



OPEN ACCESS

EDITED BY
Makon-Sébastien Njock,
University of Liège,
Belgium

REVIEWED BY
Jazmin Calyeca,
The Ohio State University,
United States
Chunbin Zou,
University of Pittsburgh,
United States

*CORRESPONDENCE
Ehsan Amiri-Ardekani
✉ ehsanamiri@sums.ac.ir
Haibing Hua
✉ hhbjytc@163.com
Yi Cheng
✉ chengyi@xinhuamed.com.cn

SPECIALTY SECTION
This article was submitted to
Pulmonary Medicine,
a section of the journal
Frontiers in Medicine

RECEIVED 03 August 2022
ACCEPTED 26 January 2023
PUBLISHED 13 February 2023

CITATION
Zhang Y, Wang C, Xia Q, Jiang W, Zhang H,
Amiri-Ardekani E, Hua H and Cheng Y (2023)
Machine learning-based prediction of
candidate gene biomarkers correlated with
immune infiltration in patients with idiopathic
pulmonary fibrosis.
Front. Med. 10:1001813.
doi: 10.3389/fmed.2023.1001813

COPYRIGHT
© 2023 Zhang, Wang, Xia, Jiang, Zhang,
Amiri-Ardekani, Hua and Cheng. This is an
open-access article distributed under the
terms of the [Creative Commons Attribution
License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is
permitted which does not comply with these
terms.

Machine learning-based prediction of candidate gene biomarkers correlated with immune infiltration in patients with idiopathic pulmonary fibrosis

Yufeng Zhang¹, Cong Wang¹, Qingqing Xia¹, Weilong Jiang¹,
Huizhe Zhang², Ehsan Amiri-Ardekani^{3*}, Haibing Hua^{4*} and
Yi Cheng^{5*}

¹Department of Pulmonary and Critical Care Medicine, Jiangyin Hospital of Traditional Chinese Medicine, Jiangyin Hospital Affiliated to Nanjing University of Chinese Medicine, Jiangyin, Jiangsu, China, ²Department of Respiratory Medicine, Yancheng Hospital of Traditional Chinese Medicine, Yancheng Hospital Affiliated to Nanjing University of Chinese Medicine, Yancheng, Jiangsu, China, ³Department of Phytopharmaceuticals (Traditional Pharmacy), Faculty of Pharmacy, Shiraz University of Medical Sciences, Shiraz, Iran, ⁴Department of Gastroenterology, Jiangyin Hospital of Traditional Chinese Medicine, Jiangyin Hospital Affiliated to Nanjing University of Chinese Medicine, Jiangyin, Jiangsu, China, ⁵Department of Respiratory Medicine, Xinhua Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

Objective: This study aimed to identify candidate gene biomarkers associated with immune infiltration in idiopathic pulmonary fibrosis (IPF) based on machine learning algorithms.

Methods: Microarray datasets of IPF were extracted from the Gene Expression Omnibus (GEO) database to screen for differentially expressed genes (DEGs). The DEGs were subjected to enrichment analysis, and two machine learning algorithms were used to identify candidate genes associated with IPF. These genes were verified in a validation cohort from the GEO database. Receiver operating characteristic (ROC) curves were plotted to assess the predictive value of the IPF-associated genes. The cell-type identification by estimating relative subsets of RNA transcripts (CIBERSORT) algorithm was used to evaluate the proportion of immune cells in IPF and normal tissues. Additionally, the correlation between the expression of IPF-associated genes and the infiltration levels of immune cells was examined.

Results: A total of 302 upregulated and 192 downregulated genes were identified. Functional annotation, pathway enrichment, Disease Ontology and gene set enrichment analyses revealed that the DEGs were related to the extracellular matrix and immune responses. COL3A1, CDH3, CEBPD, and GPIHBP1 were identified as candidate biomarkers using machine learning algorithms, and their predictive value was verified in a validation cohort. Additionally, ROC analysis revealed that the four genes had high predictive accuracy. The infiltration levels of plasma cells, M0 macrophages and resting dendritic cells were higher and those of resting natural killer (NK) cells, M1 macrophages and eosinophils were lower in the lung tissues of patients with IPF than in those of healthy individuals. The expression of the abovementioned genes was correlated with the infiltration levels of plasma cells, M0 macrophages and eosinophils.

Conclusion: COL3A1, CDH3, CEBPD, and GPIHBP1 are candidate biomarkers of IPF. Plasma cells, M0 macrophages and eosinophils may be involved in the development of IPF and may serve as immunotherapeutic targets in IPF.

KEYWORDS

gene biomarker, immune infiltration, idiopathic pulmonary fibrosis, machine learning algorithm, CIBERSORT

1. Introduction

Idiopathic pulmonary fibrosis (IPF) is not only a chronic disorder but also a progressive interstitial lung disease. The aetiology of IPF remains unclear, with its pathological presentation being usual interstitial pneumonia (UIP) (1). IPF is an infrequently diagnosed disease with an incidence of approximately 2.8–9.3 per 100,000 population in Europe and North America. Epidemiological data on IPF are scarce in China; however, its incidence has remarkably increased in recent years (2). IPF progresses gradually at the early stage, leading to diffuse fibrosis of the lungs and eventually respiratory failure and death (3). At present, a few drugs are available for treating IPF; among which, pirfenidone and nintedanib have demonstrated evident curative effects. Traditional Chinese medicine (TCM) may play a central role in managing IPF (4). Owing to the limited understanding of the pathogenesis of IPF and the lack of early intervention strategies, IPF has become a serious life-threatening disease (5). The prognosis of individuals with IPF is poor, with an estimated median survival of approximately 3 years (6). Therefore, identifying new biomarkers for the diagnosis of IPF is important for improving its treatment and prognosis.

Early and definite diagnosis of IPF is the initial step to improving the clinical treatments and survival rate of patients with IPF. To date, several biochemical markers have been associated with the occurrence of IPF and used as references for its clinical diagnosis (7, 8). However, they are inefficient for early detection of IPF owing to their limited sensitivity and specificity. Genetic factors may play a key role in the pathogenesis of IPF. IPF is a complicated and multifactorial illness that develops through the synergy of genetic and environmental factors (9, 10).

The principal processes associated with the development of IPF as a chronic lung disorder include inflammation and fibrosis. Inflammatory cytokines produced by immune cells can result in fibroblast activation, angiogenesis and connective tissue cell proliferation (11). Additionally, immune dysregulation can enhance the progression of IPF and involves numerous biomarkers associated with the prognosis of IPF (12). Studies on animals and humans have demonstrated that innate and adaptive immune processes may exacerbate the existing fibrotic responses (13).

In recent studies, microarray technology has been used in combination with machine learning algorithms to discover new genes associated with different conditions, which may serve as diagnostic and prognostic biomarkers. Additionally, scholars have suggested that immune cell infiltration, which is closely related to these disease-associated genes, plays a substantial role (14, 15). However, to date, only a few studies have employed microarray technology and machine learning algorithms to verify the role of immune cell infiltration in IPF and identify probable diagnostic markers for IPF.

In this study, three microarray datasets of IPF were extracted from the Gene Expression Omnibus (GEO) database and combined into a metadata cohort. Differentially expressed genes (DEGs) between tissues of patients with IPF and healthy individuals were identified using data from the metadata cohort. The DEGs were analysed through Gene Ontology (GO) functional annotation analysis, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis, Disease Ontology (DO) enrichment analyses and gene set enrichment analysis (GSEA). Subsequently, machine learning algorithms were used for identifying candidate gene biomarkers of IPF. The identified genes were verified in a validation cohort from the GEO database. Receiver operating characteristic (ROC) curves were plotted to assess the prognostic value of the detected biomarkers in both metadata and validation cohorts. The cell-type identification by estimating relative

subsets of RNA transcripts (CIBERSORT) algorithm was used to evaluate the proportion of immune cells in the lung tissues of patients with IPF and healthy individuals based on their gene expression data. Additionally, the correlation between the detected biomarkers and infiltrating immune cells was examined.

2. Materials and methods

2.1. Microarray data

The matrix files of the GSE21369, GSE24206 and GSE110147 datasets were acquired from the NCBI GEO database¹. Data in the GSE21369 and GSE24206 datasets were acquired based on the GPL570 platform of Affymetrix Human Genome U133 Plus 2.0 Array (16, 17), whereas data in the GSE110147 dataset were acquired based on the GPL6244 platform of Affymetrix Human Gene 1.0 ST Array (18). The GSE21369 dataset included 11 lung tissue samples from patients with IPF and 6 lung tissue samples from healthy individuals. The GSE24206 dataset included 17 lung tissue samples from patients with IPF and 6 lung tissue samples from healthy donors. The GSE110147 dataset included 22 lung tissue samples from the recipient organs of patients with IPF and 11 normal lung tissue samples from tissue flanking lung cancer resections.

Probes in all datasets were transformed to gene symbols using their probe annotation files. The probe average was determined as the final expression value of genes if more than one probe corresponded to the same gene symbol. The three datasets were combined to obtain a metadata cohort for subsequent integrative analysis.

In addition, the GSE53845 dataset based on the GPL6480 platform of the Agilent-014850 Whole Human Genome Microarray 4x44K G4112F was used as the validation cohort. It included lung tissue samples from 40 patients with IPF and 8 healthy individuals (19).

2.2. Processing of data and screening of DEGs

The 'SVA' package in R was used to pre-process data in the metadata cohort and eliminate batch effects (20). The 'limma' package in R was used for data normalisation, background correction and identification of DEGs between 50 patients with IPF and 23 healthy individuals in the metadata cohort (21). Adjusted (adj) *p*-values of <0.05 and |log₂ fold change (FC)| values of >1 were considered the threshold values for identifying significant DEGs. The 'pheatmap' package was used to construct a heatmap for demonstrating the expression levels of the identified DEGs.

2.3. Enrichment analyses of DEGs

The 'clusterProfiler', 'DOSE' and 'GSEABase' packages were used for GO functional annotation, KEGG pathway enrichment and DO enrichment analyses and GSEA to examine substantial functions of the DEGs (22–25).

¹ <http://www.ncbi.nlm.nih.gov/geo/>

GO analysis incorporates three aspects, namely, molecular functions (MFs), cellular components (CCs) and biological processes (BPs). The 'c2.cp.kegg.v7.0.symbols.gmt' gene set from the Molecular Signatures Database (MSigDB)² was used as a reference for GSEA (26, 27). The primary finding of GSEA is the enrichment score (ES), which indicates the extent to which a gene set is overexpressed at either the top or bottom of a list of ranked genes. Positive and negative ESs demonstrate gene set enrichment at the top and bottom of the ranked list, respectively. In this study, genes with |normalised ESs (NESs)| of >1, *p*-values of <0.05 and adj *p*-values of <0.25 were considered remarkably enriched.

2.4. Screening of candidate gene biomarkers

To identify remarkable predictive variables, two machine learning algorithms were used to screen for genes associated with IPF. Least absolute shrinkage and selection operator (LASSO) is an algorithm of regression analysis that uses regularisation to enhance the reliability of predictions (28). LASSO analysis was performed using the 'glmnet' package in R to identify genes associated with the diagnosis of IPF (29). Support vector machine (SVM) is a supervised and extensively used machine-learning approach that functions in not only classification but also regression (30). To alleviate overfitting, the recursive feature elimination (RFE) algorithm was used to select optimal genes from the metadata cohort (31). To identify genes with the highest discriminative power, SVM-RFE was implemented using the 'e1071' and 'kernlab' packages in R (32, 33).

The overlapping genes between the two algorithms were defined as candidate gene biomarkers. Thereafter, the expression of these genes was verified in the GSE53845 dataset.

2.5. Diagnostic value of the identified gene biomarkers in IPF

To investigate the predictive value of the identified gene biomarkers, ROC curves were plotted based on the mRNA expression data of 50 patients with IPF and 23 healthy individuals in the metadata cohort. The area under the ROC curve (AUC) was evaluated to determine the diagnostic value of the genes. The AUC value was subsequently verified in the GSE53845 dataset.

2.6. Determination of immune cell subtypes

The CIBERSORT algorithm³, a bioinformatic analytical tool, was used to evaluate the relative proportion of infiltrating immune cells based on the gene expression data of patients with IPF and healthy individuals. The CIBERSORTx tool from the Alizadeh Lab and Newman Lab is used to impute gene expression profiles and estimate the abundance of member cell types in a mixed cell population using the gene expression data (34, 35). In this study, the CIBERSORTx tool was used to evaluate the abundance of 22 types of immune cells (reference set that had 1,000 permutations in the LM22 Signature Matrix file downloaded from CIBERSORTx).

Thereafter, the 'corrplot' in R was used to assess the distribution of the abundance of the 22 types of infiltrating immune cells and the correlation among them. The 'vioplot' package in R was used to construct violin plots for demonstrating differences in immune cell infiltration between patients with IPF and healthy individuals.

2.7. Analysis of the correlation between infiltrating immune cells and candidate genes

The correlation between the expression of candidate genes and the infiltration levels of immune cells was investigated through Spearman's rank correlation analysis in the R program. The 'ggplot2' package was used to visualise the resulting relationships (36).

2.8. Statistical analysis

The R software (version: 4.0.3) was used for all statistical analyses. Continuous variables were compared between groups using two tests: The Student's *t*-test was used to compare normally distributed variables, whereas the Mann-Whitney *U* test was used to compare abnormally distributed variables. The 'glmnet' package was used for LASSO regression analysis, whereas the 'e1071' and 'kernlab' packages in R were used for SVM-RFE. ROC curves were plotted and AUC values were evaluated to assess the diagnostic efficacy of the candidate gene biomarkers. Spearman's correlation analysis was performed to examine the correlation between the expression of candidate genes and the infiltration levels of immune cells. All statistical tests were two-sided, and *p*-values of <0.05 were considered significant. For screening DEGs between patients with IPF and healthy individuals, adj *p*-values of <0.05 and |log₂ FC| values of >1 were defined as the threshold values. For GO, KEGG and DO enrichment analyses, adj *p*-values of <0.05 were considered significant. For GSEA, genes with |NESs| of >1, *p*-values of <0.05 and adj *p*-values of <0.25 were considered significantly enriched.

3. Results

3.1. Detection of DEGs

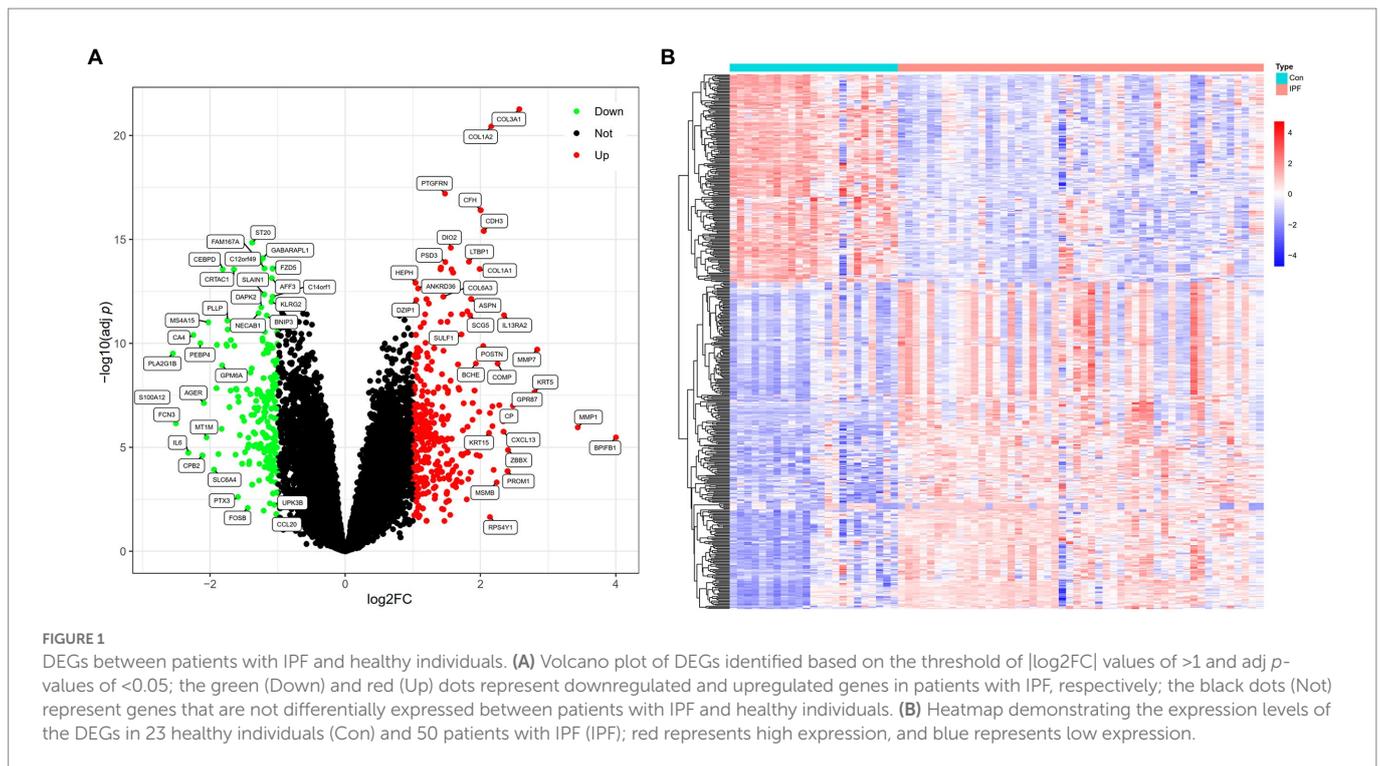
The gene expression data of 50 patients with IPF and 23 healthy individuals in the metadata cohort (GSE21369, GSE24206 and GSE110147) were retrospectively analysed (Supplementary File 1). After eliminating batch effects, DEGs between patients with IPF and healthy individuals were identified using the 'limma' package. Based on the threshold of adj *p*-values of <0.05 and |log₂FC| values of >1, 494 DEGs were identified, including 302 upregulated (log₂FC > 1) and 192 downregulated (log₂FC < -1) genes (Supplementary File 2). A volcano plot and heatmap demonstrating the expression of these DEGs are shown in Figures 1A,B, respectively.

3.2. Enrichment analyses

GO analysis revealed that the DEGs were remarkably enriched in BPs such as extracellular matrix (ECM) organisation, extracellular structure organisation, detoxification of copper ions, stress response to

2 <http://www.gsea-msigdb.org/gsea/msigdb>

3 <https://cibersortx.stanford.edu/>



copper ions, detoxification of inorganic compounds and other related processes. Additionally, the DEGs were substantially enriched in CCs such as collagen-containing ECM, endoplasmic reticulum lumen, ciliary plasm, axoneme and plasmalemma-bound cell projection cytoplasm and MFs such as ECM structural constituents, integrin binding, ECM structural constituent contributing to tensile strength, dynein light intermediate chain binding and adenosine triphosphate (ATP)-dependent/minus-end-directed microtubule motor activity (Supplementary File 3). The top 10 GO terms ranked based on their adj p -values are shown in Figure 2A.

KEGG pathway enrichment analysis revealed that the DEGs were remarkably enriched in pathways associated with mineral absorption, interleukin 17 (IL-17) signalling, advanced glycation end product (AGE) receptor (RAGE) signalling in diabetic complications, protein digestion and absorption, relaxin signalling, TNF signalling, malaria, ECM-receptor interaction and rheumatoid arthritis (Supplementary File 4). The top nine KEGG pathways ranked based on their adj p -values are shown in Figure 2B.

DO enrichment analysis was also performed to determine the functions of the DEGs. The results revealed that the DEGs were primarily associated with various illnesses (Supplementary File 5); among which, sarcoidosis, collagen disease, rheumatic disease, interstitial lung disease and pulmonary fibrosis are associated with IPF. The 20 DO terms ranked based on their adj p -values are shown in Figure 2C, and the 10 main diseases associated with IPF are shown in chord plots with the related genes in Figure 2D.

GSEA revealed that the DEGs were enriched in pathways associated with cytokine-cytokine receptor interaction, ECM-receptor interaction, Janus-activated kinase signal transducers, activators of transcription (JAK-STAT) signalling, mitogen-activated protein kinase (MAKP) signalling and focal adhesion (Supplementary File 6). The 5 gene sets enriched at the top of the ranked list ($NES > 1$) ranked based on their p -values are shown in Figure 2E, whereas the 5 gene sets enriched at the bottom of the ranked list ($NES < -1$) ranked based on their p -values are shown in Figure 2F.

3.3. Identification and validation of candidate gene biomarkers

Two algorithms were used to screen for potential diagnostic biomarkers for IPF. The DEGs were screened using the LASSO regression algorithm, resulting in the identification of 18 variables as diagnostic biomarkers (Table 1; Figure 3A). A subset of eight genes among the DEGs was determined using the SVM-RFE algorithm (Table 2; Figure 3B). The four overlapping genes between these two algorithms were eventually identified as candidate diagnostic biomarkers, including collagen type III alpha 1 chain (COL3A1), cadherin 3 (CDH3), CCAAT enhancer-binding protein delta (CEBPD) and glycosylphosphatidylinositol-anchored high-density lipoprotein-binding protein 1 (GPIHBP1) (Figure 3C).

To assess the reliability and accuracy of the four candidate genes, their expression was verified in the GSE53845 dataset (Supplementary File 7). The expression of COL3A1 and CDH3 was higher in the lung tissues of patients with IPF than in those of healthy individuals ($p < 0.05$; Figures 4A,B), whereas the expression of CEBPD and GPIHBP1 was remarkably lower in the lung tissues of patients with IPF than in those of healthy individuals ($p < 0.05$) (Figures 4C,D). These results were consistent with those of differential expression analysis in the metadata cohort. Therefore, the four genes were considered candidate diagnostic biomarkers for further analysis.

3.4. Diagnostic efficiency of the four candidate biomarkers in IPF

ROC curves were plotted to examine the efficiency of the four biomarkers in distinguishing patients with IPF from healthy individuals. The AUC values of COL3A1, CDH3, CEBPD, and GPIHBP1 were 0.996 (95% CI, 0.984–1.000) (Figure 5A), 0.980 (95% CI, 0.948–1.000)

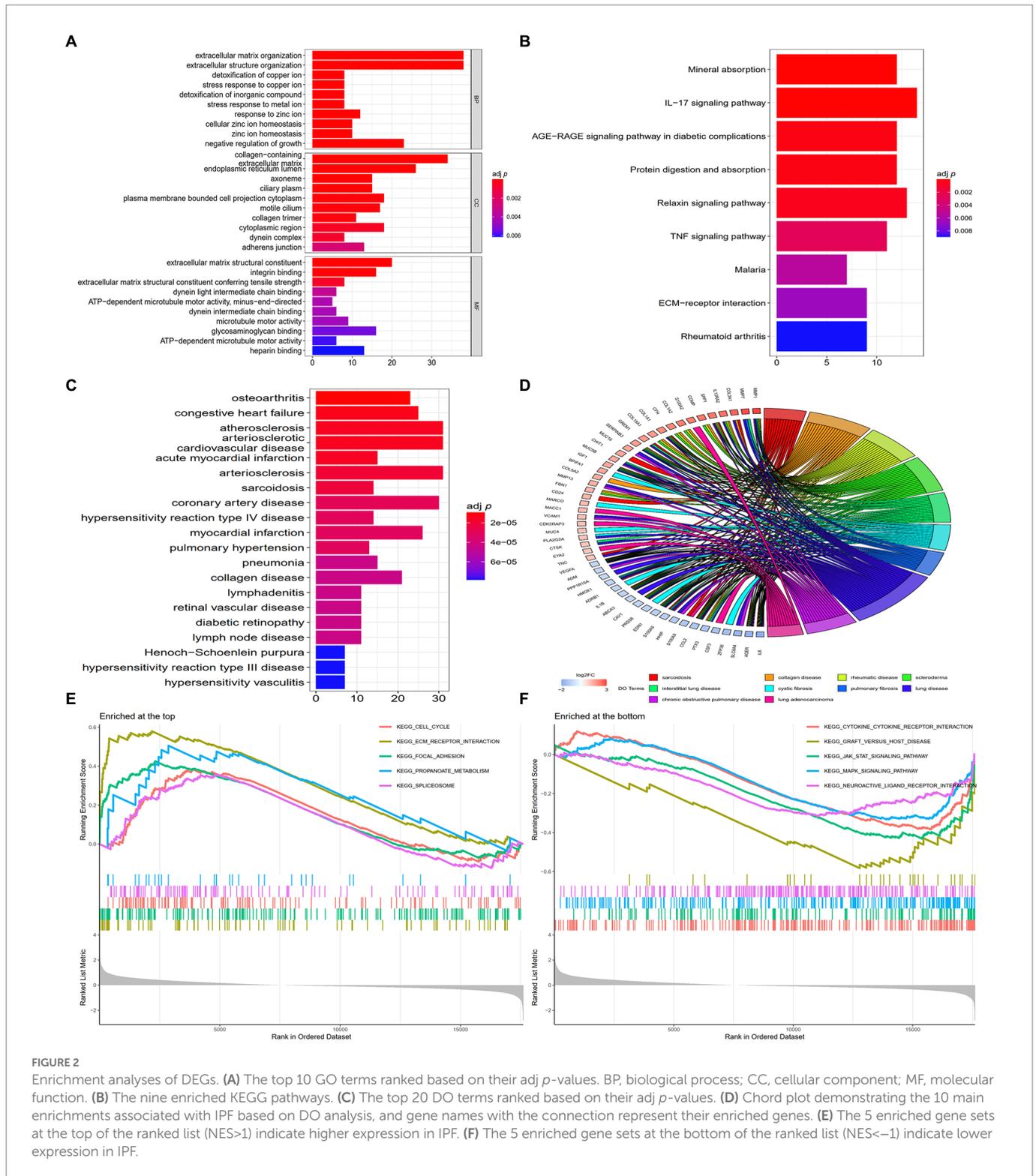


FIGURE 2 Enrichment analyses of DEGs. **(A)** The top 10 GO terms ranked based on their adj *p*-values. BP, biological process; CC, cellular component; MF, molecular function. **(B)** The nine enriched KEGG pathways. **(C)** The top 20 DO terms ranked based on their adj *p*-values. **(D)** Chord plot demonstrating the 10 main enrichments associated with IPF based on DO analysis, and gene names with the connection represent their enriched genes. **(E)** The 5 enriched gene sets at the top of the ranked list (NES>1) indicate higher expression in IPF. **(F)** The 5 enriched gene sets at the bottom of the ranked list (NES<-1) indicate lower expression in IPF.

(Figure 5B), 0.982 (95% CI, 0.952–1.000) (Figure 5C) and 0.946 (95% CI, 0.851–0.998) (Figure 5D), respectively, indicating that the four biomarkers had satisfactory diagnostic value. Additionally, the biomarkers had adequate discriminative capability in the GSE53845 dataset, with an AUC value of 0.825 (95% CI, 0.597–0.981) for COL3A1 (Figure 5E), 0.969 (95% CI, 0.897–1.000) for CDH3 (Figure 5F), 0.766 (95% CI, 0.634–0.887) for CEBPD (Figure 5G) and 0.917 (95% CI, 0.819–0.988) for GPIHBP1 (Figure 5H). These results suggest that the four candidate biomarkers have high diagnostic capability.

3.5. Immune cell infiltration

The CIBERSORT algorithm was used to evaluate the abundance of immune cells based on data extracted from the LM22 signature matrix file (Supplementary File 8). The results are shown in Supplementary File 9.

The distribution of 22 types of infiltrating immune cells in the IPF and control groups is demonstrated in Figure 6A. The correlation among the infiltration levels of 22 types of immune cells is demonstrated in

TABLE 1 Identification of 18 variables using the LASSO regression algorithm.

Gene symbol	Description
COL3A1	Collagen type III alpha 1 chain
CDH3	Cadherin 3
ST20	Suppressor of tumorigenicity 20
CEBPD	CCAAT enhancer-binding protein delta
CRTAC1	Cartilage acidic protein 1
HEPH	Hephaestin
DZIP1	DAZ-interacting zinc finger protein 1
MS4A15	Membrane spanning 4-domains a15
LOC100131541	Not applicable
GPIHBP1	Glycosylphosphatidylinositol-anchored high-density lipoprotein-binding protein 1
IRS2	Insulin receptor substrate 2
SCARNA17	Small Cajal body-specific RNA 17
LRRN1	Leucine-rich repeat neuronal 1
MYOCD	Myocardin
FNDC1	Fibronectin type III domain containing 1
CHI3L2	Chitinase 3-like 2
LYVE1	Lymphatic vessel endothelial hyaluronan receptor 1
TSPAN11	Tetraspanin 11

Figure 6B (regulatory T cells [Tregs] were not correlated with any other cell and are hence not shown).

The abundance of resting natural killer (NK) cells ($p < 0.001$), M1 macrophages ($p = 0.049$) and eosinophils ($p < 0.001$) was lower in the lung tissues of patients with IPF than in those of healthy individuals. However, the abundance of plasma cells ($p = 0.002$), M0 macrophages ($p < 0.001$) and resting dendritic cells (DCs) ($p = 0.008$) was higher in the lung tissues of patients with IPF than in those of healthy individuals (Figure 6C).

3.6. Correlation between candidate biomarkers and infiltrating immune cells

Spearman's rank correlation analysis was performed to examine and visualise the correlation between the expression of the four candidate genes and the infiltration levels of immune cells (Supplementary File 10).

COL3A1 expression was positively correlated with the infiltration levels of M0 macrophages ($r = 0.38$, $p = 0.001$), plasma cells ($r = 0.33$, $p = 0.005$) and activated NK cells ($r = 0.26$, $p = 0.024$) and negatively correlated with the infiltration levels of resting NK cells ($r = -0.48$, $p < 0.0001$), eosinophils ($r = -0.48$, $p < 0.001$), activated DCs ($r = -0.34$, $p = 0.003$), neutrophils ($r = -0.27$, $p = 0.020$) and monocytes ($r = -0.25$, $p = 0.036$). The detailed results are shown in Figure 7A.

CDH3 expression was positively correlated with the infiltration levels of M0 macrophages ($r = 0.54$, $p < 0.001$), plasma cells ($r = 0.53$, $p < 0.001$), resting DCs ($r = 0.49$, $p < 0.001$) and memory B cells ($r = 0.37$, $p = 0.002$) and negatively correlated with the infiltration levels of eosinophils ($r = -0.44$, $p < 0.001$), resting NK cells ($r = -0.44$, $p < 0.001$), M1 macrophages ($r = -0.28$, $p = 0.016$) and monocytes ($r = -0.24$, $p = 0.044$). The detailed results are shown in Figure 7B.

CEBPD expression was positively correlated with the infiltration levels of resting NK cells ($r = 0.44$, $p < 0.001$), activated DCs ($r = 0.39$,

$p < 0.001$), eosinophils ($r = 0.35$, $p = 0.002$), neutrophils ($r = 0.31$, $p = 0.009$) and monocytes ($r = 0.28$, $p = 0.018$) and negatively correlated with the infiltration levels of activated NK cells ($r = -0.41$, $p < 0.001$), M0 macrophages ($r = -0.38$, $p = 0.001$), M2 macrophages ($r = -0.36$, $p = 0.002$), resting DCs ($r = -0.35$, $p = 0.002$), memory B cells ($r = -0.26$, $p = 0.026$) and plasma cells ($r = -0.25$, $p = 0.035$). The detailed results are shown in Figure 7C.

GPIHBP1 expression was positively correlated with the infiltration levels of M1 macrophages ($r = 0.25$, $p = 0.033$) and eosinophils ($r = 0.24$, $p = 0.041$) and negatively correlated with the infiltration levels of M0 macrophages ($r = -0.49$, $p < 0.001$), resting DCs ($r = -0.29$, $p = 0.015$) and plasma cells ($r = -0.27$, $p = 0.021$). The detailed results are shown in Figure 7D.

4. Discussion

IPF is an interstitial condition characterised by UIP. At present, IPF cannot be cured and often has an unsatisfactory prognosis. Although numerous related studies have been reported, the mechanisms underlying the onset and development of IPF remain unclear (37). Epithelial-mesenchymal transition, ECM deposition and lung remodelling may be involved in the onset and progression of IPF (38–40).

Owing to the lack of biomarkers for early diagnosis of IPF, patients often miss the best opportunity for treatment, leading to progressive disease progression. Therefore, it is important to investigate the molecular mechanisms of biomarkers associated with the onset and development of IPF and identify therapeutic targets. Additionally, studies have reported that immune cell infiltration can clear ageing alveolar epithelial cells and play a role in the occurrence and development of IPF (41, 42). Therefore, the relationship between IPF-associated genes and infiltrating immune cells should be examined to improve the prognosis of IPF.

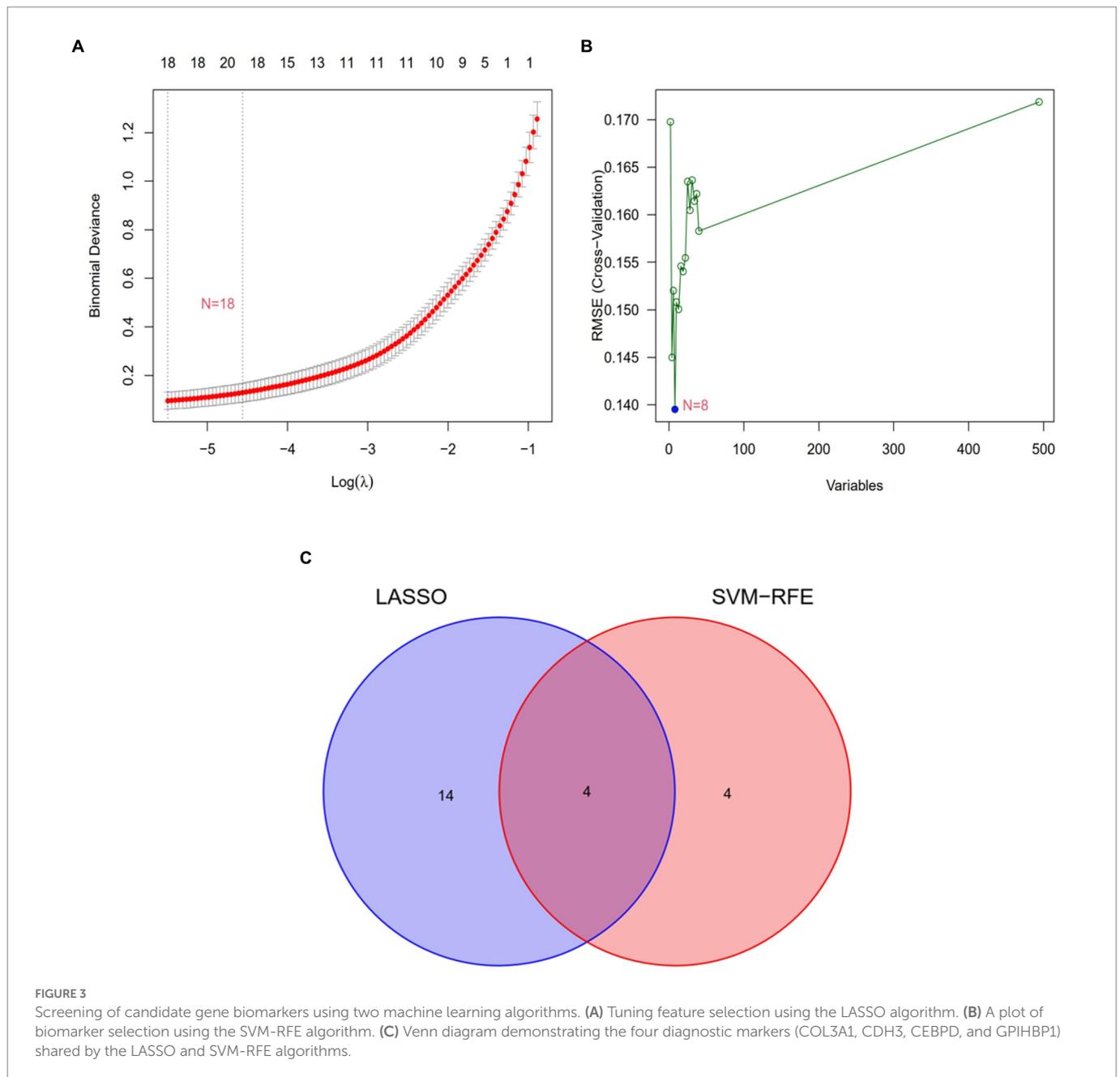
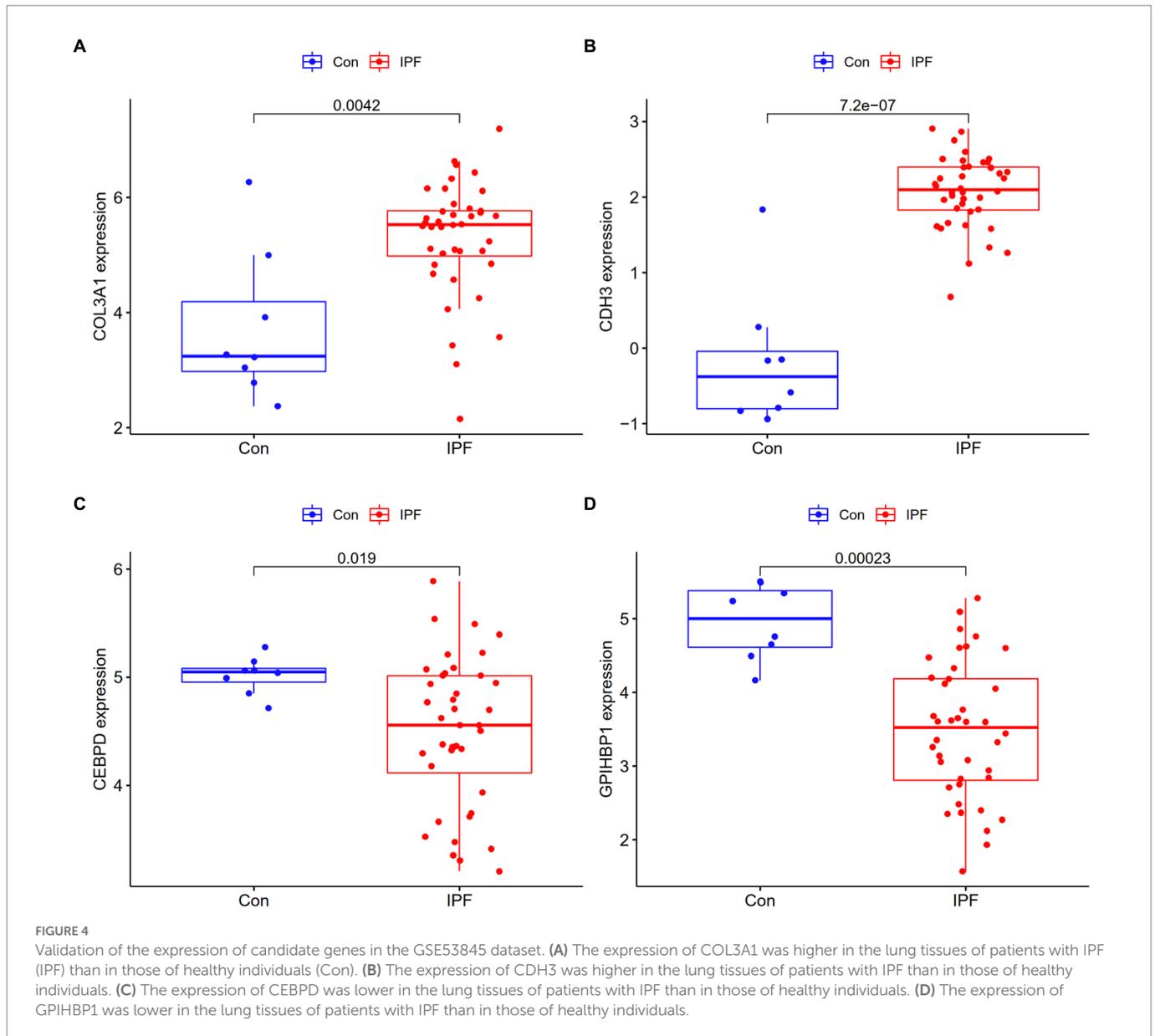


TABLE 2 Identification of eight variables using the SVM-RFE algorithm.

Gene symbol	Description
COL3A1	Collagen type III alpha 1 chain
TSHZ2	Teashirt zinc finger homeobox 2
COL1A2	Collagen type I alpha 2 chain
CDH3	Cadherin 3
PSD3	Pleckstrin and Sec7 domain-containing 3
CEBPD	CCAAT enhancer-binding protein delta
PTGFRN	Prostaglandin F2 receptor inhibitor
GPIHBP1	Glycosylphosphatidylinositol-anchored high-density lipoprotein-binding protein 1

Recent studies have reported that IPF-related microRNAs (miRNAs) play an important role in the diagnosis and treatment of IPF (43–45). In previous studies, we have constructed a modulatory network of putative

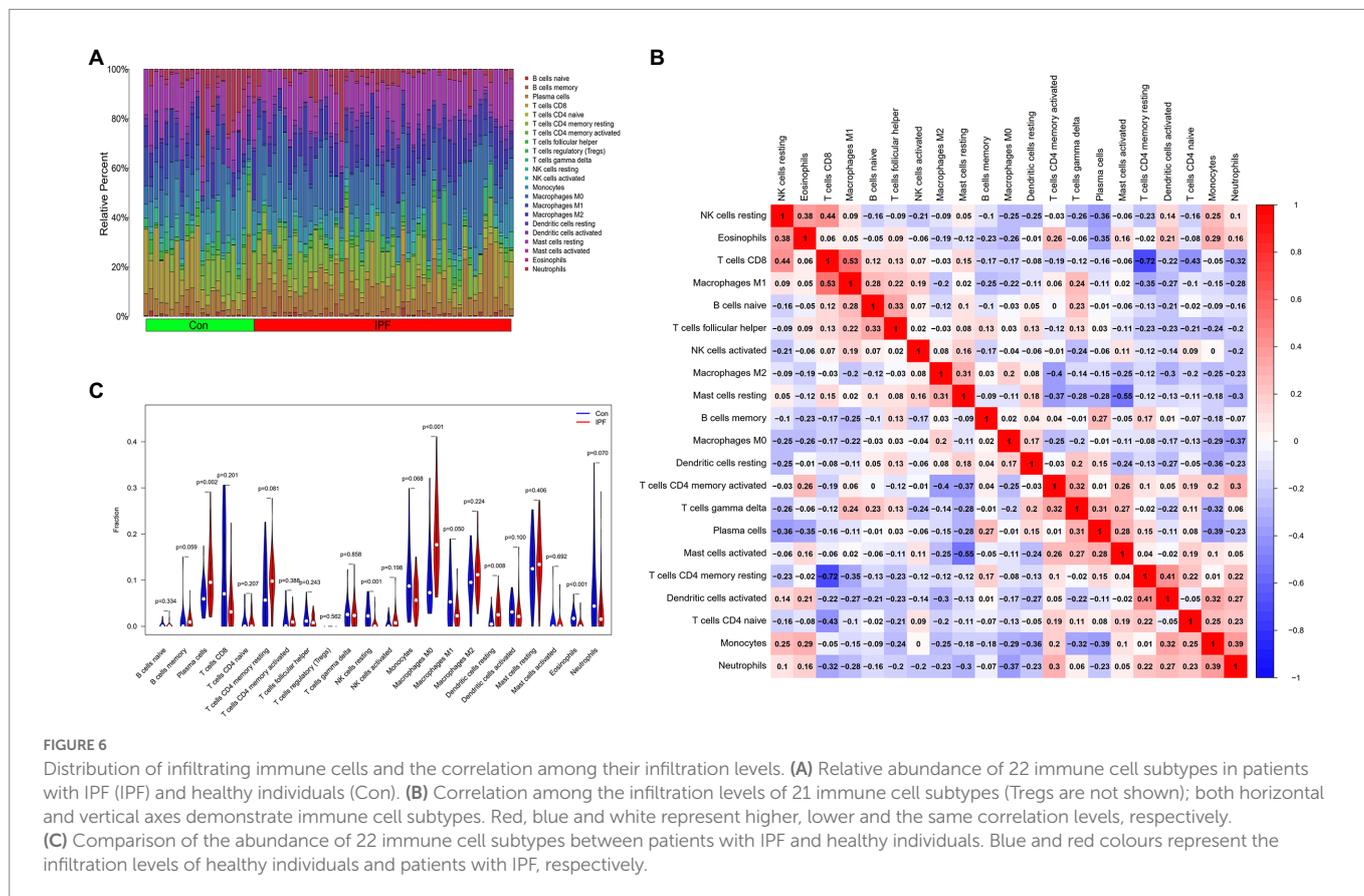
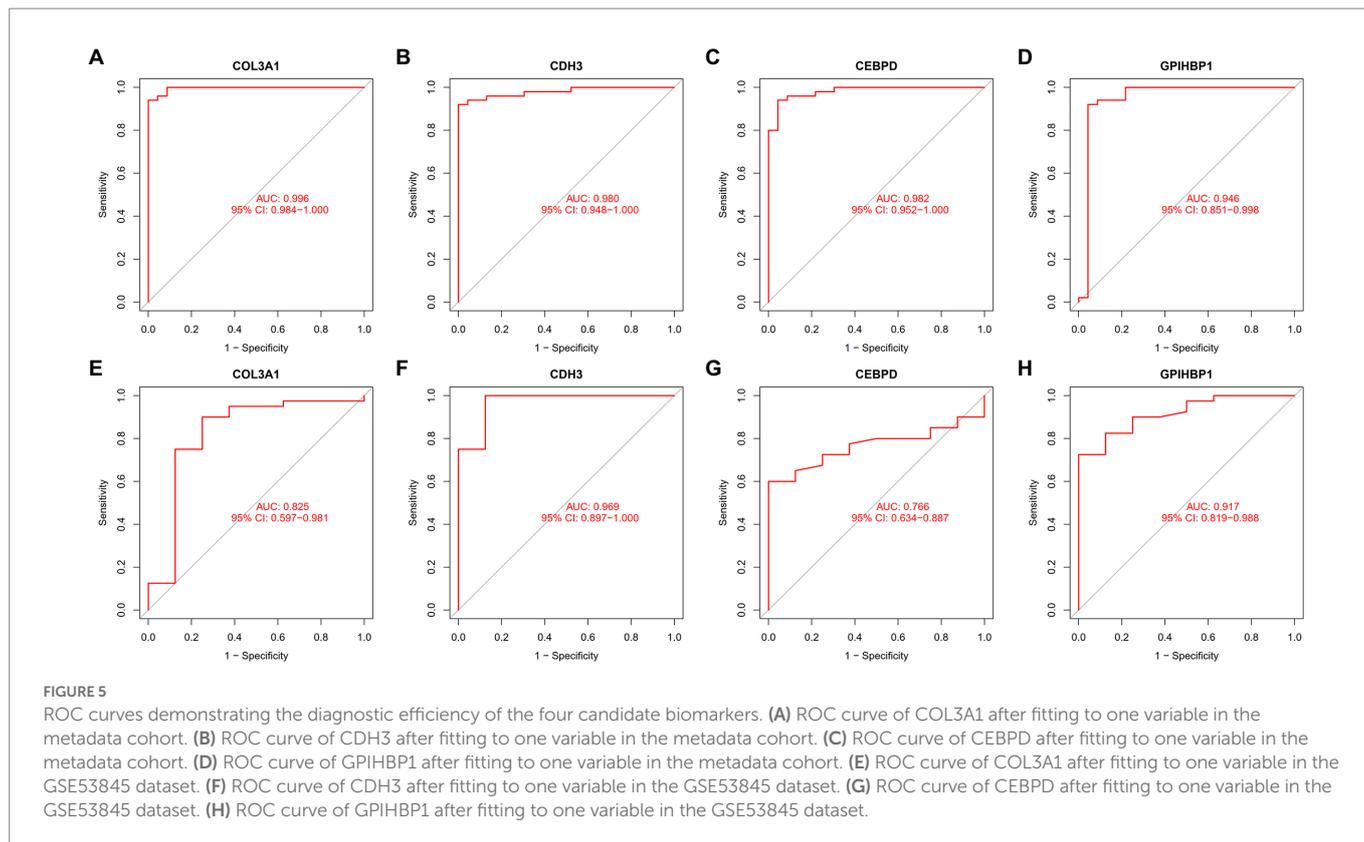
IPF-related miRNAs and messenger RNAs (mRNAs), which validates some miRNA–mRNA axes with TCM treatment of a bleomycin-induced IPF mouse model (4, 46). However, a few studies have examined the



relationship between abnormally expressed genes and immune infiltration in IPF. In this study, we identified candidate gene biomarkers for the diagnosis of IPF and examined their correlation with immune cell infiltration in IPF.

First, three microarray datasets were extracted from the GEO database and merged into a metadata cohort, which included 50 patients with IPF and 23 healthy individuals. A total of 494 DEGs were identified, including 302 upregulated and 192 downregulated genes. GO analysis revealed the DEGs were significantly enriched in BPs such as ECM organisation, extracellular structure organisation, detoxification and stress response to copper ions and detoxification of inorganic compounds; CCs such as collagen-containing ECM, endoplasmic reticulum lumen, ciliary plasm, axoneme and plasmalemma-bound cell projection cytoplasm and MFs such as ECM structural constituent, integrin binding, ECM structural constituent conferring tensile strength, dynein light intermediate chain binding and ATP-dependent/minus-end-directed microtubule motor activity. The functions of DEGs were primarily related to ECM, indicating that

the DEGs are closely related to ECM and participate in the development of IPF (38–40). KEGG analysis revealed that the DEGs were significantly enriched in pathways associated with absorption of minerals, IL-17 signalling, AGE-RAGE signalling in diabetic complications, protein digestion and absorption, relaxin signalling, TNF signalling, malaria, ECM-receptor interaction and rheumatoid arthritis. These pathways are primarily related to ECM and immune responses. DO enrichment analysis revealed that the DEGs were mainly associated with sarcoidosis, collagen disease, rheumatic disease, interstitial lung disease and pulmonary fibrosis. These diseases are associated with IPF and share some pathological characteristics with IPF. GSEA revealed that the DEGs were enriched in pathways associated with cytokine-cytokine receptor interaction, JAK-STAT signalling, ECM-receptor interaction, MAPK signalling and focal adhesion. These pathways are related to ECM, inflammation and immune responses. These findings are consistent with those of previous studies, indicating that inflammatory responses involving cytokines play a role in the pathogenesis of IPF (47–50).



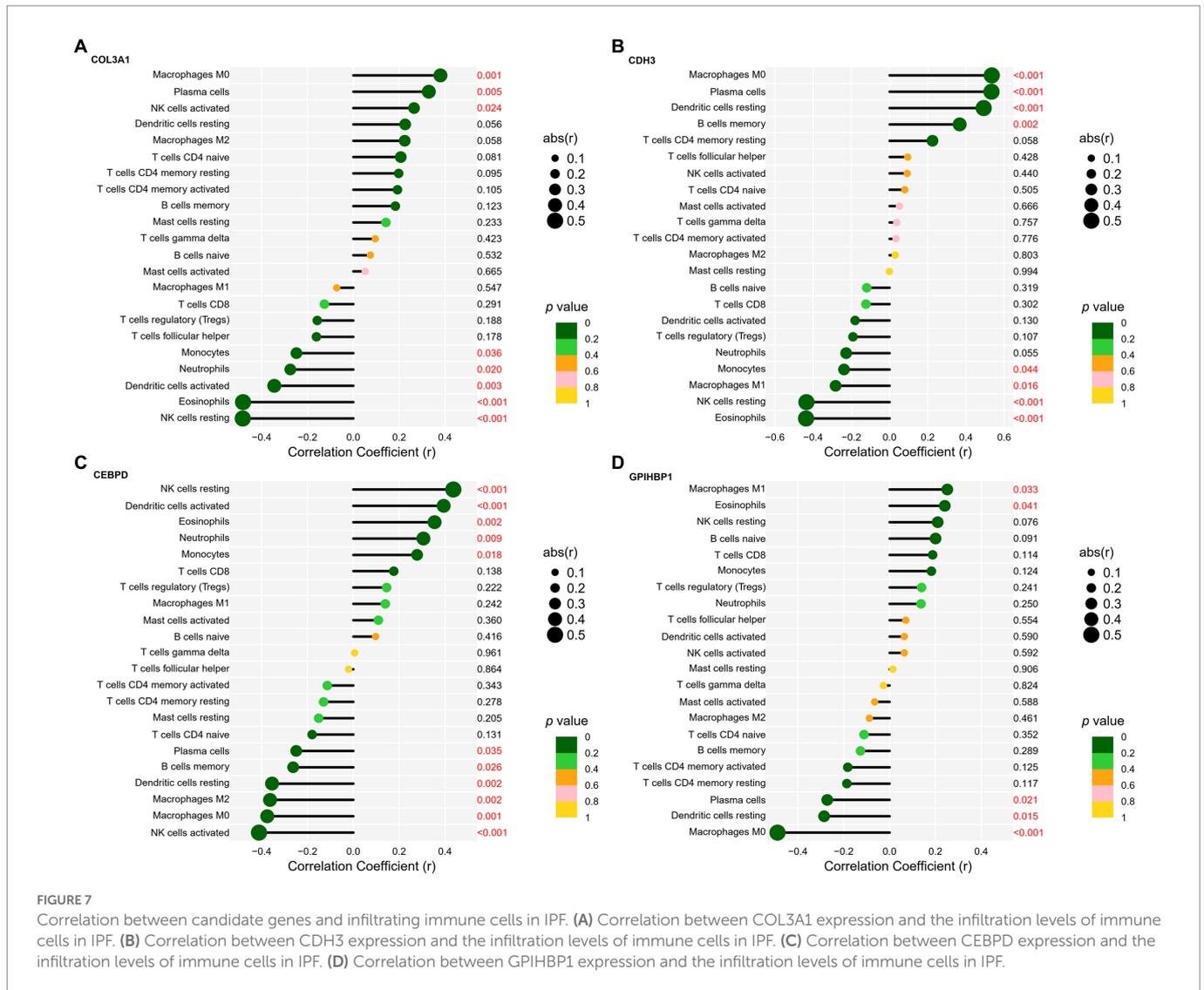


FIGURE 7

Correlation between candidate genes and infiltrating immune cells in IPF. (A) Correlation between COL3A1 expression and the infiltration levels of immune cells in IPF. (B) Correlation between CDH3 expression and the infiltration levels of immune cells in IPF. (C) Correlation between CEBPD expression and the infiltration levels of immune cells in IPF. (D) Correlation between GPIHBP1 expression and the infiltration levels of immune cells in IPF.

With the significant advancement of science and technology, machine learning algorithms are widely used for identifying gene biomarkers and predicting disease status (51, 52). The LASSO algorithm uses regularisation to enhance the predictive accuracy (53). SVM has better performance in classification and prediction and is extensively used in disease diagnosis or medical assistance. However, it is only useful for two-group classification tasks. To avoid overfitting, the RFE algorithm can be used. Therefore, the accuracy of the classification of multiclass issues may be addressed using the SVM-RFE technique (54). CIBERSORT, a bioinformatic algorithm, is widely used to measure immune cell infiltration (34, 35). In this study, the LASSO and SVM-RFE algorithms were used to determine candidate biomarkers among the DEGs, and the CIBERSORT algorithm was used to evaluate the abundance of infiltrating immune cells in IPF.

Using the two machine learning algorithms, four candidate genes associated with the diagnosis of IPF were identified, including two upregulated genes, namely, COL3A1 and CDH3, and two downregulated genes, namely, CEBPD and GPIHBP1. The expression of these genes was verified in the validation (GSE53845) cohort. Significant differences were observed in the expression of the four genes between patients with IPF and healthy individuals in the validation cohort. These results were consistent with those of differential expression analysis in the metadata

cohort. Additionally, ROC analysis revealed that the genes had a high diagnostic capability. The GSE53845 dataset contains gene expression data derived from the lung tissue samples of 40 patients with IPF and 8 healthy individuals. Because these data are derived from clinical patients, they are valid and reliable. Therefore, the abovementioned four genes were identified as candidate gene biomarkers.

COL3A1 encodes the pro-alpha 1 chains of type III collagen, which is a type of fibrillar collagen distributed in extensible connective tissues, including the skin, uterus, intestine, lung, and the vascular system, usually in association with type I collagen (55). CDH3 is a cadherin superfamily member that encodes cadherin. Multiple transcript variants are produced as a result of alternative splicing, and at least one of them encodes a preproprotein that is processed proteolytically to form a final glycoprotein. Five extracellular cadherin repeats, a greatly conserved cytoplasmic tail and a transmembrane region comprise the calcium-dependent cell-cell adhesion protein encoded by CDH3 (56). CEBPD, an intron-less gene, encodes a transcription factor with a leucine zipper domain that can attach as a homodimer to a particular DNA regulatory segment. It can also form heterodimers with CEBP-alpha, a related protein. The encoded protein plays an essential role in modulating genes involved in immune and inflammatory responses and may

be involved in the modulation of genes associated with macrophage activation and/or differentiation (57). GPIHBP1 is a protein that enhances the lipolytic digestion of triglyceride-rich lipoproteins in capillary endothelial cells. It is a glycosylphosphatidylinositol-anchored lymphocyte antigen-6 family member that plays a critical role in delivering lipoprotein lipase from the subendothelial regions to the capillary lumen (58).

Dysregulated expression of COL3A1 may affect the development of IPF through regulation of IPF-related biological processes, and the expression level of COL3A1 is correlated with the prognosis of IPF (59). COL3A1 is a potential biomarker for assessing the progression of IPF and non-small cell lung cancer (NSCLC). It may help to elucidate molecular mechanisms underlying the progression of IPF and NSCLC and serve as a potential therapeutic target for IPF (60). CEBP homologous protein (CHOP) enhances alveolar epithelial cell (AEC) senescence through the nuclear factor-kappa B (NF- κ B) pathway in pulmonary fibrosis (61). Additionally, it enhances the production of sonic hedgehog in type II AECs and stimulates the hedgehog signalling pathway in fibroblasts in pulmonary fibrosis (62). Hypoxia-inducible factor 1 alpha (HIF1A) can trigger endoplasmic reticulum stress and CHOP-mediated apoptosis in AECs, thereby playing a role in the development of IPF (63). Therefore, the four candidate genes as well as the abovementioned non-IPF-related genes warrant further intensive investigation.

CIBERSORT was used to evaluate the infiltration levels of immune cells in patients with IPF and healthy individuals. Several immune cell subtypes were found to be involved in key biological processes associated with IPF. The infiltration levels of plasma cells, M0 macrophages and resting DCs were higher and those of resting NK cells, M1 macrophages and eosinophils were lower in patients with IPF than in healthy individuals. These cells may be associated with the onset and progression of IPF.

Inflammatory and immune cells play an important role in the progression of IPF. Some results of this study are consistent with those of previous studies. The expression of FK506-binding protein (FKBP) prolyl isomerase 11 (FKBP11) is elevated in the lung tissues of patients with IPF, and FKBP11 specifically localises to antibody-producing plasma cells (64). In a study, compared with control mice, bleomycin-treated mice had an increased proportion of pulmonary IgA(+) germinal centres and plasma cells, and autoreactive IgA was identified as a diagnostic biomarker for IPF (65). M1 macrophages play a crucial role in wound healing following alveolar epithelial damage, whereas M2 macrophages are necessary for resolving inflammatory responses that develop in the lung. IPF is a pathological outcome resulting from disrupted wound healing in response to repeated injury to the lung (66). NF- κ B facilitates the production of proinflammatory cytokines to exacerbate M1 macrophage polarisation (67). Pirfenidone suppresses transforming growth factor- β , which is associated with M2 macrophage polarisation and fibroblast activation and has anti-fibrotic properties (68). Polarised M1 macrophages can be converted to M0 macrophages after 12 days of incubation in a cytokine-insufficient medium or re-differentiated into a different cell phenotype after being cultured further in a different polarising medium (69). DCs are major contributors to the pathogenesis of IPF (70). In bleomycin models, lung DCs are important proinflammatory cells that maintain pulmonary inflammation and fibrosis (71). Fms-related receptor tyrosine kinase 3 ligand is overexpressed in the serum and lung tissues of patients with IPF and may facilitate the accumulation of lung DCs during pulmonary fibrogenesis (72). The proportion of resting NK

cells is lower in the lung tissues of patients with IPF than in those of healthy individuals (73). Eosinophil is a principal source of several crucial pro-fibrogenic cytokines, especially in the initial stages of fibrosis (74).

COL3A1 may serve as a molecular biomarker for assessing prognosis and immune infiltration in pan-cancer (75). Collagen genes play an important role in regulating the immunosuppressive microenvironment and epithelial–mesenchymal transition in glioma and may serve as therapeutic targets for glioma (76). Biomarkers associated with collagen synthesis and degradation have the potential to enhance clinical trials in IPF and may be used for prognostic assessment and therapeutic decision-making in clinical settings (77). CDH3 is associated with immune infiltration in papillary thyroid carcinoma (78). CEBPD has been identified as a diagnostic biomarker for nonalcoholic fatty liver disease using machine learning algorithms and is associated with immune cell infiltration (79). In this study, the expression of COL3A1, CDH3, CEBPD and GPIHBP1 was correlated with the abundance of various immune cells including plasma cells, M0 macrophages and eosinophils. In particular, the expression of CDH3, CEBPD and GPIHBP1 was correlated with the abundance of resting DCs; the expression of COL3A1, CDH3 and CEBPD was correlated with the abundance of resting NK cells and the expression of CDH3 and GPIHBP1 was correlated with the abundance of M1 macrophages. The relationship of the four genes with these immune cells has been reported in some related studies. The infiltration of plasma cells has been associated with the expression of CDH3 and CEBPD (80, 81), whereas that of macrophages has been associated with the expression of COL3A1, CDH3 and CEBPD in multiple diseases (80, 82–84). In-depth experimental studies should be conducted to investigate the relationship between the four genes and immune cells in IPF.

Although this study was rigorous, its limitations should also be acknowledged. Although we collected as many samples as possible by combining the three datasets, the sample size of the metadata cohort is small. Additionally, the sample size of the validation cohort is also small. Because the role of the four biomarkers and infiltration of immune cells in IPF were examined using bioinformatic algorithms, in-depth studies with large sample size should be conducted to validate the findings. We will verify the results in a clinical cohort in future studies, with immunohistochemical detection of lung transplant specimens. Additionally, we will perform single-cell RNA sequencing on lung tissue and blood samples to verify whether the expression of the four genes is altered in immune cell clusters.

5. Conclusion

COL3A1, CDH3, CEBPD, and GPIHBP1 are potential biomarkers for the diagnosis of IPF. Plasma cells, M0 macrophages and eosinophils (associated with these four genes) may be involved in the development of IPF and serve as immunotherapeutic targets for the treatment of IPF.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

YZ, HZ, EA-A, HH, and YC conceived and designed the study and wrote the manuscript. YZ, CW, QX, WJ, and HZ were responsible for data collation and analysis. YZ, EA-A, HH, and YC supervised the study. YZ, HZ, and YC revised the manuscript. All authors have read and approved the final version of the manuscript.

Funding

This work was supported by the Research Grants of Jiangyin Hospital of Traditional Chinese Medicine (202013 to WJ, 202014 to YZ), Grants from the Wuxi Health Commission's Scientific Research Project (M202154 to YZ, T202130 to WJ), the ChengXing Talent Training Plan of Jiangyin Hospital of Traditional Chinese Medicine (2022 to YZ), Grants from the Traditional Chinese Medicine Science and Technology Development Plan Project of Jiangsu Province (ZT202113 to HH) and the National Natural Science Foundation of China (No. 82000039 to YC).

Acknowledgments

We would like to acknowledge the NCBI GEO database for allowing access to gene expression data. We also express our gratitude to the

References

- Raghu, G, Remy-Jardin, M, Myers, JL, Richeldi, L, Ryerson, CJ, Lederer, DJ, et al. Diagnosis of idiopathic pulmonary fibrosis. An official ATS/ERS/JRS/ALAT clinical practice guideline. *Am J Respir Crit Care Med.* (2018) 198:e44–68. doi: 10.1164/rccm.201807-1255ST
- Zhang, Y, Gu, L, Xia, Q, Tian, L, Qi, J, and Cao, M. Radix astragali and radix angelicae sinensis in the treatment of idiopathic pulmonary fibrosis: a systematic review and meta-analysis. *Front Pharmacol.* (2020) 11:415. doi: 10.3389/fphar.2020.00415
- Enomoto, N, Naoi, H, Aono, Y, Katsumata, M, Horiike, Y, Yasui, H, et al. Acute exacerbation of unclassifiable idiopathic interstitial pneumonia: comparison with idiopathic pulmonary fibrosis. *Ther Adv Respir Dis.* (2020) 14:1753466620935774. doi: 10.1177/1753466620935774
- Zhang, H, Wang, X, Shi, Y, Liu, M, Xia, Q, Jiang, W, et al. Danggui buxue decoction ameliorates idiopathic pulmonary fibrosis through MicroRNA and messenger RNA regulatory network. *Evid Based Complement Alternat Med.* (2022) 2022:3439656–19. doi: 10.1155/2022/3439656
- Biondini, D, Balestro, E, Sverzellati, N, Coconcelli, E, Bernardinello, N, Ryerson, CJ, et al. Acute exacerbations of idiopathic pulmonary fibrosis (AE-IPF): an overview of current and future therapeutic strategies. *Expert Rev Respir Med.* (2020) 14:405–14. doi: 10.1080/17476348.2020.1724096
- Kim, HJ, Perlman, D, and Tomic, R. Natural history of idiopathic pulmonary fibrosis. *Respir Med.* (2015) 109:661–70. doi: 10.1016/j.rmed.2015.02.002
- Drakopanagiotakis, F, Wujak, L, Wygrecka, M, and Markart, P. Biomarkers in idiopathic pulmonary fibrosis. *Matrix Biol.* (2018) 68–69:404–21. doi: 10.1016/j.matbio.2018.01.023
- Yoshikawa, T, Otsuka, M, Chiba, H, Ikeda, K, Mori, Y, Umeda, Y, et al. Surfactant protein a as a biomarker of outcomes of anti-fibrotic drug therapy in patients with idiopathic pulmonary fibrosis. *BMC Pulm Med.* (2020) 20:27. doi: 10.1186/s12890-020-1060-y
- Kaur, A, Mathai, SK, and Schwartz, DA. Genetics in idiopathic pulmonary fibrosis pathogenesis, prognosis, and treatment. *Front Med.* (2017) 4:154. doi: 10.3389/fmed.2017.00154
- Stainer, A, Faverio, P, Busnelli, S, Catalano, M, Della, ZM, Marruchella, A, et al. Molecular biomarkers in idiopathic pulmonary fibrosis: state of the art and future directions. *Int J Mol Sci.* (2021) 22:6255. doi: 10.3390/ijms22126255
- Jee, AS, Sahhar, J, Youssef, P, Bleasel, J, Adelstein, S, Nguyen, M, et al. Review: serum biomarkers in idiopathic pulmonary fibrosis and systemic sclerosis associated interstitial lung disease – frontiers and horizons. *Pharmacol Ther.* (2019) 202:40–52. doi: 10.1016/j.pharmthera.2019.05.014
- Harrell, CR, Sadikot, R, Pascual, J, Fellabaum, C, Jankovic, MG, Jovicic, N, et al. Mesenchymal stem cell-based therapy of inflammatory lung diseases: current understanding and future perspectives. *Stem Cells Int.* (2019) 2019:4236973–14. doi: 10.1155/2019/4236973
- Desai, O, Winkler, J, Minasyan, M, and Herzog, EL. The role of immune and inflammatory cells in idiopathic pulmonary fibrosis. *Front Med.* (2018) 5:43. doi: 10.3389/fmed.2018.00043
- Zhang, Y, Xia, R, Lv, M, Li, Z, Jin, L, Chen, X, et al. Machine-learning algorithm-based prediction of diagnostic gene biomarkers related to immune infiltration in patients with chronic obstructive pulmonary disease. *Front Immunol.* (2022) 13:740513. doi: 10.3389/fimmu.2022.740513
- Zhao, E, Xie, H, and Zhang, Y. Predicting diagnostic gene biomarkers associated with immune infiltration in patients with acute myocardial infarction. *Front Cardiovasc Med.* (2020) 7:586871. doi: 10.3389/fcvm.2020.586871
- Cho, JH, Gelinias, R, Wang, K, Etheridge, A, Piper, MG, Batte, K, et al. Systems biology of interstitial lung diseases: integration of mRNA and microRNA expression changes. *BMC Med Genet.* (2011) 4:8. doi: 10.1186/1755-8794-4-8
- Meltzer, EB, Barry, WT, D'Amico, TA, Davis, RD, Lin, SS, Onaitis, MW, et al. Bayesian probit regression model for the diagnosis of pulmonary fibrosis: proof-of-principle. *BMC Med Genet.* (2011) 4:70. doi: 10.1186/1755-8794-4-70
- Cecchini, MJ, Hosein, K, Howlett, CJ, Joseph, M, and Mura, M. Comprehensive gene expression profiling identifies distinct and overlapping transcriptional profiles in non-specific interstitial pneumonia and idiopathic pulmonary fibrosis. *Respir Res.* (2018) 19:153. doi: 10.1186/s12931-018-0857-1
- Depianto, DJ, Chandriani, S, Abbas, AR, Jia, G, N'Diaye, EN, Caplazi, P, et al. Heterogeneous gene expression signatures correspond to distinct lung pathologies and biomarkers of disease severity in idiopathic pulmonary fibrosis. *Thorax.* (2015) 70:48–56. doi: 10.1136/thoraxjnl-2013-204596
- Leek, JT, Johnson, WE, Parker, HS, Jaffe, AE, and Storey, JD. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics.* (2012) 28:882–3. doi: 10.1093/bioinformatics/bts034
- Ritchie, ME, Phipson, B, Wu, D, Hu, Y, Law, CW, Shi, W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* (2015) 43:e47. doi: 10.1093/nar/gkv007
- Oprescu, SN, Horzmann, KA, Yue, F, Freeman, JL, and Kuang, S. Microarray, IPA and GSEA analysis in mice models. *Bio Protoc.* (2018) 8:e2999. doi: 10.21769/BioProtoc.2999

researchers who have previously shared microarray datasets and to the producers of the web resource platforms and data processing software used in this study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2023.1001813/full#supplementary-material>

23. Wu, T, Hu, E, Xu, S, Chen, M, Guo, P, Dai, Z, et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovations*. (2021) 2:100141. doi: 10.1016/j.xinn.2021.100141
24. Yu, G, Wang, LG, Han, Y, and He, QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*. (2012) 16:284–7. doi: 10.1089/omi.2011.0118
25. Yu, G, Wang, LG, Yan, GR, and He, QY. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*. (2015) 31:608–9. doi: 10.1093/bioinformatics/btu684
26. Liberzon, A, Birger, C, Thorvaldsdottir, H, Ghandi, M, Mesirov, JP, and Tamayo, P. The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst*. (2015) 1:417–25. doi: 10.1016/j.cels.2015.12.004
27. Subramanian, A, Tamayo, P, Mootha, VK, Mukherjee, S, Ebert, BL, Gillette, MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. (2005) 102:15545–50. doi: 10.1073/pnas.0506580102
28. Tibshirani, R. The lasso method for variable selection in the cox model. *Stat Med*. (1997) 16:385–95. doi: 10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3
29. Engebretsen, S, and Bohlin, J. Statistical predictions with glmnet. *Clin Epigenetics*. (2019) 11:123. doi: 10.1186/s13148-019-0730-1
30. Huang, S, Cai, N, Pacheco, PP, Narrandes, S, Wang, Y, and Xu, W. Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics Proteomics*. (2018) 15:41–51. doi: 10.21873/cgp.20063
31. Escanilla, NS, Hellerstein, L, Kleiman, R, Kuang, Z, Shull, JD, and Page, D. Recursive feature elimination by sensitivity testing. *Proc Int Conf Mach Learn Appl*. (2018) 2018:40–7. doi: 10.1109/ICMLA.2018.00014
32. Scholkopf, B, Smola, AJ, Williamson, RC, and Bartlett, PL. New support vector algorithms. *Neural Comput*. (2000) 12:1207–45. doi: 10.1162/089976600300015565
33. Wang, X, Xing, EP, and Schaid, DJ. Kernel methods for large-scale genomic data analysis. *Brief Bioinform*. (2015) 16:183–92. doi: 10.1093/bib/bbu024
34. Newman, AM, Liu, CL, Green, MR, Gentles, AJ, Feng, W, Xu, Y, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods*. (2015) 12:453–7. doi: 10.1038/nmeth.3337
35. Newman, AM, Steen, CB, Liu, CL, Gentles, AJ, Chaudhuri, AA, Scherer, F, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol*. (2019) 37:773–82. doi: 10.1038/s41587-019-0114-2
36. Ito, K, and Murphy, D. Application of ggplot2 to pharmacometric graphics. *CPT Pharmacometrics Syst Pharmacol*. (2013) 2:e79. doi: 10.1038/psp.2013.56
37. Thomson, CC, Duggal, A, Bice, T, Lederer, DJ, Wilson, KC, and Raghu, G. 2018 clinical practice guideline summary for clinicians: diagnosis of idiopathic pulmonary fibrosis. *Ann Am Thorac Soc*. (2019) 16:285–90. doi: 10.1513/AnnalsATS.201809-604CME
38. James, DS, Jambor, AN, Chang, HY, Alden, Z, Tilbury, KB, Sandbo, NK, et al. Probing ECM remodeling in idiopathic pulmonary fibrosis via second harmonic generation microscopy analysis of macro/supramolecular collagen structure. *J Biomed Opt*. (2019) 25:1–13. doi: 10.1117/1.JBO.25.1.014505
39. Siekacz, K, Piotrowski, WJ, Iwanski, MA, Gorski, P, and Bialas, AJ. The role of interaction between mitochondria and the extracellular matrix in the development of idiopathic pulmonary fibrosis. *Oxidative Med Cell Longev*. (2021) 2021:9932442–12. doi: 10.1155/2021/9932442
40. Tomos, IP, Tzouvelekas, A, Aidinis, V, Manali, ED, Bouros, E, Bouros, D, et al. Extracellular matrix remodeling in idiopathic pulmonary fibrosis. It is the 'bed' that counts and not 'the sleepers'. *Expert Rev Respir Med*. (2017) 11:299–309. doi: 10.1080/17476348.2017.1300533
41. Serezani, AP, Pascoalino, BD, Bazzano, J, Vowell, KN, Tanjore, H, Taylor, CJ, et al. Multi-platform single-cell analysis identifies immune cell types enhanced in pulmonary fibrosis. *Am J Respir Cell Mol Biol*. (2022) 67:50–60. doi: 10.1165/rcmb.2021-0418OC
42. Waters, DW, Blokland, K, Pathinayake, PS, Burgess, JK, Mutsaers, SE, Prele, CM, et al. Fibroblast senescence in the pathology of idiopathic pulmonary fibrosis. *Am J Physiol Lung Cell Mol Physiol*. (2018) 315:L162–72. doi: 10.1152/ajplung.00037.2018
43. Fukunaga, S, Kakehashi, A, Sumida, K, Kushida, M, Asano, H, Gi, M, et al. Integrative analyses of miRNA and proteomics identify potential biological pathways associated with onset of pulmonary fibrosis in the bleomycin rat model. *Toxicol Appl Pharmacol*. (2015) 286:188–97. doi: 10.1016/j.taap.2015.04.014
44. Li, C, Wang, Z, Zhang, J, Zhao, X, Xu, P, Liu, X, et al. Crosstalk of mRNA, miRNA, lncRNA, and circRNA and their regulatory pattern in pulmonary fibrosis. *Mol Ther Nucleic Acids*. (2019) 18:204–18. doi: 10.1016/j.omtn.2019.08.018
45. Liu, Y, Nie, H, Ding, Y, Hou, Y, Mao, K, and Cui, Y. MiRNA, a new treatment strategy for pulmonary fibrosis. *Curr Drug Targets*. (2021) 22:793–802. doi: 10.2174/1874609813666200928141822
46. Zhang, YF, Gu, LN, Qi, J, Xia, QQ, Tian, LJ, Jiang, WL, et al. Construction of potential idiopathic pulmonary fibrosis related microRNA and messenger RNA regulatory network. *Chin Med J*. (2021) 134:584–6. doi: 10.1097/CM9.0000000000001276
47. Epstein, SG, Brook, E, Israeli-Shani, L, Edelstein, E, and Shitrit, D. Fibroblast paracrine TNF-alpha signaling elevates integrin A5 expression in idiopathic pulmonary fibrosis (IPF). *Respir Res*. (2017) 18:122. doi: 10.1186/s12931-017-0606-x
48. Qu, Y, Hao, C, Zhai, R, and Yao, W. Folate and macrophage folate receptor-beta in idiopathic pulmonary fibrosis disease: the potential therapeutic target? *Biomed Pharmacother*. (2020) 131:110711. doi: 10.1016/j.biopha.2020.110711
49. Rostami, MR, and Bradic, M. The derepression of transposable elements in lung cells is associated with the inflammatory response and gene activation in idiopathic pulmonary fibrosis. *Mob DNA*. (2021) 12:14. doi: 10.1186/s13100-021-00241-3
50. Zhang, J, Wang, D, Wang, L, Wang, S, Roden, AC, Zhao, H, et al. Profibrotic effect of IL-17A and elevated IL-17RA in idiopathic pulmonary fibrosis and rheumatoid arthritis-associated lung disease support a direct role for IL-17A/IL-17RA in human fibrotic interstitial lung disease. *Am J Physiol Lung Cell Mol Physiol*. (2019) 316:L487–97. doi: 10.1152/ajplung.00301.2018
51. Fan, Y, Han, Q, Li, J, Ye, G, Zhang, X, Xu, T, et al. Revealing potential diagnostic gene biomarkers of septic shock based on machine learning analysis. *BMC Infect Dis*. (2022) 22:65. doi: 10.1186/s12879-022-07056-4
52. Sun, YC, Qiu, ZZ, Wen, FL, Yin, JQ, and Zhou, H. Revealing potential diagnostic gene biomarkers associated with immune infiltration in patients with renal fibrosis based on machine learning analysis. *J Immunol Res*. (2022) 2022:3027200. doi: 10.1155/2022/3027200
53. Hu, JY, Wang, Y, Tong, XM, and Yang, T. When to consider logistic LASSO regression in multivariate analysis? *Eur J Surg Oncol*. (2021) 47:2206. doi: 10.1016/j.ejso.2021.04.011
54. Huang, ML, Hung, YH, Lee, WM, Li, RK, and Jiang, BR. SVM-RFE based feature selection and Taguchi parameters optimization for multiclass SVM classifier. *Sci World J*. (2014) 2014:795624. doi: 10.1155/2014/795624
55. Parkin, JD, San, AJ, Persikov, AV, Dagher, H, Dalgleish, R, Jensen, ST, et al. The collagen III fibril has a "flexi-rod" structure of flexible sequences interspersed with rigid bioactive domains including two with hemostatic roles. *PLoS One*. (2017) 12:e0175582. doi: 10.1371/journal.pone.0175582
56. Zhou, Y, Chi, Y, Bhandari, A, Xia, E, Thakur, PC, Qu, J, et al. Downregulated CDH3 decreases proliferation, migration, and invasion in thyroid cancer. *Am J Transl Res*. (2020) 12:3057–67.
57. Chi, JY, Hsiao, YW, Liu, HL, Fan, XJ, Wan, XB, Liu, TL, et al. Fibroblast CEBPD/SDF4 axis in response to chemotherapy-induced angiogenesis through CXCR4. *Cell Death Discov*. (2021) 7:94. doi: 10.1038/s41420-021-00478-0
58. Young, SG, Fong, LG, Beigneux, AP, Allan, CM, He, C, Jiang, H, et al. GPIHBP1 and lipoprotein lipase, partners in plasma triglyceride metabolism. *Cell Metab*. (2019) 30:51–65. doi: 10.1016/j.cmet.2019.05.023
59. Wan, H, Huang, X, Cong, P, He, M, Chen, A, Wu, T, et al. Identification of hub genes and pathways associated with idiopathic pulmonary fibrosis via bioinformatics analysis. *Front Mol Biosci*. (2021) 8:711239. doi: 10.3389/fmolb.2021.711239
60. Yao, Y, Li, Z, and Gao, W. Identification of hub genes in idiopathic pulmonary fibrosis and NSCLC progression: evidence from bioinformatics analysis. *Front Genet*. (2022) 13:855789. doi: 10.3389/fgene.2022.855789
61. Jing, X, Sun, W, Yang, X, Huang, H, Wang, P, Luo, Q, et al. CCAAT/enhancer-binding protein (C/EBP) homologous protein promotes alveolar epithelial cell senescence via the nuclear factor-kappa B pathway in pulmonary fibrosis. *Int J Biochem Cell Biol*. (2022) 143:106142. doi: 10.1016/j.biocel.2021.106142
62. Yang, X, Sun, W, Jing, X, Zhang, Q, Huang, H, and Xu, Z. C/EBP homologous protein promotes sonic hedgehog secretion from type II alveolar epithelial cells and activates hedgehog signaling pathway of fibroblast in pulmonary fibrosis. *Respir Res*. (2022) 23:86. doi: 10.1186/s12931-022-02012-x
63. Delbrel, E, Soumare, A, Naguez, A, Label, R, Bernard, O, Bruhat, A, et al. HIF-1alpha triggers ER stress and CHOP-mediated apoptosis in alveolar epithelial cells, a key event in pulmonary fibrosis. *Sci Rep*. (2018) 8:17939. doi: 10.1038/s41598-018-36063-2
64. Preisendorfer, S, Ishikawa, Y, Hennen, E, Winklmeier, S, Schupp, JC, Knuppel, L, et al. FK506-binding protein 11 is a novel plasma cell-specific antibody folding catalyst with increased expression in idiopathic pulmonary fibrosis. *Cells*. (2022) 11:1341. doi: 10.3390/cells11081341
65. Heukels, P, van Hulst, J, van Nimwegen, M, Boersma, CE, Melgert, BN, von der Thusen, JH, et al. Enhanced Bruton's tyrosine kinase in B-cells and autoreactive IgA in patients with idiopathic pulmonary fibrosis. *Respir Res*. (2019) 20:232. doi: 10.1186/s12931-019-1195-7
66. Zhang, L, Wang, Y, Wu, G, Xiong, W, Gu, W, and Wang, CY. Macrophages: friend or foe in idiopathic pulmonary fibrosis? *Respir Res*. (2018) 19:170. doi: 10.1186/s12931-018-0864-2
67. Mills, CD, and Ley, K. M1 and M2 macrophages: the chicken and the egg of immunity. *J Innate Immun*. (2014) 6:716–26. doi: 10.1159/000364945
68. Inomata, M, Kamio, K, Azuma, A, Matsuda, K, Kokuho, N, Miura, Y, et al. Pirfenidone inhibits fibrocyte accumulation in the lungs in bleomycin-induced murine pulmonary fibrosis. *Respir Res*. (2014) 15:16. doi: 10.1186/1465-9921-15-16
69. Tarique, AA, Logan, J, Thomas, E, Holt, PG, Sly, PD, and Fantino, E. Phenotypic, functional, and plasticity features of classical and alternatively activated human macrophages. *Am J Respir Cell Mol Biol*. (2015) 53:676–88. doi: 10.1165/rcmb.2015-00120C
70. Bocchino, M, Zanotta, S, Capitelli, L, and Galati, D. Dendritic cells are the intriguing players in the puzzle of idiopathic pulmonary fibrosis pathogenesis. *Front Immunol*. (2021) 12:664109. doi: 10.3389/fimmu.2021.664109
71. Bantsimba-Malanda, C, Marchal-Somme, J, Goven, D, Freynet, O, Michel, L, Crestani, B, et al. A role for dendritic cells in bleomycin-induced pulmonary fibrosis in mice? *Am J Respir Crit Care Med*. (2010) 182:385–95. doi: 10.1164/rccm.200907-1164OC
72. Tort, TM, Aschenbrenner, F, Maus, R, Stolper, J, Schuette, L, Knudsen, L, et al. The FMS-like tyrosine kinase-3 ligand/lung dendritic cell axis contributes to regulation of pulmonary fibrosis. *Thorax*. (2019) 74:947–57. doi: 10.1136/thoraxjnl-2018-212603

73. Wang, Z, Qu, S, Zhu, J, Chen, F, and Ma, L. Comprehensive analysis of lncRNA-associated competing endogenous RNA network and immune infiltration in idiopathic pulmonary fibrosis. *J Thorac Dis.* (2020) 12:1856–65. doi: 10.21037/jtd-19-2842
74. Gharaee-Kermani, M, and Phan, SH. The role of eosinophils in pulmonary fibrosis (review). *Int J Mol Med.* (1998) 1:43–53. doi: 10.3892/ijmm.1.1.43
75. Zhang, H, Ding, C, Li, Y, Xing, C, Wang, S, Yu, Z, et al. Data mining-based study of collagen type III alpha 1 (COL3A1) prognostic value and immune exploration in pancreatic cancer. *Bioengineered.* (2021) 12:3634–46. doi: 10.1080/21655979.2021.1949838
76. Yin, W, Zhu, H, Tan, J, Xin, Z, Zhou, Q, Cao, Y, et al. Identification of collagen genes related to immune infiltration and epithelial-mesenchymal transition in glioma. *Cancer Cell Int.* (2021) 21:276. doi: 10.1186/s12935-021-01982-0
77. Organ, LA, Duggan, AR, Oballa, E, Taggart, SC, Simpson, JK, Kang'Ombe, AR, et al. Biomarkers of collagen synthesis predict progression in the PROFILE idiopathic pulmonary fibrosis cohort. *Respir Res.* (2019) 20:148. doi: 10.1186/s12931-019-1118-7
78. Ren, H, Liu, X, Li, F, He, X, and Zhao, N. Identification of a six gene prognosis signature for papillary thyroid cancer using multi-omics methods and bioinformatics analysis. *Front Oncol.* (2021) 11:624421. doi: 10.3389/fonc.2021.624421
79. Han, N, He, J, Shi, L, Zhang, M, Zheng, J, and Fan, Y. Identification of biomarkers in nonalcoholic fatty liver disease: a machine learning method and experimental study. *Front Genet.* (2022) 13:1020899. doi: 10.3389/fgene.2022.1020899
80. Wang, H, Yu, T, and Mao, L. Placental-cadherin, a biomarker for local immune status and poor prognosis among patients with tongue squamous cell carcinoma. *Eur Arch Otorhinolaryngol.* (2021a) 279:3597–609. doi: 10.1007/s00405-021-07181-x
81. Zhang, K, Xu, Z, and Sun, Z. Identification of the key genes connected with plasma cells of multiple myeloma using expression profiles. *Onco Targets Ther.* (2015) 8:1795–803. doi: 10.2147/OTT.S80075
82. Li, S, Zhao, W, and Sun, M. An analysis regarding the association between the ISLR gene and gastric carcinogenesis. *Front Genet.* (2020) 11:620. doi: 10.3389/fgene.2020.00620
83. Loi, H, Kramar, S, Laborde, C, Marsal, D, Pizzinat, N, Cussac, D, et al. Metformin attenuates postinfarction myocardial fibrosis and inflammation in mice. *Int J Mol Sci.* (2021) 22:9393. doi: 10.3390/ijms22179393
84. Ullmann, T, Luckhardt, S, Wolf, M, Parnham, MJ, and Resch, E. High-throughput screening for CEBPD-modulating compounds in THP-1-derived reporter macrophages identifies anti-inflammatory HDAC and BET inhibitors. *Int J Mol Sci.* (2021) 22:3022. doi: 10.3390/ijms22063022