Proteomic profiling of bronchoalveolar lavage fluid uncovers protein clusters linked to survival in idiopathic forms of interstitial lung disease

Linh T. Ngo[1,12], Michaella J. Rekowski[2,12], Devin C. Koestler[3], Takafumi Yorozuya[4], Atsushi Saito[4], Imaan Azeem[1], Alexis Harrison[1], M. Kristen Demoruelle[5], Jonathan Boomer[1], Bryant R. England[6,7], Paul Wolters[8], Philip L. Molyneaux[9], Mario Castro[1], Joyce S. Lee[10], Joshua J. Solomon[11], Koji Koronuma[4], Michael P. Washburn[2] and Scott M. Matson[1]

[1]Division of Pulmonary, Critical Care and Sleep Medicine, University of Kansas School of Medicine, Kansas City, KS, USA. [2]Department of Cancer Biology, University of Kansas School of Medicine, Kansas City, KS, USA. [3]Department of Biostatistics and Data Science, University of Kansas Medical Center, Kansas City, KS, USA. [4]Department of Respiratory Medicine and Allergology, Sapporo Medical University, Sapporo, Japan. [5]Division of Rheumatology, University of Colorado School of Medicine, Aurora, CO, USA. [6]Division of Rheumatology and Immunology, University of Nebraska Medical Center, Omaha, NE, USA. [7]Veterans Affairs Nebraska-Western Iowa Health Care System, Omaha, NE, USA. [8]Division of Pulmonary and Critical Care Medicine, University of California San Francisco, San Francisco, CA, USA. [9]National Heart and Lung Institute, Imperial College London, London, UK. [10]Division of Pulmonary Sciences and Critical Care Medicine, University of Colorado, Aurora, CO, USA. [11]Division of Pulmonary, Critical Care and Sleep Medicine, National Jewish Health Hospital, Denver, CO, USA. [12]These authors contributed equally.

Corresponding author: Scott M. Matson (smatson@kumc.edu)

Shareable abstract (@ERSpublications)
Label-free quantitative proteomic analysis of bronchoalveolar lavage fluid from patients with idiopathic forms of interstitial lung disease identifies unique protein clusters that are associated with discrete survival trajectories https://bit.ly/4cfKQza

Cite this article as: Ngo LT, Rekowski MJ, Koestler DC, *et al*. Proteomic profiling of bronchoalveolar lavage fluid uncovers protein clusters linked to survival in idiopathic forms of interstitial lung disease. *ERJ Open Res* 2024; 10: 00192-2024 [DOI: 10.1183/23120541.00192-2024].

Received: 29 Feb 2024
Accepted: 3 June 2024

Abstract

Background Idiopathic interstitial pneumonias (IIPs), such as idiopathic pulmonary fibrosis and interstitial pneumonia with autoimmune features, present diagnostic and therapeutic challenges due to their heterogeneous nature. This study aimed to identify intrinsic molecular signatures within the lung microenvironment of these IIPs through proteomic analysis of bronchoalveolar lavage fluid (BALF).

Methods Patients with IIP (n=23) underwent comprehensive clinical evaluation including pre-treatment bronchoscopy and were compared with controls without lung disease (n=5). Proteomic profiling of BALF was conducted using label-free quantitative methods. Unsupervised cluster analyses identified protein expression profiles that were then analysed to predict survival outcomes and investigate associated pathways.

Results Proteomic profiling successfully differentiated IIP from controls. k-means clustering based on protein expression revealed three distinct IIP clusters, which were not associated with age, smoking history, or baseline pulmonary function. These clusters had unique survival trajectories and provided more accurate survival predictions than the Gender Age Physiology index (concordance index 0.794 *versus* 0.709). The cluster with the worst prognosis featured decreased inflammatory signalling and complement activation, with pathway analysis highlighting altered immune response pathways related to immunoglobulin production and B-cell-mediated immunity.

Conclusions The unsupervised clustering of BALF proteomics provided a novel stratification of IIP patients, with potential implications for prognostic and therapeutic targeting. The identified molecular phenotypes underscore the diversity within the IIP classification and the potential importance of personalised treatments for these conditions. Future validation in larger, multi-ethnic cohorts is essential to confirm these findings and to explore their utility in clinical decision-making for patients with IIP.

Background

Idiopathic interstitial pneumonia (IIP) encompasses a group of lung diseases with unknown causes that are distinct from those associated with autoimmune diseases or known antigen exposures. Among these, idiopathic pulmonary fibrosis (IPF) is the most common and is diagnosed using a multidisciplinary approach that integrates clinical, radiographical and histopathological criteria [1]. However, there exists a subset of IIP patients whose features diverge from the established IPF criteria, particularly those exhibiting autoimmune characteristics without meeting the criteria for a definitive autoimmune diagnosis, drawing significant interest in clinical and research settings. An international task force in 2015 defined this unique class of IIP as interstitial pneumonia with autoimmune features (IPAF) [2]. Patients with IPAF exhibit at least two autoimmune features from defined clinical, serological and morphological domains [2]. Studies have suggested that IPAF patients may have a survival advantage over those with IPF, supporting the theory that IPAF represents a distinct clinical entity [3]. Given the presumed importance of immune dysregulation in IPAF and the uncertain balance of harms and benefits between immunomodulation and antifibrotic therapy for patients with IIP, there is significant interest in determining optimal treatment approaches for these patients [4].

In the absence of predictive markers for treatment response in patients with IPAF, clinicians are forced to extrapolate therapeutic decisions from anecdotal experiences with similar patients and from conditions with similar radiographical or clinical features. For instance, patients with IPAF whose radiographical pattern is predominantly fibrotic such as usual interstitial pneumonia (UIP) [5] or pleuroparenchymal fibroelastosis (PPFE) [6, 7] are more likely to be managed like patients with IPF, whereas IPAF patients with inflammatory radiographic patterns that resemble autoimmune interstitial lung disease (ILD) such as organising pneumonia (OP) or nonspecific interstitial pneumonia (NSIP) [8, 9] are often managed like those with established connective tissue-related ILD. Given the importance of treatment considerations for these patients, it is imperative to better define the distinction across these idiopathic conditions based on intrinsic and meaningful molecular signatures.

Given the diversity of patients within the IIP classification schema, we hypothesised that lung microenvironment protein expression profiles would identify patient-specific clusters that influence clinical outcomes. We tested this hypothesis with an agnostic approach: by performing unsupervised clustering of IIP patients' protein expression profiles followed by an examination of the survival trajectories of these clusters.

Methods

IIP cohort

Patients with IPF and IPAF were collected at a single centre between June 2013 and May 2017 for a prospective research protocol at the Department of Respiratory Medicine and Allergology at the Sapporo Medical University Hospital, Japan [10]. Utilising published European Respiratory Society/American Thoracic Society IPAF research criteria [2], patients were identified from the parent registry by two board-certified pulmonologists who specialise in ILD. Patients were also evaluated upon enrolment in the parent registry by a rheumatologist who excluded the diagnosis of systemic autoimmune disease. Patients with IPF were identified from the same registry based on the most recently published international guidelines [1] *via* the same team of researchers. The Institutional Review Board of Sapporo Medical University Hospital approved this study (no. 342–201, approved on 2 September 2023).

IIP biorepository enrolment procedures

Patients enrolled in this ILD registry underwent standardised collection of physical and laboratory evaluation, pulmonary function testing (PFT) and high-resolution computed tomography (HRCT) at the time of diagnosis and enrolment. Each HRCT was interpreted by two pulmonologists specialising in ILD to determine the predominant HRCT pattern. Patients underwent bronchoalveolar lavage within 3 months of enrolment. All patients included in this subsequent analysis were treatment naive at the time of bronchoalveolar lavage fluid (BALF) sample collection. Survival status was extracted from the medical record for each patient and updated at time of data collection by the parent investigators as of November 2023.

BALF was collected following a standardised protocol with 50 mL of 0.9% sterile saline being instilled into the right middle lobe or lingula *via* a wedged bronchoscope. Lavage fluid was then collected *via* gentle suction and repeated for a total of three lavages (150 mL of 0.9% in total). The collected BALF supernatant was combined, centrifuged to remove cells and frozen at −80°C for future study. Each participant signed an informed consent for serum and BALF samples to be stored and used for future research.

PFTs at baseline (within 3 months of BALF) were extracted from the medical record as a measure of baseline disease severity, including forced vital capacity % predicted (FVC%) and diffusing capacity for carbon monoxide % predicted (D_{LCO}%).

Control subjects

Control subjects without lung disease (n=5) had BALF collected from the University of Kansas Asthma and Airway Translational Research Unit biobank. These samples were collected previously from prospective studies where subjects had research-protocol PFTs and HRCT confirming the absence of ILD and extensive collection of clinical and demographic variables. All subjects underwent informed consent for the parent studies, which included consent for banking samples for future research. BALF was obtained using similar methods as discussed above.

Protein isolation and quantification

BALF samples were thawed and centrifuged at 5000×g for 5 min. A Pierce BCA protein assay (Thermo Fisher Scientific) was used to quantify the concentration of proteins present in BALF following the manufacturer's protocol against a standard curve of BSA from 25–2000 $\mu g \cdot mL^{-1}$.

A 100-µL aliquot of BALF was reduced with TCEP (5 mM) and incubated at 37°C for 30 min. Reduced samples were alkylated with iodoacetamide (10 mM) and incubated in the dark at room temperature for 30 min. Ice cold acetonitrile (ACN) was added to each sample to a volume ratio of 3:1. Samples were incubated at −20°C overnight and were subsequently centrifuged at 14 000×g at 4°C for 30 min to pellet the proteins. The supernatant was removed and the pellet was air dried on the bench top for 10 min. The proteins were resuspended in 50 mM TEAB, pH 8, and digested with trypsin (500 ng) overnight at 37°C with shaking at 500 RPM (Thermomixer, Eppendorf). The digestion was quenched with the addition of formic acid to a final concentration of 1%. Digested samples were stored at −20°C until mass spectrometry analysis. Peptide concentration was measured using a Nanodrop spectrophotometer (Thermo Scientific) at 205 nm prior to liquid chromatography tandem mass spectrometry (LC–MS/MS) analysis.

Samples were injected using the Vanquish Neo (Thermo) nano-UPLC onto a C18 trap column (PepMap™ Neo, 0.3×5 mm, 5 µm C18) prior to elution onto the separation column (PepMap™ Neo, 2 µm, 75 µm×150 mm). Peptides were eluted with a linear gradient (supplementary table S1) from the nano-LC directly interfaced with an Orbitrap Ascend mass spectrometer (Thermo) equipped with a high-field asymmetric waveform ion mobility spectrometry source using three compensation voltages. Details can be found in the supplementary material. Raw files were searched against the human protein database downloaded from Uniprot on 5 May 2023 using SEQUEST in Proteome Discoverer 3.0 [11]. During the search, peptide groups and protein abundances were normalised by applying the normalisation of the total abundance values for each channel across all files, equalising the total abundance between different runs. Details can be found in the supplementary materials. Results were exported to Microsoft Excel for further analysis in R.

Statistical analysis

To delineate lung-based molecular phenotypes and assess their clinical relevance, we first clustered subjects IIP subjects (n=23) based on their protein expression signature. This was carried out using k-means clustering based on the log-transformed expression values of only those proteins that were detected in all 23 IIP subjects. The optimal number of clusters (k) was determined by applying the NbClust methodology [12], as implemented in the NbClust R package, to the matrix of log-transformed protein expression data, using only the proteins that were detected in all 23 IIP subjects. The NbClust framework and corresponding R package provide 30 different clustering indices (*e.g.* silhouette) to assist in the determination of the optimal number of clusters when unsupervised clustering is applied to a dataset. The resulting clusters were next examined to determine their association with various clinical and epidemiological variables, including age, gender, smoking pack-years, FVC%, D_{LCO}%, clinical diagnosis (IPF *versus* IPAF) and radiological ILD pattern (*i.e.* NSIP, OP, PPFE and UIP) and Gender Age Physiology (GAP) index [13, 14]. The association between cluster membership and continuous variables (age, smoking pack-years, FVC% and D_{LCO}%) and categorical variables (radiological patterns, clinical diagnosis and GAP index) were assessed using a series of Kruskal–Wallis and Fisher's exact tests, respectively. We also determined the specific proteins that uniquely discriminated the identified k-means clusters by conducting formal differential protein expression analyses between each pair of the identified clusters. This was carried out by fitting a series of linear regression models, modelling log-transformed expression as the dependent variable against cluster membership as the independent variable, and adjusted for subject age, gender, smoking pack-years, FVC% and subsequent treatment. Models were fit independently to each of the 302 proteins that were observed/detected across all 23 IIP subjects. Proteins

that uniquely and uniformly discriminated the identified clusters were subjected to an over-representation analysis (ORA) to determine biological pathways and Gene Ontology terms that are significantly over-represented with such proteins. The latter was carried out using the enrichGO function in the clusterProfiler Bioconductor package. To assess the potential clinical relevance of the identified k-means clusters, we next examined the association between cluster membership and overall survival using a multivariable Cox proportional hazards model adjusted for age, smoking pack-years, FVC% and clinical diagnosis. Time to death/censoring was calculated as the amount of time elapsed from BALF collection to death or censoring.

Results

The IIP cohort median (interquartile range (IQR)) age was 70 (63–75) years and predominantly male (60.9%) with a median cigarette smoking history of 27 pack-years (IQR 0.25–48 pack-years). Five patients in the IIP cohort had IPF, while 18 met IPAF criteria. The median (IQR) baseline FVC% in the IIP group was 84.5 (74.0–94.1) and the median (IQR) D_{LCO}% was 52.4 (45.3–60.4), which were both lower than the control group (table 1). Over a median (IQR) follow-up of 6.23 (1.81–29.86) years, 11 deaths occurred.

Overall heatmap

To verify that this method could differentiate IIP samples from control samples, we compared BALF protein expression across 23 IIP patients and five healthy control samples. There were 1176 proteins detected at least once in the samples (supplementary table S2). Among these, 45 proteins were considered contaminants and excluded from analysis. A heatmap of a subset of 50 proteins that were most significantly, differentially expressed (lowest p-value) demonstrated the different protein abundance profiles by disease or control (figure 1a). Further, principal-component analysis visibly discriminated between control and IIP samples (figure 1b), with the first two principal components accounting for 38.9% of the variance, indicating a distinct protein expression profile associated with the disease state.

IIP cluster analysis

Our cluster analysis was restricted to only those proteins that were detected in all 23 IIP subjects resulting in a total of 302 proteins in the k-means clustering analysis. The appropriate number of possible unique protein expression clusters across the IIP group as determined by the NbClust function [12] in R was three. A heatmap of the protein abundances by cluster is shown in figure 2. These unsupervised protein cluster memberships (cluster 1 (n=8), cluster 2 (n=5), and cluster 3 (n=10)) were then subjected to a series of Kruskal–Wallis tests to examine the association between cluster membership and selected demographic and clinical characteristics, none of which was statistically significant (smoking pack-years, p=0.86; age, p=0.71; FVC%, p=0.068; and D_{LCO}%, p=0.44) suggesting that these clinical or demographic factors do not primarily drive the clustering of patients in this study (supplementary table S3 and supplementary figure S1).

Association between cluster membership and survival

Compared with cluster 1, which had the worst survival, patients in clusters 2 and 3 were observed to have a reduced hazard of death in models adjusting for age, smoking pack-years and FVC% (hazard ratio (HR) 0.201, 95% CI 0.039–1.044; and HR 0.092, 95% CI 0.017–0.479, respectively) (figure 3). This association

TABLE 1 Patient characteristics

Variable	IIP (n=23)	Controls (n=5)	p-value
Age, years	70 (63–75)	67 (62–72)	0.88
Male	14 (60.9)	3 (60)	0.65
Smoking history, pack-years	27 (0.25–48)	35 (0.45–45)	0.59
FVC%	84.5 (74.0–94.1)	109.2 (105.1–120.3)	0.09
D_{LCO}%	52.4 (45.3–60.4)	65.0 (54.6–76.3)	0.28
GAP index score	2.9±0.92	NA	
Vital status, alive	12 (52.2)	NA	
Time to death, years	3.83 (1.81–29.86)	3.22 (3.17–3.67)	0.70

Data are presented as median (interquartile range), n (%) or mean±SD, unless otherwise stated. IIP: idiopathic interstitial pneumonia; FVC%: forced vital capacity % predicted; D_{LCO}%: diffusion capacity for carbon monoxide % predicted; GAP: Gender Age Physiology index for survival; NA: not available.

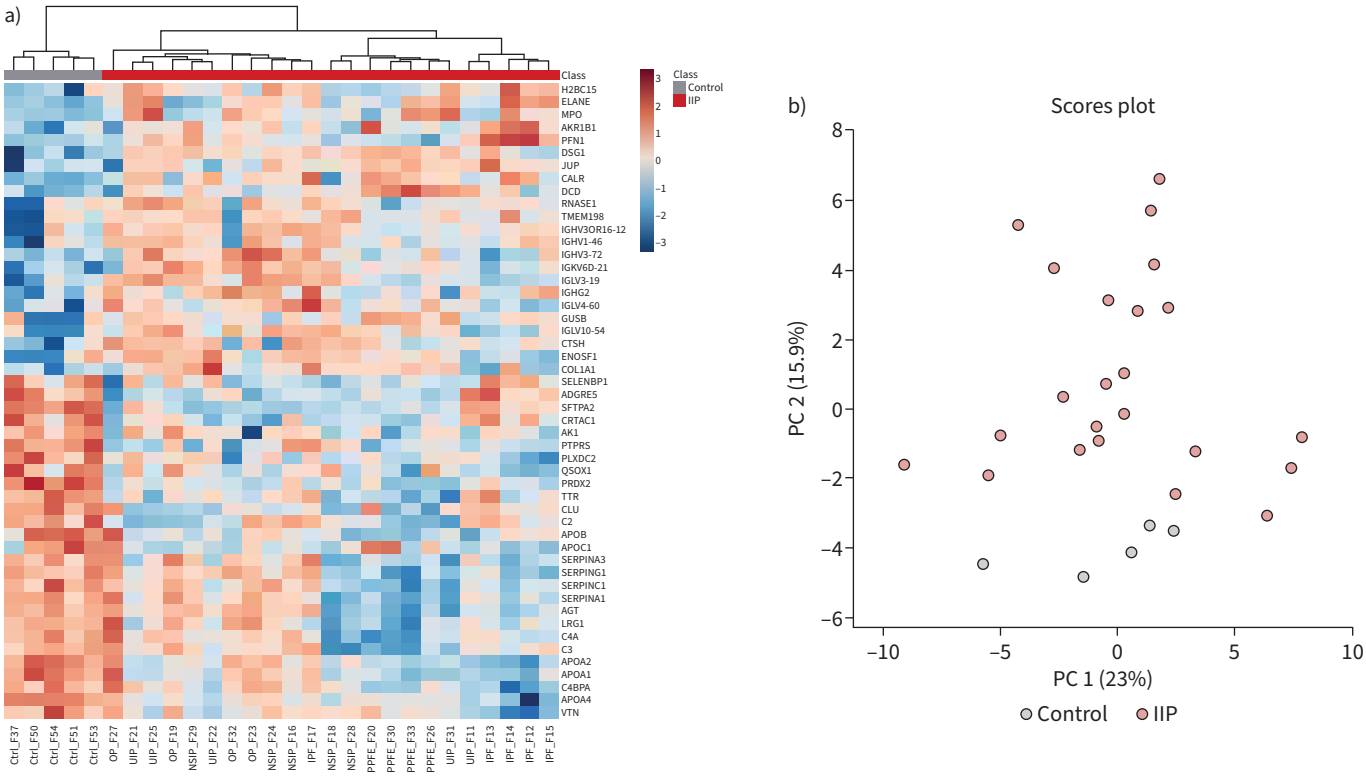

FIGURE 1 a) An abbreviated heatmap of protein expression across controls and idiopathic interstitial pneumonia (IIP) patients. The top 50 differentially expressed proteins defined by having at least two times fold change in either direction and $p<0.05$ for t-tests between IIP and control samples are shown. Grey bars on the top row indicate control samples and red bars indicate IIP patients. The dendrogram above the heatmap indicates two primary distinct branches (controls *versus* IIP). b) Principal-component analysis plot with the same colour coding for each patient bronchoalveolar lavage fluid protein profile (grey indicates controls and red indicates IIP).

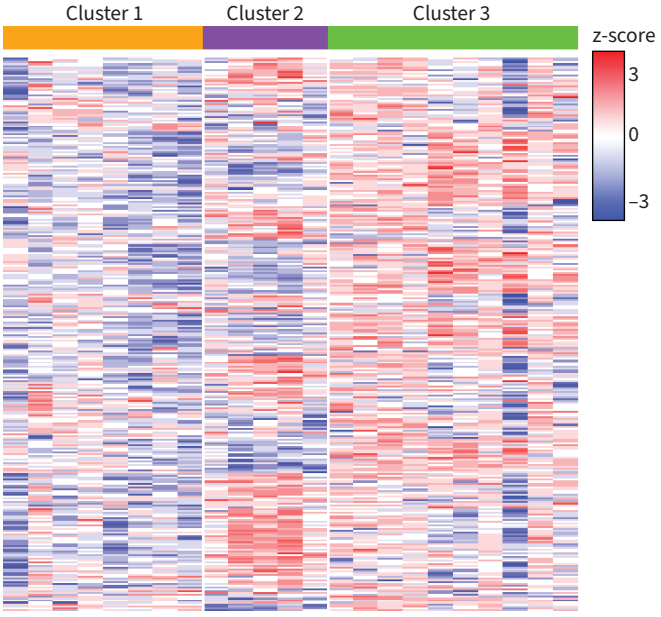

FIGURE 2 Protein expression heatmap of the 302 proteins across the 23 idiopathic interstitial pneumonia patients. k-means cluster analysis and majority rule using the NbClust function in R determined three as the appropriate number of clusters. In this heatmap the patients are represented in columns and colour-coded for cluster designation on the top row: cluster 1 in orange, cluster 2 in purple and cluster 3 in green.

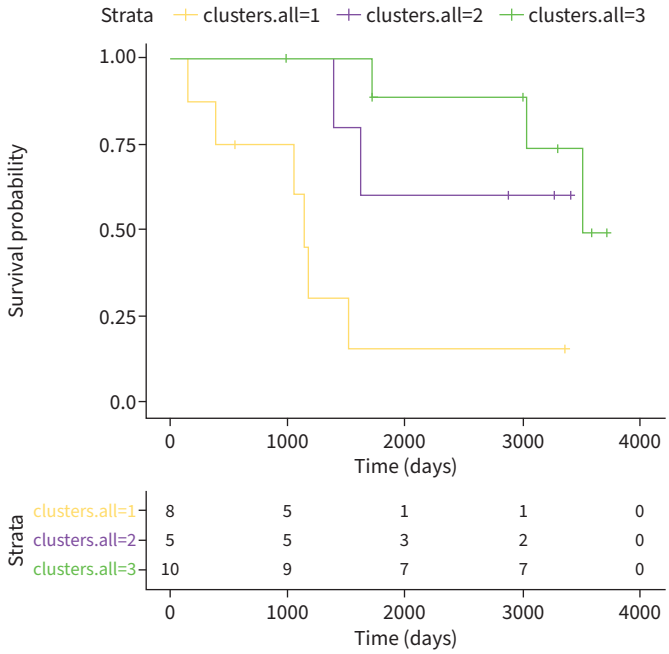

FIGURE 3 Cox proportional hazards analysis for each cluster and time to survival in days plotted on x-axis and survival probability plotted on y-axis. Cluster 1 is orange, cluster 2 is purple and cluster 3 is green.

between survival and cluster designation remained significant when adding subsequent immunosuppression treatment to the model (with pack-years, gender, age and FVC%; p=0.032).

To assess the impact of these novel, unsupervised BALF protein clusters on survival prediction we compared the concordance index (C-index) of cluster membership with different combinations of risk factors. In this analysis, cluster membership alone (C-index 0.79) outperformed models with age (C-index 0.71) or the GAP index (C-index 0.71). When cluster membership and the GAP index were combined in a single model, the C-index increased to 0.875, indicating increased prognostic performance compared with models that include only cluster membership or GAP index (figure 4).

IPF and IPAF radiographic pattern within each cluster

Cluster 1 (worst survival group) included two patients with IPAF-NSIP, all four IPAF-PPFE patients and two patients with IPAF-UIP. Cluster 2 (intermediate survival group) included four patients with IPF and one patient with IPAF-UIP. Cluster 3 (best survival group) included all four IPAF-OP patients, three patients with IPAF-NSIP, two patients with IPAF-UIP and one patient with IPF (figure 5) (supplementary table S4 for clinical and demographic characteristics based on cluster designation).

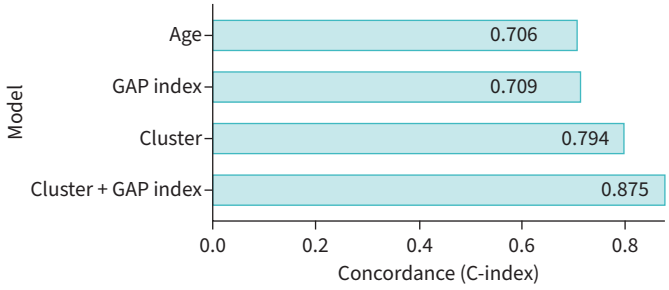

FIGURE 4 Protein cluster (cluster) identity outperforms Gender Age Physiology (GAP) index and age for concordance index (C-index) for survival. When cluster membership and the GAP index were combined in a single model, the C-index increased to 0.875.

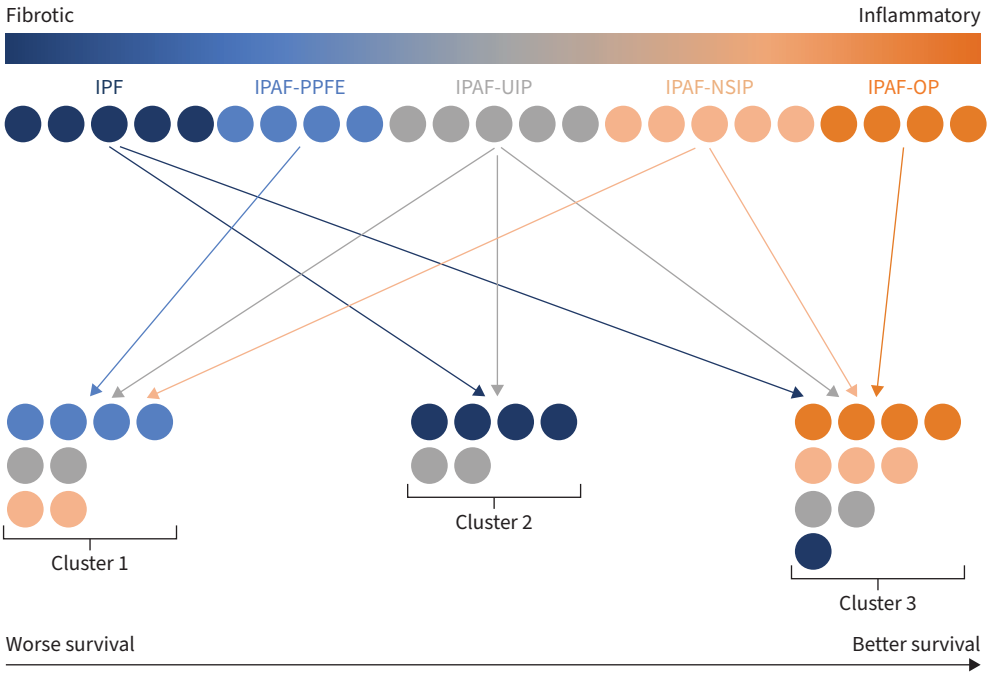

FIGURE 5 Patient diagnoses of idiopathic interstitial pneumonia (IIP) based on radiographical pattern, thought to be more fibrotic (idiopathic pulmonary fibrosis (IPF), interstitial pneumonia with autoimmune features (IPAF) with pleuroparenchymal fibroelastosis (IPAF-PPFE) and usual interstitial pneumonia (IPAF-UIP)) or more inflammatory (IPAF with nonspecific interstitial pneumonia (IPAF-NSIP) and organising pneumonia (IPAF-OP)). The dispersion of each patient in the analysis from IIP radiographic designations based on the unsupervised bronchoalveolar lavage fluid proteomics clustering data is shown by arrows and similar colour scheme.

Uniquely discriminating protein analysis of cluster 1 versus clusters 2 and 3

Given that cluster 1 was associated with the worst survival trajectory, we performed a discrimination analysis to determine proteins that uniquely discriminated cluster 1 from clusters 2 and 3. In this analysis (figure 6), there was a notable linear relationship seen across the top 10 discriminating proteins whose expression levels were decreased in cluster 1 (worst survival) compared with cluster 3 (best survival). These 10 proteins are similarly involved in inflammatory signalling *via* corticosteroid receptor activity (SERPINA6 and SERPINA1) and complement system activation (C3, F2, SERPIND1, SERPINC1, C5, C9, C8B and C8A) [15, 16].

In survival analyses adjusted for age, gender, smoking pack-years, baseline FVC% and IPF diagnosis, the abundance of each of these proteins was found to have an HR for survival <1, suggesting an inverse relationship between their expression and hazard of death (supplementary table S5). Of these 10 proteins, five had significant associations with survival after adjusting for those factors (table 2).

Uniformly discriminating proteins

A total of 19 proteins were observed that uniformly discriminated all three clusters (*i.e.* proteins that exhibited most significant differential expression across all three clusters) (figure 7 and supplementary figure S2). We incorporated these 19 proteins in a pathway analysis using an ORA. Using these 19 uniformly discriminating proteins across all three clusters, the ORA (figure 8) revealed a significant enrichment of pathways involved in the immune response, particularly those associated with immunoglobulin production, B-cell-mediated immunity and lymphocyte-mediated immunity. Pathways related to the production of molecular mediators of the immune response were also over-represented. Notably, adaptive immune responses, including somatic recombination of immune receptors built from immunoglobulin superfamily domains, were prominent, suggesting active engagement of specific and adaptive defence mechanisms ($p<0.004$ for all pathways mentioned).

Discussion

Proteomics is a high-throughput, systems biology approach that can identify and quantify proteins within a given sample, thus providing a comprehensive overview of the cellular and molecular mechanisms at play.

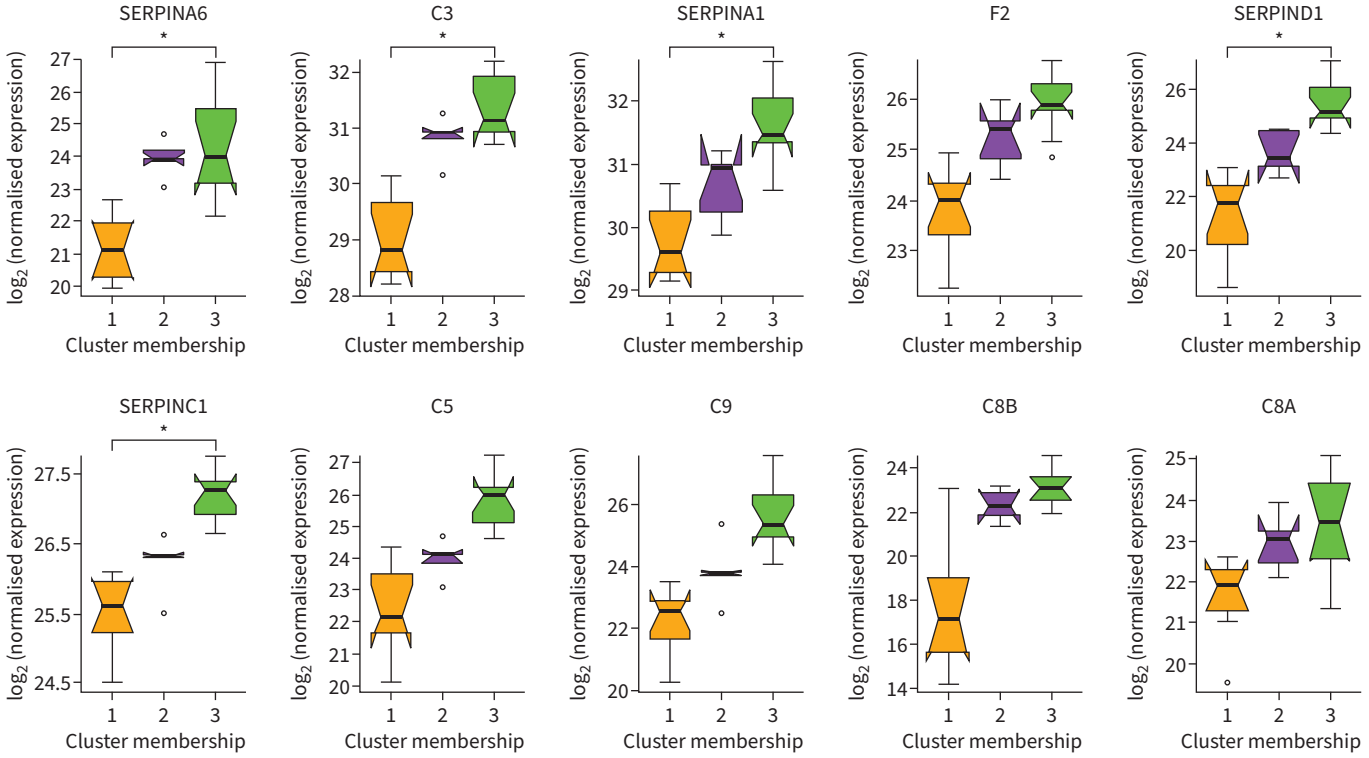

FIGURE 6 Notched box r plots of the top 10 discriminating proteins that increase in expression from worst survival cluster (cluster 1, orange), intermediate survival (cluster 2, purple) and best survival (cluster 3, green). The median normalised abundance is shown by the thick black line with the minimum and maximum indicated by the whiskers. The top and bottom lines of the box indicate the 75th percentile and 25th percentile, respectively. These 10 proteins are similarly involved in inflammatory signalling *via* corticosteroid receptor activity (SERPINA6 and SERPINA1) and complement system activation (C3, F2, SERPIND1, SERPINC1, C5, C9, C8B and C8A) [15, 16]. *: p<0.05.

In the context of lung disease, proteomics enables the characterisation of protein expression profiles that may reflect the underlying pathophysiology of heterogenous conditions such as ILD. In this study, we isolated a cohort of subjects with IPF and IPAF across a diverse array of radiographical patterns with banked BALF that was collected before treatment initiation to perform label-free, quantitative proteomics of the lung microenvironment.

BALF reflects the protein signatures of the diseased lung parenchyma more accurately than serum or plasma and offers a broader representation of lung phenotypes than isolated tissue samples which are subject to heterogeneity and sampling error. BALF has been shown to be the most direct measure of the lung microenvironment compared with other body fluids, such as serum, sputum and nasal lavage fluid [17–19]. BALF has an advantage over surgical lung biopsy due to its inherent safety [20], whereas mortality after surgical lung biopsy approaches 2% in US populations [21]. Moreover, despite the technical challenges associated with bronchoscopy, BALF remains accessible for ongoing clinical research, facilitating the

TABLE 2 Five uniquely discriminating proteins (cluster 1 *versus* clusters 2 and 3) with significant association with survival after adjustment for age, gender, smoking status and baseline severity (forced vital capacity % predicted)

Protein name	Hazard ratio	95% CI	p-value
SERPINA6	0.41	0.21–0.82	0.012
C3	0.34	0.13–0.91	0.031
SERPINA1	0.37	0.14–0.98	0.045
SERPIND1	0.67	0.46–0.97	0.032
SERPINC1	0.38	0.16–0.95	0.039

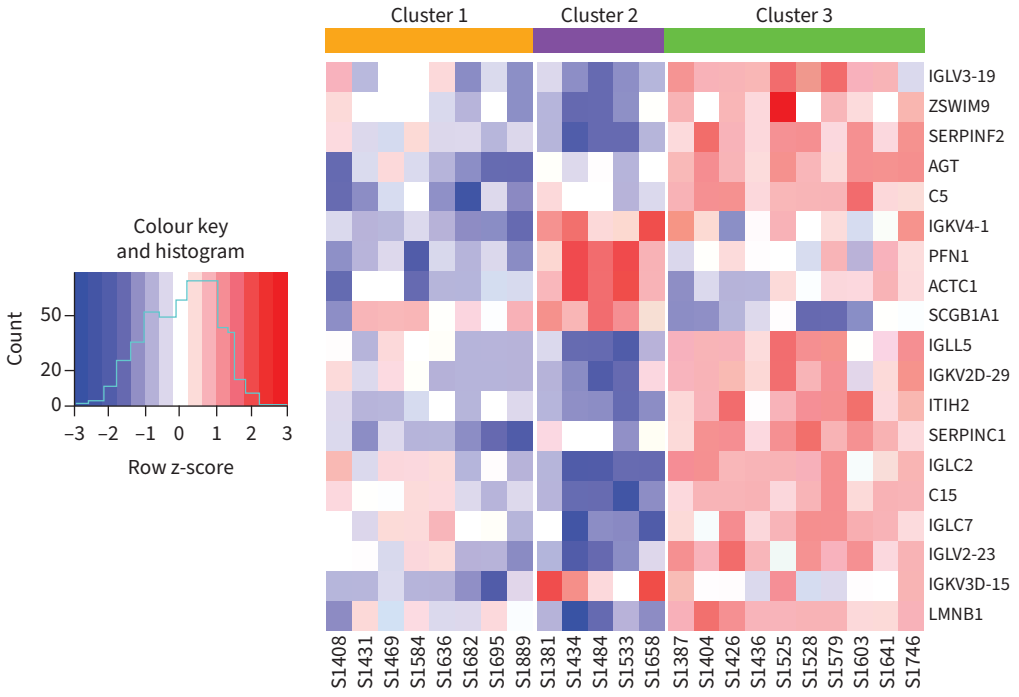

FIGURE 7 Heatmap of proteins uniformly discriminating across three clusters.

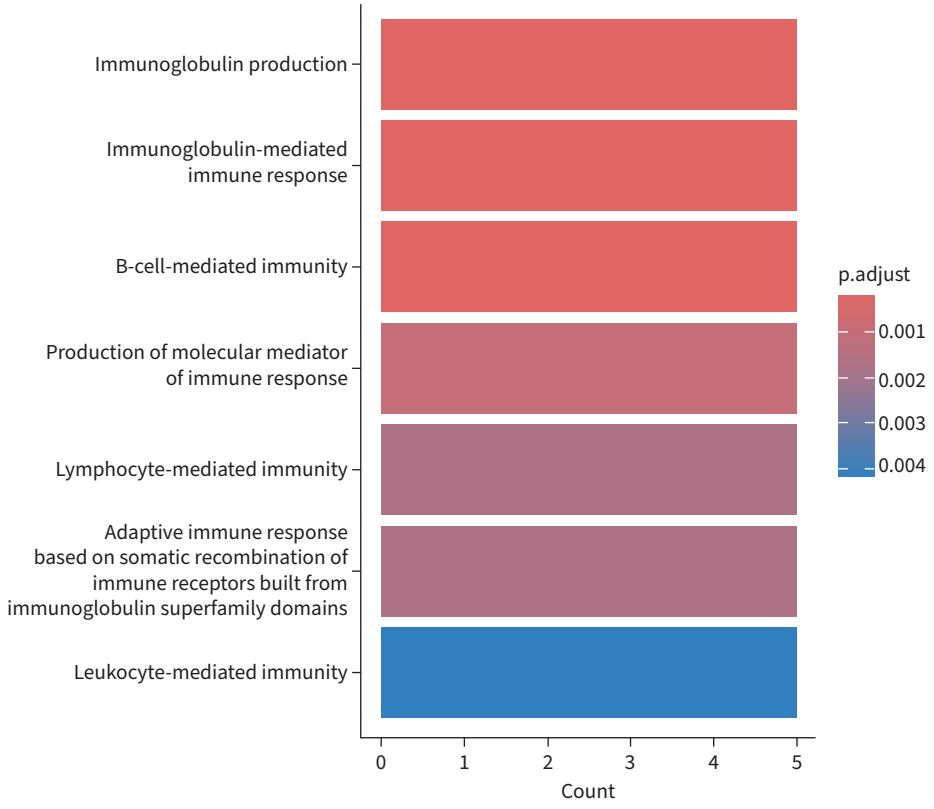

FIGURE 8 Over-representation analysis of proteins uniformly discriminating clusters 1, 2 and 3.

prospective study and validation of molecular phenotypes within the lung microenvironment. Previous studies have explored the BALF proteome in individual ILD disease states, largely focused on IPF [22–28], whereas our analysis was focused on identifying novel proteomics clusters across a group of patients with IIP.

Proteins are a major component of BALF, and those specific to pulmonary pathology such as mucins and surfactant proteins have been found in high amounts in BALF [29, 30]. BALF is known to contain endogenous proteins and peptides that exhibit biological activity specific in the distal lung [31].

In this study, we highlighted a panoramic strategy to leverage the powerful systems biology approach of proteomics to identify novel molecular profiles across a spectrum of IIP, thus enabling us to identify clusters that are driven by lung microenvironment biology as opposed to visual radiographical subtypes and subjective diagnostic criteria.

Despite the small sample size, these unsupervised pre-treatment BALF protein profile clusters largely recapitulated decades of clinical experiences. For instance, all four patients with IPAF and OP clustered together in the “inflammatory” cluster (cluster 3) with best survival. Similarly, four out of five IPF subjects grouped together in cluster 2 and all four patients with IPAF-PPFE were included in the cluster with the worst survival (cluster 1).

The current state of the art for prognostication in ILD such as IPF or IPAF remains a clinical model that combines age, gender and two physiological variables to create a risk score for mortality that has been validated in many forms of ILD (GAP index) [13, 14]. In these current data, unsupervised BALF protein cluster designation outperformed the GAP index and further improved prognostic accuracy when combined with the GAP index. This indicates a potential prognostic role for this approach that calls for further validation in larger datasets and across other ILD types.

The possible clinical utility of this approach is intriguing not only for prognostication, but there are also exciting translational implications for drug discovery and repurposing. For instance, proteins that uniquely discriminated clusters exhibited an intriguing linear relationship between inflammatory signalling and complement activation, mimicking a protective dose–response for survival. The possibility of a linear, continuous marker in BALF, agnostic to radiographical or aetiology that is associated with survival offers the first glimpse for the potential of this method to guide precision-medicine, patient-specific interventional trials.

The potential implications of this approach, if applied broadly, are perhaps most evident when considering the five patients with IPAF-UIP in this analysis. These five IPAF-UIP patients were separated across all three clusters, which led to the hypothesis that while some underlying aetiologies and radiographical patterns might have very similar BALF protein expression patterns (IPAF-PPFE and IPAF-OP), there are some patients with visually indiscernible differences in lung phenotypes that drive their clinical outcomes. The current approach to making decisions on diagnosis and treatment for individual patients with IPAF-UIP is driven largely by practice patterns and unvalidated patient characteristics. There has yet to be a clinical marker that can reliably predict treatment response or survival for these patients. These data highlighted the potential of this approach if expanded and validated across many forms of ILD to identify unique endotypes of ILD for targeted treatment.

There are important limitations to these data. Because of the small sample size, the observed associations between survival trajectory and protein expression clusters remain hypothesis generating until validated in larger, more diverse datasets. It is important to consider how this unsupervised cluster analysis would evolve with increased sample size and with a larger breadth of diagnoses, such as established autoimmune ILD or hypersensitivity pneumonitis. Other considerations that limit broad application of these results relate to the single-centre study design with important limitations to generalisability due to the single ethnicity in our IIP cohort from Japan compared with a healthy US control cohort. This lack of patient diversity limits applicability of the findings, and BALF collection methods may vary from site to site, which could limit the external validity of this approach in future multi-site studies.

There are limitations to the utility of this approach for clinical uptake based on the current lack of clinical indications for bronchoscopy in these parenchymal lung diseases. Further work to validate and extend these observations will need to assess whether less-invasive compartments such as induced sputum, exhaled breath condensate or blood could recapitulate these findings. However, one possibility of validation of these findings in larger, multi-ethnic cohorts would be increased clinical utility of bronchoscopy in these scenarios and future prospective work would need to assess and test the balance of harms and benefits in IIP.

In summary, these current results offer a potential avenue to move diagnostic and treatment choice for individual patients forward through the discovery of lung microenvironment endotypes that drive clinical outcomes. Further studies are required to realise this potential; however, diagnosis and treatment choice for patients with ILD should be guided by authentic molecular markers of disease activity rather than reliant solely on subjective visual and clinical features. These results offer a potential path towards that important effort.

Provenance: Submitted article, peer reviewed.

Ethics statement: The Institutional Review Board of Sapporo Medical University Hospital approved this study (number 342–201, approved on 2 September 2023).

Conflict of interest: L.T. Ngo has nothing to disclose. M.J. Rekowski has nothing to disclose. D.C. Koestler has nothing to disclose. T. Yorozuya has nothing to disclose. A. Saito has nothing to disclose. I. Azeem has nothing to disclose. A. Harrison has nothing to disclose. M.K. Demoruelle reports an investigator-initiated ASPIRE grant (WI227190) from Pfizer to study IIPs since the initial planning of the present study, and research funding from Pfizer and Boehringer Ingelheim for rheumatoid arthritis-associated ILD not related to the present study in the past 36 months. J. Boomer has nothing to disclose. B.R. England has nothing to disclose. P. Wolters reports grants and personal fees from Boehringer Ingelheim and Sanofi, grants from the National Institutes of Health (NIH) and Roche, a contract/grant from Pliant, consulting fees from Blade Therapeutics, and a lecture honorarium from Boehringer Ingelheim, in the past 36 months. P.L. Molyneaux reports grants or contracts from AstraZeneca, GSK, Asthma & Lung UK and Action for Pulmonary Fibrosis; consulting fees from Hoffman-La Roche, Boehringer Ingelheim, AstraZeneca, Trevi, Qureight and Endevour; payment for honoraria for lectures, presentations, speakers' bureaus, manuscript writing or educational events from Boehringer Ingelheim and Hoffman-La Roche; fees for participation on a data safety monitoring board or advisory board from United Therapeutics; stock options in Qureight; and is an associate editor of this journal. M. Castro reports grants or contracts from the NIH, ALA, PCORI, AstraZeneca, GSK, Novartis, Pulmatrix, Sanofi-Aventis and Novartis; royalties from Elsevier; consulting fees from Genentech, Teva, Sanofi-Aventis and Novartis; and payment for honoraria for lectures, presentations, speakers' bureaus, manuscript writing or educational events from AstraZeneca (unbranded, disease state only), Genentech, GSK, Regeneron, Sanofi and Teva, all in the past 36 months. J.S. Lee reports a grant from Boehringer Ingelheim; consulting fees from Blade, Avalyn Pharma, Boehringer Ingelheim, United Therapeutics and Eleven P15; participation on a data safety monitoring or advisory board for the TETON (United Therapeutics) and ATLAS (Avalyn) trials; acting as Senior Medical Advisor to the Pulmonary Fibrosis Foundation; and a research gift from Pliant Therapeutics, all outside the submitted work, all in the past 36 months. J.J. Solomon reports investigator-initiated grants from Boehringer Ingelheim and Pfizer, consulting fees for trial design from Istesso and 3+2 Pharma, payment for disease state education from Boehringer Ingelheim, payment for expert testimony from Cooper Rice and Olsen and Childs McClune, and participation on the "ATTACK my ILD" data safety monitoring board for the University of Pittsburgh, in the past 36 months. K. Koronuma has nothing to disclose. M.P. Washburn has nothing to disclose. S.M. Matson has nothing to disclose.

Support statement: S.M. Matson, L.T. Ngo, D.C. Koestler and this work were supported by NIH grant P20GM130423. The proteomic data were collected in the Mass Spectrometry and Proteomics Core facility utilising the Orbitrap Ascend Tribrid System that was purchased with funds provided by the University of Kansas Cancer Center, which is supported by the National Cancer Institute Cancer Center Support Grant P30 CA168524. Funding information for this article has been deposited with the Crossref Funder Registry.

References

1 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: e44–e68.
2 Fischer A, Antoniou KM, Brown KK, *et al.* An official European Respiratory Society/American Thoracic Society research statement: interstitial pneumonia with autoimmune features. *Eur Respir J* 2015; 46: 976–987.
3 Oldham JM, Adegunsoye A, Valenzi E, *et al.* Characterisation of patients with interstitial pneumonia with autoimmune features. *Eur Respir J* 2016; 47: 1767–1775.
4 Yamano Y, Kataoka K, Takei R, *et al.* Interstitial pneumonia with autoimmune features and histologic usual interstitial pneumonia treated with anti-fibrotic versus immunosuppressive therapy. *Respir Investig* 2023; 61: 297–305.
5 Mackintosh JA, Wells AU, Cottin V, *et al.* Interstitial pneumonia with autoimmune features: challenges and controversies. *Eur Respir Rev* 2021; 30: 210177.
6 Hashisako M, Fukuoka J. Pathology of idiopathic interstitial pneumonias. *Clin Med Insights Circ Respir Pulm Med* 2016; 9: Suppl. 1, 123–133.

7 Gudmundsson E, Zhao A, Mogulkoc N, *et al.* Pleuroparenchymal fibroelastosis in idiopathic pulmonary fibrosis: survival analysis using visual and computer-based computed tomography assessment. *EClinicalMedicine* 2021; 38: 101009.
8 Xu W, Xiao Y, Liu H, *et al.* Nonspecific interstitial pneumonia: clinical associations and outcomes. *BMC Pulm Med* 2014; 14: 175.
9 Tomassetti S, Ryu JH, Piciucchi S, *et al.* Nonspecific interstitial pneumonia: what is the optimal approach to management? *Semin Respir Crit Care Med* 2016; 37: 378–394.
10 Kameda M, Otsuka M, Chiba H, *et al.* CXCL9, CXCL10, and CXCL11; biomarkers of pulmonary inflammation associated with autoimmunity in patients with collagen vascular diseases-associated interstitial lung disease and interstitial pneumonia with autoimmune features. *PLoS One* 2020; 15: e0241719.
11 Orsburn BC. Proteome discoverer – a community enhanced data processing suite for protein informatics. *Proteomes* 2021; 9: 15.
12 Charrad M, Ghazzali N, Boiteau V, *et al.* NbClust: an R package for determining the relevant number of clusters in a data set. *J Stat Soft* 2014; 61: 1–36.
13 Ley B, Ryerson CJ, Vittinghoff E, *et al.* A multidimensional index and staging system for idiopathic pulmonary fibrosis. *Ann Intern Med* 2012; 156: 684.
14 Allen MR, Alevizos MK, Zhang D, *et al.* Performance of GAP and ILD-GAP models in predicting lung transplant or death in interstitial pneumonia with autoimmune features. *Rheumatology (Oxford)* 2023; 63: 1568–1573.
15 Dunkelberger JR, Song W-C. Complement and its role in innate and adaptive immune responses. *Cell Res* 2010; 20: 34–50.
16 Crawford AA, Bankier S, Altmaier E, *et al.* Variation in the SERPINA6/SERPINA1 locus alters morning plasma cortisol, hepatic corticosteroid binding globulin expression, gene expression in peripheral tissues, and risk of cardiovascular disease. *J Hum Genet* 2021; 66: 625–636.
17 Sabounchi-Schütt F, Mikko M, Eklund A, *et al.* Serum protein pattern in sarcoidosis analysed by a proteomics approach. *Sarcoidosis Vasc Diffuse Lung Dis* 2004; 21: 182–190.
18 Gray RD, MacGregor G, Noble D, *et al.* Sputum proteomics in inflammatory and suppurative respiratory diseases. *Am J Respir Crit Care Med* 2008; 178: 444–452.
19 Lindahl M, Ståhlbom B, Tagesson C. Two-dimensional gel electrophoresis of nasal and bronchoalveolar lavage fluids after occupational exposure. *Electrophoresis* 1995; 16: 1199–1204.
20 Molyneaux PL, Smith JJ, Saunders P, *et al.* BAL is safe and well tolerated in individuals with idiopathic pulmonary fibrosis: an analysis of the PROFILE study. *Am J Respir Crit Care Med* 2021; 203: 136–139.
21 Hutchinson JP, Fogarty AW, McKeever TM, *et al.* In-hospital mortality after surgical lung biopsy for interstitial lung disease in the United States. 2000 to 2011. *Am J Respir Crit Care Med* 2016; 193: 1161–1167.
22 Landi C, Bargagli E, Carleo A, *et al.* A system biology study of BALF from patients affected by idiopathic pulmonary fibrosis (IPF) and healthy controls. *Proteomics Clin Appl* 2014; 8: 932–950.
23 Foster MW, Morrison LD, Todd JL, *et al.* Quantitative proteomics of bronchoalveolar lavage fluid in idiopathic pulmonary fibrosis. *J Proteome Res* 2015; 14: 1238–1249.
24 Korfei M, Schmitt S, Ruppert C, *et al.* Comparative proteomic analysis of lung tissue from patients with idiopathic pulmonary fibrosis (IPF) and lung transplant donor lungs. *J Proteome Res* 2011; 10: 2185–2205.
25 Korfei M, von der Beck D, Henneke I, *et al.* Comparative proteome analysis of lung tissue from patients with idiopathic pulmonary fibrosis (IPF), non-specific interstitial pneumonia (NSIP) and organ donors. *J Proteomics* 2013; 85: 109–128.
26 Carleo A, Bargagli E, Landi C, *et al.* Comparative proteomic analysis of bronchoalveolar lavage of familial and sporadic cases of idiopathic pulmonary fibrosis. *J Breath Res* 2016; 10: 026007.
27 Norman KC, O'Dwyer DN, Salisbury ML, *et al.* Identification of a unique temporal signature in blood and BAL associated with IPF progression. *Sci Rep* 2020; 10: 12049.
28 Vantaggiato L, Shaba E, Cameli P, *et al.* BAL proteomic signature of lung adenocarcinoma in IPF patients and its transposition in serum samples for less invasive diagnostic procedures. *Int J Mol Sci* 2023; 24: 925.
29 Cheng G, Ueda T, Numao T, *et al.* Increased levels of surfactant protein A and D in bronchoalveolar lavage fluids in patients with bronchial asthma. *Eur Respir J* 2000; 16: 831–835.
30 Sepper R, Prikk K, Metsis M, *et al.* Mucin5B expression by lung alveolar macrophages is increased in long-term smokers. *J Leukoc Biol* 2012; 92: 319–324.
31 Vento G, Tirone C, Lulli P, *et al.* Bronchoalveolar lavage fluid peptidomics suggests a possible matrix metalloproteinase-3 role in bronchopulmonary dysplasia. *Intensive Care Med* 2009; 35: 2115–2124.