Serum proteomics identify biomarkers associated with the pathogenesis of idiopathic pulmonary fibrosis

Lan Wang, Minghui Zhu, Yan Li, Peishuo Yan, Zhongzheng Li, Xiuping Chen, Juntang Yang, Xin Pan, Huabin Zhao, Shenghui Wang, Hongmei Yuan, Mengxia Zhao, Xiaogang Sun, Ruyan Wan, Fei Li, Xiaobo Wang, Hongtao Yu, Ivan Rosas, Chen Ding, Guoying Yu

PII: S1535-9476(23)00034-8

DOI: https://doi.org/10.1016/j.mcpro.2023.100524

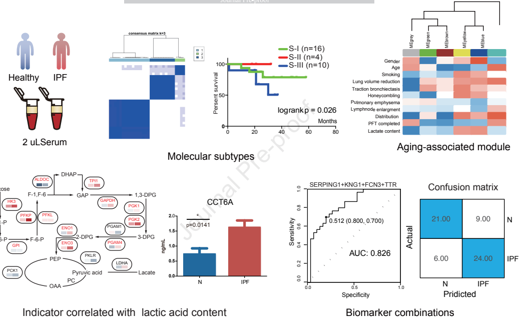
Reference: MCPRO 100524

To appear in: *Molecular & Cellular Proteomics*

Received Date: 7 July 2022

Revised Date: 31 January 2023

Accepted Date: 28 February 2023



Molecular subtypes

Aging-associated module

Indicator correlated with lactic acid content

Biomarker combinations

Serum proteomics identify biomarkers associated with the pathogenesis of idiopathic pulmonary fibrosis

Lan Wang[1#], Minghui Zhu[2#], Yan Li[3#], Peishuo Yan[1#], Zhongzheng Li[1], Xiuping Chen[3], Juntang Yang[1], Xin Pan[1], Huabin Zhao[1], Shenghui Wang[1], Hongmei Yuan[1], Mengxia Zhao[1], Xiaogang Sun[1], Ruyan Wan[1], Fei Li[2], Xiaobo Wang[2], Hongtao Yu[2], Ivan Rosas[4], Chen Ding[3*], Guoying Yu[1*]

1. State Key Laboratory of Cell Differentiation and Regulation, Henan International Joint Laboratory of Pulmonary Fibrosis, Henan center for outstanding overseas scientists of pulmonary fibrosis, College of Life Sciences, Institute of Biomedical Science, Henan Normal University, Xinxiang, Henan 453007, China
2. Henan Provincial Chest Hospital, Zhengzhou, Henan 450003, China
3. State Key Laboratory of Genetic Engineering, Human Phenome Institute, Institutes of Biomedical Sciences, and School of Life Sciences, Zhongshan Hospital, Fudan University, Shanghai 200433, China.
4. Division of Pulmonary, Critical Care and Sleep Medicine, Baylor College of Medicine, Houston, TX 77030, USA

*Corresponding authors

#Contribute equally

Running title: Serum proteomics identify biomarkers of IPF

Address for correspondence:

Guoying Yu, PhD
Henan Normal University, 46 Jianshe Road, Xinxiang, Henan 453007, China
Email: guoyingyu@htu.edu.cn

Abstract

The heterogeneity of idiopathic pulmonary fibrosis (IPF) limits its diagnosis and treatment. The association between the pathophysiological features and the serum protein signatures of IPF currently remains unclear. The present study analyzed the specific proteins and patterns associated with the clinical parameters of IPF based on a serum proteomic dataset by Data-Independent Acquisition (DIA) using mass spectrometry. Differentiated proteins in sera distinguished in IPF patients into three subgroups in signal pathways and overall survival. Aging-associated signatures by WGCNA coincidently provided clear and direct evidence that aging is a critical risk factor for IPF rather than a single biomarker. LDHA and CCT6A expression, which were associated with glucose metabolic reprogramming, were correlated with high serum lactic acid content in the patients with IPF. Cross-model analysis and machine learning showed that a combinatorial biomarker accurately distinguished IPF patients from healthy subjects with an AUC of 0.848 (95% CI = 0.684–0.941) and validated from another cohort and ELISA assay. This serum proteomic profile provides rigorous evidence that enables understanding of the heterogeneity of IPF and protein alterations that could help in its diagnosis and treatment decisions.

Keywords: Serum proteome, Molecular subtype, machine learning, indicator panel, combinatorial biomarker

Introduction

IPF is a chronic and fatal progressive fibrotic lung disease with a reported median survival of 3–5 years (1) (2). The heterogeneity of IPF and the various pathophysiological mediators involved in its clinical progression limit its diagnosis and treatment. Aging is one of the critical risk factors for IPF, with increasing evidence highlighting the important role of senescence in IPF(3). Cellular senescence leads to DNA damage, cell cycle arrest, telomere shortening(4), mitochondrial dysfunction, metabolic reprogramming, resistance to apoptosis, and deficient autophagy. Mitochondrial dysfunction, including the leakage of high energy electrons from the electron transport chain (ETC), disrupted cristae, and a diminished capacity for oxidative phosphorylation, establish a close link between senescence and IPF(5). Metabolic dysfunction alters processes during lung tissue repair, as well as crucial metabolic pathways such as augmented glycolysis and increased fatty acid oxidation, which are important drivers of fibroblast activation(6). In particular, altered lactate metabolism may be an underlying feature of IPF and a novel clinical diagnostic marker(7, 8).

The use of machine learning tool did not reach a formal recommendation in American Thoracic Society（ATS）/European Respiratory Society（ERS）/Japanese Respiratory Society（JRS)/ Latin American Thoracic Society（ALAT) clinical practice guideline, but more of a consideration in specific circumstances at certain centers to identify diagnostic markers and to combine these molecular markers with current diagnostic modalities in the multidisciplinary diagnosis of IPF. Novel biomarkers integrated into clinical diagnosis can include circulating markers or molecular signatures obtained from less invasive sampling(9). To date, most biomarkers are the molecules abundant enriched and associated with pathophysiological process in a specific disease. Proteomic strategies have allowed extensive assessment of larger patient cohorts and the identification of novel biomarkers, while reducing the need for invasive acquisition and analysis of blood and body fluids(10). Improvements in deep proteomes may result in the identification of individual biomarkers or biomarker panels that may not be directly involved in the disease pathophysiology and may only be associated with it. These biomarkers may have the potential to better understand the pathophysiology of IPF, not only for diagnostic but also for therapeutic purposes.

Previous studies found that aberrations in complement activation and oxidative damage, haptoglobin-related protein were identified as candidate marker in IPF using the label-free plasma proteomics(11). Here, we wished to gain further insights into the changed serum proteomic of IPF patients, to obtain the proteins associated with the disease pathophysiology. A global correlation network related to clinical traits was constructed, and machine learning was used to identify a combinatorial biomarker.

Materials and methods

Experimental Design and Statistical Rationale

The purpose of this study to identify signatures associated with the pathogenesis of idiopathic pulmonary fibrosis in serum from IPF patients. The workflow is depicted in Figure S1. Serum samples were collected from 30 IPF patients as a cohort, IPF was diagnosed based on ATS/ERS/JRS/ALAT Clinical Practice Guidelines (12). Subjects were obtained at diagnosis and followed by physicians according to institutional practices, including by high-resolution computed tomography (HRCT) and pulmonary function tests (PFTs). All patients with IPF underwent routine blood tests, including measurements of lactic acid concentrations and some antibodies. None of the included patients had evidence of autoimmune syndromes, malignancy, infections, or drug or occupational exposures associated with lung fibrosis. Serum samples from 30 healthy volunteers were collected as a control group, of which all participants underwent a full medical examination prior to inclusion in the study. The validation cohort consisted of an additional patient with IPF for the ELISA. The study was approved by the Henan Provincial Chest Hospital Medical Research Ethics Committee (No. 2020-03-06). Oral and written informed consent was obtained from all participants of this study. All samples used in this study were collected at Henna Provincial Chest Hospital according to the guidelines in the Declaration of Helsinki. The demographic and clinical characteristics of the cohorts are provided, which including the summary data with statistics on age, sex, smoking status in Table 1 and other characteristics in the Table S1. A public IPF cohort (PRIDE project PXD010965) that included 19 healthy individuals and 17 IPF patients was used to validate the accuracy of the machine-learning-based classification of IPF. The animal handling procedures followed the Henan Normal University Institutional Animal Care and Use Committee (IACUC, SMKX-2019S002) guidelines, which coordinate with the Association of Animal Behavior and National Regulations.

Serum sample preparation

Blood samples from IPF patients and healthy volunteers were taken from a vein in the cubital fossa. The blood collection was done into commercial Monovette tubes containing tripotassium ethylenediaminetetraacetic acid as the anticoagulant and whole blood glass tubes with anticoagulant. The samples were centrifuged for serum separation (2000 rpm for 10 min, +4 ◦C) immediately after collection. The supernatant was frozen at −80 ◦C before liquid chromatography-mass spectrometry (LC-MS) analysis.

The 14 most abundant serum proteins were removed from each sample using commercial depletion kits (High-Select™ Top14 Abundant Protein Depletion Mini Spin Columns), according to the manufacturer's instructions. Following depletion, the proteins were denatured, reduced, alkylated, digested into peptides, and desalted using a C-18 column for LC-MS/MS analysis.

High-pressure liquid chromatography and mass spectrometry

Samples were subjected to LC-MS/MS, consisting of an EASY-nLC 1200 system coupled to a nano-electrospray ion source and a Fusion Lumos Orbitrap (Thermo Fisher Scientific). Purified peptides were separated on 150 µm I.D. × 15 cm columns (C18, 1.9 µm, 120Å, Dr. Maisch GmbH). Each column was loaded with about 0.5 µg peptides in buffer A (0.1% formic acid), followed by elution at a flow rate of 450 nL/min with a linear gradient of 3–30% of buffer B (0.1% formic acid, 80% (v/v) acetonitrile) for 35 min, 75% buffer B for 7 min, 98% buffer B for 1 min, and a wash with 98% buffer B for 2 min. The column temperature was maintained at 60°C using a Peltier element containing an oven developed in house.

MS spectra were acquired with a Data-Independent Acquisition (DIA) method. The DIA-MS method consisted of an MS1 scan from 300 to 1,400 m/z range (AGC target of 4×10^5, maximum injection time of 50 ms) at a resolution of 60,000 and 30 DIA segments (AGC target of 5×10^4, maximum injection time of 22 ms) at a resolution of 15,000.

Library-based DIA data analysis and quality control

To build the spectral library, we acquired 128 DDA files on a Fusion Lumos Orbitrap mass spectrometer in DDA mode, which was used as reference spectra libraries. A library was built by Skyline-daily (22.2.1.278, University of Washington, USA) for DIA analysis, which were composed of various body fluids and organ tissue samples from 64 individuals, covering blood, hydrothorax, joint effusion, bile, ascites, cerebrospinal, urine, etc., with a deep fractionation ranging from 7 to 31. For Skyline library building, carbamidomethyl (C) was set as the fixed modification, and acetyl (protein N-term) and oxidation (M) were set as the variable modifications. Two missed trypsin cleavages were allowed. Precursor ion score charges were limited to +2, +3, and +4. The precursor and fragment tolerance were set as dynamic. Finally, a library containing 68,781 peptides and 4,437 proteins was built. In our previous research, the DIA library has been used for blood molecular markers for the pathophysiology and clinical progress of COVID-19 (13). For Skyline analysis, the default setting was used for library-based DIA analysis according to the standard workflow in Skyline (https://skyline.ms/_webdav/home/software/Skyline/@files/tutorials). A total of 60 raw files' reports were exported by Skyline DIA analysis, and were merged into an integrated expression matrix including the expression of each single protein, of which all identified distinct peptides were used for the corresponding protein quantification. The detection Q value was set to 5% at the peptide and protein levels. Proteome qualification was performed as previously reported with the iBAQ algorithm(14), followed by normalization to the fraction of the total (FOT), defined as a protein's iBAQ divided by the total iBAQ of all identified proteins within one sample, thus representing the normalized abundance of a particular protein across samples. Finally, the FOT values were further multiplied by 10^5 for ease of presentation, and missing values were replaced by the minimal value.

The quality of proteomic data was ensured at multiple levels. Instrument performance was evaluated using a whole cell extract of HEK293T cells. To avoid carryover, blank samples (buffer A) were run after every five injections. The consistency of sample collection and handling was validated by assessing the abundance of the quality markers FGA, FGB, and FGC.

Differential protein analysis

The differential expression of proteins in IPF patients and healthy controls was also analyzed by Student's t-tests. Proteins differentially expressed with p-values < 0.05 and fold changes > 1.5 or < 2/3 were visualized using an R package heatmap. Between-group analysis of DEPs was performed using paired two-class analysis of the same R package with an FDR threshold of 0.05.

Pathway enrichment analysis and functional annotation

The biological characteristics of the three IPF subtypes and the proteins differentially expressed by IPF patients and healthy controls were determined by pathway enrichment analysis with Reactome. The statistical significance of pathway enrichment was determined by Fisher's exact test and pathways with an FDR threshold of 0.05 were regarded as being significantly regulated.

Proteome molecular subtyping of IPF

Prior to clustering analysis, proteins that were expressed in more than 25% of patient samples were selected (n = 1190) (Table S4). The serum proteomic subtypes of IPF were identified by consensus clustering (R package Consensus Cluster Plus v.1.48.0) (15). A total of 1190 proteins were subjected to k-means clustering with up to six clusters. The consensus matrix of k = 3 showed clear among-cluster separation (Figure S3A), and the cumulative distribution function of the consensus matrix for each k-value was measured. Clustering by k = 3 resulted in the lowest proportion of ambiguous clustering. To determine the correlations between proteomic subtypes and clinical features, categorical variables, including age, gender, smoke status, and HRCT characteristics, were assessed by Fisher's exact tests.

WGCNA analysis

To identify differentially co-expressed gene modules, WGCNA was applied to the proteins that were expressed in more than 67% of patient samples (n =687). WGCNA was performed in R (R Core Team, 2019) using a WGCNA package (16). Module eigenproteins were calculated as the first principal components of the co-expressed genes in the module (17, 18). The eigengenes of each module were used to measure the association between a module and clinical information. The eigengene-based connectivity (kME) was used to represent the strength of a gene's correlation with other gene module members.

Machine-learning-based selection of biomarker combinations of IPF

Biomarker combinations were identified using the random forest method, a machine learning method that can predict the value of a response variable. Data with coefficients of variation (CV) less than 0.5 were selected as candidate reservoirs, with no more than four proteins randomly selected to form the potential optimal biomarker combination (OBC), and 5,000 potential OBCs were prepared. Each candidate OBC was subjected to 5-fold cross-validation, with the original dataset randomly divided 4:1 into a training set and a verification set. The training set was used to train the model, and the verification set was used to evaluate the model. In penalized logistic regression (PLR) , the weights of four proteins were optimized iteratively using the least shrinkage and selection operator (Lasso, L1 regularization) penalty and the ridge regression (L2 regularization) penalty. The combination with the highest AUC value was selected. To simplify OBC, sets of any three of the four proteins were selected, resulting in four combinations, and the AUC values of these combinations were compared with the AUC value of OBC. The combination with an AUC value closest to that of OBC was selected as the final combination. The PLR algorithm was implemented in R 4.1.2 with the glmnet package.

Survival analysis

Univariate Cox regression analysis was conducted to determine the relationship between the expression of proteins and prognosis of IPF patients. Proteins with a p-value < 0.05 were regarded as prognostic proteins. After that, patients were divided into high-risk and low-risk groups by setting the median value of risk scores as cut-off value. The overall survival (OS) of these two groups was calculated by the Kaplan-Meier method with log-rank test. All statistical analyses were performed using Prism 8 software and the R package “survival”, with statistical significance defined as $p < 0.05$.

Cell culture

The human lung fibroblast cell line (MRC-5) was purchased from the ATCC (CCL-171). Cells were cultured in DMEM supplemented with 10% fetal bovine serum and a 1% antibiotic-antimycotic solution at 37°C in 5% CO2.

Plasmids RNA interference and transfection

The human CCT6A gene was cloned into the pCDNA3.1 plasmid (Generay Biotech, CN). Fibroblasts grown to 80–90% confluence were transfected with this plasma using Lipofectamine 3000 reagent according to the manufacturer’s protocol. The CCT6A siRNA transfection target sequence, 5′-GTGTCATTAGAGTATGAGA-3′, and a negative control were purchased from RiboBio. The siRNAs (75 nM) were transfected into cells using INVI DNA RNA Transfection Reagent (Invigentech) according to the manufacturer’s instructions.

Protein extraction and western blot analysis

Mouse lung tissue samples and cultured cells were lysed in RIPA lysis buffer. Equal amounts of protein were separated on SDS-PAGE and transferred to PVDF-membranes, which were hybridized overnight with appropriate primary antibodies. The membranes were washed and incubated with horseradish peroxidase-conjugated secondary antibodies, followed by visualization using the Odyssey Fc Dual-Mode Imaging System (LI-COR, USA), according to the manufacturer’s instructions.

Immunofluorescence staining

Transfected fibroblasts were fixed with 4% paraformaldehyde and permeabilized with 0.3% Triton X100/PBS. Cells were incubated with primary antibodies at 4°C overnight followed by incubation with fluorescent-labeled secondary antibodies for 30 min at 37℃. Images were visualized using an Axio Imager D2 (Zeiss, GER).

Extracellular flux technology

The extracellular acidification rate (ECAR) of fibroblasts was measured using a Seahorse XF96 Extracellular Flux Analyzer (Seahorse Bioscience, USA). All assays were performed using a seeding density of 30,000 cells/well in 200 μL DMEM in an XF96 cell culture microplate (Seahorse Bioscience). ECAR was measured after sequential addition of glucose, oligomycin, and 2-DG, to reach working concentrations of 10 mM, 1 μM, and 50 mM, respectively.

LDH activity

LDH activity was assessed using LDH activity assay kits, according to the manufacturer's instructions. Briefly, extract was added to the transfected cells, and the cells were disrupted by ultrasound and centrifuged at 8,000 g for 10 min at 4°C. LDH activity was evaluated by measuring the amount of pyruvate produced.

Lactate assay

The intracellular and tissue concentrations of lactate were determined using Lactate Assay Kits, according to the manufacturer's instructions. Tissues or cells were homogenized in four volumes of Lactate Assay Buffer and centrifuged at 13,000g for 10 minutes to remove insoluble material. The samples were deproteinized with a 10 kDa MWCO spin filter to remove lactate dehydrogenase, and the absorbance of the soluble fraction at 570 nm was measured.

Immunoassays

Serum protein concentrations were measured using commercially available ELISA kits, as described by the manufacturer. Measure the absorbance of each sample at 450nm with Microplate Reader（Thermo Fisher）. For immunohistochemistry (IHC) staining, paraffin-embedded tissue sections(5 µm thick) were de-paraffinized and dehydrated, followed by antigen retrieval according to standard procedures. Tissue samples were incubated with specific antibodies, with images captured by AxioScan.Z1 (Zeiss).

Statistical analysis

GraphPad Prism 8.0 and R was used for statistical analysis. The details of experiments can be found in the methods and figure legends. Genes with p-values < 0.05 and fold changes > 1.5 or other thresholds were visualized using R package heatmaps. Between-group analysis of DEPs was performed using paired two-class of the same R package with an FDR threshold of 0.05. Pathway enrichment to identify pathway alterations was analyzed using Reactome. Differential analysis of samples with different phenotypes was performed using Fisher's exact t-tests, with DEPs compared in groups of patients with IPF and healthy controls. Spearson rank analysis was used to analyze the correlation. GraphPad Prism 8.0 was used to analyze the quantitative results of the cell / animal experiments and ELISA results. Significant differences between groups were evaluated using the student's t-test or analysis of variance (ANOVA). $p< 0.05$ was considered to be statistically significant.

Results

Serum proteome profiling of IPF

The serum proteomic landscape was investigated in 30 patients with IPF and 30 healthy subjects differing in demographic and clinical characteristics, including by gender, age, smoking status, features of HRCT, and others (Table 1, Table S1). A data-independent acquisition (DIA) strategy was adopted (Figure S1), and the consistency of the MS performance of the whole HEK293T cell extract was assessed using Spearman correlation coefficients (average correlation coefficient; R = 0.89) (Figure S2A). The abundance profiles of the quality markers FGA, FGB, and FGG indicated that the collection and handling of the samples were regular(19) (Figure S2B). About 2,383 gene products were collected from the 30 healthy subjects and the 30 patients with IPF (Figure 1A), with the number of proteins per sample ranging from 703 to 1,014 (median 892) (Figure 1B, Table S2). The abundance of the identified proteins varied widely, with APOA1 being most abundant and ATP6V1A being the least abundant (Figure 1C). Sixty-seven significantly differentially expressed proteins (DEPs) ($P < 0.05$ and a differential expression ratio [IPF/N] >1.5 or <0.67) were identified (Figure S2C, Table S3). Of the DEPs 3.7% upregulated, whereas 3.8% of the significantly downregulated proteins in patients with IPF (Figure S2D)

Three molecular subtypes of IPF and their association with clinical features

Consensus Cluster Plus (Table S4, Figure S3A) analysis of the top 1,190 DEPs identified three distinct patient clusters (S-I,S-II,S-III) with differences in survival (Figure 1D). The 30 patients with IPF were followed-up for a median 27.9 months (range, 1–58 months). Association analysis between IPF subtypes and OS demonstrated that OS was longest in the S-II and shortest in the S-III (log-rank *P* = 0.026, Figure 1E). IPF patients in the three proteomic subgroups showed distinct molecular features, including differences in subgroup-specific pathways and expression of representative proteins (Figure 1F, G, Figure S3B, Table S5). Higher expression of BMP2K, which has been implicated in endocytosis and cell differentiation(20), was associated with a longer OS in the S-I; and a high level of PI16, a shear stress and inflammation-regulated inhibitor of MMP2(21), increased OS in the S-II. By contrast, elevated expression of ATP5A1, a subunit of mitochondrial ATP synthase, was associated with a poorer OS in the S-III (Figure 1H). These specific protein signatures may enable classification of these IPF subgroups. The associations between proteomic subtypes and clinical features were examined using Fisher's exact tests for categorical data and Wilcoxon rank-sum tests for continuous data. We found that younger age was closely associated with longer OS in the S-II (Figure 1I), indicating that age affects the survival of patients with IPF(22).

Aging-associated signatures highlighted in the sera of IPF patients

Weighted gene correlation network analysis (WGCNA) of a single dataset composed of samples from all 30 IPF patients with 686 proteomic variables and ten clinical traits yielded the global correlation network heatmap shown in Figure 2A (Table S6). Module-trait relationships analysis showed that the module MEturquoise was positively associated with age patterns (Figure 2B). The signatures correlated with age were clustered, and the top altered proteins in this module mainly belonged to the S-III subgroup with elder patients (Figure 2C, D). Cellular senescence-associated proteins, such as KL (Klotho), HSP90AB1, and SERPINE1; mitochondrial dysfunction-associated proteins, including HSPD1, ATP5A1, and SDPR; and several other proteins associated with DNA repair and the cell cycle, such as HIST2H2BE, NCK1, S100A8, and CDK10, were significantly upregulated in S-III subgroup. In line with our findings, VCAM1 and POSTN expression correlated positively with age(23), whereas UBA, CD14, ORM1, and ORM2, which are involved in inflammatory responses, and CREM and CAMKK1, which are involved in cell apoptosis, correlated negatively with age in the S-III subgroup. SERPINA4, an age-related marker in lung disease(24), was decreased in the S-III subgroup (Figure 2D). Moreover, increased expression of HSP90AB1 and reduced expression of CAMKK1 were associated with poor survival in patients with IPF (Figure 2E).These protein correlation profiles reflect the complex relationships between age and cellular senescence, mitochondrial dysfunction, DNA repair and replication, inflammatory response, and cell apoptosis.

Integration of specific molecular markers with high level of lactic acid for multidisciplinary diagnosis of IPF

Increased glycolysis contributes to IPF by regulating glucose metabolic enzymes; these enzymes are secreted and can be measured in blood. HK1, PFKP, ENO1/3, GAPDH, LDHA, and ALDOB were significantly differentially expressed in the IPF and control groups (Figure 3A, B). Lactate dehydrogenase (LDH) converts pyruvate to lactic acid during glycolysis, with human LDH, consisting of two subunits, LDHA and LDHB, being a key glycolytic terminal enzyme that catalyzes the interconversion of pyruvate and lactate in the anaerobic glycolytic pathway. Compared with controls, LDHA and LDHB were altered in the sera of patients with

IPF (Figure 3C), with survival analysis showing that LDHA may be a significant predictor of poor prognosis in these patients (Figure 3D). Proteomics data showed that the level of serum LDHA was upregulated in IPF patients with high lactate content (>1.7 mmol/L, Table S1,Table S2), based on routine blood tests by ELISA. In addition, the expression of CCT6A, which was predicted to act through an interactive network of signaling pathways with LDHA, was increased in the serum of patients with IPF (Figure 3E). To explore the association of CCT6A with high serum lactic acid content, we measured the levels of CCT6A in IPF patients and the bleomycin model of lung fibrosis in mice. ELISA analysis confirmed that the level of CCT6A was higher in IPF patients in an independent cohort (Figure 3F, Table S7), and the increases were in accordance with MS data (Figure S4). IHC staining of lung tissue from patients with IPF showed that CCT6A was mainly expressed by macrophages and the alveolar epithelium surrounding the fibrotic interstitium, but was weakly expressed in normal alveolar epithelium (Figure 3G). CCT6A expression was also significantly increased in the bleomycin model of lung fibrosis in mice (Figure 3H, I). Moreover, the downregulation of GAPDH observed in the sera of patients with IPF (Figure 3 B) was also observed in fibrotic mouse lungs (Figure 3J). The increased levels of serum CCT6A in patients with IPF were associated with elevated lactic acid concentrations, which may lead to pulmonary fibrosis.

To demonstrate that the changes of CCT6A have a direct effect on fibroblast phenotype, CCT6A was overexpressed or knocked down in MRC-5 cells. Overexpression of CCT6A significantly enhanced the expression of α-SMA in MRC-5 cells (Figure 4A-C), whereas knockdown of CCT6A reduced the levels of FN-1 and Col1A1 (Figure 4D, E), indicating that CCT6A promotes the development of lung fibrosis. To further clarify the association of increased

CCT6A with the high content of lactic acid in the sera of patients with IPF, real-time extracellular acidification rate (ECAR) was measured using the Seahorse XFe96 Analyser (Agilent Technologies). Overexpression of CCT6A was associated with significant increases in glycolysis rate and glycolytic capacity (Figure 4F), as was lactate production in the supernatants of MRC-5 cells and in the lungs of bleomycin-treated mice (Figure 4G, H). Cells overexpressing CCT6A also showed significant upregulation of the expression of LDHA in mRNA and protein level, and decreased production of pyruvate (Fig 4I-L). Collectively, these results show that CCT6A plays an important role in glycolysis through regulation of LDHA and drives pulmonary fibrosis.

Machine-learning-based selection of combinatorial biomarkers for classification of IPF

A machine-learning algorithm involving potential combinatorial biomarkers was developed to classify IPF patients and healthy subjects (Figures S5A). Candidate biomarkers were selected from the significantly differentially expressed proteins using PLR for model training and parameter optimization. This process generated a set of combinatorial biomarkers, including serpin G1 (SERPING1), kininogen 1 (KNG1), ficolin 3 (FCN3), and transthyretin (TTR). The 5-fold cross-validation AUC value of this four-protein combinatorial that differentiated IPF patients and healthy individuals was 0.826 (95% confidence interval [CI] = 0.700–0.800) (Figure 5A, B). The corresponding matrix demonstrated that the training model could correctly classify different samples with high accuracy (Figures 5C). The accuracy of the machine-learning-based classification of IPF was validated in a public IPF cohort (PRIDE project PXD010965) that included 19 healthy individuals and 17 IPF patients. The AUC value for the diagnosis of IPF was 0.848 (95% CI = 0.684–0.941) (Figure 5 D, E), with the data matrix showing promising accuracy in this independent cohort (Figures 5F). The combinatorial biomarkers predicted poorer, but not significantly different, OS in our cohort (Figures 5G). Lack of survival information prevented determination of the ability of the combinatorial biomarkers to predict OS in the public dataset, but these markers exhibited significant performance based on their relative abundances (Figures 5H, Figures S5B).

Our previous study showed that thyroid hormone inhibits lung fibrosis in mice(25). Because TTR transports thyroid hormones in plasma and cerebrospinal fluid, the serum concentrations of TTR were measured by ELISA in an independent cohort. Serum TTR concentrations were significantly lower in IPF patients than in normal controls (Figure 5I). Although low transthyretin levels were reported to correlate with age and stroke(26), serum TTR level did not significantly correlate with age in our patient cohort (Figure 5J).

Discussion

Poor molecular understanding of the heterogeneity of IPF can impede determination of its pathogenesis, leading to inefficient treatment and an inability to predict its occurrence. To address this problem, we sought to determine the serum protein profile in patients with IPF. Analysis showed that IPF could be classified into three subtypes, which exhibited its heterogeneity and diversity. This study also found that CCT6A was associated with the elevated levels of lactic acid in IPF. A global correlation network was developed to identify the indicators of senescence associated with IPF; a combinatorial predictive biomarker that can be used to distinguish patients with IPF from healthy subjects.

Molecular subtyping can stratify patients into subtypes associated with clinical features, responses to treatment, and biological characteristics(27). IPF could be classified into three subtypes based on serum proteomes, with these proteomic subtypes differing in signaling pathways and clinical outcome. Specifically, patients with the S-III subtype had a poorer prognosis. In addition, a functional module related to senescence was found to be associated with the S-III subtype. Because aging is a multifactorial series of molecular alterations that result in progressive reduction of lung tissue function, the involvement of proteins associated with various physiological processes related to aging was not surprising, serum proteins may be candidate markers of aging. The altered-senescence-associated protein patterns in S-III were related to aging rather than to a single biomarker, providing clear and direct evidence that aging is a critical risk factor for IPF.

Coupling of altered proteins under defined conditions could exploit the information content of serum and identify biomarkers likely to be of clinical value. MS-based proteomics can enable assessment of the roles of blood proteins in clinical diagnoses, as well as identifying new biomarkers and biomarker panels. Analysis of serum proteomes can result in the detection of secreted metabolic enzymes, including those involved in enhancing glycolysis, upregulation of the key metabolic enzyme LDHA was indicative of poorer clinical outcomes. Therefore, the presence of high levels of CCT6A and LDHA and high serum lactic acid concentrations may be diagnostic of IPF.

Use of machine learning to explore the ability of combined biomarkers to predict disease outcomes and prognosis is a promising strategy to improve the accuracy of diagnostic performance. Intriguingly, SERPING1 itself is a candidate biomarker in patients with tuberculosis(28), downregulation of KNG1 expression was observed in patients with sepsis-induced ALI(29), and TTR is a specific biomarker for the clinical diagnosis of non-small cell lung carcinoma(30). In the present study, these three proteins were selected by the machine-learning algorithm as the most important indicators for classification of IPF, showing high specificity and sensitivity in two independent patient cohorts. Furthermore, the combinatorial biomarker panel and clinical data was found to be prognostic in this patient cohort.

The present study had several limitations. The number of patients included in the study cohort was small, as were the numbers in each of the subgroups, suggesting the need for studies in larger patient cohorts, as well as validation of these biomarkers by methods other than serum proteome analysis. Moreover, the kits used to process serum samples can lead to the depletion of highly abundant proteins. For example, EDTA could interfere with the precise determination of MMPs, such as MMP7 and CCL18, previously shown to be markers of IPF(31). Taken together, our data characterized the molecular subtypes of IPF and identified a biomarker panel associated with the pathophysiology of IPF. These results strongly suggest that measuring CCT6A and LDHA, along with high serum levels of lactic acid, could be diagnostic of IPF. Additional studies in larger patient cohorts are needed to determine whether the combination of these three biomarkers could accurately predict IPF.

Data and materials availability

The raw mass spectrometry (MS) proteomics data generated in this study have been deposited in the ProteomeXchange Consortium via the iProX partner repository (http://www.iprox.cn/)(32) under Project ID IPX0004334000, and can be accessed with a direct link https://www.iprox.cn/page/PSV023.html;?url=1664089052598znXd with the password: ASQd.

Acknowledgments

We greatly appreciate Naftali Kaminski (Yale University) for reviewing the manuscript. The work was supported by a Startup package (GY), Henan Normal University, the 111 Project “State Innovation Base for Pulmonary Fibrosis”, the Ministry of Science and Technology, PR China, 2019YFE0119500, and Henan Province Science and Technology Project, 212102310894.

Author Contributions

G.Y. and C.D.: Designed the research plan. L. W.: Data curation, Writing- Original draft preparation, L.W. Y. L., X. Ch.: Proteomics experiments, Z. L.,H.Z: Statistical analysis, Data analysis and data visualization, S.Y.: Performed cell and mouse assay and related data visualization, IHC staining of tissue samples, J.Y and X.P.: Performed ELISA assay, H.Y. and M.Z. :Consulted on clinical questions. I.R. Writing – review & editing. All authors discussed the results and commented on the manuscript.

References

1. Lederer, D. J., and Martinez, F. J. (2018) Idiopathic Pulmonary Fibrosis. *N Engl J Med* 378, 1811-1823

2. Hutchinson, J., Fogarty, A., Hubbard, R., and McKeever, T. (2015) Global incidence and mortality of idiopathic pulmonary fibrosis: a systematic review. *Eur Respir J* 46, 795-806

3. Selman, M., Lopez-Otin, C., and Pardo, A. (2016) Age-driven developmental drift in the pathogenesis of idiopathic pulmonary fibrosis. *European Respiratory Journal* 48, 538-552

4. Chakravarti, D., LaBella, K. A., and DePinho, R. A. (2021) Telomeres: history, health, and hallmarks of aging. *Cell* 184, 306-322

5. Su, Y. J., Wang, P. W., and Weng, S. W. (2021) The Role of Mitochondria in Immune-Cell-Mediated Tissue Regeneration and Ageing. *International Journal of Molecular Sciences* 22

6. Henderson, N. C., Rieder, F., and Wynn, T. A. (2020) Fibrosis: from mechanisms to medicines. *Nature* 587, 555-566

7. Le, A., Cooper, C. R., Gouw, A. M., Dinavahi, R., Maitra, A., Deck, L. M., Royer, R. E., Vander Jagt, D. L., Semenza, G. L., and Dang, C. V. (2010) Inhibition of lactate dehydrogenase A induces oxidative stress and inhibits tumor progression. *Proc Natl Acad Sci U S A* 107, 2037-2042

8. Kottmann, R. M., Kulkarni, A. A., Smolnycki, K. A., Lyda, E., Dahanayake, T., Salibi, R., Honnons, S., Jones, C., Isern, N. G., Hu, J. Z., Nathan, S. D., Grant, G., Phipps, R. P., and Sime, P. J. (2012) 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* 186, 740-751

9. Raghu, G., Remy-Jardin, M., Myers, J. L., Richeldi, L., Ryerson, C. J., Lederer, D. J., Behr, J., Cottin, V., Danoff, S. K., Morell, F., Flaherty, K. R., Wells, A., Martinez, F. J., Azuma, A., Bice, T. J., Bouros, D., Brown, K. K., Collard, H. R., Duggal, A., Galvin, L., Inoue, Y., Jenkins, R. G., Johkoh, T., Kazerooni, E. A., Kitaichi, M., Knight, S. L., Mansour, G., Nicholson, A. G., Pipavath, S. N. J., Buendia-Roldan, I., Selman, M., Travis, W. D., Walsh, S., Wilson, K. C., Soc, A. T., Soc, E. R., Soc, J. R., and Soc, L. A. T. (2018) Diagnosis of Idiopathic Pulmonary Fibrosis An Official ATS/ERS/JRS/ALAT Clinical Practice Guideline. *Am J Resp Crit Care* 198, E44-

E68

10. Geyer, P. E., Holdt, L. M., Teupser, D., and Mann, M. (2017) Revisiting biomarker discovery by plasma proteomics. *Mol Syst Biol* 13

11. Saraswat, M., Joenväärä, S., Tohmola, T., Sutinen, E., Vartiainen, V., Koli, K., Myllärniemi, M., and Renkonen, R. (2020) Label-free plasma proteomics identifies haptoglobin-related protein as candidate marker of idiopathic pulmonary fibrosis and dysregulation of complement and oxidative pathways. *Sci Rep* 10, 7787

12. Raghu, G., Remy-Jardin, M., Myers, J. L., Richeldi, L., Ryerson, C. J., Lederer, D. J., Behr, J., Cottin, V., Danoff, S. K., Morell, F., Flaherty, K. R., Wells, A., Martinez, F. J., Azuma, A., Bice, T. J., Bouros, D., Brown, K. K., Collard, H. R., Duggal, A., Galvin, L., Inoue, Y., Jenkins, R. G., Johkoh, T., Kazerooni, E. A., Kitaichi, M., Knight, S. L., Mansour, G., Nicholson, A. G., Pipavath, S. N. J., Buendia-Roldan, I., Selman, M., Travis, W. D., Walsh, S., Wilson, K. C., American Thoracic Society, E. R. S. J. R. S., and Latin American Thoracic, S. (2018) Diagnosis of Idiopathic Pulmonary Fibrosis. An Official ATS/ERS/JRS/ALAT Clinical Practice Guideline. *Am J Respir Crit Care Med* 198, e44-e68

13. Chen, Y. M., Zheng, Y., Yu, Y., Wang, Y., Huang, Q., Qian, F., Sun, L., Song, Z. G., Chen, Z., Feng, J., An, Y., Yang, J., Su, Z., Sun, S., Dai, F., Chen, Q., Lu, Q., Li, P., Ling, Y., Yang, Z., Tang, H., Shi, L., Jin, L., Holmes, E. C., Ding, C., Zhu, T. Y., and Zhang, Y. Z. (2020) Blood molecular markers associated with COVID-19 immunopathology and multi-organ damage. *Embo j* 39, e105896

14. Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011) Global quantification of mammalian gene expression control. *Nature* 473,

337-342

15. Wilkerson, M. D., and Hayes, D. N. (2010) ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* 26, 1572-1573

16. Langfelder, P., and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9, 559

17. Zhang, B., and Horvath, S. (2005) A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* 4, Article17

18. Langfelder, P., and Horvath, S. (2007) Eigengene networks for studying the relationships between co-expression modules. *Bmc Syst Biol* 1

19. Niu, L., Geyer, P. E., Wewer Albrechtsen, N. J., Gluud, L. L., Santos, A., Doll, S., Treit, P. V., Holst, J. J., Knop, F. K., Vilsboll, T., Junker, A., Sachs, S., Stemmer, K., Muller, T. D., Tschop, M. H., Hofmann, S. M., and Mann, M. (2019) Plasma proteome profiling discovers novel proteins associated with non-alcoholic fatty liver disease. *Mol Syst Biol* 15, e8793

20. Cendrowski, J., Kaczmarek, M., Mazur, M., Kuzmicz-Kowalska, K., Jastrzebski, K., Brewinska-Olchowik, M., Kominek, A., Piwocka, K., and Miaczynska, M. (2020) Splicing variation of BMP2K balances abundance of COPII assemblies and autophagic degradation in erythroid cells. *Elife* 9

21. Hazell, G. G. J., Peachey, A. M. G., Teasdale, J. E., Sala-Newby, G. B., Angelini, G. D., Newby, A. C., and White, S. J. (2016) PI16 is a shear stress and inflammation-regulated inhibitor of MMP2. *Sci Rep-Uk* 6

22. Thannickal, V. J. (2013) Mechanistic links between aging and lung fibrosis. *Biogerontology* 14, 609-615

23. O'Dwyer, D. N., and Moore, B. B. (2017) The role of periostin in lung fibrosis and airway remodeling. *Cell Mol Life Sci* 74, 4305-4314

24. Kim, Y. I., Ahn, J. M., Sung, H. J., Na, S. S., Hwang, J., Kim, Y., and Cho, J. Y. (2016) Meta-markers for the differential diagnosis of lung cancer and lung disease. *J Proteomics* 148, 36-43

25. Yu, G., Tzouvelekis, A., Wang, R., Herazo-Maya, J. D., Ibarra, G. H., Srivastava, A., de Castro, J. P. W., DeIuliis, G., Ahangari, F., Woolard, T., Aurelien, N., Arrojo, E. D. R., Gan, Y., Graham, M., Liu, X., Homer, R. J., Scanlan, T. S., Mannam, P., Lee, P. J., Herzog, E. L., Bianco, A. C., and Kaminski, N. (2018) Thyroid hormone inhibits lung fibrosis in mice by improving epithelial mitochondrial function. *Nat Med* 24, 39-49

26. Qiu, H., Song, J., Hu, J., Wang, L., Qiu, L., Liu, H., Lin, G., Luan, X., Liu, Y., and He, J. (2022) Low serum transthyretin levels predict stroke-associated pneumonia. *Nutr Metab Cardiovasc Dis* 32, 632-640

27. Ge, S., Xia, X., Ding, C., Zhen, B., Zhou, Q., Feng, J., Yuan, J., Chen, R., Li, Y., Ge, Z., Ji, J., Zhang, L., Wang, J., Li, Z., Lai, Y., Hu, Y., Li, Y., Li, Y., Gao, J., Chen, L., Xu, J., Zhang, C., Jung, S. Y., Choi, J. M., Jain, A., Liu, M., Song, L., Liu, W., Guo, G., Gong, T., Huang, Y., Qiu, Y., Huang, W., Shi, T., Zhu, W., Wang, Y., He, F., Shen, L., and Qin, J. (2018) A proteomic landscape of diffuse-type gastric cancer. *Nat Commun* 9, 1012

28. Lubbers, R., Sutherland, J. S., Goletti, D., de Paus, R. A., Dijkstra, D. J., van Moorsel, C. H. M., Veltkamp, M., Vestjens, S. M. T., Bos, W. J. W., Petrone, L., Malherbe, S. T., Walzl, G., Gelderman, K. A., Groeneveld, G. H., Geluk, A., Ottenhoff, T. H. M., Joosten, S. A., and Trouw, L. A. (2020) Expression and production of the SERPING1-encoded endogenous complement

regulator C1-inhibitor in multiple cohorts of tuberculosis patients. *Mol Immunol* 120, 187-195

29. Hu, Q., Wang, Q., Han, C. G., and Yang, Y. (2020) Sufentanil attenuates inflammation and oxidative stress in sepsis-induced acute lung injury by downregulating KNG1 expression. *Mol Med Rep* 22, 4298-4306

30. Wang, D. B., Li, X., Lu, X. K., Sun, Z. Y., Zhang, X., Chen, X., Ma, L., and Xia, H. G. (2021) Transthyretin Suppressed Tumor Progression in Nonsmall Cell Lung Cancer by Inactivating MAPK/ERK Pathway. *Cancer Biother Radio*

31. Hamai, K., Iwamoto, H., Ishikawa, N., Horimasu, Y., Masuda, T., Miyamoto, S., Nakashima, T., Ohshimo, S., Fujitaka, K., Hamada, H., Hattori, N., and Kohno, N. (2016) Comparative Study of Circulating MMP-7, CCL18, KL-6, SP-A, and SP-D as Disease Markers of Idiopathic Pulmonary Fibrosis. *Dis Markers* 2016, 4759040

32. Ma, J., Chen, T., Wu, S. F., Yang, C. Y., Bai, M. Z., Shu, K. X., Li, K. L., Zhang, G. Q., Jin, Z., He, F. C., Hermjakob, H., and Zhu, Y. P. (2019) iProX: an integrated proteome resource. *Nucleic Acids Res* 47, D1211-D1217

Figure legends

Fig. 1 Proteomic features of the IPF subgroups. Molecular subtyping of IPF was based on altered proteomes and their correlations with clinical features.

A. Cumulative number of proteins identified in serum samples from 30 healthy controls (blue dots) and 30 patients with IPF (red dots).

B. Numbers of identified proteins in serum samples from 30 healthy controls (blue dots) and 30 IPF patients (red dots).

C. Relative abundance of 2,314 serum proteins. Several proteins ranged widely in abundance (black dots).

D. Consensus clustering analysis of the proteomic profiling identifying three subtypes in the IPF cohort.

E. Kaplan–Meier analyses of overall survival (OS) of patients in the S-I (n=16), S-II (n=4), and S-III (n=10) subgroups. (P-values calculated by two-sided log-rank tests).

F. Heat map of the over-represented proteins in the three IPF subtypes.

G. Proteins differentially expressed in the three IPF subtypes.

H. Associations between expression of BMP2K, PI16, and ATP5A1 proteins, and overall survival (Kaplan–Meier analysis, P-value from log-rank test, high means IPF/N >median value).

I. Age with the three IPF proteomic subtypes (P-values calculated by Fisher's exact tests).

Fig. 2 WGCNA identification of modules of highly correlated genes and assessment of their relationships to clinical variables.

A. Heatmap of the weighted gene co-expression network. The plot indicates the TOM among all genes analyzed. Genes in columns and their corresponding rows are hierarchically clustered by cluster dendrograms, which are presented along the top and left side of the plot.

B. Module-trait relationships between six modules and ten clinical traits.

C. Heatmap of the change in genes in the module of age.

D. Heatmap of the age-related genes in the three subgroups.

E. Associations of HSP90AB and CAMKK1 expression with clinical outcomes in 30 IPF patients.

Fig. 3 Aberrantly expressed metabolic enzymes involved in enhanced glycolysis in serum

proteomes of patients with IPF.

A. Pathway schematic showing DEPs (t-test, $p < 0.05$) mapped onto glucose metabolism pathways.

B. Boxplots showing proteins differentially expressed by IPF patients with normal and above-normal levels of serum lactate (P-values calculated by t-test).

C. Violin plots of LDHA and LDHB expression in 30 healthy controls (blue dots) and 30 IPF patients (red dots).

D. Associations of LDHA expression with clinical outcomes in IPF patients (p-values calculated by log-rank tests).

E. Violin plots of CCT6A expression in 30 healthy controls (blue dots) and 30 IPF patients (red dots).

F. ELISA validation of CCT6A expression in IPF patients (P-values calculated by t-tests).

G. IHC staining showing CCT6A expression in lungs from healthy controls and IPF patients.

H. IHC staining showing CCT6A expression in the bleomycin model of lung fibrosis in mice.

I. Representative immunoblots of whole lung lysates of mice incubated with antibodies against CCT6A and GAPDH.

J. Western blots of CCT6A expression normalized to β-actin. * $P < 0.05$, as determined by ANOVA.

Fig. 4 Association of changes in CCT6A expression and high lactic acid concentrations with the fibroblast phenotype.

A. Representative immunoblots showing CCT6A and α-SMA expression in MCR5 cells transfected with control plasmid and plasmid overexpressing CCT6A.

B. Western blots of CCT6A expression normalized to β-actin. * $P < 0.05$, as determined by ANOVA.

C. Representative images of α-SMA immunofluorescence staining of MRC5 cells. Original magnification, ×100. Scale bars: 5 μm.

D. Representative immunoblots showing CCT6A, COLA1, and FN expression in MCR5 cells transfected with control and CCT6A siRNAs.

E. Western blots of CCT6A expression normalized to β-actin. * $P < 0.05$, ** $P < 0.01$, as determined by ANOVA.

F. ECAR of control and CCT6A-overexpressing MRC5 cells.

G-H. Lactate production in the supernatants of MRC5 cells and in the lungs of bleomycin mice.

I. Pyruvate production in MRC5 cells.

J. Expression of LDHA mRNA in MRC5 cells overexpressing CCT6A.

K. Representative immunoblots showing LDHA expression in MRC5 cells overexpressing CCT6A.

L. Western blots of LDHA expression normalized to β-actin. * $P < 0.05$, as determined by ANOVA.

Fig. 5 Machine-learning-based selection of biomarker combinations for classification of IPF.

A. Receiver operating characteristic (ROC) curve for the classification model. Calculation of AUC values in the patient cohort by 5-fold cross-validation. Confusion matrix of the four-protein combination in the patient cohort.

B. ROC curve for the test model Calculation of AUC values in the public cohort by 5-fold cross-validation. Confusion matrix of the four-protein combination in the public cohort.

C. Associations between the protein combinations and clinical outcomes in 30 IPF patients of the classification model.

D. Heatmap of the combination biomarkers in the public cohort (PRIDE project PXD010965).

E. ELISA determination of TTR expression in an independent cohort. (P-values calculated by t-tests).

F. Correlation between TTR expression and patient age in the study cohort.

Supplemental figures legends

Figure S1 Schematic of the proteomic analyses of serum samples from 30 IPF patients and 30 healthy controls.

Figure S2 Profiling of serum proteomics of IPF patients and healthy controls.

A. Quality control of mass spectrometry using a tryptic digest of HEK293T cells. The bottom-left half of the panel shows the pairwise Spearman’s correlation coefficients of the samples, and the top-right half of the panel depicts the pairwise scatter plots from the same comparisons.

B. Assessment of study quality by analysis of the protein markers FGA, FGB, and FGG.

C. Heatmap of the altered proteins in healthy controls and IPF samples.

D. Volcano plot of differentially expressed genes differing significantly in non-IPF and IPF samples. The log2 differential expression ratio and the -log10 (p-value) were plotted for each gene. Proteins with differential expression ratios >2 or <2 were defined as those significantly up- and downregulated, respectively.

Figure S3 Proteomic subtypes of IPF with their molecular characteristics.

A. Consensus clustering plus identification of three serum proteomic subtypes of IPF samples. The panel shows a consensus matrix of 30 IPF samples from k=2 to k=6, with k=3 considered the ideal value based on visual inspection of the consensus matrix and the change in area under the CDF.

B. Top 30 exclusively expressed proteins in S-I, S-II, and S-III patients.

Figure S4 Correlation of FOT value for CCT6A(MS data) with ELISA value of CCT6A in the same sera of IPF patients; Pearson correlation, P = 0.0056. n = 17 human samples.

Figure S5 Identification of combination of biomarkers by machine-learning.

A. Workflow of the machine-learning.

B. TTR, KNG1 and FCN3 expression in the public cohort (left) and the study cohort (right).

Table 1 Information of IPF cohort and healthy control cohort

Characteristics	Control	IPF
Number	30	30
Age	61.07 ± 9.85	64.50 ± 10.58
Gender		
Male	17	22
Female	13	8
Smoking	8	11

Figure 1

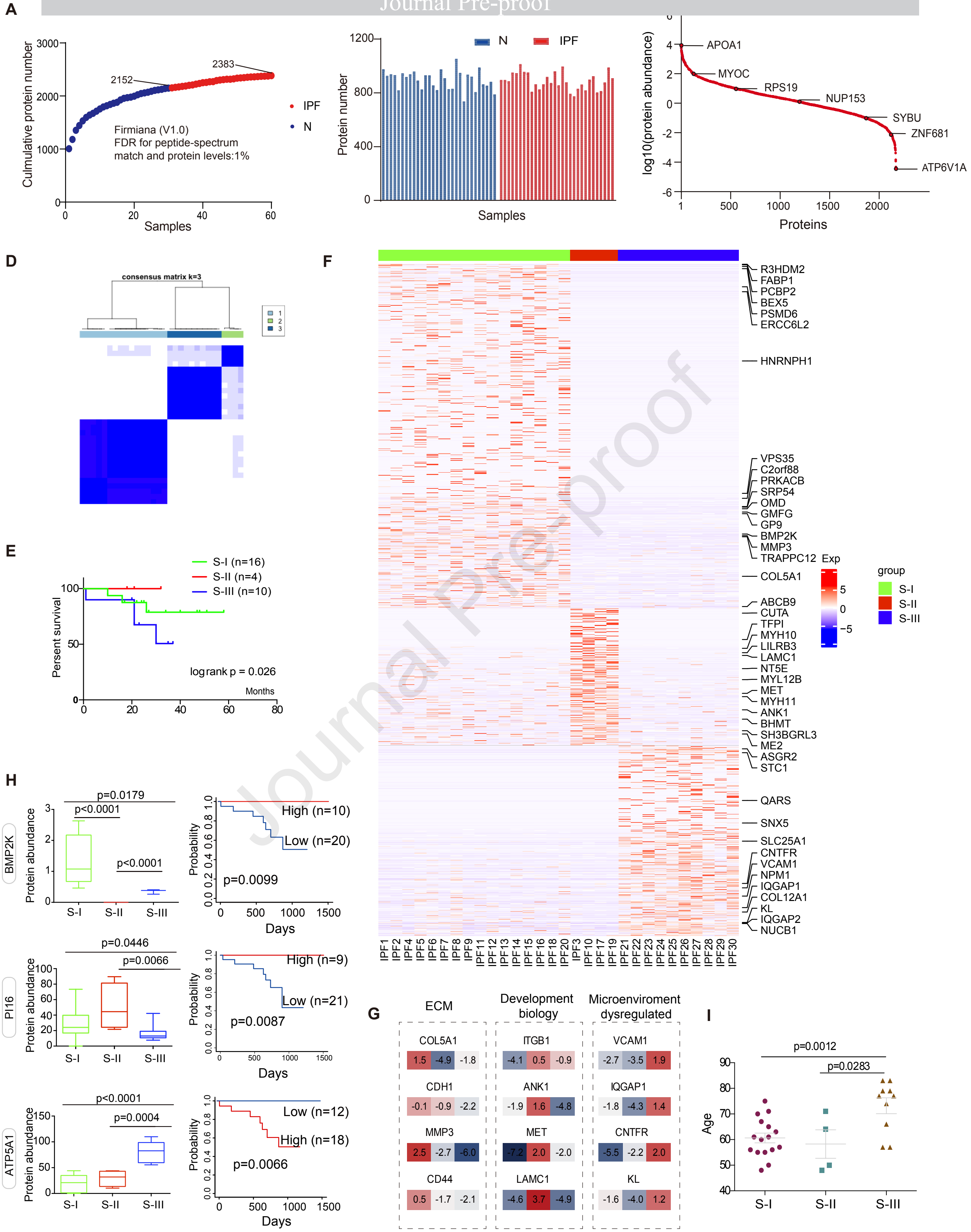

A
Culmulative protein number
3000
2000
1000
0
0
20
40
60
Samples
2152
2383
Firmiana (V1.0)
FDR for peptide-spectrum
match and protein levels:1%
IPF
N
Protein number
1200
800
400
0
N
IPF
Samples
log10(protein abundance)
4
2
0
-2
-4
-6
1
500
1000
1500
2000
Proteins
APOA1
MYOC
RPS19
NUP153
SYBU
ZNF681
ATP6V1A
D
consensus matrix k=3
1
2
3
F
R3HDM2
FABP1
PCBP2
BEX5
PSMD6
ERCC6L2
HNRNPH1
VPS35
C2orf88
PRKACB
SRP54
OMD
GMFG
GP9
BMP2K
MMP3
TRAPPC12
COL5A1
ABCB9
CUTA
TFPI
MYH10
LILRB3
LAMC1
NT5E
MYL12B
MET
MYH11
ANK1
BHMT
SH3BGRL3
ME2
ASGR2
STC1
QARS
SNX5
SLC25A1
CNTFR
VCAM1
NPM1
IQGAP1
COL12A1
KL
IQGAP2
NUCB1
Exp
5
0
−5
group
S-I
S-II
S-III
IPF1
IPF2
IPF4
IPF5
IPF6
IPF7
IPF8
IPF9
IPF11
IPF12
IPF13
IPF14
IPF15
IPF16
IPF18
IPF20
IPF3
IPF10
IPF17
IPF19
IPF21
IPF22
IPF23
IPF24
IPF25
IPF26
IPF27
IPF28
IPF29
IPF30
E
S-I (n=16)
S-II (n=4)
S-III (n=10)
Persent survival
100
50
0
logrank p = 0.026
Months
0
20
40
60
80
H
BMP2K
Protein abundance
p=0.0179
p<0.0001
p<0.0001
S-I
S-II
S-III
Probability
High (n=10)
Low (n=20)
p=0.0099
Days
0
500
1000
1500
PI16
Protein abundance
p=0.0446
p=0.0066
S-I
S-II
S-III
Probability
High (n=9)
Low (n=21)
p=0.0087
Days
ATP5A1
Protein abundance
p<0.0001
p=0.0004
S-I
S-II
S-III
Probability
Low (n=12)
High (n=18)
p=0.0066
Days
G
ECM
Development biology
Microenviroment dysregulated
COL5A1
1.5 -4.9 -1.8
CDH1
-0.1 -0.9 -2.2
MMP3
2.5 -2.7 -6.0
CD44
0.5 -1.7 -2.1
ITGB1
-4.1 0.5 -0.9
ANK1
-1.9 1.6 -4.8
MET
-7.2 2.0 -2.0
LAMC1
-4.6 3.7 -4.9
VCAM1
-2.7 -3.5 1.9
IQGAP1
-1.8 -4.3 1.4
CNTFR
-5.5 -2.2 2.0
KL
-1.6 -4.0 1.2
I
Age
p=0.0012
p=0.0283
90
80
70
60
50
40
S-I
S-II
S-III

Figure 2

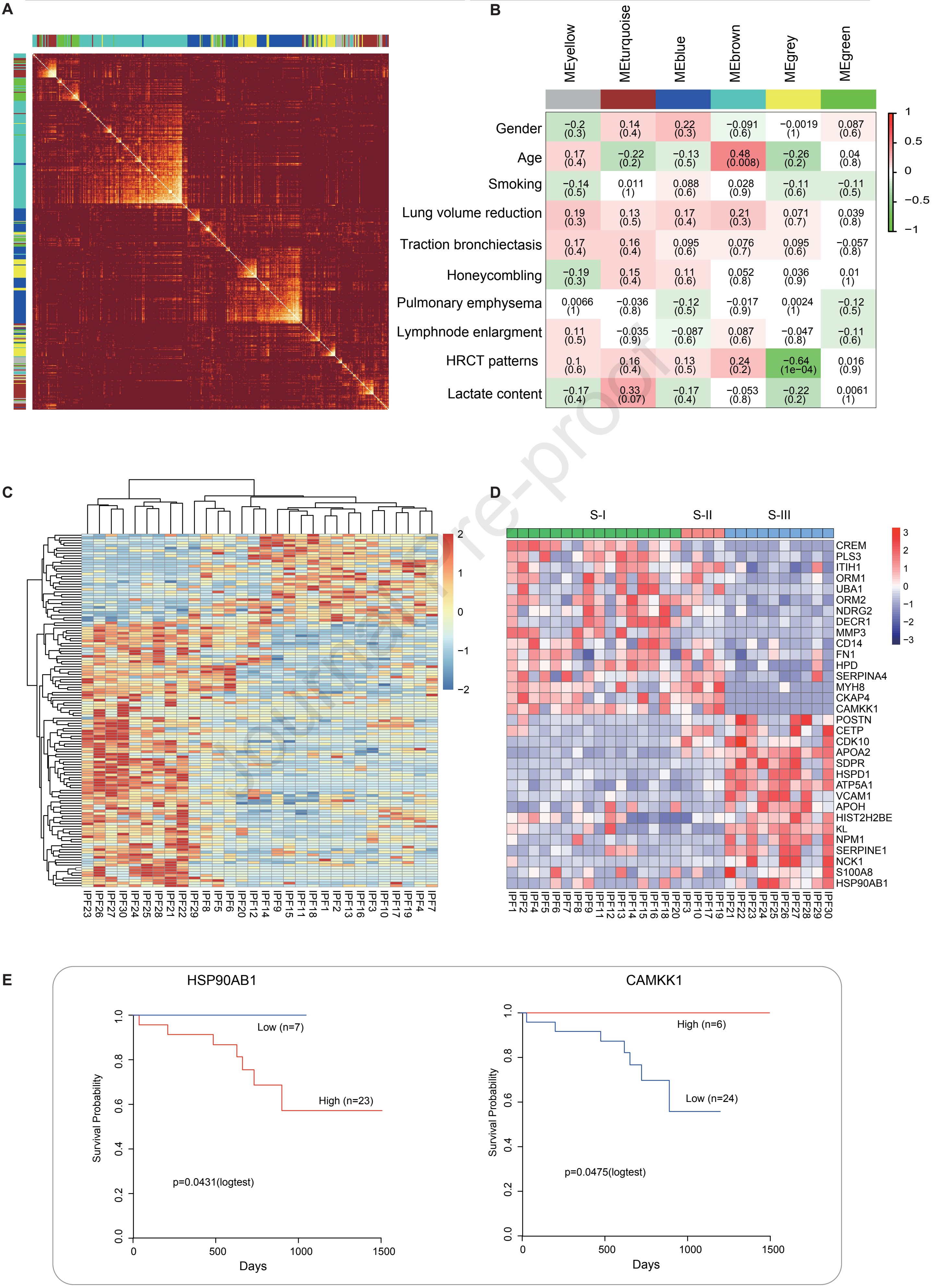
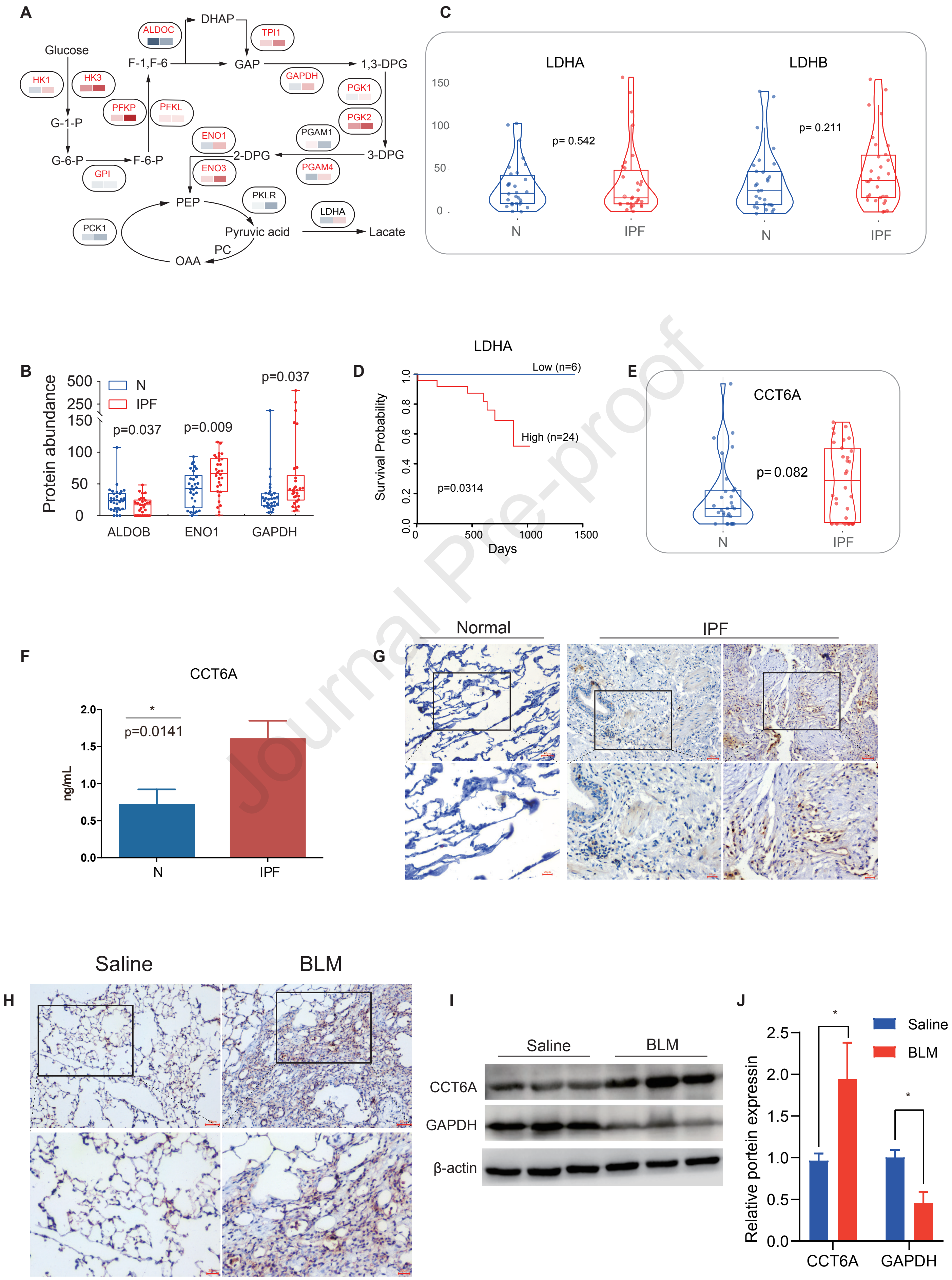

A
B
MEyellow
MEturquoise
MEblue
MEbrown
MEgrey
MEgreen
Gender -0.2 (0.3) 0.14 (0.4) 0.22 (0.3) −0.091 (0.6) −0.0019 (1) 0.087 (0.6)
Age 0.17 (0.4) −0.22 (0.2) −0.13 (0.5) 0.48 (0.008) −0.26 (0.2) 0.04 (0.8)
Smoking −0.14 (0.5) 0.011 (1) 0.088 (0.6) 0.028 (0.9) −0.11 (0.6) −0.11 (0.5)
Lung volume reduction 0.19 (0.3) 0.13 (0.5) 0.17 (0.4) 0.21 (0.3) 0.071 (0.7) 0.039 (0.8)
Traction bronchiectasis 0.17 (0.4) 0.16 (0.4) 0.095 (0.6) 0.076 (0.7) 0.095 (0.6) −0.057 (0.8)
Honeycombling −0.19 (0.3) 0.15 (0.4) 0.11 (0.6) 0.052 (0.8) 0.036 (0.9) 0.01 (1)
Pulmonary emphysema 0.0066 (1) −0.036 (0.8) −0.12 (0.5) −0.017 (0.9) 0.0024 (1) −0.12 (0.5)
Lymphnode enlargment 0.11 (0.5) −0.035 (0.9) −0.087 (0.6) 0.087 (0.6) −0.047 (0.8) −0.11 (0.6)
HRCT patterns 0.1 (0.6) 0.16 (0.4) 0.13 (0.5) 0.24 (0.2) −0.64 (1e−04) 0.016 (0.9)
Lactate content −0.17 (0.4) 0.33 (0.07) −0.17 (0.4) −0.053 (0.8) −0.22 (0.2) 0.0061 (1)
1
0.5
0
−0.5
−1
C
2
1
0
−1
−2
IPF23 IPF26 IPF27 IPF30 IPF24 IPF25 IPF28 IPF21 IPF22 IPF29 IPF8 IPF5 IPF6 IPF20 IPF12 IPF14 IPF9 IPF15 IPF11 IPF18 IPF1 IPF2 IPF13 IPF16 IPF3 IPF10 IPF17 IPF19 IPF4 IPF7
D
S-I
S-II
S-III
CREM
PLS3
ITIH1
ORM1
UBA1
ORM2
NDRG2
DECR1
MMP3
CD14
FN1
HPD
SERPINA4
MYH8
CKAP4
CAMKK1
POSTN
CETP
CDK10
APOA2
SDPR
HSPD1
ATP5A1
VCAM1
APOH
HIST2H2BE
KL
NPM1
SERPINE1
NCK1
S100A8
HSP90AB1
3
2
1
0
−1
−2
−3
IPF1 IPF2 IPF4 IPF5 IPF6 IPF7 IPF8 IPF9 IPF11 IPF12 IPF13 IPF14 IPF15 IPF16 IPF18 IPF20 IPF3 IPF10 IPF17 IPF19 IPF21 IPF22 IPF23 IPF24 IPF25 IPF26 IPF27 IPF28 IPF29 IPF30
E
HSP90AB1
Low (n=7)
High (n=23)
p=0.0431(logtest)
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
Days
CAMKK1
High (n=6)
Low (n=24)
p=0.0475(logtest)
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
Days

Figure 3



A
Glucose
HK1
HK3
G-1-P
G-6-P
GPI
F-6-P
PFKP
PFKL
F-1,F-6
ALDOC
DHAP
TPI1
GAP
GAPDH
1,3-DPG
PGK1
PGK2
3-DPG
PGAM1
PGAM4
2-DPG
ENO1
ENO3
PEP
PKLR
Pyruvic acid
LDHA
Lacate
PCK1
PC
OAA
C
LDHA
LDHB
150
100
50
0
p= 0.542
p= 0.211
N
IPF
N
IPF
B
Protein abundance
500
250
150
100
50
0
N
IPF
p=0.037
p=0.009
p=0.037
ALDOB
ENO1
GAPDH
D
LDHA
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
Low (n=6)
High (n=24)
p=0.0314
0
500
1000
1500
Days
E
CCT6A
p= 0.082
N
IPF
F
CCT6A
ng/mL
2.0
1.5
1.0
0.5
0.0
*
p=0.0141
N
IPF
G
Normal
IPF
H
Saline
BLM
I
Saline
BLM
CCT6A
GAPDH
β-actin
J
Relative portein expressin
2.5
2.0
1.5
1.0
0.5
0.0
Saline
BLM
*
*
CCT6A
GAPDH

Figure 4

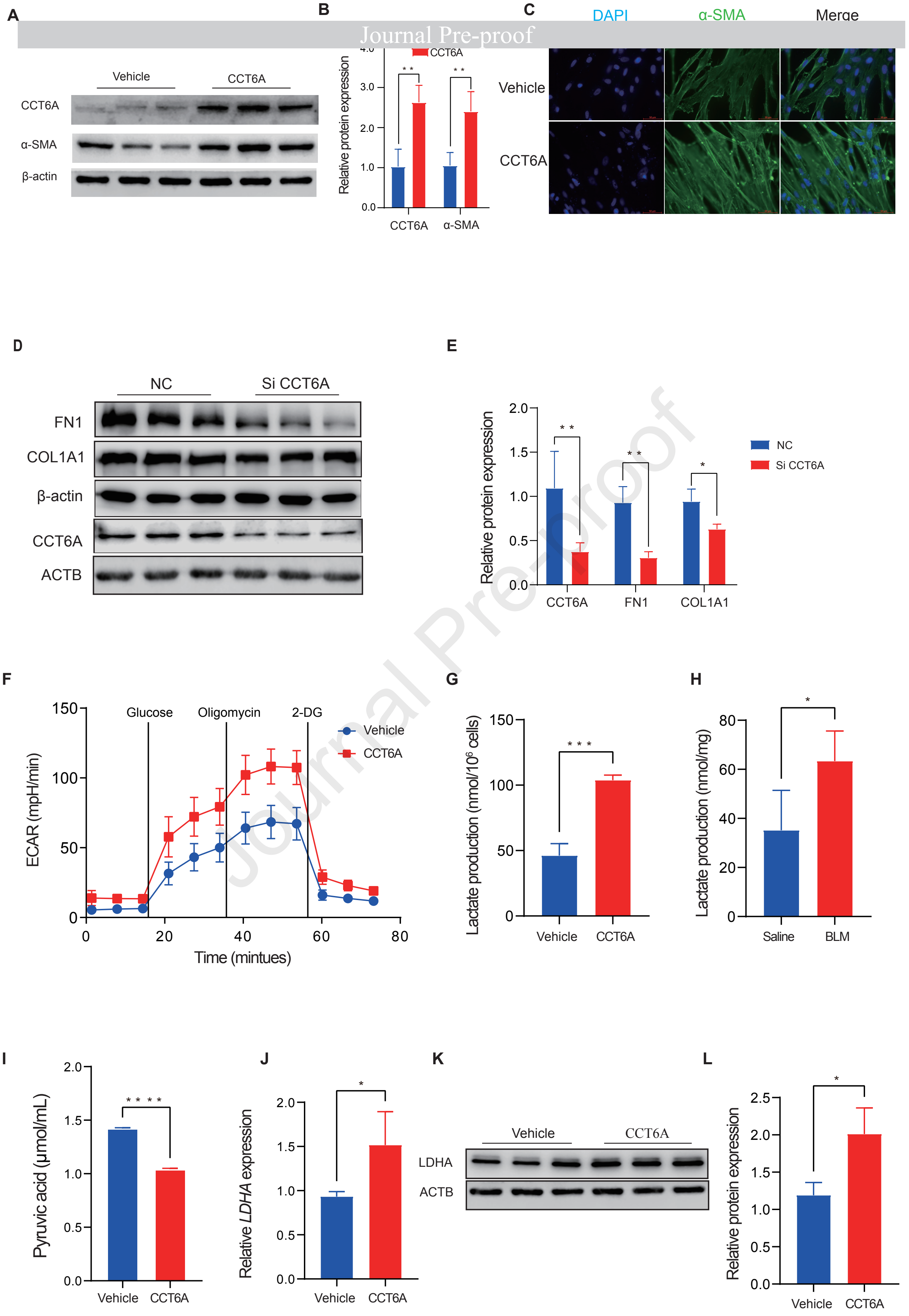
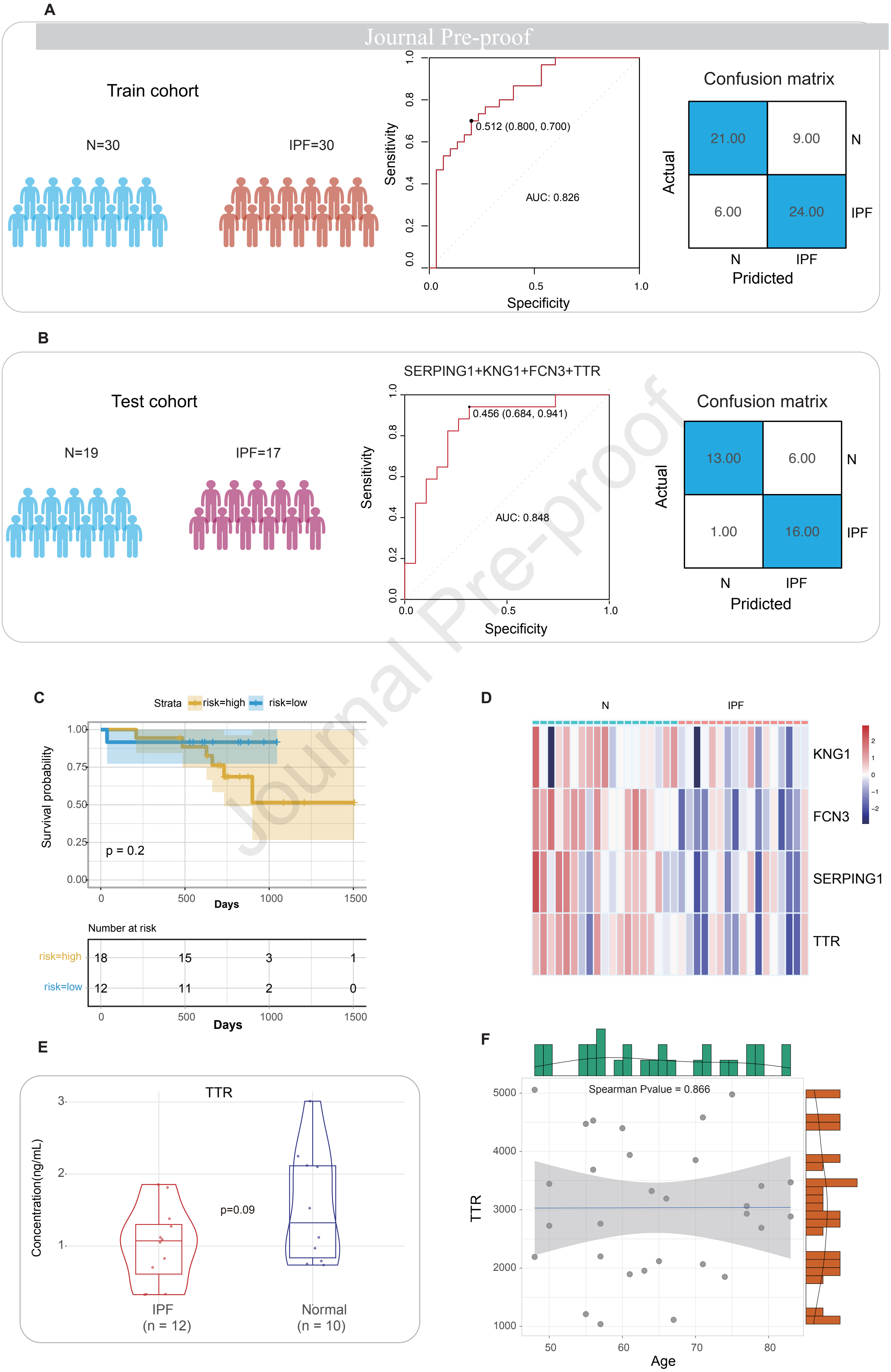

A
Vehicle
CCT6A
CCT6A
α-SMA
β-actin
B
CCT6A
Relative protein expression
4.0
3.0
2.0
1.0
0.0
* *
* *
CCT6A
α-SMA
C
DAPI
α-SMA
Merge
Vehicle
CCT6A
D
NC
Si CCT6A
FN1
COL1A1
β-actin
CCT6A
ACTB
E
Relative protein expression
2.0
1.5
1.0
0.5
0.0
* *
* *
*
CCT6A
FN1
COL1A1
NC
Si CCT6A
F
ECAR (mpH/min)
150
100
50
0
Glucose
Oligomycin
2-DG
Vehicle
CCT6A
0
20
40
60
80
Time (mintues)
G
Lactate production (nmol/10^6 cells)
150
100
50
0
* * *
Vehicle
CCT6A
H
Lactate production (nmol/mg)
80
60
40
20
0
*
Saline
BLM
I
Pyruvic acid (μmol/mL)
2.0
1.5
1.0
0.5
0.0
* * * *
Vehicle
CCT6A
J
Relative *LDHA* expression
2.0
1.5
1.0
0.5
0.0
*
Vehicle
CCT6A
K
Vehicle
CCT6A
LDHA
ACTB
L
Relative protein expression
2.5
2.0
1.5
1.0
0.5
0.0
*
Vehicle
CCT6A

Figure 5


A
Train cohort
N=30
IPF=30
Sensitivity
Specificity
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
0.512 (0.800, 0.700)
AUC: 0.826
Confusion matrix
Actual
Pridicted
21.00
9.00
6.00
24.00
N
IPF
N
IPF
B
Test cohort
N=19
IPF=17
SERPING1+KNG1+FCN3+TTR
Sensitivity
Specificity
0.456 (0.684, 0.941)
AUC: 0.848
Confusion matrix
Actual
Pridicted
13.00
6.00
1.00
16.00
N
IPF
N
IPF
C
Strata
risk=high
risk=low
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0.2
0
500
1000
1500
Days
Number at risk
risk=high
18
15
3
1
risk=low
12
11
2
0
0
500
1000
1500
Days
D
N
IPF
KNG1
FCN3
SERPING1
TTR
2
1
0
−1
−2
E
TTR
Concentration(ng/mL)
3
2
1
p=0.09
IPF
(n = 12)
Normal
(n = 10)
F
Spearman Pvalue = 0.866
TTR
5000
4000
3000
2000
1000
Age
50
60
70
80

Highlights

● A serum proteome profiling by DIA-MS identified 2833 gene products from IPF and normal subjects, three subgroups were distinguished in IPF patients in signal pathways and overall survival.

● Aging-associated signatures in module MEturquoise were identified by WGCNA coincidently falling into S-III which provided clear and direct evidence that aging is a critical risk factor for IPF rather than to a single biomarker

● LDHA and CCT6A expression, which were associated with glucose metabolic reprogramming, were correlated with high serum lactic acid content in the patients with IPF.

● Cross-model analysis and machine learning showed that a combinatorial biomarker accurately distinguished IPF patients from healthy subjects and validated from another cohort and ELISA assay.

AUTHOR CONTRIBUTION STATEMENT

G.Y. and C.D.: Designed the research plan. L. W.: Data curation, Writing- Original draft preparation, L.W. Y. L., X. Ch.: Proteomics experiments, Z. L.,H.Z: Statistical analysis, Data analysis and data visualization, S.Y.: Performed cell and mouse assay and related data visualization, IHC staining of tissue samples, J.Y and X.P.: Performed ELISA assay, H.Y. and M.Z. :Consulted on clinical questions. I.R. Writing – review & editing.

All authors discussed the results and commented on the manuscript.

Brief

Wang et el (2022) performed serum proteomics by DIA-MS and identified 2833 gene products from IPF and normal subjects, and distinguished in IPF patients into three subgroups in signal pathways and overall survival. Aging-associated signatures by WGCNA coincidently provided clear and direct evidence that aging is a critical risk factor for IPF rather than to a single biomarker. LDHA and CCT6A expression, which were associated with glucose metabolic reprogramming, were correlated with high serum lactic acid content in the patients with IPF. Cross-model analysis and machine learning showed that a combinatorial biomarker accurately distinguished IPF patients from healthy subjects and validated from another cohort and ELISA assay.

Declaration of interests

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

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: