows the regulation of genes (up-regulated and down-regulated) for GSE117261.

4.3 Common DEGs identifications for further molecular analysis and ensure the efficiency of predictive drugs

For SARS-CoV-2 infection responses to human lung epithelial cells observation, the DEGs of dataset GSE147507 is identified. Regarding the analysis, a total of 108 DEGs was found. Notably, 93 DEGs show upregulation, and the remaining 15 DEGs show downregulation. However, comparison analysis between PAH lung and healthy controls for GSE117261 shows a total of 59 DEGs of which 27 DEGs show upregulation and another 32 DEGs show downregulation. Comparing SARS-CoV-2 infection responses and PAH samples, 6 DEGs (SAA2, S100A9, S100A8, SAA1, S100A12, and EDN1) manifest concordance, which is used for identifying GO terms and pathway results, PPIs network, hub nodes, and module identification and TF-miRNA regulation and prediction of drug compounds. The concordance produced from the comparison between these two datasets is visualized using a Venn diagram (Figure 5A). The heat map regarding the log fold change for the shared common genes between SARS-CoV-2 and PAH showed unparalleled transcriptional signature impelled upon SARS-CoV-2 infection (Figure 5B). The gene validation is provided according to the risk groups of the genes in a heat map that provides information regarding S100A9 and S100A8 that are highly prone to inflammation (Figure 6A). The boxplot of the risk group comparison also indicates that S100A9 and S100A8 are highly risked prone (Figure 6B).

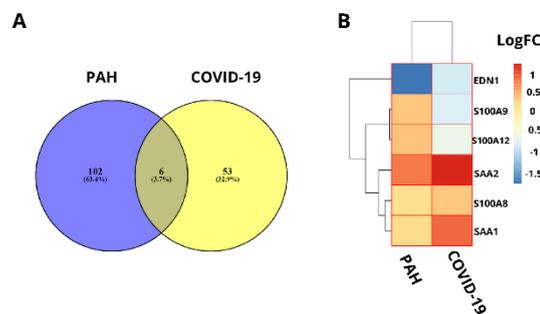


Figure 5: (A) Concordant gene identification between GSE147507 and GSE117261 dataset that provide evidence of 6 common differentially expressed genes in between 108 genes of GSE147507 (COVID-19) and 59 genes of GSE117261 (PAH) dataset.

(B) Heat map according to the log fold changes for the shared common DEGs between COVID-19 dataset and PAH dataset.

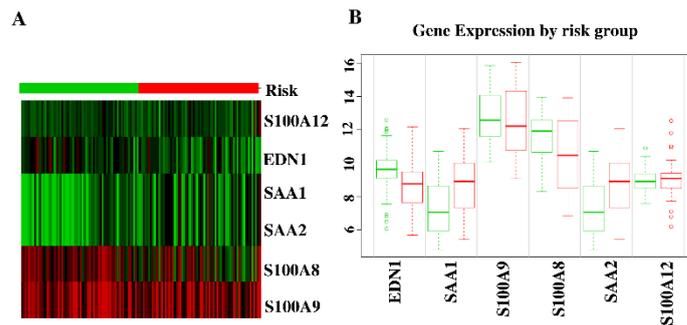


Figure 6: (A) Heat map for the identification of highly risk prone nature of S100A9 and S100A8 genes. (B) Risk group comparisons between the shared common genes of SARS-CoV-2 and PAH.

4.4 Gene ontology and pathway analysis based on the similar DEGs

NetworkAnalyst After the identification of unique DEGs aligned with SARS-CoV-2 infection profile to lung epithelial cells, a number of databases (KEGG, Reactome, WikiPathways, BioCarta, and The Gene Ontology) were utilized to identify gene ontology terms and cell informative pathways. Among all the GO terms, the top 10 biological processes, cellular components, and molecular functions were predicted (Table 1). Analysis of biological processes provides neutrophil chemotaxis, granulocyte chemotaxis, and regulation of inflammatory responses to SARS-CoV-2 infections according to the number of genes interaction. Molecular function regarded studies show enrichment of calcium ion binding, zinc ion binding, transition metal ion binding, and metal ion binding factors. Cytoplasmic vesicle lumen cellular component factor is significantly involved with the corresponding identified DEGs which eventually refer to SARS-CoV-2 infection responses to the human lung. Notably, top pathways based on the DEGs were allied in the current study (Table 2). IL-17 signaling pathway, TNF signaling pathway, Vitamin B12 Metabolism are among the top pathways that were identified through the analysis of the curated databases. The comparison of GO terms is represented in Figure 7(A) and the comparison of pathways from numerous databases is provided in Figure 7(B).

Table 1: The association of concordant genes in GO terms and GO pathways and the proportional p-values.

Category	GO ID	Term	P-value	Genes
GO Biological Process	GO:0030593	neutrophil chemotaxis	6.563e-10	SAA1, S100A12, S100A9, S100A8
	GO:0071621	granulocyte chemotaxis	8.230e-10	SAA1, S100A12, S100A9, S100A8
	GO:1990266	neutrophil migration	9.506e-10	SAA1, S100A12, S100A9, S100A8
	GO:0050832	defense response to fungus	1.018e-8	S100A12, S100A9, S100A8
	GO:0050727	regulation of inflammatory response	6.777e-8	SAA1, S100A12, S100A9, S100A8
	GO:0051091	positive regulation of sequence-specific DNA binding transcription factor activity	1.915e-7	EDN1, S100A12, S100A9, S100A8
	GO:0050729	positive regulation of inflammatory response	9.257e-7	S100A12, S100A9, S100A8
	GO:0031349	positive regulation of defense response	9.647e-7	S100A12, S100A9, S100A8
	GO:0070486	leukocyte aggregation	0.000001574	S100A9, S100A8
	GO:0032103	positive regulation of response to external stimulus	0.000001745	S100A12, S100A9, S100A8
GO Molecular Function	GO:0050786	RAGE receptor binding	1.259e-9	S100A12, S100A9, S100A8
	GO:0035325	Toll-like receptor binding	0.000002697	S100A9, S100A8
	GO:0005509	calcium ion binding	0.00005490	S100A12, S100A9, S100A8
	GO:0008270	zinc ion binding	0.00006592	S100A12, S100A9, S100A8
	GO:0046914	transition metal ion binding	0.0001507	S100A12, S100A9, S100A8
	GO:0046872	metal ion binding	0.0002040	S100A12, S100A9, S100A8

	GO:0008017	microtubule binding	0.001383	S100A9, S100A8
	GO:0015631	tubulin binding	0.002348	S100A9, S100A8
	GO:0005507	copper ion binding	0.01224	S100A12
GO Cellular Component	GO:0060205	cytoplasmic vesicle lumen	2.453e-8	SAA1, S100A12, S100A9, S100A8
	GO:0071682	endocytic vesicle lumen	0.005388	SAA1
	GO:0005881	cytoplasmic microtubule	0.01135	SAA1
	GO:0034774	secretory granule lumen	0.00007614	S100A12, S100A9, S100A8
	GO:0045111	intermediate filament cytoskeleton	0.02111	S100A8
	GO:0005856	cytoskeleton	0.0003296	S100A12, S100A9, S100A8
	GO:0030139	endocytic vesicle	0.03197	SAA1
	GO:0005874	microtubule	0.06138	SAA1

Table 2: The association of concordant genes in KEGG, WikiPathways, Reactome and Biocarta databases and the proportional p-values.

Databases	Pathways	P-value	Genes
KEGG	IL-17 signaling pathway	0.0003170	S100A9, S100A8
	Renin secretion	0.02052	EDN1
	Hypertrophic cardiomyopathy (HCM)	0.02523	EDN1
	AGE-RAGE signaling pathway in diabetic complications	0.02963	EDN1
	HIF-1 signaling pathway	0.02963	EDN1
	Melanogenesis	0.02992	EDN1

	TNF signaling pathway	0.03255	EDN1
	Relaxin signaling pathway	0.03838	EDN1
	Vascular smooth muscle contraction	0.03896	EDN1
	Fluid shear stress and atherosclerosis	0.04099	EDN1
WikiPathways	Vitamin B12 Metabolism WP1533	0.00009129	SAA1, SAA2
	Folate Metabolism WP176	0.0001595	SAA1, SAA2
	IL1 and megakaryocytes in obesity WP2865	0.007179	S100A9
	Physiological and Pathological Hypertrophy of the Heart WP1528	0.007477	EDN1
	Selenium Micronutrient Network WP15	0.0002711	SAA1, SAA2
	Endothelin Pathways WP2197	0.009860	EDN1
	Photodynamic therapy-induced HIF-1 survival signaling WP3614	0.01105	EDN1
	Melatonin metabolism and effects WP3298	0.01105	EDN1
	Prostaglandin Synthesis and Regulation WP98	0.01343	EDN1
	Vitamin D Receptor Pathway WP2877	0.001206	S100A9, S100A8
Reactome	Advanced glycosylation endproduct receptor signaling Homo sapiens R-HSA-879415	0.000005841	SAA1, S100A12
	DEx/H-box helicases activate type I IFN and inflammatory cytokines production Homo sapiens R-HSA-3134963	0.000005841	SAA1, S100A12
	Scavenging by Class B Receptors Homo sapiens R-HSA-3000471	0.001499	SAA1
	RIP-mediated NFkB activation via ZBP1 Homo sapiens R-	0.00001571	SAA1, S100A12

	HSA-1810476		
	TRAF6 mediated NF-kB activation Homo sapiens R-HSA-933542	0.00002064	SAA1, S100A12
	ZBP1(DAI) mediated induction of type I IFNs Homo sapiens R-HSA-1606322	0.00002430	SAA1, S100A12
	TAK1 activates NFkB by phosphorylation and activation of IKKs complex Homo sapiens R-HSA-445989	0.00002430	SAA1, S100A12
	Formyl peptide receptors bind formyl peptides and many other ligands Homo sapiens R-HSA-444473	0.002398	SAA1
	Cytosolic sensors of pathogen-associated DNA Homo sapiens R-HSA-1834949	0.0001595	SAA1, S100A12
	TRAF6 Mediated Induction of proinflammatory cytokines Homo sapiens R-HSA-168180	0.0001899	SAA1, S100A12
BioCarta	G-Protein Signaling Through Tubby Proteins Homo sapiens h tubbyPathway	0.002997	EDN1
	Activation of PKC through G-protein coupled receptors Homo sapiens h pkcPathway	0.003296	EDN1
	Hypoxia-Inducible Factor in the Cardiovascular System Homo sapiens h hifPathway	0.004791	EDN1
	Cystic fibrosis transmembrane conductance regulator (CFTR) and beta 2 adrenergic receptor (b2AR) pathway Homo sapiens h cftrPathway	0.005986	EDN1
	Corticosteroids and cardioprotection Homo sapiens h gcrPathway	0.007477	EDN1
	Beta-arrestins in GPCR Desensitization Homo sapiens h bArrestinPathway	0.008372	EDN1
	Activation of cAMP-dependent protein kinase, PKA Homo sapiens h gsPathway	0.008670	EDN1

	Role of Beta-arrestins in the activation and targeting of MAP kinases Homo sapiens h barr-mapkPathway	0.008967	EDN1
	Role of EGF Receptor Transactivation by GPCRs in Cardiac Hypertrophy Homo sapiens h cardiacegfPathway	0.009860	EDN1
	Roles of Beta-arrestin-dependent Recruitment of Src Kinases in GPCR Signaling Homo sapiens h bArrestin-srcPathway	0.01016	EDN1

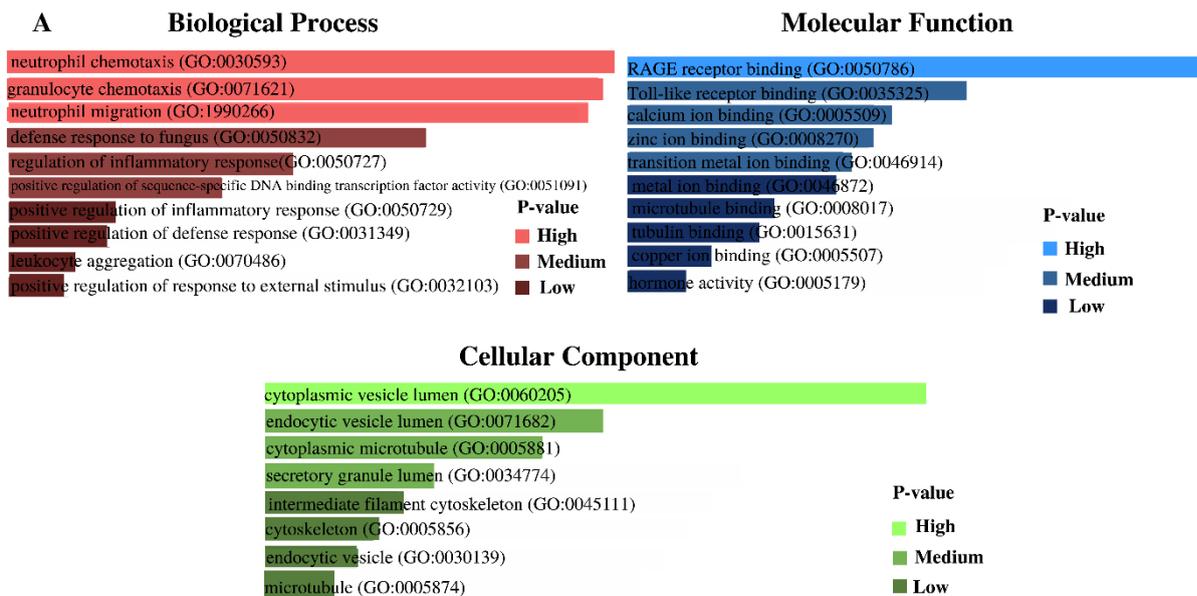


Figure 7(A): GO terms regarding biological process, molecular function and cellular component according to the associative p-values.

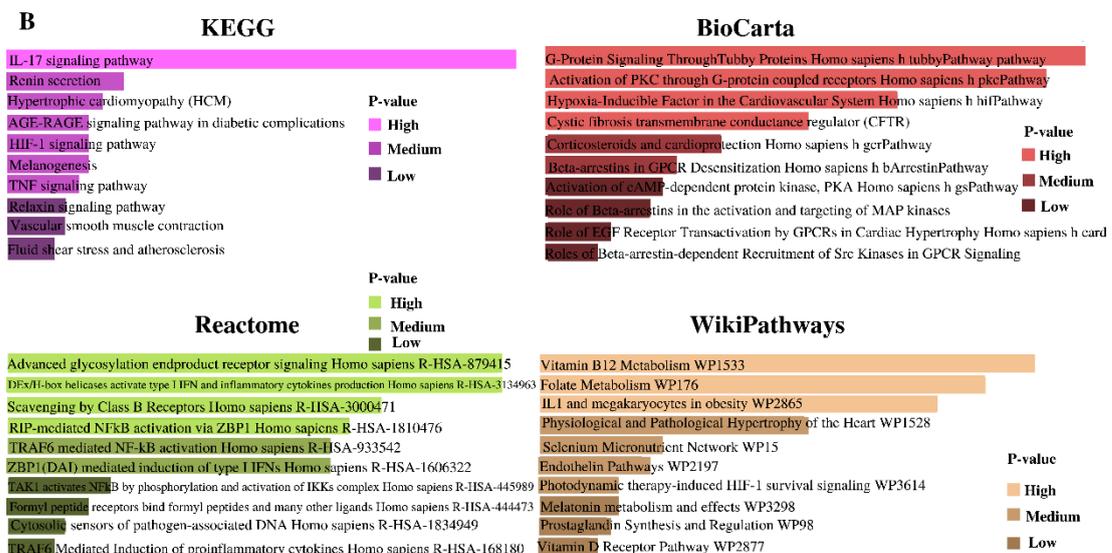


Figure 7(B): Cell informative pathways (KEGG, BioCarta, Reactome and WikiPathways) analysis result regarding associative p-values.

4.5 PPIs network construction to perceive hub nodes

Topological Using the NetworkAnalyst platform, 6 DEGs (SAA2, S100A9, S100A8, SAA1, S100A12, and EDN1) were provided as input and the generated network file was further customized in Cytoscape. The representation of the PPIs network show immense interaction of S100A9 and S100A8 genes and the interaction reveals the evidence of enrichment of S100A9 and S100A8 genes to SARS-CoV-2 responses in the human lung. Hub gene identification, module analysis, and prediction of effective drug compounds are mainly concerned with the corresponding PPIs network. The PPIs

network is represented in Figure 8 with customized visualization that contains 125 nodes and 136 edges.

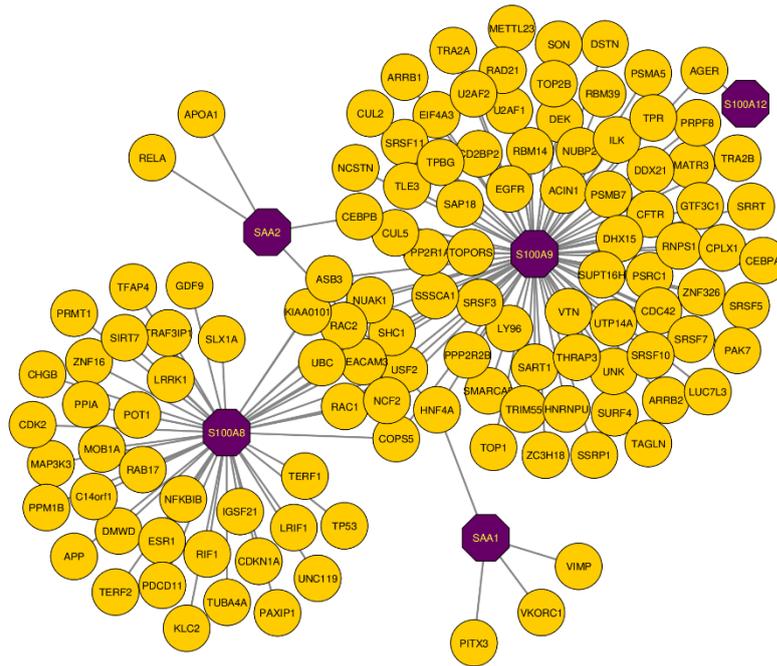


Figure 8: PPIs network for identified common DEGs that refers to SARS-CoV-2 infections in human lung and PAH lung. The common genes are highlighted with purple node (SAA2, S100A9, S100A8, SAA1 and S100A12). The network consists of 125 nodes and 136 edges.

4.6 Hub nodes identification based on the topological analyses and module detection from the PPIs network

Among the similar DEGs, hub nodes from the PPIs network are identified using cytohubba. The identified top 3 hub nodes are S100A9, S100A8 and SAA1. The Degree algorithm was used for the identification purpose and the degree algorithm shows the highest number of interaction in a specific network. The highlighted hub genes in a hub node identification network are presented in Figure 9 and the network consists of 124 nodes and 135 edges. The regions where the hub nodes are established in the PPIs network are considered as the prominent modules. Module analysis network is represented in Figure 10 that consists of 13 nodes and 13 edges. Topological analysis results for the top 3 hub genes are presented in Table 3.

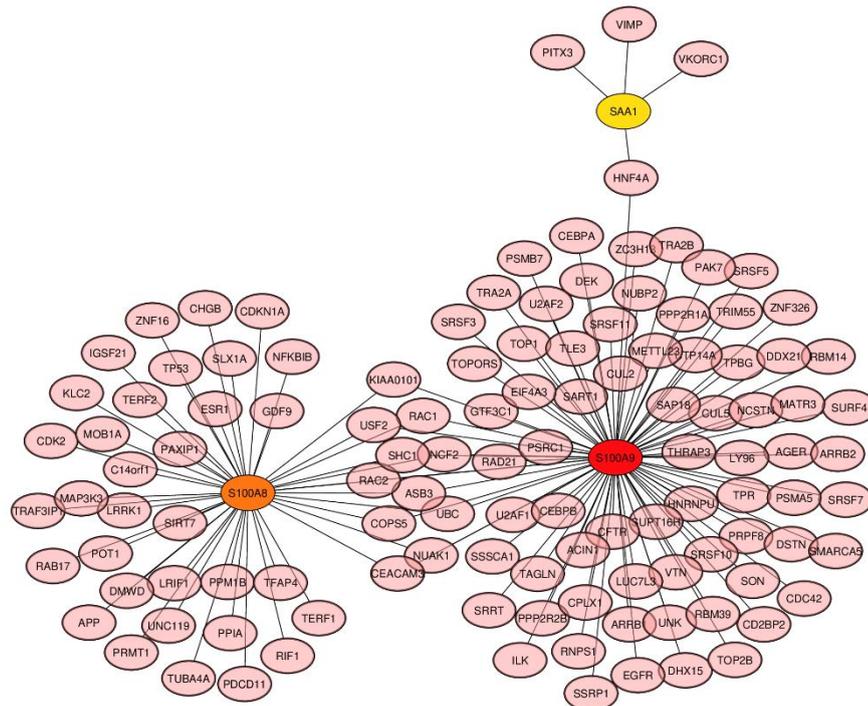


Figure 9: Hub gene detection from the similar DEGs based on the PPIs network. The highlighted nodes S100A9 (red), S100A8 (orange) and SAA1 (yellow) are regarded as highly interconnected nodes considered as hub nodes. The network is made up of 124 nodes and 135 edges.

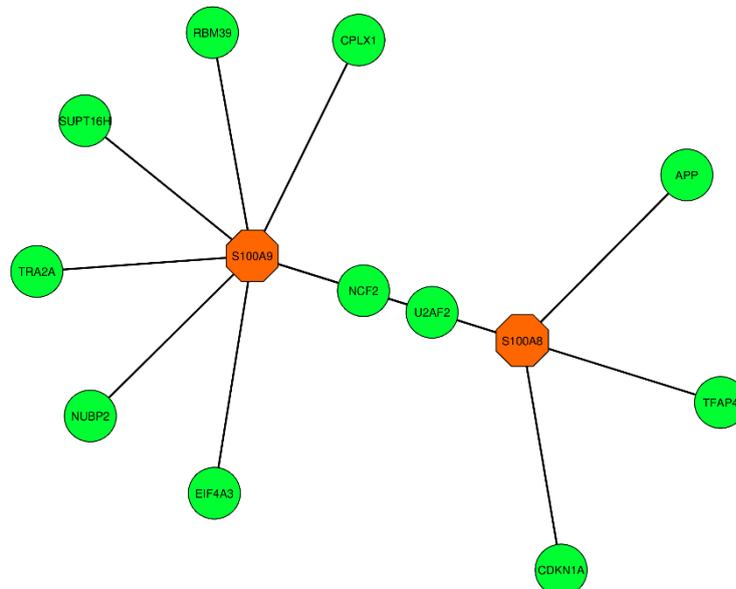


Figure 10: Highly interconnected regions (module) identification network that consists of 13 nodes and 13 edges. The hub genes S100A9 (orange) and S100A8 (orange) are visualized in the corresponding module network.

4.8 Predictive drug compounds

The drug compounds were proposed from the DSigDB database using the Enrichr web platform. The drug compounds were predicted according to identified 6 DEGs (SAA2, S100A9, S100A8, SAA1, S100A12, and EDN1). The results were accomplished based on adjusted p-value and p-value scores. MIGLITOL CTD 00002031 and metoprolol HL60 UP are the two prominent drug compounds with which a significant amount of genes are connected. Besides, among the top hub genes, S100A9 is interconnected with both the drug compounds, which makes the drug compounds even more eminent in terms of the efficiency of the drugs. The predictive drug compounds are presented in Table 4.

Table 4: Predictive drug compounds of the concordant genes of SARS-CoV-2 and PAH

Name of drugs	P-value	Adjusted p-value	Genes
MIGLITOL CTD 00002031	0.000004943	0.01990	S100A12, S100A9
Bosentan CTD 00003071	0.003296	0.5529	EDN1
Coenzyme Q10 CTD 00001167	0.003595	0.5789	EDN1
metoprolol HL60 UP	0.00007383	0.04954	S100A12, S100A9
9-(2-Phosphonomethoxypropyl)adenine CTD 00003259	0.004193	0.5821	EDN1
(+)-chelidone HL60 DOWN	0.00009129	0.05250	S100A9, S100A8
sildenafil CTD 00003367	0.004492	0.6028	EDN1
norepinephrine CTD 00006417	0.00009879	0.04972	S100A9, S100A8
dydrogesterone CTD 00005882	0.004791	0.6028	EDN1
1,3-Dimethylthiourea CTD 00001818	0.004791	0.5845	EDN1

5.1 Findings:

In this study, biological domains, regulatory elements, and identified biomarkers had been discussed in brief that is expected to accelerate the pace of therapeutics development against the ongoing COVID-19 pandemic. The superiority of our study can be considered as it is by far the largest genomic and transcriptomic study on SARS-CoV-2. We provided multiple ways of analyses including comparative genomic differences of SARS-CoV and SARS-CoV-2, and the difference has been made to look for transcriptomic analyses on SARS-CoV-2 and its PAH comorbidity condition. Phylogenetic analyses of this research have produced genomic differences between SARS-CoV and SARS-CoV-2. We have identified the concordant genes between SARS-CoV-2 and PAH that produce further molecular results and show the association of the differentially expressed genes in SARS-CoV-2 affected human lung epithelial cells and PAH patients' lung. A different type of transcriptional response was found due to the SARS-CoV-2 infection in human lung epithelial cells, which is enriched in inflammatory responses and neutrophil chemotaxis. The predicted drug compounds show activity against inflammatory responses against RNA viruses.

Recommendation for Future works:

It is necessary to understand that the infected genes are correlated with all these diseases in order to generate a medicine for more than one disease. It is also important to know the connection between the genes and the associated diseases in order to reach the destination. The current study carried out almost all the investigations with the support of bioinformatics tools which allow more calculations in the field of bioinformatics using bioinformatics software. Researchers who are willing to improve further can work to create a standard medicine for diseases using a microarray or microRNA data set for analysis.

REFERENCES

- [1] Coronaviridae Study Group of the International Committee on Taxonomy of Viruses, 2020. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology*, 5(4), p.536.
- [2] Walls, A.C., Park, Y.J., Tortorici, M.A., Wall, A., McGuire, A.T. and Velesler, D., 2020. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell*, 181(2), pp.281-292.
- [3] Chi, X., Yan, R., Zhang, J., Zhang, G., Zhang, Y., Hao, M., Zhang, Z., Fan, P., Dong, Y., Yang, Y. and Chen, Z., 2020. A neutralizing human antibody binds to the N-terminal domain of the Spike protein of SARS-CoV-2. *Science*, 369(6504), pp.650-655.
- [4] Oany, A.R., Mia, M., Pervin, T., Junaid, M., Hosen, S.Z. and Moni, M.A., 2020. Design of novel viral attachment inhibitors of the spike glycoprotein (S) of severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) through virtual screening and dynamics. *International journal of antimicrobial agents*, 56(6), p.106177.
- [5] Al-Awadhi, A.M., Al-Saifi, K., Al-Awadhi, A. and Alhamadi, S., 2020. Death and contagious infectious diseases: Impact of the COVID-19 virus on stock market returns. *Journal of Behavioral and Experimental Finance*, 27, p.100326.
- [6] Zulkar Nain, Humayan Kabir Rana, Pietro Liò, Sheikh Mohammed Shariful Islam, Matthew A Summers, Mohammad Ali Moni, Pathogenetic profiling of COVID-19 and SARS-like viruses, *Briefings in Bioinformatics*, <https://doi.org/10.1093/bib/bbaa173>
- [7] Mackenzie, J.S. and Smith, D.W., 2020. COVID-19: a novel zoonotic disease caused by a coronavirus from China: what we know and what we don't. *Microbiology Australia*, 41(1), pp.45-50.
- [8] Ferretti, L., Wymant, C., Kendall, M., Zhao, L., Nurtay, A., Abeler-Dörner, L., Parker, M., Bonsall, D. and Fraser, C., 2020. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science*, 368(6491).
- [9] Schermuly, R.T., Ghofrani, H.A., Wilkins, M.R. and Grimminger, F., 2011. Mechanisms of disease: pulmonary arterial hypertension. *Nature Reviews Cardiology*, 8(8), p.443.
- [10] Lai, Y.C., Potoka, K.C., Champion, H.C., Mora, A.L. and Gladwin, M.T., 2014. Pulmonary arterial hypertension: the clinical syndrome. *Circulation research*, 115(1), pp.115-130.
- [11] Natarajan, R., 2011. Recent trends in pulmonary arterial hypertension. *Lung India: Official Organ of Indian Chest Society*, 28(1), p.39.
- [12] Horn, E.M., Chakinala, M., Oudiz, R., Joseloff, E. and Rosenzweig, E.B., 2020. Could pulmonary arterial hypertension patients be at a lower risk from severe COVID-19?. *Pulmonary Circulation*, 10(2), p.2045894020922799.
- [13] Chen, L., Li, X., Chen, M., Feng, Y. and Xiong, C., 2020. The ACE2 expression in human heart indicates new potential mechanism of heart injury among patients infected with SARS-CoV-2. *Cardiovascular research*, 116(6), pp.1097-1100.
- [14] Guzzi, P.H., Mercatelli, D., Ceraolo, C. and Giorgi, F.M., 2020. Master regulator analysis of the SARS-CoV-2/human Interactome. *Journal of clinical medicine*, 9(4), p.982.
- [15] Wichert, S., Fokianos, K. and Strimmer, K., 2004. Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics*, 20(1), pp.5-20.
- [16] Soon, W.W., Hariharan, M. and Snyder, M.P., 2013. High-throughput sequencing for biology and

medicine. *Molecular systems biology*, 9(1), p.640.

[17] Irigoyen, N., Firth, A.E., Jones, J.D., Chung, B.Y.W., Siddell, S.G. and Brierley, I., 2016. High-resolution analysis of coronavirus gene expression by RNA sequencing and ribosome profiling. *PLoS pathogens*, 12(2), p.e1005473.

[18] Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J. and Thompson, J.D., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology*, 7(1), p.539.

[19] Letunic, I. and Bork, P., 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23(1), pp.127-128.

[20] Clough, E. and Barrett, T., 2016. The gene expression omnibus database. In *Statistical genomics* (pp. 93-110). Humana Press, New York, NY.

[21] Edgar, R., Domrachev, M. and Lash, A.E., 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research*, 30(1), pp.207-210.

[22] Blanco-Melo D, Nilsson-Payant BE, Liu WC, Uhl S et al. Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19. *Cell* 2020 May 28;181(5):1036-1045.e9.

[23] Smyth GK: Limma: linear models for microarray data. *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Edited by: Gentleman R, Care V, Dudoit S, Irizarry R, Huber W. 2005, New York: Springer, 397-420.

[24] Love, M.I., Huber, W. and Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), p.550.

[25] Davis, S. and Meltzer, P.S., 2007. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, 23(14), pp.1846-1847.

[26] Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J. and Hornik, K., 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10), p.R80.

[27] Benjamini, Y. and Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1), pp.289-300.

[28] Subramanian, A., Kuehn, H., Gould, J., Tamayo, P. and Mesirov, J.P., 2007. GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics*, 23(23), pp.3251-3253.

[29] Gene Ontology Consortium, 2019. The gene ontology resource: 20 years and still GOing strong. *Nucleic acids research*, 47(D1), pp.D330-D338.

[30] Podder, N.K., Rana, H.K., Azam, M.S., Rana, M.S., Akhtar, M.R., Rahman, M.R., Rahman, M.H. and Moni, M.A., 2020. A system biological approach to investigate the genetic profiling and comorbidities of type 2 diabetes. *Gene Reports*, 21, p.100830.

[31] Doms, A. and Schroeder, M., 2005. GoPubMed: exploring PubMed with the gene ontology. *Nucleic acids research*, 33(suppl_2), pp.W783-W786.

[32] Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A. and McDermott, M.G., 2016. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, 44(W1), pp.W90-W97.

[33] Kanehisa, M. and Goto, S., 2000. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1), pp.27-30.

- [34] Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B. and Milacic, M., 2018. The reactome pathway knowledgebase. *Nucleic acids research*, 46(D1), pp.D649-D655.
- [35] Slenter, D.N., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., Nunes, N., Mélius, J., Cirillo, E., Coort, S.L., Digles, D. and Ehrhart, F., 2018. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic acids research*, 46(D1), pp.D661-D667.
- [36] Šikić, M., Tomić, S. and Vlahoviček, K., 2009. Prediction of protein–protein interaction sites in sequences and 3D structures by random forests. *PLoSComput Biol*, 5(1), p.e1000278.
- [37] Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Stümpflen, V., Mewes, H.W. and Ruepp, A., 2005. The MIPS mammalian protein–protein interaction database. *Bioinformatics*, 21(6), pp.832-834.
- [38] Chowdhury, U.N., Islam, M.B., Ahmad, S. and Moni, M.A., 2020. Network-based identification of genetic factors in ageing, lifestyle and type 2 diabetes that influence to the progression of Alzheimer's disease. *Informatics in Medicine Unlocked*, p.100309.
- [39] Breuer, K., Foroushani, A.K., Laird, M.R., Chen, C., Sribnaia, A., Lo, R., Winsor, G.L., Hancock, R.E., Brinkman, F.S. and Lynn, D.J., 2013. InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. *Nucleic acids research*, 41(D1), pp.D1228-D1233.
- [40] Xia, J., Gill, E.E. and Hancock, R.E., 2015. NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. *Nature protocols*, 10(6), pp.823-844.
- [41] Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11), pp.2498-2504.
- [42] Hsing, M., Byler, K.G. and Cherkasov, A., 2008. The use of Gene Ontology terms for predicting highly-connected hub nodes in protein-protein interaction networks. *BMC systems biology*, 2(1), p.80.
- [43] Chin, C.H., Chen, S.H., Wu, H.H., Ho, C.W., Ko, M.T. and Lin, C.Y., 2014. cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC systems biology*, 8(S4), p.S11.
- [44] Wang, J., Zhong, J., Chen, G., Li, M., Wu, F.X. and Pan, Y., 2014. ClusterViz: a cytoscape APP for cluster analysis of biological network. *IEEE/ACM transactions on computational biology and bioinformatics*, 12(4), pp.815-822.
- [45] Liu, Z.P., Wu, C., Miao, H. and Wu, H., 2015. RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database*, 2015.
- [46] Zhou, G., Soufan, O., Ewald, J., Hancock, R.E., Basu, N. and Xia, J., 2019. NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. *Nucleic acids research*, 47(W1), pp.W234-W241.
- [47] Al Mustanjid, M., Mahmud, S.H., Royel, M.R.I., Rahman, M.H., Islam, T., Rahman, M.R. and Moni, M.A., 2020. Detection of molecular signatures and pathways shared in inflammatory bowel disease and colorectal cancer: A bioinformatics and systems biology approach. *Genomics*, <https://doi.org/10.1016/j.ygeno.2020.06.001>
- [48] Yoo, M., Shin, J., Kim, J., Ryall, K.A., Lee, K., Lee, S., Jeon, M., Kang, J. and Tan, A.C., 2015. DSigDB: drug signatures database for gene set analysis. *Bioinformatics*, 31(18), pp.3069-3071.
- [49] Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R. and Ma'ayan, A., 2013. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics*, 14(1), p.1.

