2023 26th International Conference on Computer and Information Technology (ICCIT) 13-15 December, Cox's Bazar, Bangladesh

Identification of Molecular Biomarkers and Key Pathways among Idiopathic pulmonary fibrosis (IPF), Chronic Obstructive Pulmonary Disease (COPD) and Lung cancer

2023 26th International Conference on Computer and Information Technology (ICCIT) | 979-8-3503-5901-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCIT60459.2023.10441016

Mst. Farjana Yasmin[b], Md. Faruk Hosen[a,b,*], Md. Abul Basar[a,c], Khairul Alam Shadhin[d], Arifa Ferdoushi Trisha[d], Muhammad Shahin Uddin[a]

[a]*Department of Information and Communication Technology, Mawlana Bhashani Science and Technology University, Tangail 1902, Bangladesh.*

[b]*Department of Computing and Information System, Daffodil International University (DIU), Ashulia, Dhaka, Bangladesh.*

[c]*Department of Computer Science and Engineering, Green University of Bangladesh (GUB), Purbachal American City, Kanchon 1460, Bangladesh.*

[d]*Department of Software Engineering, Daffodil International University (DIU), Ashulia, Dhaka, Bangladesh.*

*Corresponding author: faruk.cis@diu.edu.bd; farukictmbstu@gmail.com

***Abstract*—Lung cancer (LC), idiopathic pulmonary fibrosis (IPF) and chronic obstructive pulmonary disease (COPD) are the most fatal disorders in the globe, generating frequent human issues. Having IPF and COPD are the risk factors of the LC, but the molecular mechanisms that underlie among IPF, COPD, and LC are not yet elucidated. In this study, we looked for shared molecular indicators and pathways that might explain how individuals with IPF, COPD, and LC are related to one another. GSE24206, GSE76925 and GSE18842 microarray datasets are utilized for IPF, COPD and LC samples. The preprocessing of datasets has done using R language, and then concordant differentially expressed genes (DEGs) are discovered. After that the protein-protein interactions (PPIs) are built using the similar DEGs, and the hub genes are determined using topological analysis. ETS1, MSH2, SORD, RORA and NEDD9 are the PPI network's top 5 hub genes. The pathways of KEGG demonstrated that the concordant DEGs are related to the colorectal cancer and pathways in cancer. Future work for this project will focus on miRNA, TF, and gene ontology (GO) analyses, as well as module analysis networks. Finally, based on the concordant DEGs, a number of potential medications have been suggested.**

***Index Terms*—Lung cancer, IPF, Differential expressed genes, Protein-protein network, Hub gene, COPD, enrichment analysis, Drug compounds.**

I. Introduction

A majority common cause of death due to cancer globally is lung cancer. Typically, lung cancer may develop due to mutations in oncogenes leading to in the growth of altered cells that eventually lead to the development of lung tumors [1]. Numerous studies have shown that prevalent diseases including Lung Cancer, IPF and COPD. Three diseases were selected for research in this paper. With the intention of identifying their relationships with each other, three separate disease diseases were chosen. The risk is significantly lowered when a patient has just one disease. The risk multiplies tenfold when a person has multiple diseases. Our research aims to address the existing gap in identifying distinct molecular biomarkers and pathways among IPF, COPD, and lung cancer, providing crucial insights into their unique pathophysiological mechanisms. IPF, COPD, and lung cancer are the three diseases. COPD is a severe and only partially reversible condition which involves airflow restriction and aberrant inflammatory responses to environmental contaminants [2]. The fourth-leading cause of mortality worldwide is COPD [3]. Furthermore, COPD can raise the chance of lung cancer development [4]. According to certain epidemiological research, smokers with COPD are five times more probable to get lung cancer compared smokers with adequate lung function [5]. Poor lung function is a significant factor in the progression of lung cancer, and forced expiratory volume in one second (FEV1) has been established as a biomarker of overall respiratory risk from smoking, implying that there may be a link between COPD and lung cancer [6]. Idiopathic pulmonary fibrosis (IPF) is the other chronic lung condition having age and smoking as danger indicators, however it is distinguished by lung parenchymal scarring on histology and visualization, as well as lung function testing limitations [7]. A deterioration of the alveolar tissue or the lungs' airbags may result from this severe version of the disease [8]. IPF is classified as a potentially cancerous lung disease as people with IPF sometimes acquire concurrent lung cancer, while individuals with IPF have an approximate 3.34-fold increased chance of acquiring main lung cancer compared to ordinary

979-8-3503-5901-5/23/$31.00 ©2023 IEEE

people [9]. Despite the fact that lung cancer is regarded an advanced manifestation of IPF, the histological kinds of lung cancer associated with IPF still unknown, with contradictory results published in studies [9]. A newly published genomic sequencing study discovered that IPF and lung cancer share several somatic mutations [10]. The study utilizes study based on genes to clarify the link among IPF, COPD and Lung cancer to find novel approaches in order to cure of the disease. The performance of high throughput approaches has substantially increased due to Examining microarray information as well as the details extracted from A collection of expressions. 8 shared genes identified as shared among datasets were found after analyzing genes from the GSE24206, GSE18842, and GSE76925 datasets. The PPI network will be the subject of our next analysis since it is the key element for the current study. Then, to do this, we build a PPI network to better illustrate how related these DEGs are. Making use of a degree topological technique, the PPI network recognized and prioritized hub genes. In many bioinformatics investigations, similar DEGs have been combined to find specific medicinal compounds based on those DEGs. In addition, examination of common DEGs, particularly is included in this research, may establish gene ontology (GO) and other biological pathways. Microarray data contain molecular information that may be discovered through computer examination, assisting biological researchers. Finding the molecular link among IPF, COPD and Lung Cancer is the main goal of this research, along with identifying suitable biomarkers in accordance with the findings of gene-based analysis. Differentially expressed genes (DEGs) must be found to be able to ascertain which genes cause IPF, COPD and Lung Cancer. The interaction of the DEGs then comes to displayed by utilizing a PPIs network that is created. To evaluate the fundamental activities of the biological system, KEGG pathway analysis is then carried out. Drug candidates for the shared DEGs among IPF, COPD and Lung Cancer are suggested following the discovery of hub genes. The contribution of these research are

- Identifying distinct molecular signatures specific to IPF, COPD, and lung cancer.
- Unveiling key biomarkers facilitating early disease detection and precise treatment strategies.
- Unraveling novel pathways elucidating the unique pathophysiological mechanisms of these diseases.

Figure 1 shows an example of procedure for the present research.

II. Methodology

A. Dataset Collections

The GEO database's data was used to create the GSE24206, GSE18842, and GSE76925 datasets, accordingly [11]. This GSE24206 dataset contains entire lung samples taken from 11 IPF patients who had diagnostic surgical biopsy or lung transplantation. Six control samples were taken at the same time as the IPF patients' transplanted lungs from healthy donors. The GPL570 platforms were utilized to examine GSE24206 dataset. Out of 91 samples in GSE18842 dataset, 45 controls and 46 malignancies were examined. All samples aside from three are matched. The GPL570 platforms were used to analyze GSE18842 dataset. In the GSE76925 dataset, which included 40 nonsmokers as controls and 111 COPD patients, 214 genes showed differential expression. The GPL10558 platforms were used to analyze GSE76925 dataset.

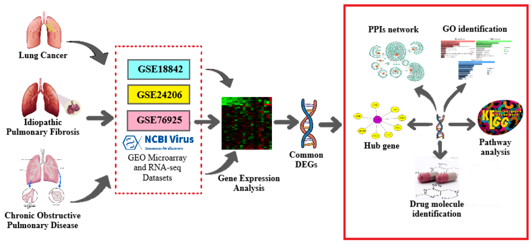

Fig. 1. *Figure of the information flow showing the inquiry and analysis process. Sample data (both healthy and infected cells) were obtained by analyzing data from GSE24206, GSE18842, and GSE76925. This GSE24206 dataset contains entire lung samples taken from 11 IPF patients who had diagnostic surgical biopsy or lung transplantation. In all, 91 non-small cells, In the GSE18842 dataset, samples of lung cancer (NSCLC). In the GSE76925 dataset, which included 40 nonsmokers as controls and 111 COPD patients, 214 genes showed differential expression. R programming languages were used to identify common DEGs from these two datasets. The PPIs network is recognized by common DEGs, KEGG pathways and pharmacological signatures.*

B. Identifying DEGs and reciprocal DEGs among IPF and COPD and Lung Cancer

If a result is statistically significant variation among several tests' circumstances at the transcription level, a gene is said to be expressed differentially [8]. Finding DEGs for the datasets GSE24206, GSE18842, and GSE76925 is the main goal of this study. The most common DEGs could potentially be found with the use of a programming language for computers called R. To find statistically significant DEGs in each of the datasets, a pair of threshold parameters were used: for up-regulated and down-regulated consider $|logFC|>1$, and an Adjusted p-value<0.05. Using the online venn analysis program, the common DEGs of GSE24206, GSE18842, and GSE76925 were obtained.

C. Examining the protein-protein interaction network

PPI interaction is widely acknowledged as the main area of the biology of the cell study and as a condition for studying system biology. PPIs (interplay between proteins) shed light on how proteins fulfill their biological functions [12]. PPIs are used to describe the molecular interactions composed of two or more proteins and supported by biochemical, hydrophobic, and electrostatic elements. PPI networks offer a plethora of novel knowledge on how proteins operate. With the help of NetworkAnalyst, we created PPI networks Utilizing the DEG proteins' physical links found from the String database [13]. Making use of tools like Cytoscpae (https://cytoscape.org/), the PPI network may be seen more clearly. PPI analysis employing topological properties was used to find significantly interacting hub proteins at the degree (greater than 13°).

D. Hub gene identification

A hub gene is one that has the most links to other genes, as well as a description of a protein-protein interaction network is one that displays the interconnections between the proteins by the use of edges and nodes. the genes that are functioning in this research project with the aid of the degree topological method. PPI networks are currently being studied using Cytoscape. Researchers are able to using the cytoHubba plugin for Cytoscape (http://apps.cytoscape.org/apps/cytohubba) to find the genes that make up the particular PPI network [14].

E. Pathway Analysis

Gene set enrichment analysis is focused on gene sets with specific chromosomal regions and high biological function. Understanding metabolic pathways as well as having a significant influence on gene annotation is the KEGG pathway [15]. Enrichr, a web application available online, provided all the pathways about the shared genes identified in the first step (https://amp.pharm.mssm.edu/Enrichr/). For the investigation of cellular pathways, the KEGG databases are used. Additionally, the Enrichr [16] platform is used to examine the results from the databases.

F. Identification of drug signatures

Currently conducted research is fundamentally dependent on the identification of new therapeutic molecules. The 22527 gene sets in the DSigDB [17] database are used to create the therapeutic compound. DSigDB considers each gene set while taking a chemical into account and largely depends drug predictions are made using gene expression-based databases.

III. Results

A. The location of a gene common to IPF, COPD and lung cancer, as well as the detection of DEGs

With the help of the computer language R, we examined the 693 genes for IPF, the 2338 genes for lung cancer, and the 597 genes for COPD and found 8 common DEGs. In the shape of a Venn diagram, Figure 2 analyzes the volume of DEGs in three separate data sets.

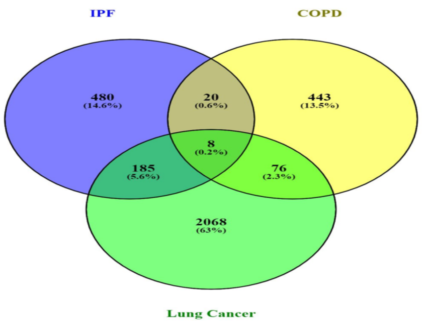

Fig. 2. *DEGs associated with IPF, COPD and lung cancer sample sets are shown graphically. Differentially expressed genes (DEGs) are present in all of the IPF, COPD, and lung cancer datasets; the IPF dataset has 693 DEGs, the COPD dataset contains 597 DEGs, and the lung cancer dataset contains 2338 DEGs; 8 genes were determined to be similar. Only 8 genes were shared by all three datasets out of the 3628 genes exhibiting differential expression.*

B. PPIs network analysis

The input used by NetworkAnalyst was the common DEGs. In this specific study, the hub genes and PPIs network are both analyzed. The NetworkAnalyst web-based application and the network diagram's Simple Interaction Format (SIF) files are created using the STRING interatomic database. The PPIs network is composed of common DEGs and consists totaling 344 nodes, 366 edges, and 8 seeds. Figure 3 shows a visual representation of the PPI network.

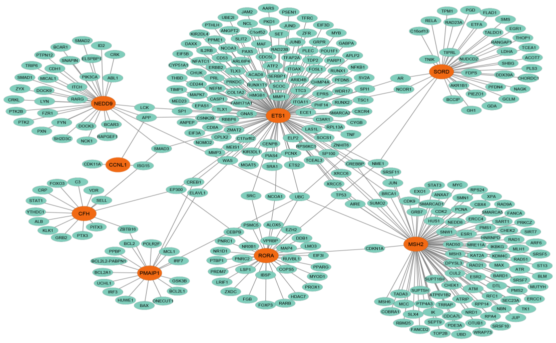

Fig. 3. *PPI network for identified concordant DEGs shared by IPF, COPD, and LC.*

C. Hub genes' findings

Simply described, hub nodes are the nodes in a network with the greatest number of connections. Following a cytohubba study of the PPIs network for further investigation, the top eight active genes are identified. The top eight genes are CFH, MSH2, SORD, NEDD9, CCNL1, RORA, ETS1, and PMAIP1. ETS1 sticks out from the rest it has the most degrees, hence of any DEG in the network. The PPI network is investigated using The Network Analyzer on Cytoscape displays its topological properties. Table I gives the outcomes for the top five topological characteristics. of the genes.

TABLE I
WE USED CYTOSCAPE TO EVALUATE THE TOPOLOGICAL PROPERTIES OF THE FIVE MAIN CRITICAL HUB GENES.

Hub gene	*Degree*	*Stress*	*Closeness Centrality*	*Betweenness Centrality*
ETS1	133.0	39372.0	204.16667	72812.13461
MSH2	95.0	305904.0	176.6667	51293.265
SORD	35.0	49354.0	136.6667	20606.21815
RORA	34.0	59786.0	137.65	19531.40852
NEDD9	33.0	48032.0	137.5	19569.67527

D. Pathway Analysis

In order to understand more about the biological processes involved, KEGG databases have been used to identify DEGs shared by IPF, COPD, and lung cancer. The Enrichr uses the p-value and the log of the z-score to create a composite

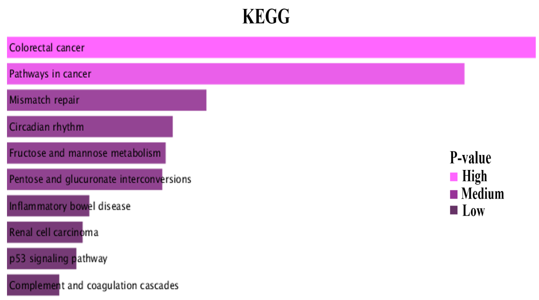

Fig. 4. *Effective biochemical data result from KEGG pathway P-value analysis.*

score. The KEGG pathway database is shown in Figure 4 is organized according to p-value.

E. Drug Compound Identification

The drug compounds are obtained from the DSigDB database via the online Enrichr program. Based on the p-value and changed p-value, it was anticipated that the following suggested drugs will be taken. The following table contains common DEGs that can be applied as therapeutic agents that are identified in IPF, COPD, and lung cancer. Based on the most popular DEGs, TABLE II displays the most potent medication molecules.

TABLE II
THE COMMON DEGS OF IPF, COPD, AND LUNG CANCER-RELATED DRUG SUGGESTED COMPOUNDS.

Name of drugs	*p-value*	*Adjusted p-value*	*Genes*
astemizole MCF7 UP	0.0000345	0.00951077	NEDD9; PMAIP1; CCNL1
calmidazolium MCF7 UP	0.0000401	0.00951077	NEDD9; PMAIP1; CCNL1
ivermectin MCF7 UP	0.0000926	0.009510477	NEDD9; PMAIP1
DICHLOROMETHANE CTD 00006313	0.0001631	0.009510477	NEDD9; PMAIP1
ETHYLBENZENE CTD 00000178	0.0001699	0.009510477	NEDD9; PMAIP1

IV. Discussion

IPF, COPD, and lung cancer (LC) have a considerable influence on global health. Lung cancer's widespread lethality, COPD's airflow restriction, and IPF's deadly lung deterioration put a burden on healthcare systems and economies. A better understanding of their molecular causes offers promise for more accurate diagnosis, individualized care, and general management, reducing their severe ramifications. When we filtered the genes in GSE24206, GSE18842, and GSE76925, we discovered 8 DEGs that were related. Following the discovery of the shared genes, a PPIs network analysis was carried out. The PPIs network that this study is looking at will get explained next. This study's inquiry of the PPIs network produced the discovery of hub genes, which in turn led to the discovering of the top 5 hub genes. Based on that finding, the PPIs network used a degree topological look at to discover 10 key hub genes. Several bioinformatics studies have been performed to locate targeted pharmaceuticals and cluster identical DEGs. With the indicated medications, IPF, COPD, and lung cancer therapy studies should be effective.

V. Conclusion

Transcriptome analysis has not been previously investigated for IPF, COPD, or lung cancer. To pinpoint the crucial pathways and biomolecules implicated in IPF, COPD, and lung cancer, we applied a system biology and functional enrichment method. By evaluating the IPF, COPD, and lung cancer microarray datasets GSE24206, GSE18842, and GSE76925 appropriately, we were able to pinpoint DEGs that were both up-regulated and down-regulated. The filtering of genes, one of the essential processes in systems biology, is carried out through a number of bioinformatics techniques.The identification of regular therapy drugs for IPF, COPD, and lung cancer is thus possible through gene comparison. We developed PPI networks according to common DEGs and, considering their degree values, found 5 hub genes (ETS1, MSH2, SORD, RORA, and NEDD9). We may deduce that the hub genes are particularly harmful when three disorders have identical differentially expressed genes (DEGs), which in turn aids in identifying the most effective pharmaceutical compounds for therapy. We provide predictions about the top five biomarkers that could facilitate the creation of therapeutic substances linked to IPF, COPD, and lung cancer. Unique biological markers for IPF, COPD and lung cancer have been found, and management guidelines have been provided.

ACKNOWLEDGMENT

There are no plans to publish this paper anywhere else, nor has it been submitted to any publications. All of the participants in this study are overjoyed to have the chance to contribute both their time and their knowledge.

References

[1] Fathinavid, A., Mousavian, Z., Najafi, A., Nematzadeh, S., Salimi, M. and Masoudi-Nejad, A., 2022. Identifying common signatures and potential therapeutic biomarkers in COPD and lung cancer using miRNA-mRNA co-expression networks. Informatics in Medicine Unlocked, 34, p.101115.

[2] Yang, I.A., Relan, V., Wright, C.M., Davidson, M.R., Sriram, K.B., Savarimuthu Francis, S.M., Clarke, B.E., Duhig, E.E., Bowman, R.V. and Fong, K.M., 2011. Common pathogenic mechanisms and pathways in the development of COPD and lung cancer. Expert opinion on therapeutic targets, 15(4), pp.439-456.

[3] Zhang, F., Chen, X., Wei, K., Liu, D., Xu, X., Zhang, X. and Shi, H., 2017. Identification of key transcription factors associated with lung squamous cell carcinoma. Medical science monitor: international medical journal of experimental and clinical research, 23, p.172.

[4] Gagnat, A.A., Gjerdevik, M., Lie, S.A., Gulsvik, A., Bakke, P. and Nielsen, R., 2020. Acute exacerbations of COPD and risk of lung cancer in COPD patients with and without a history of asthma. European Clinical Respiratory Journal, 7(1), p.1799540.

[5] Wang, H., Yang, L., Zou, L., Huang, D., Guo, Y., Pan, M., Tan, Y., Zhong, H., Ji, W., Ran, P. and Zhong, N., 2012. Association between chronic obstructive pulmonary disease and lung cancer: a case-control study in Southern Chinese and a meta-analysis.

[6] Young, R.P., Hopkins, R. and Eaton, T.E., 2007. Forced expiratory volume in one second: not just a lung function test but a marker of premature death from all causes. European Respiratory Journal, 30(4), pp.616-622.

[7] Ghosh, A.J., Hobbs, B.D., Yun, J.H., Saferali, A., Moll, M., Xu, Z., Chase, R.P., Morrow, J., Ziniti, J., Sciurba, F. and Barwick, L., 2022. Lung tissue shows divergent gene expression between chronic obstructive pulmonary disease and idiopathic pulmonary fibrosis. Respiratory research, 23(1), pp.1-14.

[8] Mahmud, S.H., Al-Mustanjid, M., Akter, F., Rahman, M.S., Ahmed, K., Rahman, M.H., Chen, W. and Moni, M.A., 2021. Bioinformatics and system biology approach to identify the influences of SARS-CoV-2 infections to idiopathic pulmonary fibrosis and chronic obstructive pulmonary disease patients. Briefings in Bioinformatics, 22(5), p.bbab115.

[9] Leng, D., Yi, J., Xiang, M., Zhao, H. and Zhang, Y., 2020. Identification of common signatures in idiopathic pulmonary fibrosis and lung cancer using gene expression modeling. BMC cancer, 20(1), pp.1-15.

[10] Hwang, J.A., Kim, D., Chun, S.M., Bae, S., Song, J.S., Kim, M.Y., Koo, H.J., Song, J.W., Kim, W.S., Lee, J.C. and Kim, H.R., 2018. Genomic profiles of lung cancer associated with idiopathic pulmonary fibrosis. The Journal of pathology, 244(1), pp.25-35.

[11] Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P. and Jensen, L.J., 2016. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic acids research, p.gkw937.

[12] Basar, M.A., Hosen, M.F., Al Amin, M., Bithi, N.I. and Paul, B.K., 2022, December. A bioinformatics and system biology technique to identify candidate biomarkers and functional pathways among stress and depression. In 2022 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE) (pp. 143-146). IEEE.

[13] 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.

[14] 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(4), pp.1-7.

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

[16] M. v. Kuleshov et al., “Enrichr: a comprehensive gene set enrichment analysis web server 2016 update,” Nucleic Acids Res, vol. 44, no. 1, pp. W90–W97, Jul. 2016, doi: 10.1093/nar/gkw377.

[17] 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.