Contents lists available at ScienceDirect

International Immunopharmacology

journal homepage: www.elsevier.com/locate/intimp

Shared biomarkers and mechanisms in idiopathic pulmonary fibrosis and non-small cell lung cancer

Xiaorui Ding [a], Huarui Liu [a], Qinghua Xu [a], Tong Ji [a], Ranxun Chen [a], Zhengcheng Liu [b,*], Jinghong Dai [a,**]

[a] *Department of Pulmonary and Critical Care Medicine, Nanjing Drum Tower Hospital, Affiliated Hospital of Medical School, Nanjing University, Nanjing 210008, Jiangsu, China*
[b] *Department of Thoracic Surgery, Nanjing Drum Tower Hospital, Affiliated Hospital of Medical School, Nanjing University, Nanjing 210008, Jiangsu, China*

ARTICLE INFO

Keywords:
Idiopathic pulmonary fibrosis
Non-small-cell lung cancer
Extracellular matrix
Bioinformatics

ABSTRACT

Background: Epidemiological evidence has indicated the occurrence of idiopathic pulmonary fibrosis (IPF) with coexisting lung cancer is not a coincidence. The pathogenic mechanisms shared between IPF and non-small cell lung cancer (NSCLC) at the transcriptional level remain elusive and need to be further elucidated.
Methods: IPF and NSCLC datasets of expression profiles were obtained from the GEO database. Firstly, to detect the shared dysregulated genes positively correlated with both IPF and NSCLC, differentially expressed analysis and WGCNA analysis were carried out. Functional enrichment and the construction of protein–protein network were employed to reveal pathogenic mechanisms related to two diseases mediated by the shared dysregulated genes. Then, the LASSO regression was adopted for screening critical candidate biomarkers for two disorders. Moreover, ROC curves were applied to evaluate the diagnostic value of the candidate biomarkers in both IPF and NSCLC.
Results: The 20 shared dysregulated genes positively correlated with both IPF and NSCLC were identified after intersecting differentially expressed analysis and WGCNA analysis. Functional enrichment revealed the 20 shared genes mostly enriched in extracellular region, which is critical in the organization of extracellular matrix. The protein–protein networks unrevealed the interaction of the 11 shared genes involving in collagen deposition and the connection between PYCR1 with PSAT1. PSAT1, PYCR1, COL10A1 and KIAA1683 were screened by the LASSO regression. ROC curves comprising area under the curve (AUC) verified the potential diagnostic value of PSAT1 and COL10A1 in both IPF and NSCLC.
Conclusions: We revealed dysregulated extracellular matrix through aberrant expression of the relevant genes, which provided further understanding for the common molecular mechanisms predisposing the occurrence of both IPF and NSCLC.

1. Introduction

Idiopathic pulmonary fibrosis (IPF), the most common idiopathic interstitial pneumonia, is a chronic and fatal fibrotic lung disease with unknown etiology, with higher incidence in older adults whose median age at diagnosis is about 65 years and the median survival time ranging 3–5 years after diagnosis [1,2]. Patients often appear respiratory failure at the end stage. In recent years, the incidence and mortality of IPF have risen rapidly worldwide [2,3]. Several comorbidities can occur [4–7], among which patients with IPF have a risk nearly five times as high as that of the general population to develop lung cancer (LC) [8,9] primarily developing in periphery adjacent to fibrotic areas of patients with IPF. Additionally, epidemiological evidence suggests that the incidence of lung cancer developed in patients with IPF ranges from 3 % to 22 % [10,11], indicating there is no doubt about the association of IPF and LC.

The occurrence of LC is associated with the poor prognosis of IPF

* Corresponding author at: Department of Thoracic surgery, Nanjing Drum Tower Hospital, Affiliated Hospital of Medical School, Nanjing University, Nanjing 210008, Jiangsu, China.
** Corresponding author at: Department of Pulmonary and Critical Care Medicine, Nanjing Drum Tower Hospital, Affiliated Hospital of Medical School, Nanjing University, Nanjing 210008, Jiangsu, China
E-mail addresses: lzclzc0928@foxmail.com (Z. Liu), daijinghong@nju.edu.cn (J. Dai).

https://doi.org/10.1016/j.intimp.2024.112162
Received 25 March 2024; Received in revised form 21 April 2024; Accepted 24 April 2024
Available online 3 May 2024

with a reduced median survival time (1.6–1.7 years) compared with IPF patients without LC [12]. There is no consensus statement raising the management and treatment for patients with this disease combination, thus early diagnosis through appropriate screening algorithms and detective methods is vital. High-resolution CT (HRCT) suggested by the comment to monitor the progression of IPF is difficult to detect precancerous lesion [10,11]. The awareness of the pathogenic mechanisms of the coexistence of IPF and LC is essential to raise for excavating effective management and therapy of patients with both disorders. While recent studies have illustrated that multiple-overlapping mechanisms, including common genetic mutants, epigenetic alterations and activation of signal transduction pathways, involve the occurrence of IPF and LC [11,13–15], how abnormal interstitial lung tissues produce cancerous lesion remains enigmatic. More identifications of molecular mechanisms promoting carcinogenesis within fibrotic lung are needed to further elucidate the coexistence of two lethal lung diseases.

In this context, we investigated links between IPF and NSCLC, comprising biological mechanisms involved by differentially expressed genes in two lung disorders' tissue compared with normal tissues, which was based on comprehensive bioinformatics analysis and machine learning of gene expression profile from lung disorders' tissue sample.

2. Materials and methods

2.1. Data collection and processing

The workflow chart of this study is presented in Fig. 1. We searched for IPF and NSCLC patients' gene expression profiles from Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/) [16] using the term "Idiopathic pulmonary fibrosis" or "non-small cell lung cancer". The inclusion criteria were as follows: (1) Homo sapiens; (2) Expression profiling by array and profiles should consist of case and control groups; (3) Samples should be derived from lung tissue samples; (4) The number of samples in each dataset must be greater than 30. Ultimately, GSE10667 (comprising human lung tissue from 15 normal individuals and 31 patients with idiopathic pulmonary fibrosis) and GSE21933 (comprising human lung tissue from 21 normal individuals and 21 patients with non-small cell lung cancer) were chosen as training sets for the next research. GSE53845 (comprising human lung tissue from 8 normal individuals and 40 patients with idiopathic pulmonary fibrosis) and GSE18842 (comprising human lung tissue from 45 normal individuals and 46 patients with non-small cell lung cancer) were chosen as validating sets for further research. We downloaded the four GEO datasets' series matrix files using "GEOquery" packages [17] in R software (version 4.3.1) for the next analysis. The detailed description of datasets was shown in Table 1.

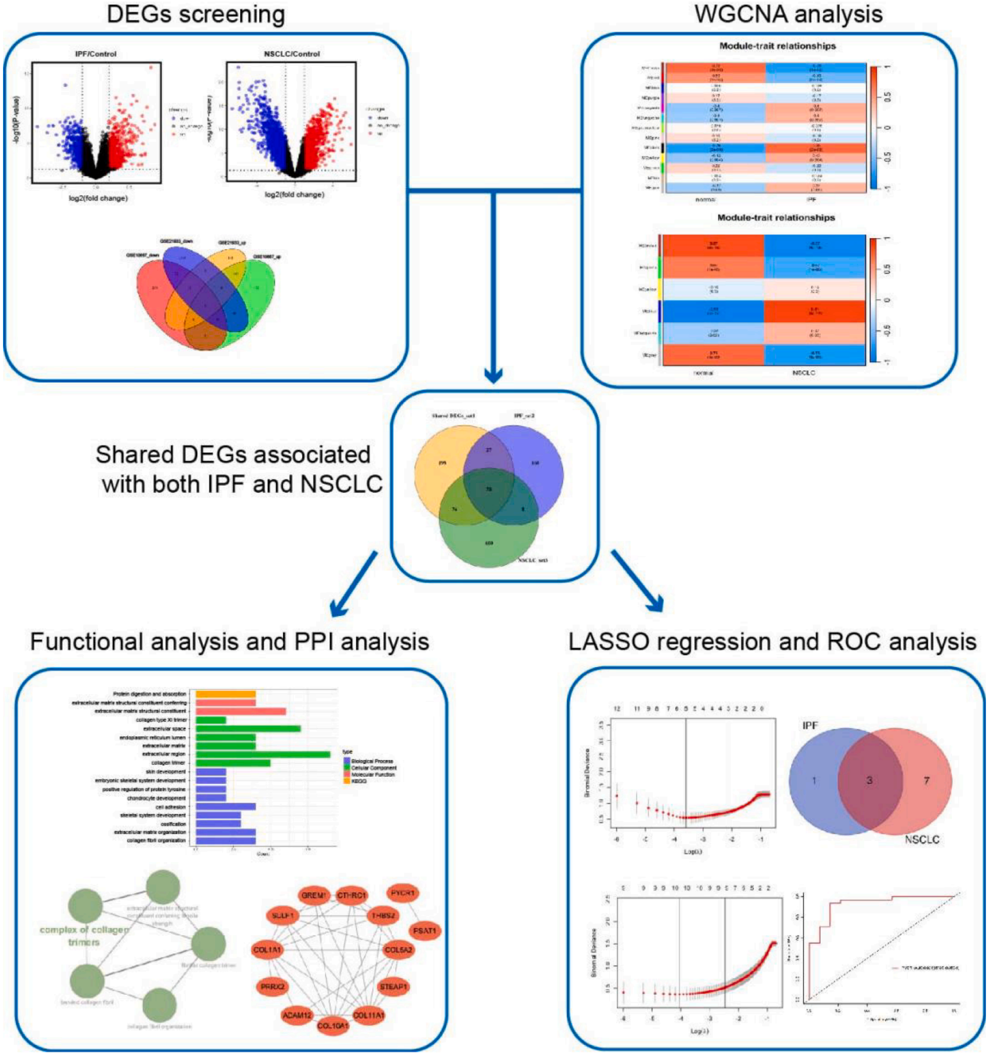

Fig. 1. The workflow chart of this study.

Table 1
Information of GEO datasets containing the IPF/NSCLC patients.

GSE number	Platform	Samples	Disease	Groups
GSE10667	GPL4133	31 patients and 15 controls	IPF	Training set
GSE21933	GPL6254	21 patients and 21 controls	NSCLC	Training set
GSE53845	GPL6480	40 patients and 8 controls	IPF	Validating set
GSE18842	GPL570	46 patients and 45 controls	NSCLC	Validating set

Table 2
GO and KEGG analysis terms of 20 shared genes.

Term	Description	Count	P Value	Gene symbol
GO:0030199	collagen fibril organization	4	3.77E−05	GREM1, COL1A1, COL11A1, COL5A2
GO:0030198	extracellular matrix organization	4	5.33E−04	COL1A1, COL11A1, COL5A2, COL10A1
GO:0007155	cell adhesion	4	0.012	COL1A1, CGREF1, ADAM12, THBS2
GO:0061098	positive regulation of protein tyrosine kinase activity	2	0.026	GREM1, DOK7
GO:0005201	extracellular matrix structural constituent	6	7.81E−08	COL1A1, COL11A1, COL5A2, COL10A1, THBS2, CTHRC1
GO:0030020	extracellular matrix structural constituent conferring tensile strength	4	6.86E−06	COL1A1, COL11A1, COL5A2, COL10A1
hsa04974	Protein digestion and absorption	4	1.39E−04	COL1A1, COL11A1, COL5A2, COL10A1

2.2. Differentially expressed genes (DEGs) analysis

The identification of DEGs for the datasets GSE10667 and GSE21933 was carried out using “limma” packages [18] in R software (version 4.3.1) after normalization between arrays. Cutoff criteria (P-value $<$ 0.05 and $|\text{log2FC}| > 1$) was applied to detect significant DEGs from the two datasets. Subsequently, the expression patterns of DEGs were visualized in the form of volcano plots and the top20 genes heatmaps with the “ggplot2” package and “pheatmap” package in R software, respectively. The shared DEGs associated with IPF and NSCLC were acquired using online Venn diagram (https://bioinformatics.psb.ugent.be/webtools/Venn/) and visualized with “VennDiagram” package [19].

2.3. Weighted gene co-expression network analysis

Weighted gene co-expression network analysis (WGCNA), a systematic biology method, can be used for detecting modules of correlated genes and for relating gene sets to samples’ phenotype [20]. In this study, key modules of correlated genes mostly associated with IPF or NSCLC were found with “WGCNA” package in R software. Firstly, we checked the missing values for removing the offending genes and samples and clustered all samples for judging and detecting outlier samples. Secondly, the optimal soft threshold power was established in accordance with the scale-free topology criterion in this experiment. Flowing this, the topological overlap matrix (TOM) was constructed to detect gene connectivity, and modules expressing similarly were merged on the basis of mergeCutHeight = 0.25 and 30 as the gene dendrogram’s minimum size. Then, we analyzed and established the correlation of modules and traits for screening key modules associated with disease. Finally, using Venn diagram to take intersections of genes of key modules and shared DEGs between IPF with NSCLC to screen shared DEGs mostly both related to the two diseases.

2.4. Functional enrichment analysis

Gene ontology (GO) can functionally categorize and annotate genes according to the fundamental roles of cellular component (CC), biological process (BP) and molecular function (MF) in organism [21]. The Kyoto Encyclopedia of Genes and Genomes (KEGG) integrates a large amount of functional information about biological pathways and allows pathway enrichment analysis of genes [22]. To explore the shared pathways of IPF and NSCLC, we imported gene list screened through forward analysis into DAVID database (https://david.ncifcrf.gov/) [23] to carry out GO and KEGG analysis. The P-value $<$ 0.05 was considered as the significant term. Using ClueGO plugin in Cytoscape (3-10-1) to categorize the significant GO terms and KEGG pathways.

2.5. Protein-protein interaction network analysis

The foundation and analysis of Protein-protein interaction networks, composed of proteins that participate in various aspects of biological processes such as biological signaling, regulation of gene expression, energy and material metabolism, and cell cycle regulation, can form the full understanding of the cellular machinery and biological phenomena. The STRING (version 11.0) (https://string-db.org/) [24], which can construct a comprehensive PPI network from the aspects of physical as well as functional interactions, was used to form the PPI network of the shared genes screened from forward steps. The confidence score threshold was kept at 0.4. The PPI networks were visualized using Cytoscape.

2.6. Machine learning and receiver operating characteristic (ROC) curves of hub genes

To further screen the candidate genes of patients with IPF and NSCLC, Least Absolute Shrinkage and Selection Operator (LASSO) was employed, a logistic regression approach to facilitate variable selection, which was conducted using the R package “glmnet” [25,26]. The receiver operating characteristic curve (ROC) was carried with R package “pROC” [27] to assess the diagnostic values of hub genes for IPF and NSCLC, respectively. Area under the curve (AUC) was calculated to determine the sensitivity and accuracy of hub genes in the diagnosis of diseases. In addition, ROC analysis was performed on the validating sets to further verify the diagnostic values of hub genes.

3. Results

3.1. Data processing and identification of shared DEGs associated with IPF and NSCLC

The datasets GSE10667 and GSE21933 were firstly normalized between arrays, respectively, which were visualized in Fig. 2A–D. Then, as shown in Fig. 3A, differential analysis between IPF patients and normal samples revealed 658 upregulated and 428 downregulated genes with cutoff criterion of P-value $<$ 0.05 and $|\text{log2FC}| > 1$. 1341 upregulated and 1570 downregulated genes between NSCLC patients and normal samples were shown in Fig. 3B. To better demonstrate the differences between the normal and disease groups, we show the top20 DEGs between two groups with heatmaps (Fig. 3C and D). The shared DEGs associated with IPF and NSCLC, containing 148 upregulated and 173 downregulated genes, were exhibit with the Venn diagram (Fig. 3E).

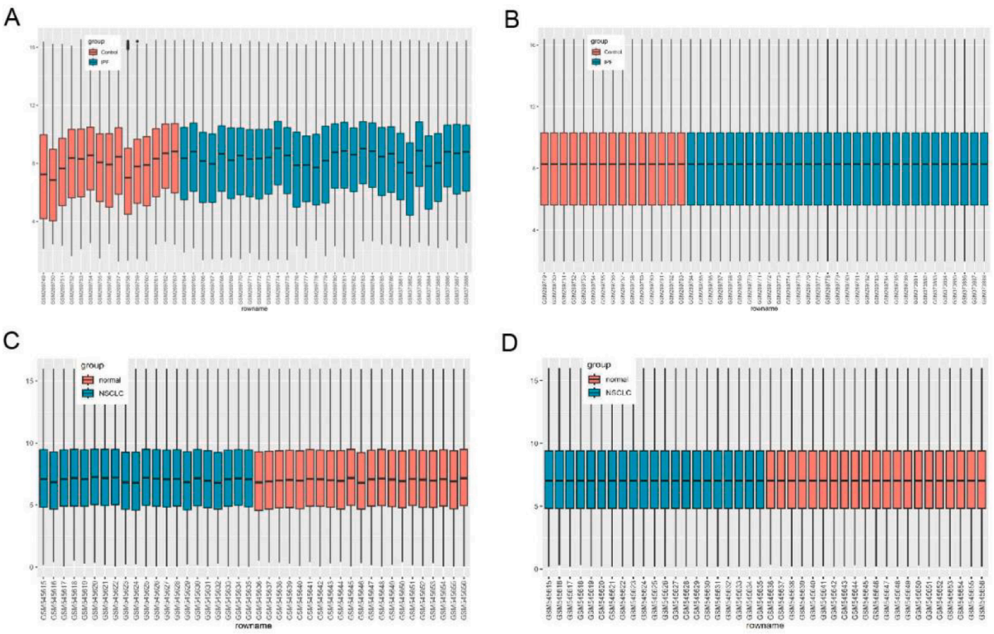

Fig. 2. The box-plots of two datasets. (A) The box-plot of GSE10667 before normalizing between arrays. (B) The box-plot of GSE10667 after normalizing between arrays. (C) The box-plot of GSE21933 before normalizing between arrays. (D) The box-plot of GSE21933 after normalizing between arrays.

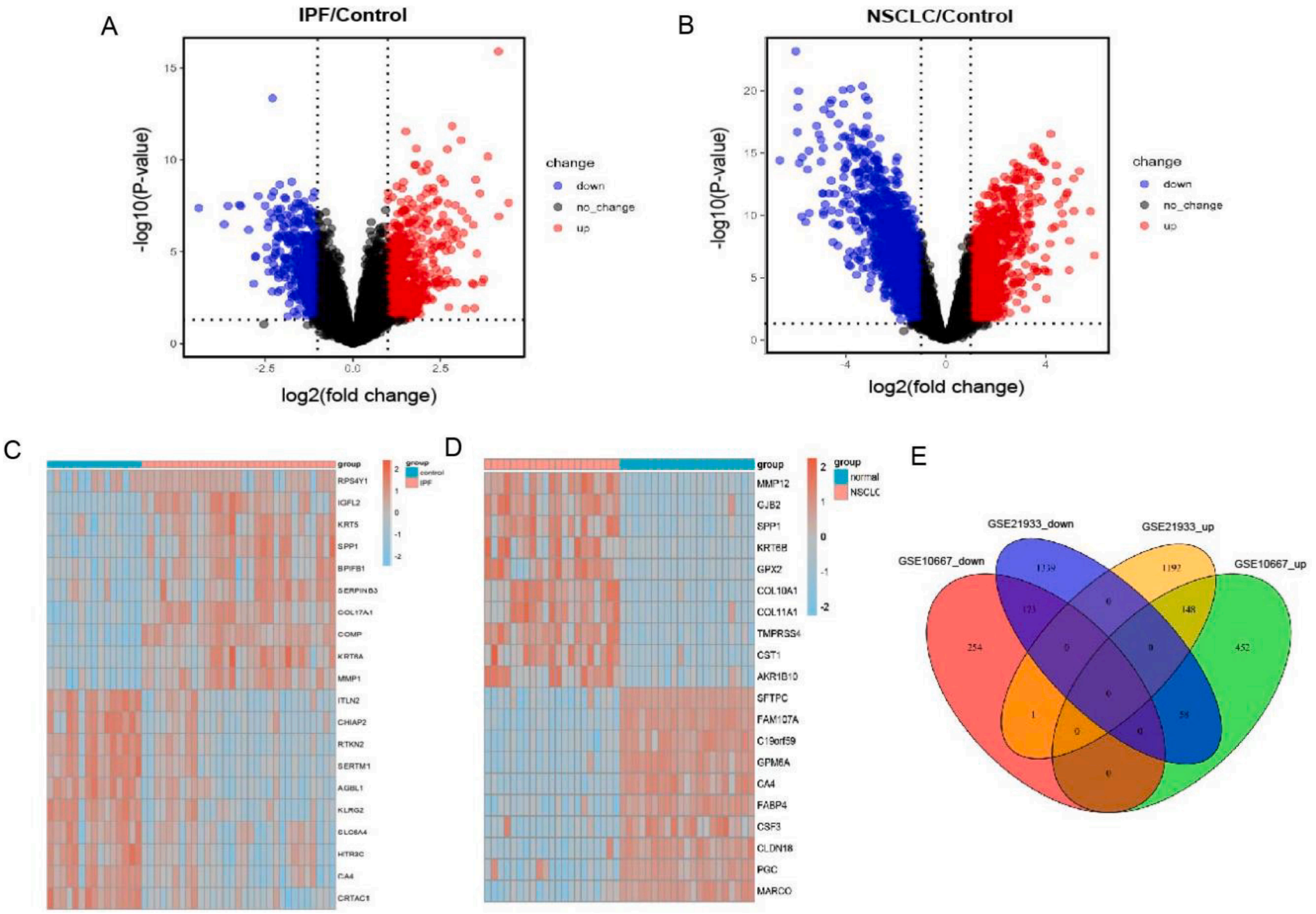

Fig. 3. Screening of shared dysregulated genes of IPF and NSCLC. (A) The volcano plot presenting IPF DEGs in GSE10667. (B) The volcano plot presenting NSCLC DEGs in GSE21933. (C) The heatmap presenting top20 IPF DEGs in GSE10667. (D) The heatmap presenting top20 NSCLC DEGs in GSE21933. (E) The Venn diagram presenting the shared dysregulated genes including 148 upregulated genes and 173 downregulated genes of IPF and NSCLC.

3.2. The construction of weighted gene co-expression network and the identification of key modules in IPF and NSCLC

The weighted gene co-expression network analysis (WGCNA) was carried out to identify the key modules most relevant with IPF or NSCLC, respectively. As shown in Fig. 4A and B, after removing outlier samples and choosing the soft-thresholding power of 6, totally 13 modules associated with IPF were generated and presented using the cluster dendrogram. The correlation between modules with IPF was performed with the heatmap and the Spearman's correlation coefficient. The MEblack module with 223 genes was found to be the most positively relevant to IPF ($r = 0.76$, $p = 2e{-}09$). The MEbrown module with 649 genes was found to be the most negatively correlated with IPF ($r = -0.73$, $p = 1e{-}08$) (Fig. 4C). The MEblack was selected as key module of GSE10667 for further analysis. No samples in GSE21933, comprising 21 NSCLC samples and normal samples, were found to be outliers (Fig. 4D and E). Similarly, after the soft-thresholding power of 14 was chosen, 6 modules associated with NSCLC were generated. The MEblue module with 563 genes was most positively correlated with NSCLC ($r = 0.91$, $p = 5e{-}17$). In contrast, the MEbrown module with 502 genes was found to be the most negatively with NSCLC ($r = -0.87$, $p = 6e{-}14$) (Fig. 4F). The MEblue module of GSE21933 was used to further study. In order to identify the shared DEGs mostly both related to two diseases, we take the intersection of the MEblack module of GSE10667 and the MEblue module of GSE21933 with the shared DEGs. Lastly, 20 shared DEGs associated with IPF and NSCLC were obtained, which was consisted of 17 upregulated and 3 downregulated genes (Fig. 5A and Fig. 7A and B).

3.3. The enrichment analysis of shared DEGs from IPF and NSCLC

To explore the shared pathways of IPF and NSCLC, we imported 20 shared DEGs associated with IPF and NSCLC into DAVID online database to take GO and KEGG analysis, which can identify the potential common pathogenesis of two diseases. The results revealed that biological progress (BP) of GO enrichment in several main terms, such as "collagen fibril organization", "extracellular matrix organization", "cell adhesion", "chondrocyte development" and "positive regulation of protein tyrosine kinase activity". Cellular component (CC) of GO analysis indicated that 20 shared genes were mainly located in "collagen trimer", "extracellular region", "extracellular matrix", "endoplasmic reticulum lumen", "extracellular space", "collagen type XI trimer". Molecular function (MF) showed that "extracellular matrix structural constituent" and "extracellular matrix structural constituent conferring tensile strength" were significantly meaningful terms enriched by the screened genes (Fig. 5B and C). KEGG analysis revealed that "protein digestion and absorption" was the mostly related pathways enriched by the screened genes (Fig. 5B).

3.4. The construction of protein–protein interaction network of shared DEGs from IPF and NSCLC

To better understand the interactions between the proteins encoded by shared DEGs positively related to IPF and NSCLC, we imported these genes into the STRNG online database, and obtained the file of the protein–protein interaction network with a medium confidence score of >0.4. Then, the interaction network of the proteins was visualized by Cytoscape, which showed 13 nodes and 34 lines (Fig. 5D). These nodes represented proteins encoded by 13 upregulated genes in IPF and NSCLC, and lines indicated interaction between proteins.

3.5. The identification of biomarkers both IPF and NSCLC via LASSO regression algorithm and the evaluation of diagnostic value via ROC curves

To further identify and narrow down the crucial shared DEGs from

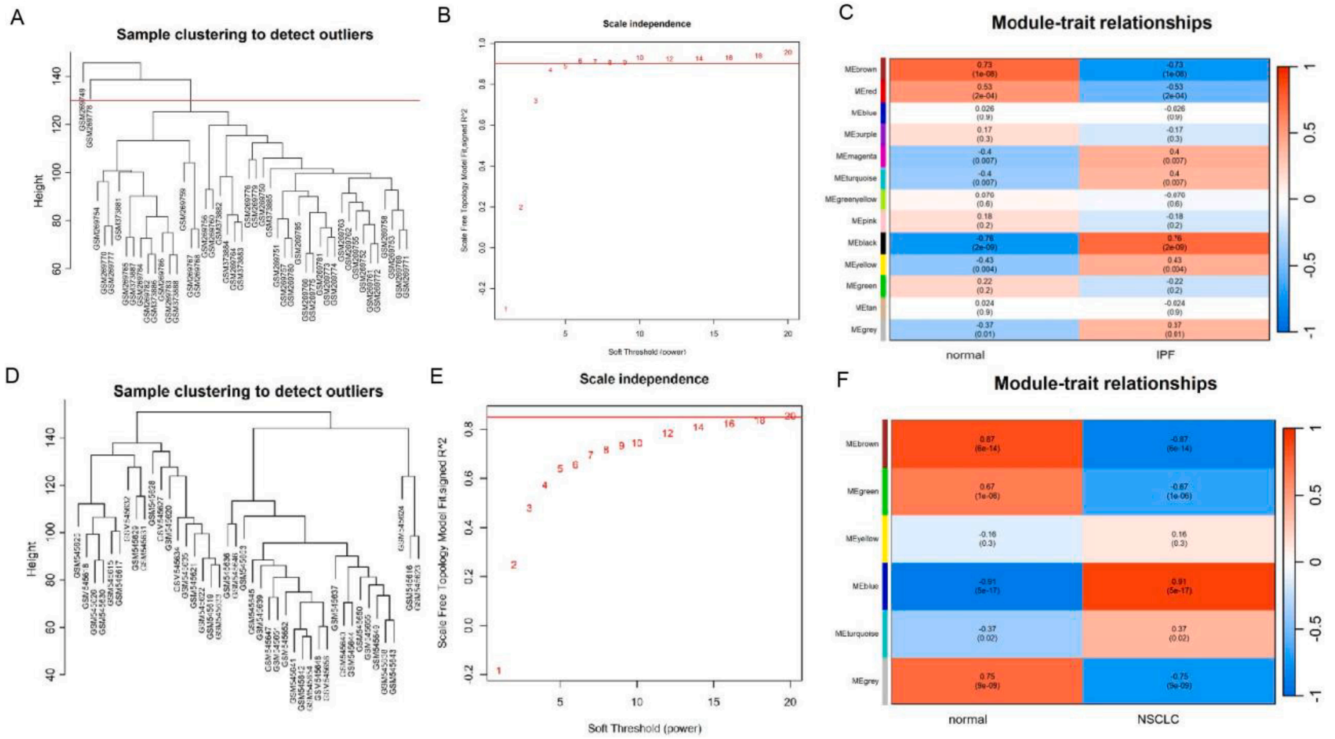

Fig. 4. WGCNA analysis for screening genes and modules positively correlated with IPF and NSCLC. (A) The sample clustering tree for detecting outliers of IPF. (B) The scale-free topology model was utilized to identify the best β value, and the $\beta = 6$ was chosen as the soft threshold based on the average connectivity and scale independence. (C) The heatmap revealing the relationship between module-traits relationship of IPF. (D) The sample clustering tree for detecting outliers of NSCLC. (E) The scale-free topology model was utilized to identify the best β value, and the $\beta = 14$ was chosen as the soft threshold based on the average connectivity and scale independence. (F) The heatmap revealing the relationship between module-traits relationship of NSCLC.

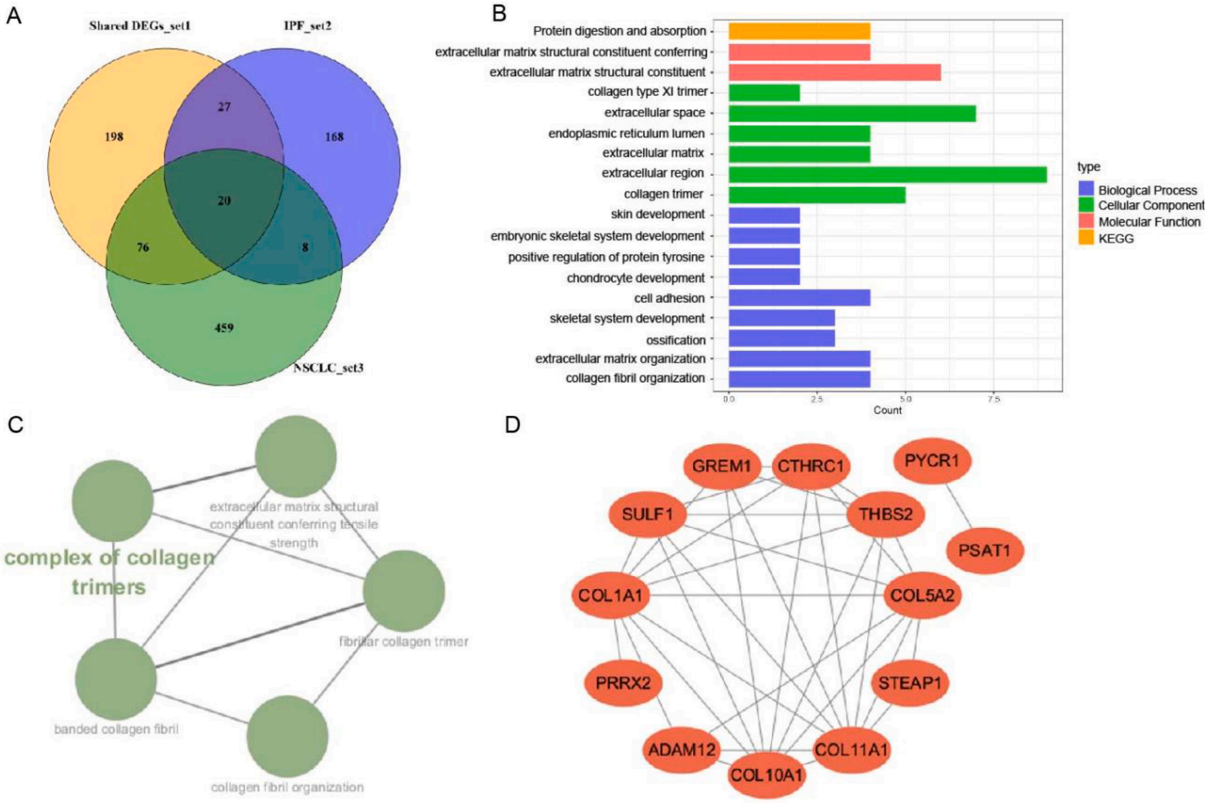

Fig. 5. The identification of shared dysregulated genes mostly associated with IPF and NSCLC, the enrichment analysis of the screened genes and the construction of PPI network. (A) The venn diagram of intersection of DEGs and WGCNA analysis. (B) The plot presenting the enrichment analysis of the screened genes. (C) The GO terms enriched by the screened genes visualized be Cytoscape. (D) The protein–protein network of the screened genes.

20 screened genes, the LASSO regression algorithm was applied. As shown in Fig. 6, four biomarkers from GSE10667 and ten biomarkers from GSE21933 were identified. Lastly, total three biomarkers were screened through taking the intersection of biomarkers from both datasets, which were PSAT1, COL10A1, KIAA1683 (Fig. 7C). PSAT1 and COL10A1 expressed highly in IPF and NSCLC. In contrast, KIAA1683 was downregulated in IPF and NSCLC (Fig. 7A and B). When performing the LASSO regression algorithm, we found that PSAT1 had the largest correlation coefficient with IPF and PYCR1 had the largest correlation coefficient with NSCLC (Table 3). In forward PPI network, we also found that PYCR1 interacted with PSAT1 (Fig. 5D). Therefore, PYCR1 was included in biomarkers for further analysis. Because the GO terms and KEGG analysis of 20 shared DEGs hadn't revealed the pathway and function involved by PYCR1 and PSAT1 (Table 2), to understand it, we imported two genes into DAVID online database again. The result showed that biological progress was "amino-acid biosynthesis" and KEGG pathway was "biosynthesis of amino acid".

To evaluate diagnostic value of four biomarkers for IPF and NSCLC, we used ROC curves based on GSE10667 and GSE21933 datasets at first. The AUCs of four hub genes demonstrated significant diagnostic value for IPF and NSCLC in two training sets (Fig. 8A and B). To further confirm the diagnostic efficacy of hub genes, we validated them using GSE53845 and GSE18842 datasets, validating sets of IPF and NSCLC, respectively. The AUC greater than 0.7 was considered having credible diagnostic values. The results were demonstrated on Fig. 8A–D, which illustrated that PSAT1(AUC = 0.921 in GSE10667, 0.791 in GSE53845, 0.961 in GSE21933 and 0.998 in GSE18842) and COL10A1 (AUC = 0.906 in GSE10667, 0.982 in GSE21933, 0.978 in GSE53845 and 0.954 in GSE18842) could act as potential diagnostic biomarkers for both IPF and NSCLC.

4. Discussion

The similarities between idiopathic pulmonary fibrosis with lung cancer, coupled with poor prognosis, have been established by abundant researches in multiple respects more than tobacco exposure [28]. However, epidemiological evidences are difficult to prove causality in both disorders, additional studies are needed to demonstrate the association through comprehension of their pathogenetic commonalities. Considering non-small cell lung cancer accounting for approximately 85 % patients with lung cancer is the most common histological subtype of lung cancer [29], this study aims to provide more common molecular mechanisms involving in incidence and development of IPF and NSCLC.

In this study, underlining the GEO datasets, 17 upregulated genes and 3 downregulated genes associated with both diseases, most of which were enriched in extracellular matrix and relevant to the organization of that, were screened through the integration of differential expression analysis and WGCNA analysis. In order to further identify hub genes, we analyzed 20 shared genes through the LASSO regression and attained 4 meaningful genes, whose diagnostic efficacy on IPF and NSCLC were assessed by ROC curves subsequently. The AUC of PSAT1 and COL10A1 in two training datasets and validating datasets were greater than 0.7 showing their potential diagnostic value for the condition of IPF and NSCLC.

Extracellular matrix (ECM), mediated by matrix-degrading enzymes, is a complex and dynamic structure composed of collagen, elastin and several glycoproteins, deregulation of which is associated with the progression of multiple diseases [30]. While the precise pathogenesis of idiopathic pulmonary fibrosis still remains elusive, one of major mechanisms already illustrated is the composition and deregulation of ECM secreted by activated myofibroblasts through serine-threonine kinase and tyrosine kinase pathways, causing structural lung damage and remodeling [1,31,32]. IPF lung myofibroblasts presenting with cancer-

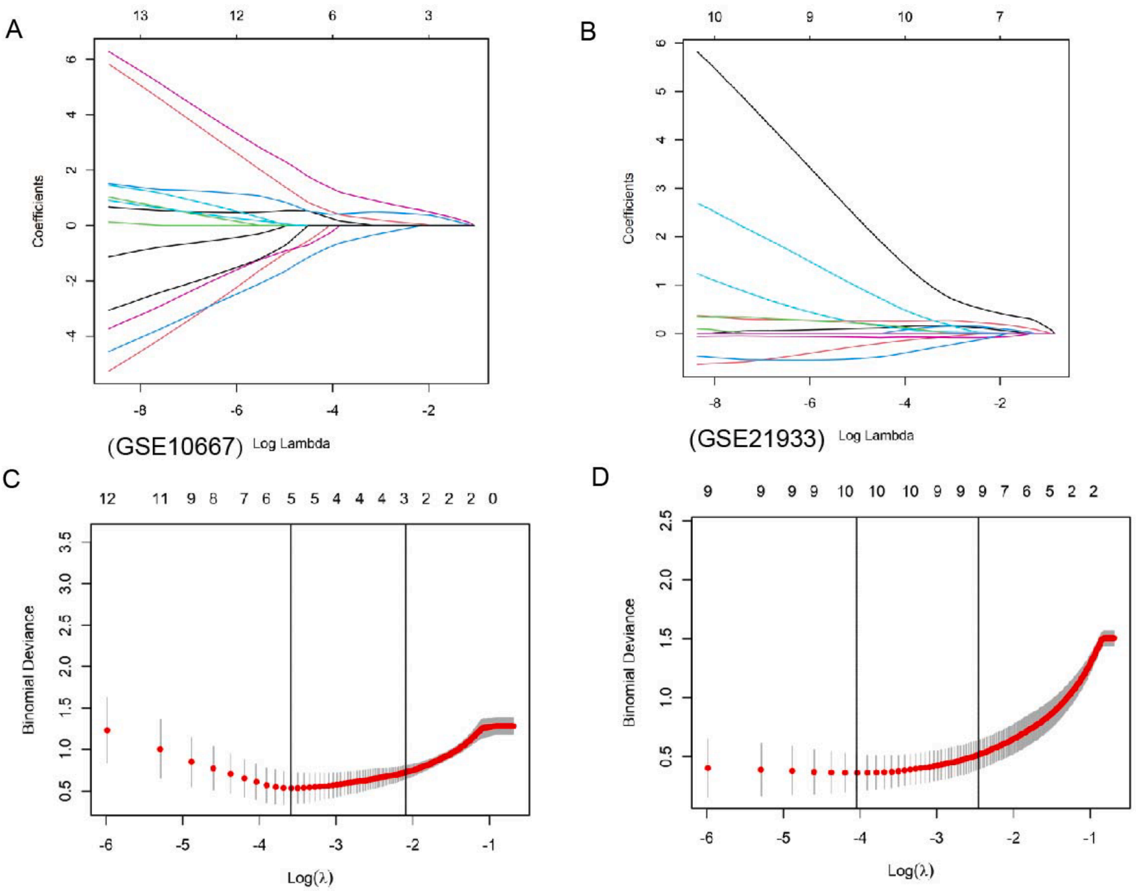

Fig. 6. LASSO regression model was established to the identification of hub genes. (A) and (B) LASSO coefficient profiles were calculated in GSE10667 and GSE21933. (C) and (D) Candidate characteristic genes from the univariate Cox regression analysis were filtered by the LASSO algorithm in GSE10667 and GSE 21933.

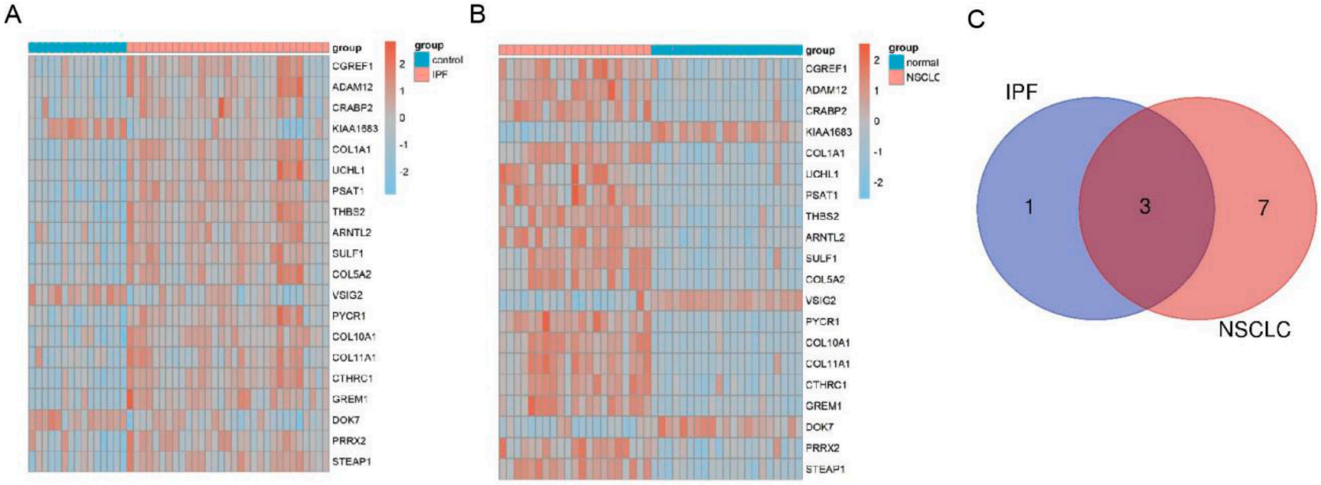

Fig. 7. The expression of 20 shared genes and the identification of hub genes. (A) The heatmap presenting the expression of the 20 screened genes in IPF dataset. (B) The heatmap presenting the expression of the 20 screened genes in NSCLC dataset. (C) The Venn diagram showing the intersected hub genes between IPF and NSCLC datasets after LASSO regression.

like characteristics including stemness potential and resistance to apoptosis, similar to cancer associated fibroblasts in lung cancer, can produce abundant ECM, which might contribute to the initiation and growth of cancer through construction of the tumor stroma [11,33–35]. Consistent with forward studies' opinion that "dysregulation of extracellular matrix" is the hallmark of the aging lung diseases [33], we found screened genes associated with ECM overexpressed in both two disorders, indicating the increased production of ECM.

Interestingly, 3 meaningful hub genes screened finally in our study, PSAT1, PYCR1 and COL10A1, are also tightly associated with the production of collagen, component of the ECM. PSAT1 and PYCR1, meaningful upregulated genes identified by LASSO regression and key

Table 3
Coefficients of genes associated with diseases through LASSO regression.

Gene symbol	Coefficients	Disease
CTHRC1	0.4384	IPF
PSAT1	1.0880	IPF
COL10A1	0.3164	IPF
KIAA1683	−0.5299	IPF
CRABP2	0.1469	NSCLC
KIAA1683	−0.1456	NSCLC
PSAT1	0.0729	NSCLC
ARNTL2	0.1000	NSCLC
VSIG2	−0.0721	NSCLC
PYCR1	1.4612	NSCLC
COL10A1	0.2548	NSCLC
GREM1	0.1360	NSCLC
DOK7	−0.4065	NSCLC
PRRX2	0.4984	NSCLC

enzymes for serine synthesis and proline synthesis [36,37], respectively, are directly linked in the PPI network, which can show the potential interaction in amino acid synthesis. Seeing that essential function of amino acid synthesis for the production of collagen by myofibroblasts, especially proline and glycine synthesis pathways, aberrant composition of collagen due to the upregulated enzymes encoded by genes PSAT1 and PYCR1, may finally induce deregulation of ECM leading to tumor development and fibrosis [30,38–40]. Additionally, there is study showing the upregulated COL10A1 remodels the ECM and promotes tumor invasion and metastasis [41,42]. To sum up, the deregulation of ECM through aberrant expression of genes encoding components and synthesis enzymes of it, plays an important role in both IPF and NSCLC, normalization of which will be a potential strategy for the treatment of this coexistence.

Metabolic reprogramming in response to the activation of lung fibroblasts and cancer cells for cell growth, proliferation and synthesis has been elucidated, in which non-essential amino acids synthesis plays a key role [37,43]. Phosphoserine aminotransferase (PSAT1), can generate 3-phosphopyruvate, an intermediate of de no serine and glycine synthesis required for collagen protein production by myofibroblasts [36]. The expression of PSAT1 was evaluated in TGF-β stimulated lung fibroblasts and Vitamin D3 can inhibit the activation of mitogen-activated protein kinase (MAPK) pathway via targeting PSAT1, which alleviated pulmonary fibrosis through decreasing collagen production [44]. Also, the serine biosynthesis pathway was identified as an important source of metabolic processes providing materials for rapid cell growth and proliferation of tumors. As the enzyme of serine biosynthesis, silencing the expression of PSAT1 can degrade cyclin D1 and block cell division in the G0/G1 phase in NSCLC, inhabiting proliferation of NSCLC cells [45]. The above studies established that PSAT1 plays an important role in the development of IPF or

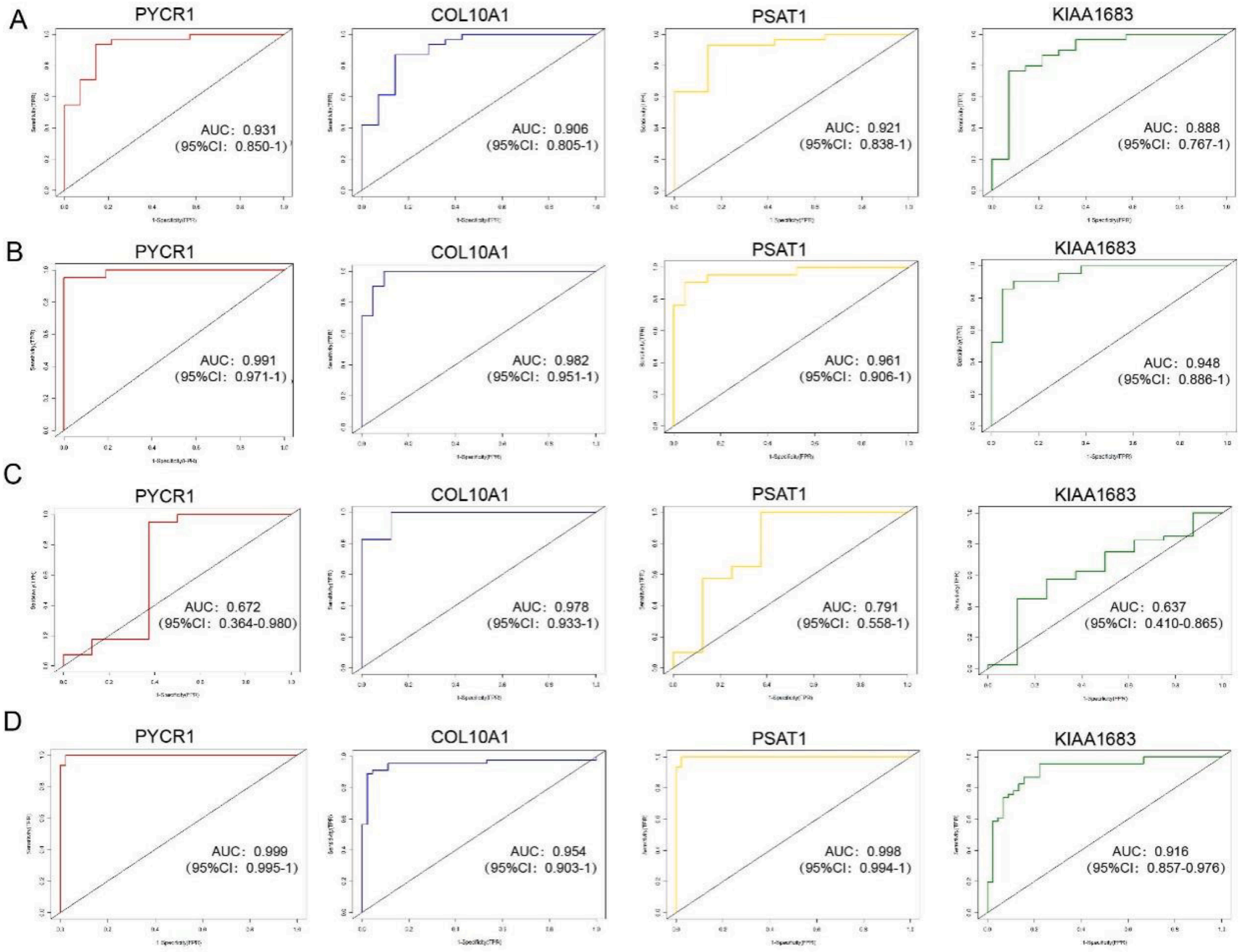

Fig. 8. ROC curves of hub genes on training and validating sets to assess and validate the diagnostic values for IPF and NSCLC. (A) The ROC curves of four hub genes on GSE10667. (B) The ROC curves of four hub genes on GSE21933. (C) The ROC curves of four hub genes on GSE53845. (D) The ROC curves of four hub genes on GSE18842.

NSCLC mainly relying on enzymatic function of serine synthesis to collagen production and cell proliferation, which would be supposed to the mechanism of the enzyme leading to the occurrence and development of IPF and IPF-based tumorigenesis.

Likewise, Pyrroline-5-carboxylate reductase 1 (PYCR1) is the key enzyme for proline synthesis. Current studies suggested that PYCR1 with its binding partner kindlin-2 significantly overexpressed in the lung tissues of patients with pulmonary fibrosis or lung adenocarcinoma, the complex composed of which can promote the collagen matrix synthesis and induce the tumor growth through proline synthesis [46,47]. Therefore, we suggested the overexpression of PSAT1 and PYCR1 leading to the progression of IPF and NSCLC may be related to metabolic reprogramming.

The silence of collagen α-1(X) chain (COL10A1), a member of the collagen family and component of extracellular matrix, could inhibit cell proliferation and promote cell apoptosis and autophagy in NSCLC cells [48]. Seldom studies indicated how COL10A1 plays role in the progression of fibrotic diseases. Consistent with previous studies, our results also demonstrated the upregulated expression of COL10A1 associated with progression of NSCLC. The expression of KIAA1683, known as IQ Motif Containing N (IQCN) located in mitochondrion [49], was downregulated in both IPF and NSCLC in our study. However, the mechanisms of KIAA1683 in fibrotic diseases and cancers have not been elucidated yet. Further studies are needed to show its potential mechanisms in pathogenesis of IPF or NSCLC.

There are inherent limitations in this study. Firstly, the validation of the screened genes' expression was limited possibly due to low universality of their dysregulated expression among the general samples. Furthermore, the validation of hub genes' potential diagnostic value in tumorigenesis during the progression of IPF was limited to datasets usually related to individual disease, which is necessarily established by further datasets with the coexistence of two diseases. In conclusion, considering the limitations of this investigation, we can only find that the dysregulation of ECM is a main common mechanism of both diseases, normalization of which might provide therapeutic promise for the patients with IPF and NSCLC.

5. Conclusion

Our study showed the dysregulation of ECM induced by screened dysregulated genes is an important shared mechanism involving in the progression of IPF and NSCLC. Considering the inherent limitations of this study, further experimental validations for hub genes including large samples are urgently needed.

6. Fundng

The study was supported by the National Natural Science Foundation of China (82170076).

7. Ethics approval and consent to participate

Not applicable.

8. Consent for publication

Not applicable.

CRediT authorship contribution statement

Xiaorui Ding: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Huarui Liu:** Writing – review & editing, Visualization, Methodology, Formal analysis, Data curation, Conceptualization. **Qinghua Xu:** Writing – review & editing, Visualization, Methodology, Data curation, Conceptualization. **Tong Ji:** Writing – review & editing, Visualization, Validation, Methodology, Data curation, Conceptualization. **Ranxun Chen:** Writing – review & editing, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Zhengcheng Liu:** Writing – review & editing, Resources, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Jinghong Dai:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

We are grateful to all research scientists and patients who participated in the GEO cohorts.

References

[1] L. Richeldi, H.R. Collard, M.G. Jones, Idiopathic pulmonary fibrosis, Lancet 389 (10082) (2017) 1941–1952, https://doi.org/10.1016/s0140-6736(17)3086 6-8.

[2] T.M. Maher, et al., Global incidence and prevalence of idiopathic pulmonary fibrosis, Respir. Res. 22 (1) (2021) 197, https://doi.org/10.1186/s12931-021-01791-z.

[3] J. Hutchinson, et al., Global incidence and mortality of idiopathic pulmonary fibrosis: a systematic review, Eur. Respir. J. 46 (3) (2015) 795–806, https://doi.org/10.1183/09031936.00185114.

[4] A. Caminati, et al., Comorbidities in idiopathic pulmonary fibrosis: an underestimated issue, Eur. Respir. Rev. 28 (153) (2019), https://doi.org/10.1183/16000617.0044-2019.

[5] B. Bade, et al., Comorbidity and life expectancy in shared decision making for lung cancer screening, Semin. Oncol. (2022), https://doi.org/10.1053/j.seminoncol.2022.07.003.

[6] J.H. Lee, et al., Epidemiology and comorbidities in idiopathic pulmonary fibrosis: a nationwide cohort study, BMC Pulm. Med. 23 (1) (2023) 54, https://doi.org/10.1186/s12890-023-02340-8.

[7] J. Zhu, et al., A causal atlas on comorbidities in idiopathic pulmonary fibrosis: a bidirectional Mendelian randomization study, Chest 164 (2) (2023) 429–440, https://doi.org/10.1016/j.chest.2023.02.038.

[8] Y. Ozawa, et al., Cumulative incidence of and predictive factors for lung cancer in IPF, Respirology 14 (5) (2009) 723–728, https://doi.org/10.1111/j.1440-1843.2009.01547.x.

[9] L. Carobene, et al., Lung cancer and interstitial lung diseases: the lack of prognostic impact of lung cancer in IPF, Intern. Emerg. Med. 17 (2) (2022) 457–464, https://doi.org/10.1007/s11739-021-02833-6.

[10] A. Tzouvelekis, et al., Patients with IPF and lung cancer: diagnosis and management, Lancet Respir. Med. 6 (2) (2018) 86–88, https://doi.org/10.1016/s2213-2600(17)30478-2.

[11] A. Tzouvelekis, et al., Common pathogenic mechanisms between idiopathic pulmonary fibrosis and lung cancer, Chest 156 (2) (2019) 383–391, https://doi.org/10.1016/j.chest.2019.04.114.

[12] S.A. Whittaker Brown, et al., Outcomes of older patients with pulmonary fibrosis and non-small cell lung cancer, Ann. Am. Thorac. Soc. 16 (8) (2019) 1034–1040, https://doi.org/10.1513/annalsats.201808-510oc.

[13] B. Ballester, J. Milara, J. Cortijo, Idiopathic pulmonary fibrosis and lung cancer: mechanisms and molecular targets, Int. J. Mol. Sci. 20 (3) (2019), https://doi.org/10.3390/ijms20030593.

[14] T. Kinoshita, T. Goto, Molecular mechanisms of pulmonary fibrogenesis and its progression to lung cancer: a review, Int. J. Mol. Sci. 20 (6) (2019), https://doi.org/10.3390/ijms20061461.

[15] Y. Yoneshima, et al., Paired analysis of tumor mutation burden for lung adenocarcinoma and associated idiopathic pulmonary fibrosis, Sci. Rep. 11 (1) (2021) 12732, https://doi.org/10.1038/s41598-021-92098-y.

[16] T. Barrett, et al., NCBI GEO: archive for functional genomics data sets–update, Nucleic Acids Res. 41 (database issue) (2013) D991–D995, https://doi.org/10.1093/nar/gks1193.

[17] S. Davis, P.S. Meltzer, GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor, Bioinformatics 23 (14) (2007) 1846–1847, https://doi.org/10.1093/bioinformatics/btm254.

[18] M.E. Ritchie, et al., Limma powers differential expression analyses for RNA-sequencing and microarray studies, Nucleic Acids Res. 43 (7) (2015) e47.

[19] A. Jia, L. Xu, Y. Wang, Venn diagrams in bioinformatics, Brief Bioinform. 22 (2021) 5, https://doi.org/10.1093/bib/bbab108.

[20] P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network analysis, BMC Bioinformat. 9 (2008) 559, https://doi.org/10.1186/1471-2105-9-559.

[21] Gene Ontology Consortium: going forward, Nucleic Acids Res. 43 (Database issue) (2015) D1049–D1056. https://doi.org/10.1093/nar/gku1179.

[22] H. Ogata, et al., KEGG: Kyoto Encyclopedia of Genes and Genomes, Nucleic Acids Res. 27 (1) (1999) 29–34, https://doi.org/10.1093/nar/27.1.29.

[23] B.T. Sherman, et al., DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update), Nucleic Acids Res. 50 (W1) (2022) W216–W221, https://doi.org/10.1093/nar/gkac194.

[24] D. Szklarczyk, et al., STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets, Nucleic Acids Res. 47 (D1) (2019) D607–D613, https://doi.org/10.1093/nar/gky1131.

[25] Y. Jiang, et al., Outlier detection and robust variable selection via the penalized weighted LAD-LASSO method, J. Appl. Stat. 48 (2) (2021) 234–246, https://doi.org/10.1080/02664763.2020.1722079.

[26] S. Engebretsen, J. Bohlin, Statistical predictions with glmnet, Clin. Epigenetics 11 (1) (2019) 123, https://doi.org/10.1186/s13148-019-0730-1.

[27] X. Robin, et al., pROC: an open-source package for R and S+ to analyze and compare ROC curves, BMC Bioinformat. 12 (2011) 77, https://doi.org/10.1186/1471-2105-12-77.

[28] T. Karampitsakos, et al., Lung cancer in patients with idiopathic pulmonary fibrosis, Pulm. Pharmacol. Ther. 45 (2017) 1–10, https://doi.org/10.1016/j.pupt.2017.03.016.

[29] R.S. Herbst, D. Morgensztern, C. Boshoff, The biology and management of non-small cell lung cancer, Nature 553 (7689) (2018) 446–454, https://doi.org/10.1038/nature25183.

[30] A.D. Theocharis, et al., Extracellular matrix structure, Adv. Drug Deliv. Rev. 97 (2016) 4–27, https://doi.org/10.1016/j.addr.2015.11.001.

[31] D. Guillotin, et al., Transcriptome analysis of IPF fibroblastic foci identifies key pathways involved in fibrogenesis, Thorax 76 (1) (2021) 73–82, https://doi.org/10.1136/thoraxjnl-2020-214902.

[32] D.J. Lederer, F.J. Martinez, Idiopathic pulmonary fibrosis, N. Engl. J. Med. 378 (19) (2018) 1811–1823, https://doi.org/10.1056/nejmra1705751.

[33] S. Meiners, O. Eickelberg, M. Königshoff, Hallmarks of the ageing lung, Eur. Respir. J. 45 (3) (2015) 807–827, https://doi.org/10.1183/09031936.001869 14.

[34] C. Wang, J. Yang, Mechanical forces: The missing link between idiopathic pulmonary fibrosis and lung cancer, Eur. J. Cell Biol. 101 (3) (2022) 151234, https://doi.org/10.1016/j.ejcb.2022.151234.

[35] M. Kreus, et al., Extracellular matrix proteins produced by stromal cells in idiopathic pulmonary fibrosis and lung adenocarcinoma, PLoS One 16 (4) (2021) e0250109.

[36] R.B. Hamanaka, et al., Glutamine metabolism is required for collagen protein synthesis in lung fibroblasts, Am. J. Respir. Cell Mol. Biol. 61 (5) (2019) 597–606, https://doi.org/10.1165/rcmb.2019-0008oc.

[37] C. D'Aniello, et al., Proline metabolism in tumor growth and metastatic progression, Front. Oncol. 10 (2020) 776, https://doi.org/10.3389/fonc.2020.00776.

[38] Z. Yuan, et al., Extracellular matrix remodeling in tumor progression and immune escape: from mechanisms to treatments, Mol. Cancer 22 (1) (2023) 48, https://doi.org/10.1186/s12943-023-01744-8.

[39] R. Nigdelioglu, et al., Transforming Growth Factor (TGF)-β promotes de novo serine synthesis for collagen production, J. Biol. Chem. 291 (53) (2016) 27239–27251, https://doi.org/10.1074/jbc.m116.756247.

[40] K. Chen, L. Guo, C. Wu, How signaling pathways link extracellular mechano-environment to proline biosynthesis: a hypothesis: PINCH-1 and kindlin-2 sense mechanical signals from extracellular matrix and link them to proline biosynthesis, Bioessays 43 (9) (2021) e2100116 https://doi.org/10.1002/bies.202100116.

[41] Y. Liang, et al., Upregulated collagen COL10A1 remodels the extracellular matrix and promotes malignant progression in lung adenocarcinoma, Front. Oncol. 10 (2020) 573534 https://doi.org/10.3389/fonc.2020.573534.

[42] F. Andriani, et al., Diagnostic role of circulating extracellular matrix-related proteins in non-small cell lung cancer, BMC Cancer 18 (1) (2018) 899, https://doi.org/10.1186/s12885-018-4772-0.

[43] M. Bueno, et al., Mitochondria dysfunction and metabolic reprogramming as drivers of idiopathic pulmonary fibrosis, Redox Biol. 33 (2020 Jun) 101509, https://doi.org/10.1016/j.redox.2020.101509.

[44] W. Zhu, et al., Vitamin D3 alleviates pulmonary fibrosis by regulating the MAPK pathway via targeting PSAT1 expression in vivo and in vitro, Int. Immunopharma Col. 101 (Pt B) (2021) 108212, https://doi.org/10.1016/j.intimp.2021.108212.

[45] Y. Yang, et al., PSAT1 regulates cyclin D1 degradation and sustains proliferation of non-small cell lung cancer cells, Int. J. Cancer. 136 (4) (2015 Feb 15) E39–E50, https://doi.org/10.1002/ijc.29150.

[46] P. Zhang, et al., Kindlin-2 acts as a key mediator of lung fibroblast activation and pulmonary fibrosis progression, Am. J. Respir. Cell Mol. Biol. 65 (1) (2021 Jul) 54–69, https://doi.org/10.1165/rcmb.2020-0320oc.

[47] Y. Gao, et al., PYCR1 knockdown inhibits the proliferation, migration, and invasion by affecting JAK/STAT signaling pathway in lung adenocarcinoma, Mol. Carcinog. 59 (5) (2020) 503–511, https://doi.org/10.1002/mc.23174.

[48] Q. Guo, et al., MiR-384 induces apoptosis and autophagy of non-small cell lung cancer cells through the negative regulation of Collagen α-1(X) chain gene, Biosci. Rep. 39 (2) (2019) BSR20181523, https://doi.org/10.1042/bsr20181523.

[49] Y. Wang, et al., Loss-of-function mutations in IQCN cause male infertility in humans and mice owing to total fertilization failure, Mol Hum Reprod. 29 (7) (2023) gaad018, https://doi.org/10.1093/molehr/gaad018.