Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Integrative Single-Cell Transcriptomics and Machine Learning Identify Pentose Phosphate Pathway-Related Core Biomarkers for Diagnosis of Idiopathic Pulmonary Fibrosis

Yong qi Cheng
Wuxi People's Hospital
Shi qing Yuan
Jiangnan University
Ming feng Zheng
zhengmf68@sina.com

Wuxi People's Hospital

Research Article

Keywords: Idiopathic pulmonary fibrosis, Pentose phosphate pathway, Metabolic reprogramming, SPP1 + alveolar macrophage, Machine learning

Posted Date: June 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9927320/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Integrative Single-Cell Transcriptomics and Machine Learning Identify Pentose Phosphate Pathway-Related Core Biomarkers for Diagnosis of Idiopathic Pulmonary Fibrosis

Yongqi Cheng[1,#], Shiqing Yuan[2,#], Mingfeng Zheng[1,*]

[1] Department of Thoracic Surgery, The Affiliated Wuxi People's Hospital of Nanjing Medical University, Wuxi People's Hospital, Wuxi Medical Center, Nanjing Medical University, Wuxi 214023, China

[2] School of Wuxi Medicine, Jiangnan University, Wuxi, Jiangsu 214000, China

#These authors contributed equally to this work and should be considered co-first authors.

***Corresponding author:** Mingfeng Zheng, Department of Thoracic Surgery, The Affiliated Wuxi People's Hospital of Nanjing Medical University, Wuxi People's Hospital, Wuxi Medical Center, Nanjing Medical University, Wuxi 214023, China

E-mail: zhengmf68@sina.com

Abstract

Background: Idiopathic pulmonary fibrosis (IPF) carries a poor prognosis, and metabolic reprogramming is a key driver of its progression. However, the cell-type-specific activation patterns and key regulatory genes of the pentose phosphate pathway (PPP) in IPF remain poorly defined.

Methods: Single-cell transcriptomic data of IPF were analyzed to profile PPP metabolic activity using ssGSEA and UCell. Bulk phenotypes were mapped to the single-cell level via the Scissor algorithm. Core diagnostic genes were identified by integrating hdWGCNA co-expression network analysis with five machine learning algorithms, followed by construction of a nomogram diagnostic model. Immune microenvironment and pathway characteristics were dissected using CIBERSORT and GSVA. Monocle 3 and CellChat were applied to delineate the functional trajectory and intercellular communication of alveolar macrophage subsets. Targeted therapeutic compounds were predicted using DGIdb. Finally, human IPF lung tissues and paired normal lung tissues were collected for protein-level validation by Western blot.

Results: Alveolar macrophages exhibited the highest baseline PPP activity among all lung cell types, with Scissor+ IPF-phenotype cells predominantly enriched in this population. Independent validation using UCell further confirmed global glucose metabolic remodeling in IPF alveolar macrophages. Four core diagnostic genes—BPIFB1, CCL18, IGFBP2, and WFDC2—were identified at the intersection of all five machine learning algorithms. The constructed nomogram achieved an area under the receiver operating characteristic curve (AUC) of 0.889. Immune infiltration analysis revealed enrichment of M0/M2 macrophages and mast cells in IPF lungs. Secondary clustering of alveolar macrophages identified an SPP1+ profibrotic subset with markedly expanded proportions in IPF. CellChat analysis demonstrated significantly enhanced cell–cell communication between SPP1+ macrophages and fibroblasts, driven by ECM-related signaling pathways.

Conclusion: This study delineates the landscape of PPP metabolism in IPF at single-cell resolution, establishes a high-performance diagnostic model based on four core genes, and identifies SPP1+ alveolar macrophages as the key cellular subset exhibiting the most active PPP metabolism and the densest profibrotic intercellular communication in the IPF microenvironment. These findings provide novel insights into the mechanisms underlying metabolism–immunity–stroma crosstalk in IPF.

Keywords: Idiopathic pulmonary fibrosis; Pentose phosphate pathway; Metabolic reprogramming; SPP1+ alveolar macrophage; Machine learning

1. Introduction

Idiopathic pulmonary fibrosis (IPF) is a chronic progressive interstitial lung disease of unknown etiology, characterized by diffuse alveolar epithelial injury, fibroblastic foci formation, and irreversible scarring, with a histopathological pattern of usual interstitial pneumonia (UIP) [1,6]. The global incidence of IPF ranges from 2.8 to 9.3 per 100,000 individuals and continues to rise annually [2]. Without treatment, the median survival is only 3–5 years, and the 5-year survival rate is lower than that of most malignant tumors [1]. Although the currently approved antifibrotic agents pirfenidone and nintedanib can slow lung function decline, they cannot reverse fibrosis progression, and many patients experience treatment failure or acute exacerbation [4,5]. The pathogenesis of IPF involves recurrent microinjury of the alveolar epithelium, aberrant repair responses, immune dysregulation, and persistent activation of myofibroblasts [6,8]. Elucidating the core molecular mechanisms underlying IPF and identifying highly specific

diagnostic biomarkers and therapeutic targets remain critical unmet needs in respiratory medicine [7].

Recent advances in single-cell RNA sequencing (scRNA-seq) have enabled high-resolution dissection of cellular heterogeneity in IPF lung tissues. Reyfman et al. first revealed aberrant epithelial transitions, fibroblast heterogeneity, and profibrotic macrophage subsets in IPF via scRNA-seq [9]. Morse et al. further demonstrated that SPP1+/MERTK+ macrophages represent a pathognomonic profibrotic subset with sustained proliferative capacity in IPF [10]. Adams and Habermann et al. identified a KRT5−/KRT17+ aberrant basal-like cell population and established the spatial colocalization of SPP1+ macrophages with myofibroblasts [11,12]. Strunz et al. showed that a Krt8+ transitional stem cell state persists in human pulmonary fibrosis [13], whereas Tsukui et al. constructed a collagen-producing cell atlas and identified multiple profibrotic fibroblast subtypes [14]. Collectively, these studies highlight that dissecting IPF heterogeneity at single-cell resolution is pivotal for discovering novel diagnostic biomarkers and therapeutic targets.

In parallel with cellular profiling, metabolic reprogramming in the tissue microenvironment has emerged as a key driver of fibrosis progression [15]. In fibrotic lungs, energy metabolism shifts from oxidative phosphorylation toward aerobic glycolysis, accompanied by lactate accumulation, which activates TGF-β1 in a pH-dependent manner, thereby establishing a positive profibrotic feedback loop [16,17]. Glycolytic intermediates can be shunted into the pentose phosphate pathway (PPP), a central glucose metabolic branch with glucose-6-phosphate dehydrogenase (G6PD) as the rate-limiting enzyme. The PPP primarily generates NADPH and ribose-5-phosphate, supporting redox homeostasis and nucleotide synthesis, respectively [18,19]. Metabolomic analyses have confirmed significantly elevated PPP intermediate levels in IPF lung tissues compared with healthy controls [20,21]. Hamanaka et al. reported that fibroblast activation is accompanied by coordinated upregulation of the PPP, one-carbon metabolism, and glutamine metabolism [22]. Furthermore, a study employing a small-molecule G6PD inhibitor established the essential role of the PPP in the inflammatory effector functions of lymphocytes and macrophages [23].

Although an association between the PPP and pulmonary fibrosis has been suggested, clinically specific biomarkers reflecting PPP activity in IPF remain lacking, and the cell-type-specific mechanisms and key regulatory genes driving PPP dysregulation in IPF are poorly defined. In the present study, we integrated scRNA-seq data and the clinical cohort GSE47460 to systematically profile PPP activity in IPF lungs. Using an integrative framework combining Scissor phenotype mapping [24], hdWGCNA network analysis [25], and five machine learning algorithms (LASSO, SVM-RFE, random forest, Boruta, and XGBoost), we identified core hub genes associated with PPP dysregulation

and constructed a nomogram diagnostic model. At single-cell resolution, secondary clustering of alveolar macrophages, Monocle 3 pseudotime analysis [26], and CellChat communication analysis [27] revealed the differentiation trajectory of SPP1+ profibrotic macrophages and their metabolism–stroma crosstalk with fibroblasts. Finally, we validated core biomarker protein expression in human IPF lung tissues by Western blot and predicted targeted therapeutic compounds, laying a molecular foundation for early screening and metabolism-driven personalized intervention in IPF.

2. Materials and Methods

2.1 Data Acquisition and Processing

All data were retrieved from the NCBI Gene Expression Omnibus (GEO) database. The scRNA-seq dataset GSE128033, comprising 12 IPF and 6 healthy control samples, was used for single-cell PPP activity assessment, Scissor phenotype mapping, and secondary clustering of alveolar macrophages. Bulk RNA-seq datasets were divided into a training set and an independent validation set: the training set GSE47460 included 193 IPF and 91 healthy control samples and was used for differential expression analysis, hdWGCNA co-expression network construction, machine learning-based feature selection, and nomogram development with internal validation; the independent validation set GSE53845 contained 40 IPF and 8 healthy control samples and was used for external validation of core gene diagnostic performance. All bulk RNA-seq data were $\log_2(X+1)$-normalized, and batch effects were corrected using the sva package. For scRNA-seq data, high-quality cells with a mitochondrial gene proportion <20% were retained, and genes expressed in fewer than 3 cells were excluded. Data integration followed the standard Seurat v4 workflow [28], with principal component analysis (PCA) for dimensionality reduction, batch effect removal using Harmony [29], and cell clustering using the FindClusters function. Cell type annotation combined automated annotation by SingleR with manual curation based on canonical marker expression patterns.

2.2 Gene Set Scoring and Metabolic Activity Analysis

Single-cell pathway activity scores for the PPP, glycolysis/gluconeogenesis, galactose metabolism, fructose and mannose metabolism, and the tricarboxylic acid (TCA) cycle were computed using the ssGSEA algorithm [30]. For robustness, the UCell algorithm [31], which is optimized for sparse single-cell data, was used for independent validation, with results cross-referenced against ssGSEA scores.

2.3 Scissor Algorithm and Differential Expression Analysis

The Scissor algorithm [24] was applied to integrate bulk disease phenotypes with scRNA-seq data via an optimized network-regularized sparse regression model, identifying cell subsets significantly associated with the IPF phenotype. Differential expression analysis of bulk RNA-seq data was performed using the limma package [32], with thresholds of adjusted P-value (P.adj) <0.05 and $|\log_2$ fold-change (logFC)| >0.585. Multiple testing correction was performed using the Benjamini–Hochberg procedure.

2.4 High-Dimensional Weighted Gene Co-Expression Network Analysis

The R package hdWGCNA [25] was used to construct scale-free co-expression networks at the single-cell level, reducing dropout noise through metacell matrix construction for specific cell types. The ssGSEA-derived PPP score was used as the primary trait variable, with IPF case/control status serving as an auxiliary trait for module–trait correlation analysis. Genes within modules most strongly correlated with PPP scores were included in downstream feature selection.

2.5 Machine Learning and Diagnostic Model Construction

Five complementary machine learning algorithms were employed for feature selection: LASSO regression with 5-fold cross-validation [34], SVM-RFE [35], random forest [36], Boruta [37], and XGBoost [38]. The intersection of genes identified by all five algorithms was defined as the set of core diagnostic biomarkers for IPF. A clinical nomogram model was constructed using the R package rms. Diagnostic performance was evaluated by the area under the receiver operating characteristic curve (AUC), and model accuracy and clinical utility were assessed using calibration curves and decision curve analysis, respectively.

2.6 Immune Infiltration, GSVA, and Drug Prediction

The relative proportions of 22 immune cell subtypes in the GSE47460 dataset were estimated using CIBERSORT [39]. Gene set variation analysis (GSVA) [40] was performed based on MSigDB KEGG pathways to compare pathway activity between IPF and control groups, with a false discovery rate (FDR) <0.05 considered statistically significant (Benjamini–Hochberg correction). Targeted compounds for core diagnostic genes were predicted using the DGIdb database [41].

2.7 Secondary Clustering, Pseudotime, and Cell Communication Analysis of Alveolar Macrophages

Alveolar macrophages were extracted and subjected to secondary clustering, yielding four transcriptionally distinct subtypes. Differentiation trajectories were reconstructed using Monocle 3 [26] based on UMAP embeddings, with

distribution differences between healthy control (HC) and IPF groups compared. Cell–cell communication networks were inferred using CellChat [27] based on the CellChatDB ligand–receptor database, with communication probabilities quantified via mass-action modeling, and interaction differences between HC and IPF groups analyzed.

2.8 Human Lung Tissue Collection

IPF lung tissues were obtained from end-stage IPF patients (n = 6) undergoing lung transplantation at Wuxi People's Hospital, with diagnosis confirmed per the 2018 ATS/ERS/JRS/ALAT clinical practice guidelines [3] and postoperative histological confirmation of the UIP pattern. Control tissues were adjacent normal lung tissues (n = 6) obtained from patients undergoing surgical resection for lung cancer, collected >5 cm from the tumor margin, from donors without chronic pulmonary disease. Tissues were harvested within 30 minutes of resection, rinsed with phosphate-buffered saline (PBS) to remove residual blood, snap-frozen in liquid nitrogen, and stored at −80°C until analysis. All patients provided written informed consent prior to enrollment. The study protocol was approved by the Ethics Committee of Wuxi People's Hospital (Approval No. ky25140) and conducted in accordance with the principles of the Declaration of Helsinki.

2.9 Western Blot

Lung tissues were homogenized in RIPA lysis buffer supplemented with protease inhibitors, and total protein concentration was quantified using the BCA protein assay. A total of 20 μg of protein per sample was separated by 10% sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE) and transferred to polyvinylidene fluoride (PVDF) membranes. Membranes were blocked with 5% non-fat dry milk in Tris-buffered saline with Tween-20 (TBST) for 1 hour at room temperature, followed by overnight incubation at 4°C with primary antibodies against BPIFB1, CCL18, IGFBP2, WFDC2, and GAPDH (detailed antibody information is provided in Supplementary Table S1). After washing, membranes were incubated with horseradish peroxidase (HRP)-conjugated secondary antibodies for 1 hour at room temperature. Protein bands were visualized using enhanced chemiluminescence (ECL). Band intensities were quantified using ImageJ (v1.53) and normalized to GAPDH as the loading control. Each sample was analyzed in triplicate.

2.10 Statistical Analysis

All bioinformatic analyses were performed using R v4.4.1, with $P < 0.05$ considered statistically significant. Normally distributed continuous data were compared using independent two-sample t-tests, and non-normally distributed data

were compared using the Mann–Whitney U test. Western blot results are presented as mean ± standard deviation. Between-group comparisons of Western blot data were performed using independent t-tests in GraphPad Prism v9.0.

3. Results

3.1 Single-Cell Landscape and PPP Metabolic Heterogeneity in IPF Lungs

Following quality control, dimensionality reduction, and unsupervised clustering via the standard Seurat workflow, nine major lung cell types were identified based on canonical marker gene expression (AGER and CAV1 for AT1 cells; SFTPD and SFTPB for AT2 cells; FOXJ1 for ciliated cells; PECAM1 for endothelial cells; COL1A1 for fibroblasts; MARCO for alveolar macrophages; KIT for mast cells; CD3D for T cells; CD19 for B cells; Figures 1A–1B).

ssGSEA-based single-cell scoring revealed substantial heterogeneity in PPP activity across cell lineages, with alveolar macrophages exhibiting the highest PPP scores, accompanied by significant activation of glycolysis and the TCA cycle (Figure 1C). UMAP and violin plots of PPP scores further confirmed alveolar macrophages as the cell type with the globally highest PPP activity (Figures 1D–1E). Scissor mapping of bulk IPF phenotypes onto the single-cell atlas showed that Scissor+ cells were most significantly enriched in alveolar macrophages (4,440 cells), far exceeding all other cell types (Figures 1F–1G).

Independent validation using UCell confirmed robust PPP activation in IPF alveolar macrophages, with coordinated upregulation of glycolysis, the TCA cycle, fructose/mannose metabolism, and galactose metabolism, indicating global glucose metabolic remodeling in this population (Supplementary Figure S1A). Cell-type-specific comparison of PPP scores between IPF and control groups further verified that PPP activation in IPF is cell-type specific (Supplementary Figure S1B), consistent with ssGSEA results.

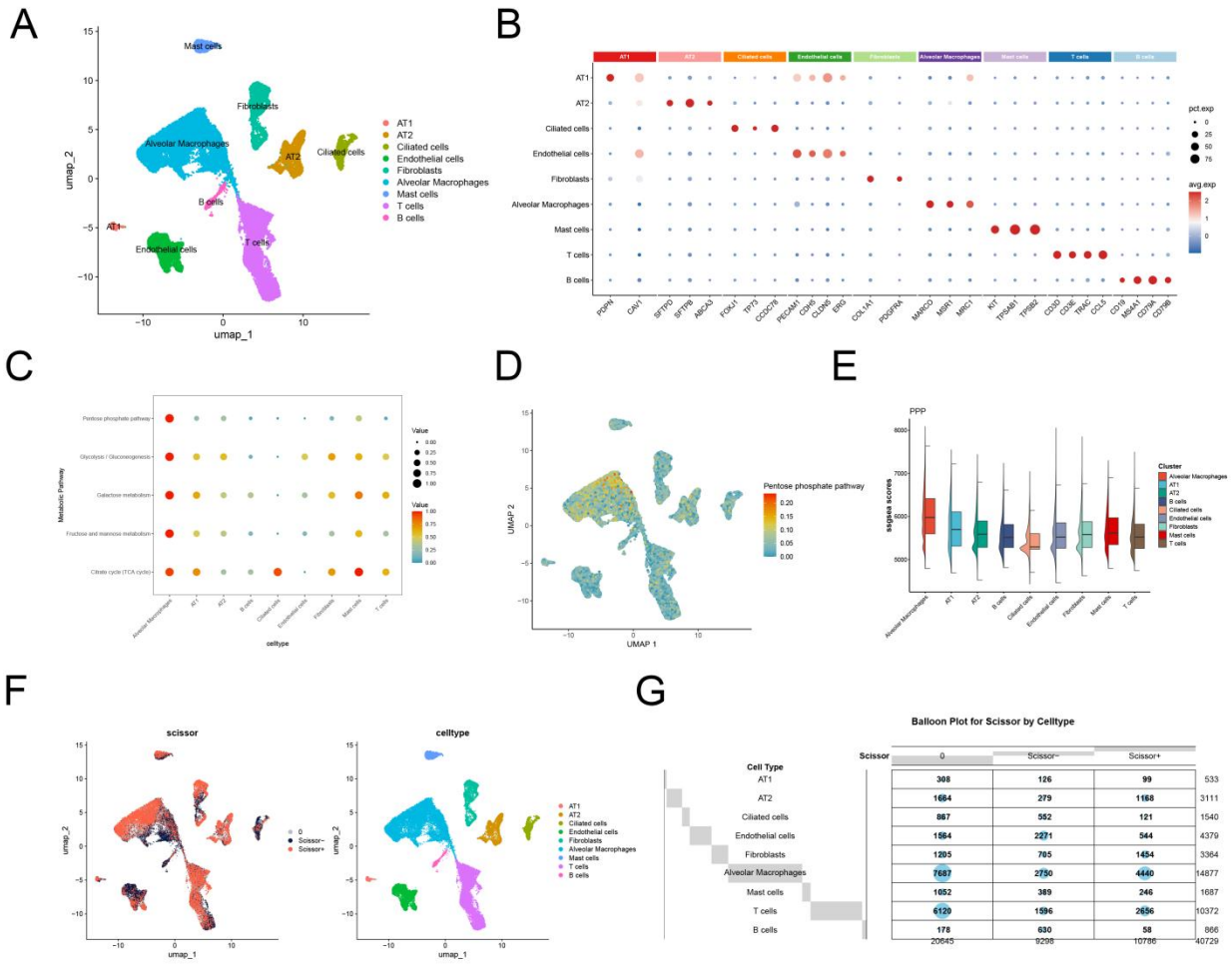


Fig 1. Single-cell landscape and PPP metabolism in IPF lungs. (A) UMAP distribution of nine identified cell types. (B) Bubble plot of canonical marker genes per cell type. (C) ssGSEA scores for five glucose metabolic pathways across cell types. (D) UMAP visualization of PPP activity scores. (E) Violin plot of PPP scores across cell types. (F) UMAP distribution of Scissor+/Scissor− cells. (G) Bar plot of Scissor+ cell counts per cell type.

3.2 Multi-Omics Integration Identifies PPP-Related Core Genes

Differential expression analysis of the training set GSE47460 identified numerous differentially expressed genes (DEGs) in IPF lungs. Among the top upregulated genes were IL13RA2, IGFL2, COL17A1, MMP1, and KRT5; among the top downregulated genes were CSF3, SLC6A4, FOSB, ITLN2, and SERTM1 (Figures 2A–2B). hdWGCNA partitioned the transcriptome into multiple co-expression modules (Figure 2C). Module–trait correlation analysis revealed that the brown and yellow modules were most strongly correlated with PPP scores in alveolar macrophages

and with IPF status (Figure 2D). Intersection of DEGs from GSE47460 with genes within the key hdWGCNA modules yielded 11 core overlapping candidate genes (Figure 2E).

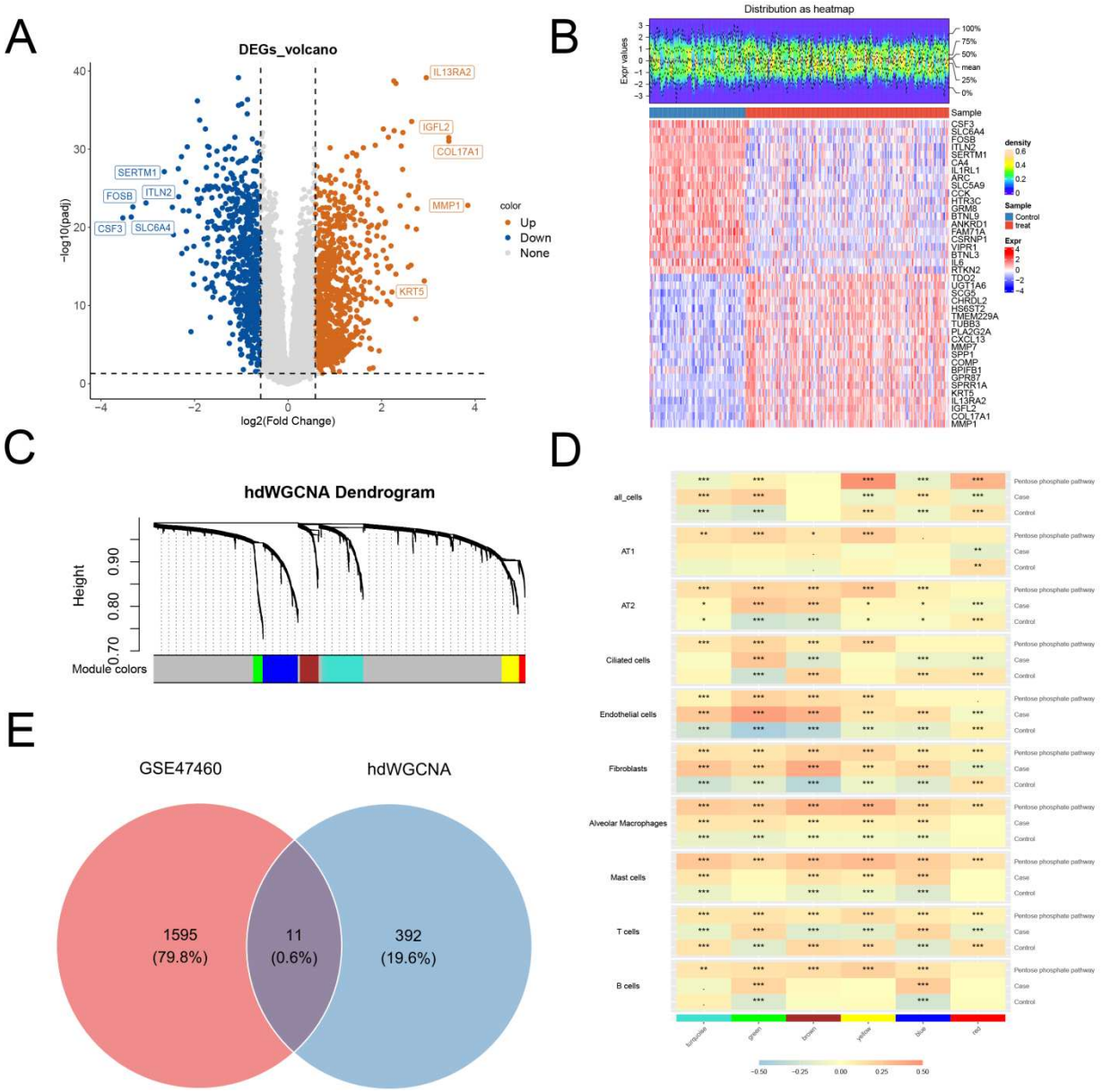


Fig 2. Multi-omics integration identifies PPP-related core genes. (A) Volcano plot of DEGs in the training set GSE47460. (B) Heatmap of the top 50 DEGs. (C) hdWGCNA dendrogram and module assignment. (D) Module–trait correlation heatmap. (E) Venn diagram of DEGs and hdWGCNA module genes.

3.3 Five Machine Learning Algorithms Identify Four Core Diagnostic Biomarkers

LASSO regression with 5-fold cross-validation determined the optimal penalization parameter lambda (Figures 3A–3B). SVM-RFE achieved peak classification accuracy (0.8378) and minimal error rate (0.1622) with six retained features (Figures 3C–3D). The random forest out-of-bag error curve stabilized at approximately 200 trees (Figure 3E), with the top eight genes ranked by Mean Decrease in Gini index: WFDC2, CCL18, IGFBP2, PLA2G7, AGR3, AGR2, TSPAN1, and BPIFB1 (Figure 3F). XGBoost ranked the top eight genes by Gain score: WFDC2, IGFBP2, CCL18, PLA2G7, SCGB1A1, AGR3, AGR2, and BPIFB1 (Figure 3G). Boruta identified significant features including CCL18, WFDC2, IGFBP2, PLA2G7, AGR3, AGR2, TSPAN1, and BPIFB1 (Figure 3H). The intersection of genes selected by all five algorithms yielded four core diagnostic biomarkers: BPIFB1, CCL18, IGFBP2, and WFDC2 (Figure 3I).

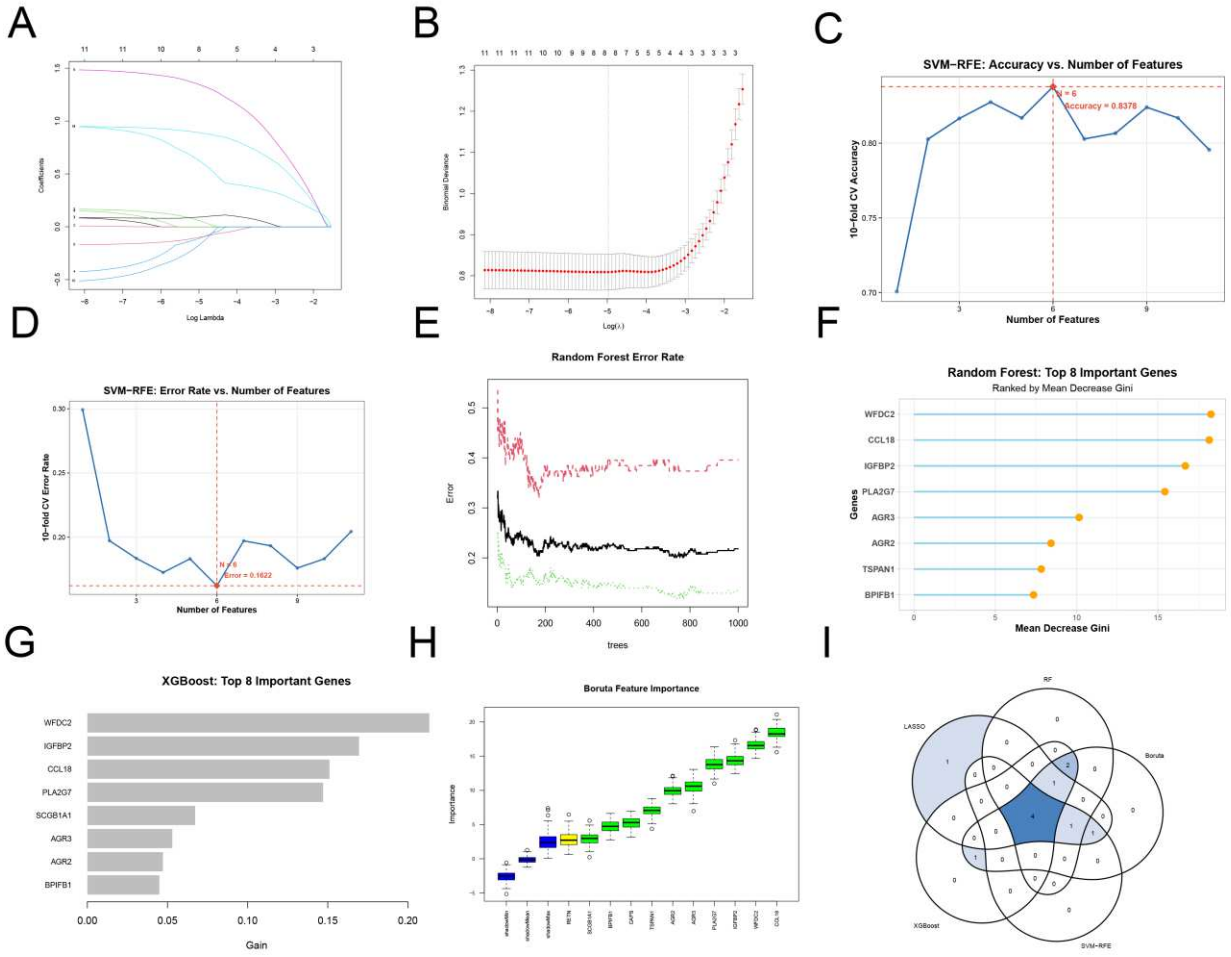


Fig 3. Five machine learning algorithms identify core diagnostic genes. (A–B) LASSO coefficient paths and cross-validation error curve. (C–D) SVM-RFE accuracy and error curves. (E) Random forest out-of-bag error curve. (F)

Top eight genes ranked by Mean Decrease Gini (random forest). (G) Top eight genes ranked by Gain (XGBoost). (H) Boruta feature importance boxplot. (I) Venn diagram of genes selected by all five algorithms.

3.4 Diagnostic Performance and External Validation of Core Genes

In the training set GSE47460, ROC analysis demonstrated that WFDC2 had the highest individual AUC (0.799), followed by IGFBP2 (0.777), CCL18 (0.758), and BPIFB1 (0.712) (Figure 4A). All four genes were significantly upregulated in IPF samples relative to controls ($P < 0.05$; Figure 4B). External validation in the independent dataset GSE53845 (40 IPF/8 controls) confirmed robust diagnostic performance: IGFBP2 AUC = 0.872, BPIFB1 AUC = 0.812, WFDC2 AUC = 0.812, and CCL18 AUC = 0.787 (Figure 4C). All four genes were also significantly upregulated in the validation cohort (Figure 4D). The consistent results across independent datasets confirm the reliability and reproducibility of these four core genes as diagnostic biomarkers for IPF.

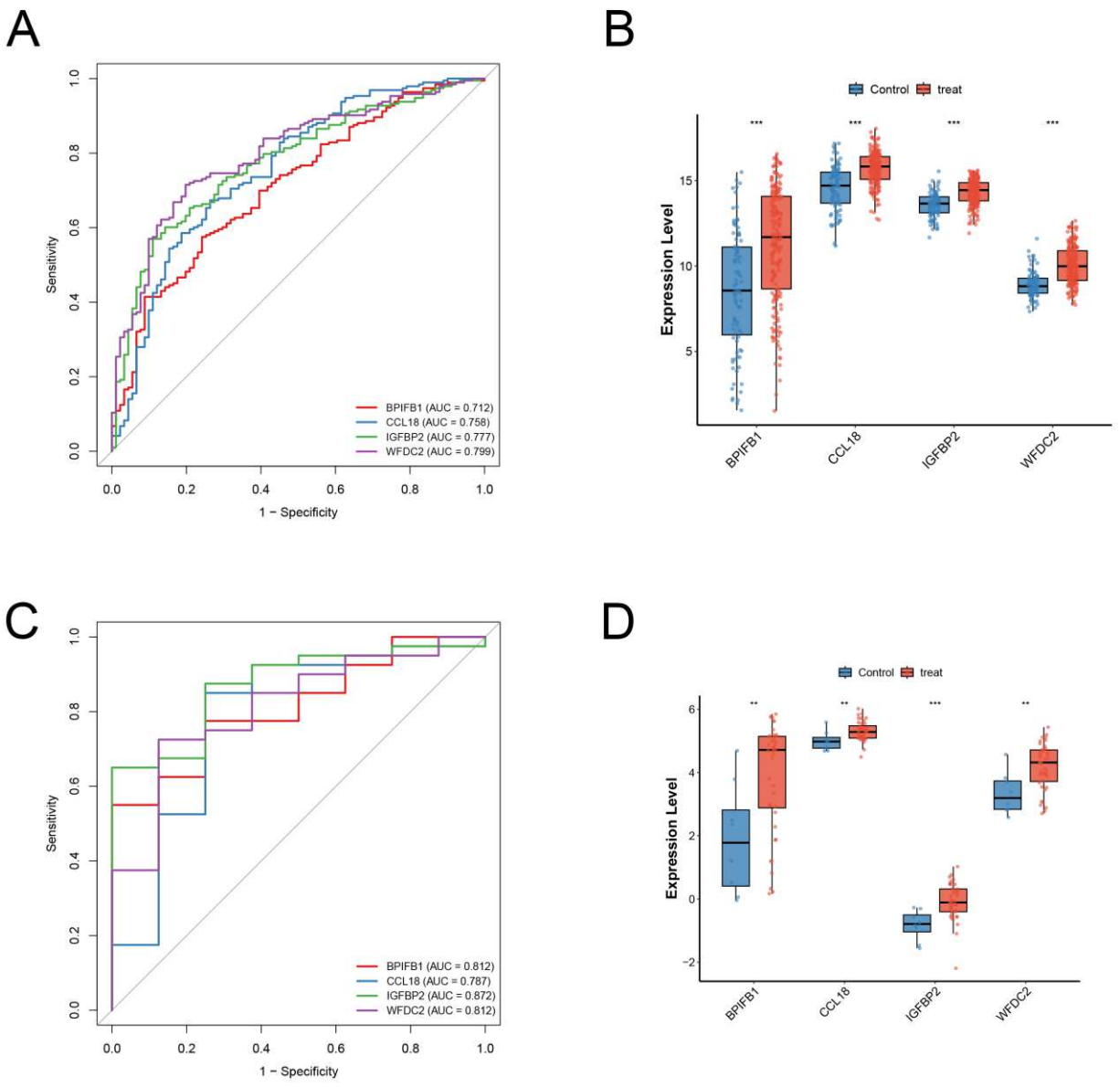


Fig 4. Diagnostic performance and external validation of core genes. (A) ROC curves in the training set GSE47460. (B) Expression boxplots in GSE47460. (C) ROC curves in the independent validation set GSE53845. (D) Expression boxplots in GSE53845.

3.5 Construction and Evaluation of an IPF Diagnostic Nomogram

A nomogram model integrating the four core genes was constructed based on the training set GSE47460, with higher regression weights assigned to IGFBP2 and CCL18 relative to BPIFB1 and WFDC2 (Figure 5A). The bias-corrected

calibration curve closely approximated the ideal diagonal, with a mean absolute error of 0.013, indicating excellent agreement between predicted and observed probabilities (Figure 5B). Decision curve analysis demonstrated that the nomogram provided superior clinical net benefit across a wide range of clinically relevant threshold probabilities compared with single-gene models and the all-positive or all-negative strategies (Figure 5C). The integrated nomogram achieved a composite AUC of 0.889, outperforming each individual gene (Figure 5D).

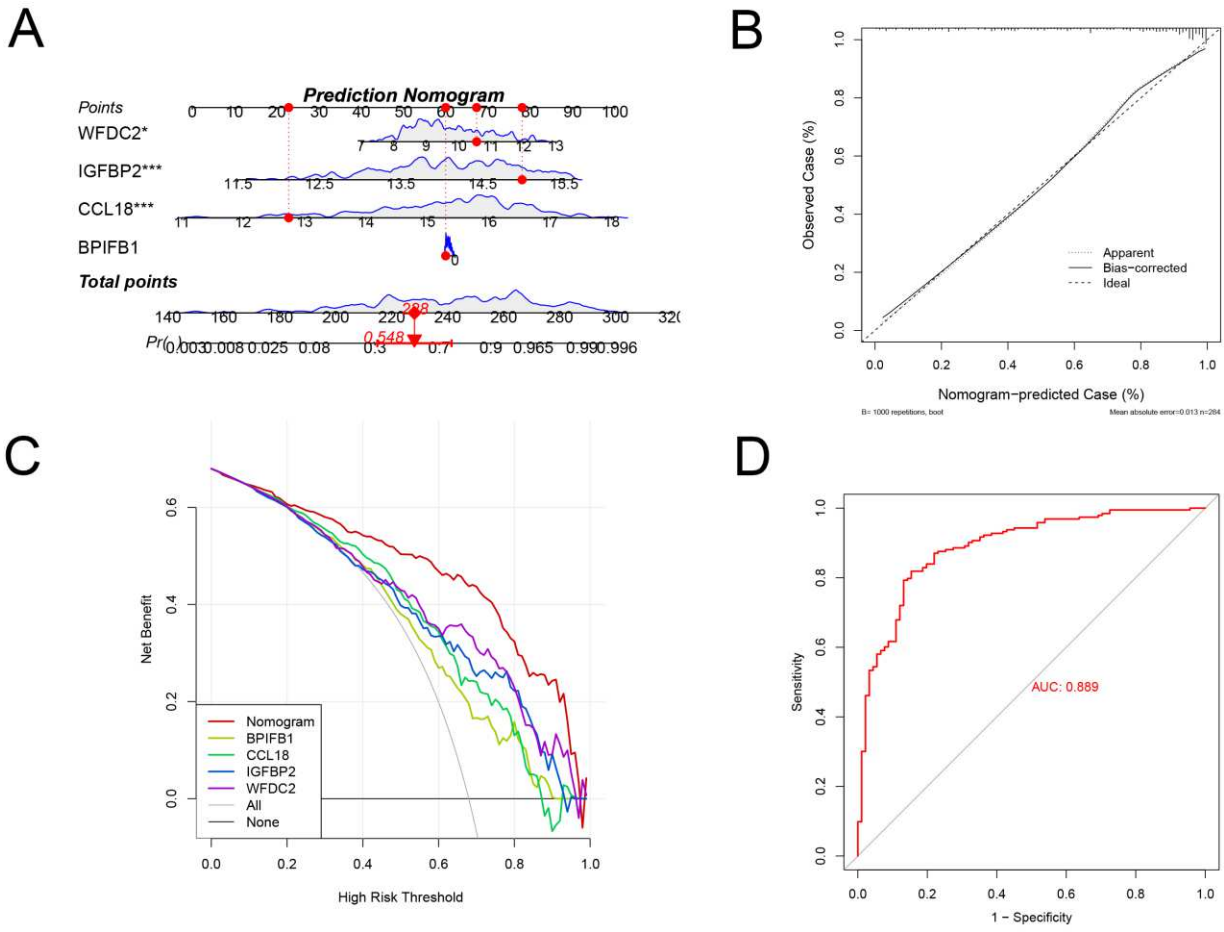


Fig 5. IPF diagnostic nomogram. (A) Nomogram based on four core genes. (B) Bias-corrected calibration curve. (C) Decision curve analysis comparing the nomogram to single-gene models. (D) ROC curve of the integrated nomogram .

3.6 Immune Microenvironment and Pathway Characteristics in IPF

CIBERSORT analysis revealed significant differences in immune cell composition between IPF and control groups

(Figure 6A). Compared with controls, IPF lungs showed increased proportions of M0 and M2 macrophages, activated mast cells, plasma cells, and selected T-cell subsets, with correspondingly reduced M1 macrophages, classical monocytes, resting CD4+ memory T cells, activated natural killer (NK) cells, and neutrophils (Figure 6B).

GSVA identified significantly upregulated pathways in IPF, including alanine/aspartate/glutamate metabolism, N-glycan biosynthesis, protein export, homologous recombination, non-homologous end joining, nucleotide excision repair, p53 signaling, base excision repair, one-carbon folate metabolism, and DNA replication. Significantly downregulated pathways included Wnt signaling, MAPK signaling, tight junction assembly, adherens junction, endocytosis, and leukocyte transendothelial migration (Figures 6C–6D). DGIdb-based drug prediction identified multiple potential therapeutic compounds targeting the four core genes (Figure 6E).

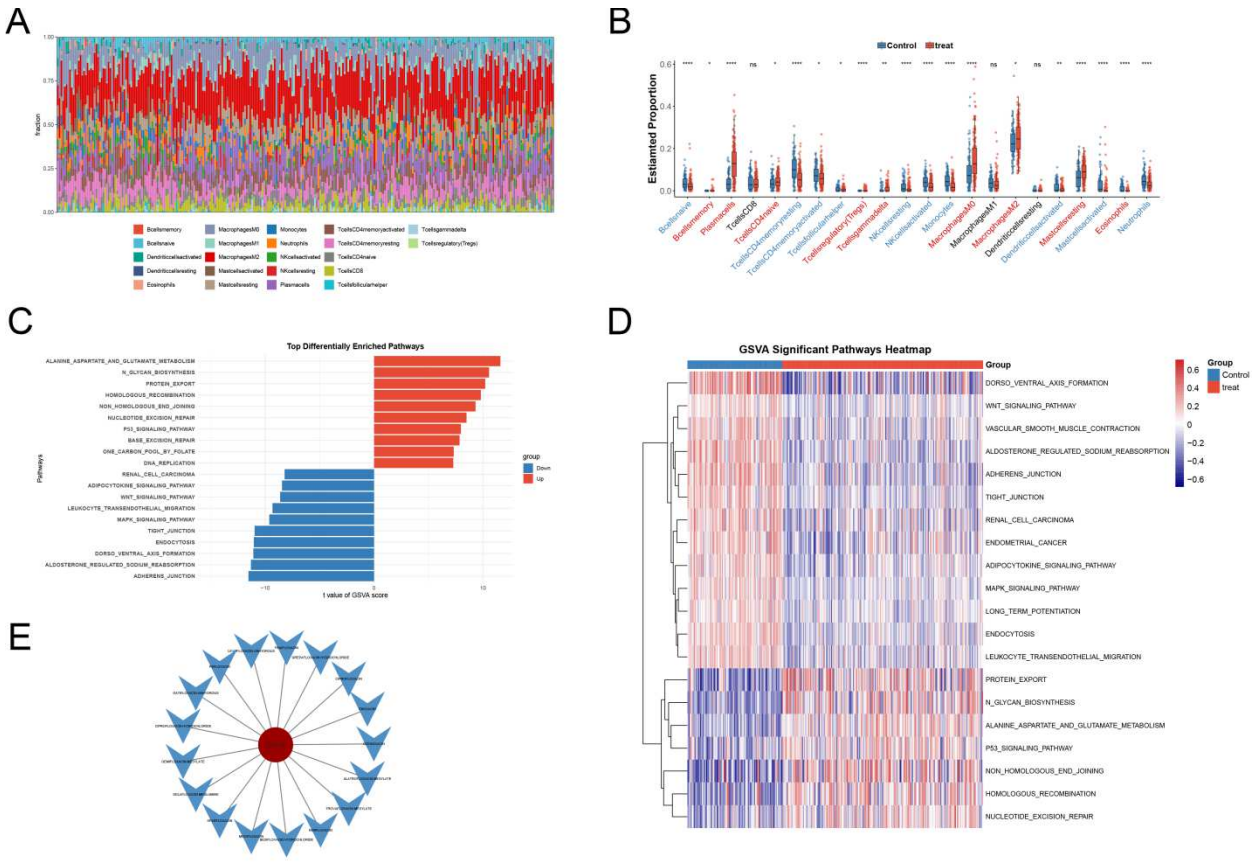


Fig 6. Immune microenvironment, pathway characteristics, and drug targets in IPF. (A) Stacked bar plot of 22 immune cell proportions across samples. (B) Boxplots of immune cell proportion differences between IPF and control groups. (C) Bar plot of significantly altered GSVA pathways. (D) Heatmap of GSVA pathway activity scores. (E) Core gene–drug interaction network from DGIdb.

3.7 Secondary Clustering, Cell Communication, and Pseudotime Analysis of Alveolar Macrophages

Secondary clustering of alveolar macrophages identified four transcriptionally distinct subsets: FABP4+ tissue-resident macrophages (TRMs), SPP1+ macrophages, FCN1+ macrophages, and SELENOP+ macrophages (Figures 7A–7B). IPF lungs showed expanded proportions of SPP1+ and SELENOP+ macrophages, with a concomitant reduction in FABP4+ TRMs (Figure 7C).CellChat analysis revealed a marked increase in the total number of cell–cell interactions in IPF (518 vs. 338 in HC) and in overall interaction strength (10.032 vs. 5.947 in HC) (Figures 7D–7F). Differential communication strength heatmaps identified enhanced signaling output from endothelial cells, fibroblasts, and SPP1+ macrophages in IPF (Figure 7G). Ligand–receptor bubble plots revealed activation of ECM-related signaling pathways—including SPP1–CD44, COLLAGEN, and FN1–CD44—from macrophage subsets to fibroblasts in IPF (Figure 7H). SPP1+ macrophages formed dense ligand–receptor interaction networks with fibroblasts, endothelial cells, and AT1 cells (Figure 7I). Information flow analysis confirmed upregulation of profibrotic signaling pathways (COLLAGEN, LAMININ, FN1, SPP1, MIF, and GALECTIN) in IPF (Figure 7J).Monocle 3 pseudotime analysis revealed continuous differentiation trajectories of alveolar macrophages (Figures 7K–7L). HC cells were predominantly localized to early trajectory stages (mainly FABP4+ TRMs), whereas IPF cells accumulated preferentially in mid-to-late trajectory branches (Figure 7M). Pseudotime density distribution confirmed FABP4+ TRMs in early pseudotime, SELENOP+ macrophages in mid-pseudotime, and SPP1+/FCN1+ macrophages in late pseudotime (Figure 7N), indicating that the IPF microenvironment drives progressive differentiation of FABP4+ tissue-resident macrophages toward SPP1+ profibrotic macrophages.

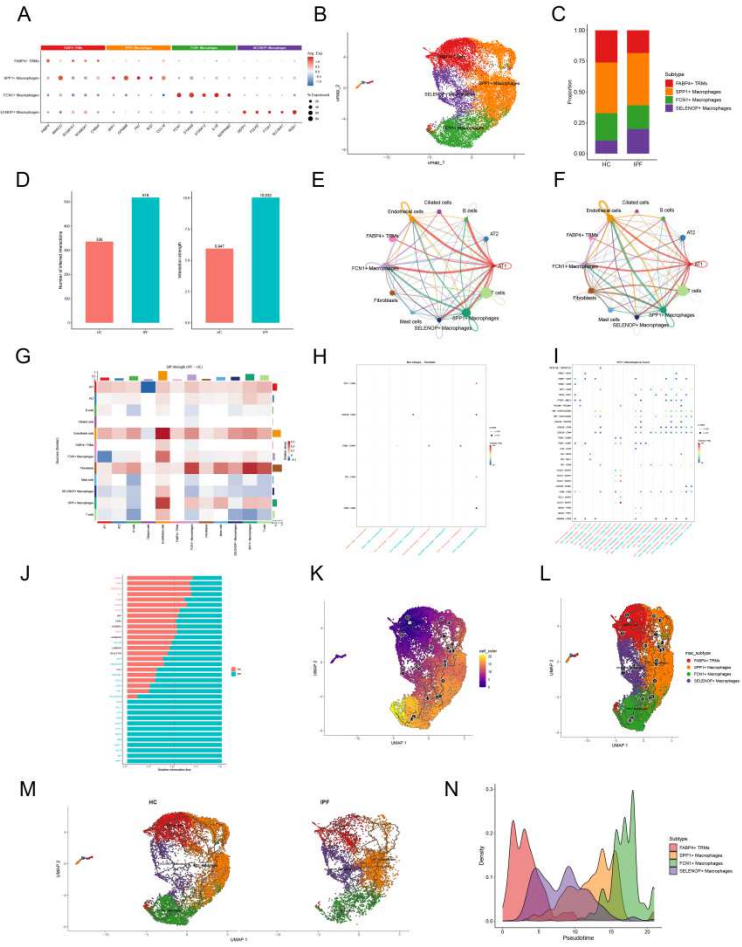


Fig 7. Alveolar macrophage subsets, cell communication, and pseudotime analysis. (A) Bubble plot of marker genes for each macrophage subset. (B) UMAP distribution of four macrophage subsets. (C) Stacked bar plot of macrophage subset proportions in HC and IPF. (D–F) Number and overall strength of cell–cell interactions in HC and IPF. (G) Differential communication strength heatmap. (H) Ligand–receptor bubble plot of interactions from macrophage subsets to fibroblasts. (I) Ligand–receptor bubble plot of interactions from SPP1+ macrophages. (J) Relative information flow of signaling pathways in HC vs. IPF. (K–L) Monocle 3 pseudotime trajectory overlaid with cell type identity and pseudotime score. (M) Cell distribution of HC and IPF along the pseudotime axis. (N) Pseudotime density distribution of four macrophage subsets.

3.8 Validation of Core Gene Protein Expression in Human IPF Lungs

Western blot analysis was performed on human IPF lung tissues (n = 6) and paired adjacent normal lung tissues (n = 6) to validate protein-level expression of the four core diagnostic genes. Compared with controls, IPF tissues demonstrated significantly elevated protein levels of BPIFB1, CCL18, IGFBP2, and WFDC2 (Figures 8A–8D). BPIFB1 exhibited the most pronounced upregulation (P <0.001), while CCL18, IGFBP2, and WFDC2 were also significantly elevated (P <0.05), consistent with their mRNA expression trends in the transcriptomic datasets. These results validate the translational relevance of the four core genes as protein-level diagnostic biomarkers for IPF.

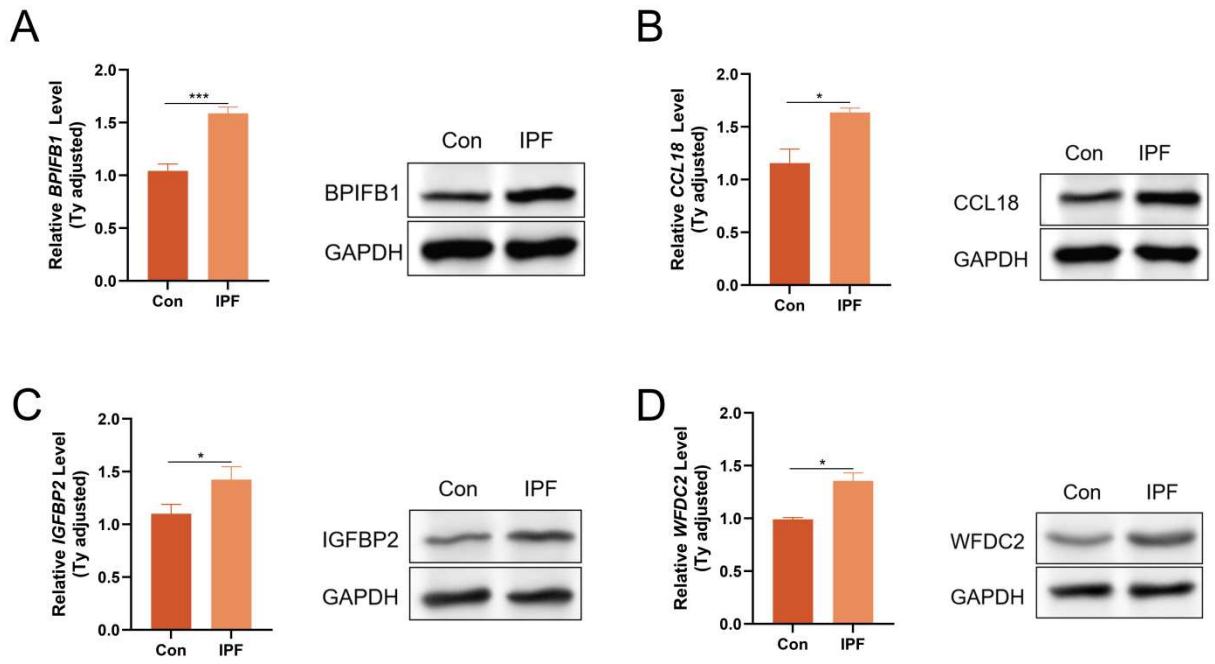


Fig 8. Western blot validation of core gene protein expression in human IPF lungs. (A–D) Representative Western blot bands and quantified protein levels of BPIFB1, CCL18, IGFBP2, and WFDC2 in IPF (n = 6) and control (n = 6) lung tissues. GAPDH was used as the internal loading control. Data are presented as mean ± SD. *P <0.05; ***P <0.001.

4. Discussion

IPF is a chronic progressive interstitial lung disease with rising incidence, substantial disease burden, and dismal prognosis [1,2]. Currently approved antifibrotic agents, pirfenidone and nintedanib, remain insufficient to meet the full spectrum of clinical needs [4,5]. Metabolic reprogramming in the tissue microenvironment has increasingly been

recognized as a key driver of chronic fibrosis progression [15,22]. Glycolytic intermediates shunted into the PPP supply myofibroblasts and profibrotic macrophages with NADPH and ribose-5-phosphate to support proliferation, biosynthesis, and redox defense [18,19,22]. In the present study, we integrated scRNA-seq and bulk RNA-seq data, combining hdWGCNA with five machine learning algorithms to develop a PPP-based gene screening framework. This approach identified four core regulatory genes—BPIFB1, CCL18, IGFBP2, and WFDC2—as hub biomarkers for assessing IPF disease status.

Alveolar macrophages were identified as the focal point of PPP–IPF crosstalk, supported by three independent lines of evidence. First, cross-cell-type comparison confirmed alveolar macrophages as the cell type with the highest PPP activity in the lung, suggesting that the PPP supports their functional homeostasis and effector activity [22,23]. Second, ssGSEA and UCell independently verified global glucose metabolic remodeling in IPF alveolar macrophages, with coordinated upregulation of the PPP, glycolysis, and the TCA cycle, reflecting systemic metabolic adaptation to the IPF microenvironment. Third, Scissor phenotype mapping identified alveolar macrophages as the primary cell population enriched with IPF phenotype-associated cells. Furthermore, secondary clustering revealed that alveolar macrophages comprise four functionally distinct subsets—FABP4+ TRMs, SPP1+, FCN1+, and SELENOP+ macrophages—with significant expansion of SPP1+ profibrotic macrophages in IPF, underscoring the importance of dissecting metabolic heterogeneity at single-cell resolution.

CCL18, the most mechanistically characterized core biomarker, is predominantly secreted by alveolar macrophages and induces fibroblast collagen synthesis and activation; its levels are elevated in the serum and bronchoalveolar lavage fluid of IPF patients [42]. Prasse et al. first demonstrated that serum CCL18 correlates with lung function decline and survival in IPF [42,43]. Post-hoc analyses of the CAPACITY and ASCEND trials subsequently confirmed CCL18 as one of the most stable prognostic biomarkers among 12 evaluated plasma proteins [44]. In the present study, CCL18 exhibited consistent diagnostic performance across training and validation sets. When combined with Scissor and PPP score analyses, CCL18-expressing alveolar macrophages were identified as among the most PPP-active cells, suggesting that PPP-derived NADPH supports the sustained activation and extracellular matrix (ECM) secretory activity of CCL18+ macrophages through redox homeostasis maintenance.

WFDC2 (HE4), a member of the WAP four-disulfide core protein family, is highly expressed in the respiratory epithelium [45]. Tian et al. reported elevated serum HE4 levels in IPF patients, particularly during acute exacerbation,

with inverse correlations with forced vital capacity percentage (FVC%) and diffusing capacity for carbon monoxide percentage (DLCO%), and demonstrated independent prognostic value for survival [46]. Nishiyama et al. further established serum HE4 as a prognostic biomarker for progressive fibrosing interstitial lung disease [47]. Mechanistically, WFDC2 inhibits matrix metalloproteinase activity, thereby promoting collagen deposition and ECM remodeling [45].

IGFBP2, a member of the insulin-like growth factor-binding protein family, regulates IGF signaling and organ fibrosis. Guiot et al. reported elevated serum IGFBP2 levels in IPF patients, with a reduction following antifibrotic therapy [48]. Chin et al. demonstrated that loss of IGFBP2 drives alveolar type 2 (AT2) cell senescence and promotes pulmonary fibrosis, while transgenic overexpression confers antifibrotic protection, indicating that IGFBP2 functions both as a disease activity biomarker and as an endogenous antifibrotic factor [49].

BPIFB1 (LPLUNC1) is an airway epithelial secreted protein expressed in submucosal glands. Bingle et al. reported its upregulation in cystic fibrosis airways, associated with airway remodeling [50]. Yang et al. recently identified BPIFB1 as a prognostic biomarker for IPF via bioinformatic analysis, with elevated expression in IPF lungs, enrichment in ciliated and MUC5B+ cells, and inverse correlations with FVC% and DLCO% [51]. In the present study, BPIFB1 was significantly upregulated in both training and validation datasets, suggesting that it reflects alveolar epithelial barrier disruption and enhanced profibrotic secretory activity in IPF.

Immune infiltration analysis and secondary clustering of alveolar macrophages collectively demonstrated the dominance of SPP1+ profibrotic macrophages in the IPF microenvironment. First characterized by Reyfman et al. [9], SPP1+ macrophages proliferate in fibrotic regions of IPF lungs and colocalize with myofibroblasts via secretion of SPP1/osteopontin [10]. Adams and Habermann et al. documented progressive SPP1 upregulation and coordinated ECM remodeling gene expression along profibrotic macrophage differentiation trajectories [11,12]. Our pseudotime and CellChat analyses confirmed dense SPP1–CD44, COLLAGEN, LAMININ, and FN1 ligand–receptor crosstalk between SPP1+ macrophages and fibroblasts, with significantly increased interaction numbers and interaction strength in IPF. This pattern of dense communication suggests that SPP1+ macrophages, as the most PPP-active cell population, drive persistent fibroblast activation through ECM-related signaling, forming a core profibrotic axis in IPF lungs. Consistent with this interpretation, GSVA identified upregulation of nucleotide excision repair, one-carbon folate metabolism, and N-glycan biosynthesis pathways in IPF—all directly linked to PPP-derived ribose-5-phosphate

and NADPH—further supporting the PPP as a key molecular driver of IPF progression.

This study has several limitations that warrant consideration. First, the transcriptomic analyses relied on publicly available GEO datasets, which lack detailed clinical metadata such as disease severity scores, lung function decline rates, and antifibrotic treatment history. Although batch effects were corrected, prospective multicenter cohorts with standardized sample collection protocols are needed to fully eliminate potential confounders. Second, transcriptomic associations between gene expression and the PPP do not directly reflect metabolite abundance or enzymatic activity; functional metabolomic and enzymatic studies are required to establish causal relationships. Third, Western blot validation was limited to a relatively small human sample size (n = 6 per group), and control tissues were adjacent normal lung tissues rather than lungs from healthy donors, which—despite collection >5 cm from tumor margins from donors without chronic pulmonary disease—may not fully recapitulate the microenvironment of truly healthy lungs. Future studies should incorporate large-scale prospective clinical cohorts, spatial transcriptomics, high-resolution metabolomics, and conditional knockout animal models to directly establish causal relationships between PPP dysregulation and SPP1+ profibrotic macrophage differentiation.

5. Conclusion

this study delineates the single-cell landscape of PPP metabolism in IPF, identifies a synergistic diagnostic network of four core genes (BPIFB1, CCL18, IGFBP2, and WFDC2), and constructs a high-performance clinical nomogram with an AUC of 0.889. Integrated alveolar macrophage secondary clustering, pseudotime trajectory analysis, and CellChat communication analysis establish SPP1+ profibrotic macrophages as the key cellular mediators of PPP-driven IPF progression. These findings provide novel mechanistic insights into the metabolism–immunity–stroma crosstalk in IPF and lay a theoretical foundation for developing metabolic reprogramming-targeted therapeutic interventions.

Declarations

Clinical trial number

Not applicable.

Ethics Approval and Consent to Participate

All subjects participating in this study agreed to and signed an informed consent form. The study protocol was approved by the Ethics Committee of Wuxi People's Hospital (Approval No. ky25140) and conducted in accordance with the principles of the Declaration of Helsinki.

Consent for Publication

Not applicable.

Declaration of Interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Key Research and Development Program of Jiangsu Province (BE2022697) and the Taihu Talent Program of Wuxi – International and Domestic Top Medical Expert Team (2019-THRCTD-1).

Data Statement

The scRNA-seq and bulk RNA-seq datasets analyzed during the current study are publicly available in the NCBI Gene Expression Omnibus (GEO) database (accession numbers: GSE128033, GSE47460, GSE53845). Additional data are available from the corresponding author on reasonable request.

Authors' Contributions

All authors contributed to the study conception and design. Yongqi Cheng: Writing – original draft preparation, formal analysis and investigation; Shiqing Yuan: Writing – original draft preparation, methodology; Mingfeng Zheng: Conceptualization, funding acquisition, resources, supervision, writing – review and editing. All authors read and approved the final manuscript.

Acknowledgments

Not applicable.

References

[1] Lederer DJ, Martinez FJ. Idiopathic pulmonary fibrosis. N Engl J Med. 2018;378(19):1811–1823.

[2] Maher TM, Bendstrup E, Dron L, et al. Global incidence and prevalence of idiopathic pulmonary fibrosis.

Respir Res. 2021;22(1):197.

[3] Raghu G, Remy-Jardin M, Myers JL, et al. Diagnosis of idiopathic pulmonary fibrosis: an official ATS/ERS/JRS/ALAT clinical practice guideline. Am J Respir Crit Care Med. 2018;198(5):e44–e68.

[4] Richeldi L, du Bois RM, Raghu G, et al. Efficacy and safety of nintedanib in idiopathic pulmonary fibrosis. N Engl J Med. 2014;370(22):2071–2082.

[5] King TE Jr, Bradford WZ, Castro-Bernardini S, et al. A phase 3 trial of pirfenidone in patients with idiopathic pulmonary fibrosis. N Engl J Med. 2014;370(22):2083–2092.

[6] Martinez FJ, Collard HR, Pardo A, et al. Idiopathic pulmonary fibrosis. Nat Rev Dis Primers. 2017;3:17074.

[7] Moss BJ, Ryter SW, Rosas IO. Pathogenic mechanisms underlying idiopathic pulmonary fibrosis. Annu Rev Pathol. 2022;17:515–546.

[8] Allen RJ, Guillen-Guio B, Oldham JM, et al. Genome-wide association study of susceptibility to idiopathic pulmonary fibrosis. Am J Respir Crit Care Med. 2020;201(5):564–574.

[9] Reyfman PA, Walter JM, Joshi N, et al. Single-cell transcriptomic analysis of human lung provides insights into the pathobiology of pulmonary fibrosis. Am J Respir Crit Care Med. 2019;199(12):1517–1536.

[10] Morse C, Tabib T, Sembrat J, et al. Proliferating SPP1/MERTK-expressing macrophages in idiopathic pulmonary fibrosis. Eur Respir J. 2019;54(2):1802441.

[11] Adams TS, Schupp JC, Poli S, et al. Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. Sci Adv. 2020;6(28):eaba1983.

[12] Habermann AC, Gutierrez AJ, Bui LT, et al. Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis. Sci Adv. 2020;6(28):eaba1972.

[13] Strunz M, Simon LM, Ansari M, et al. Alveolar regeneration through a Krt8+ transitional stem cell state that persists in human lung fibrosis. Nat Commun. 2020;11(1):3559.

[14] Tsukui T, Sun KH, Wetter JB, et al. Collagen-producing lung cell atlas identifies multiple subsets with distinct localization and relevance to fibrosis. Nat Commun. 2020;11(1):1920.

[15] Selvarajah B, Platé M, Chambers RC. Pulmonary fibrosis: emerging diagnostic and therapeutic strategies. Mol Aspects Med. 2023;94:101227.

[16] Kottmann RM, Kulkarni AA, Smolnycki KA, et al. Lactic acid is elevated in idiopathic pulmonary fibrosis and induces myofibroblast differentiation via pH-dependent activation of transforming growth factor-β. Am J Respir Crit Care Med. 2012;186(8):740–751.

[17] Xie N, Tan Z, Banerjee S, et al. Glycolytic reprogramming in myofibroblast differentiation and lung fibrosis. Am J Respir Crit Care Med. 2015;192(12):1462–1474.

[18] Stanton RC. Glucose-6-phosphate dehydrogenase, NADPH, and cell survival. IUBMB Life. 2012;64(5):362–369.

[19] Patra KC, Hay N. The pentose phosphate pathway and cancer. Trends Biochem Sci. 2014;39(8):347–354.

[20] Zhao YD, Yin L, Archer S, et al. Metabolic heterogeneity of idiopathic pulmonary fibrosis: a metabolomic study. BMJ Open Respir Res. 2017;4(1):e000183.

[21] Kang YP, Lee SB, Lee JM, et al. Metabolic profiling regarding pathogenesis of idiopathic pulmonary fibrosis. J Proteome Res. 2016;15(5):1717–1724.

[22] Hamanaka RB, Mutlu GM. Metabolic requirements of pulmonary fibrosis: role of fibroblast metabolism. FEBS J. 2021;288(22):6331–6352.

[23] Ghergurovich JM, García-Cañaveras JC, Wang J, et al. A small molecule G6PD inhibitor reveals immune dependence on pentose phosphate pathway. Nat Chem Biol. 2020;16(7):731–739.

[24] Sun D, Guan X, Moran AE, et al. Identifying phenotype-associated subpopulations by integrating bulk and single-cell sequencing data. Nat Biotechnol. 2022;40(4):527–538.

[25] Morabito S, Reese F, Rahimzadeh N, et al. hdWGCNA identifies co-expression networks in high-dimensional transcriptomics data. Cell Rep Methods. 2023;3(6):100498.

[26] Cao J, Spielmann M, Qiu X, et al. The single-cell transcriptional landscape of mammalian organogenesis. Nature. 2019;566(7745):496–502.

[27] Jin S, Guerrero-Juarez CF, Zhang L, et al. Inference and analysis of cell-cell communication using CellChat. Nat Commun. 2021;12(1):1088.

[28] Stuart T, Butler A, Hoffman P, et al. Comprehensive integration of single-cell data. Cell. 2019;177(7):1888–1902.e21.

[29] Korsunsky I, Millard N, Fan J, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019;16(12):1289–1296.

[30] Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–15550.

[31] Andreatta M, Carmona SJ. UCell: robust and scalable single-cell gene signature scoring from single-cell expression data. Comput Struct Biotechnol J. 2021;19:3796–3798.

[32] Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.

[33] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.

[34] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1–22.

[35] Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines. Mach Learn. 2002;46(1–3):389–422.

[36] Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.

[37] Kursa MB, Rudnicki WR. Feature selection with the Boruta package. J Stat Softw. 2010;36(11):1–13.

[38] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM; 2016:785–794.

[39] Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12(5):453–457.

[40] Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14:7.

[41] Cannon M, Stevenson J, Stahl K, et al. DGIdb 5.0: rebuilding the drug-gene interaction database for precision medicine and drug discovery platforms. Nucleic Acids Res. 2024;52(D1):D1227–D1235.

[42] Prasse A, Pechkovsky DV, Toews GB, et al. A vicious circle of alveolar macrophages and fibroblasts perpetuates pulmonary fibrosis via CCL18. Am J Respir Crit Care Med. 2006;173(7):781–792.

[43] Prasse A, Probst C, Bargagli E, et al. Serum CC-chemokine ligand 18 concentration predicts outcome in idiopathic pulmonary fibrosis. Am J Respir Crit Care Med. 2009;179(8):717–723.

[44] Neighbors M, Cabanski CR, Ramalingam TR, et al. Prognostic and predictive biomarkers for patients with idiopathic pulmonary fibrosis treated with pirfenidone: post-hoc assessment of the CAPACITY and ASCEND trials. Lancet Respir Med. 2018;6(8):615–626.

[45] Bingle L, Cross SS, High AS, et al. WFDC2 (HE4): a potential role in the innate immunity of the oral cavity and respiratory tract and the development of adenocarcinomas of the lung. Respir Res. 2006;7(1):61.

[46] Tian Y, Li R, Liu Y, et al. Elevated serum human epididymis protein 4 is associated with disease severity and worse survival in idiopathic pulmonary fibrosis: a cohort study. Ann Transl Med. 2022;10(18):1010.

[47] Nishiyama N, Masuo M, Nukui Y, et al. Human epididymis protein 4 is a new biomarker to predict the prognosis of progressive fibrosing interstitial lung disease. Respir Investig. 2021;59(1):90–98.

[48] Guiot J, Bondue B, Henket M, et al. Raised serum levels of IGFBP-1 and IGFBP-2 in idiopathic pulmonary fibrosis. BMC Pulm Med. 2016;16(1):86.

[49] Chin C, Ravichandran R, Sanborn K, et al. Loss of IGFBP2 mediates alveolar type 2 cell senescence and promotes lung fibrosis. Cell Rep Med. 2023;4(2):100945.

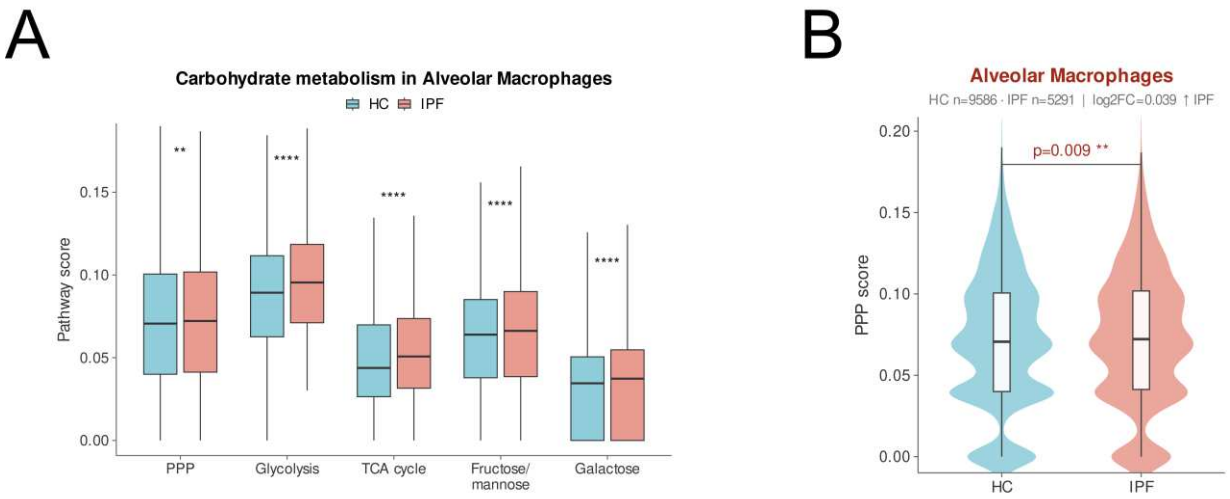
[50] Bingle L, Wilson K, Musa M, et al. BPIFB1 (LPLUNC1) is upregulated in cystic fibrosis lung disease. Histochem Cell Biol. 2012;138(5):749–758.

[51] Yang J, Ruan X, He J. BPIFB1 is a prognostic biomarker and mediated collagen synthesis in idiopathic pulmonary fibrosis. Medicine. 2025;104(22):e42671.

[52] DePianto DJ, Chandriani S, Abbas AR, et al. Heterogeneous gene expression signatures correspond to distinct lung pathologies and biomarkers of disease severity in idiopathic pulmonary fibrosis. Thorax. 2015;70(1):48–56.

Supplementary Materials

Supplementary Figure S1. Independent validation of PPP activity using UCell. (A) UCell scores for five glucose metabolic pathways in HC and IPF alveolar macrophages. (B) PPP scores in HC and IPF alveolar macrophages..



515

516 **Supplementary Table S1.** Western blot antibody information .

Target	Catalon No	supplier
BPIFB1	ab219098	Abcam
CCL18	ab300057	Abcam
IGFBP2	ab188200	Abcam
WFDC2	ab200828	Abcam
GAPDH	ab181602	Abcam

517

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- figureS1.tif
- Table1.docx