






OPEN ACCESS

Original research

# Cluster analysis of transcriptomic datasets to identify endotypes of idiopathic pulmonary fibrosis

Luke M Kraven ,<sup>1,2</sup> Adam R Taylor,<sup>2</sup> Philip L Molyneaux,<sup>3,4</sup> Toby M Maher,<sup>3,4,5</sup> John E McDonough,<sup>6</sup> Marco Mura ,<sup>7</sup> Ivana V Yang,<sup>8</sup> David A Schwartz,<sup>8</sup> Yong Huang,<sup>9</sup> Imre Noth,<sup>9</sup> Shwu Fan Ma,<sup>9</sup> Astrid J Yeo,<sup>2</sup> William A Fahy,<sup>2</sup> R Gisli Jenkins ,<sup>4,10</sup> Louise V Wain<sup>1,11</sup>

► Additional supplemental material is published online only. To view, please visit the journal online (<http://dx.doi.org/10.1136/thoraxjnl-2021-218563>).

For numbered affiliations see end of article.

## Correspondence to

Dr Louise V Wain, Health Sciences, University of Leicester, Leicester, Leicestershire, UK; [lvw1@le.ac.uk](mailto:lvw1@le.ac.uk)

LMK and ART contributed equally.  
AJY, WAF, RGJ and LVW contributed equally.

Received 8 December 2021  
Accepted 14 April 2022  
Published Online First 9 May 2022

## ABSTRACT

**Background** Considerable clinical heterogeneity in idiopathic pulmonary fibrosis (IPF) suggests the existence of multiple disease endotypes. Identifying these endotypes would improve our understanding of the pathogenesis of IPF and could allow for a biomarker-driven personalised medicine approach. We aimed to identify clinically distinct groups of patients with IPF that could represent distinct disease endotypes.

**Methods** We co-normalised, pooled and clustered three publicly available blood transcriptomic datasets (total 220 IPF cases). We compared clinical traits across clusters and used gene enrichment analysis to identify biological pathways and processes that were over-represented among the genes that were differentially expressed across clusters. A gene-based classifier was developed and validated using three additional independent datasets (total 194 IPF cases).

**Findings** We identified three clusters of patients with IPF with statistically significant differences in lung function ( $p=0.009$ ) and mortality ( $p=0.009$ ) between groups. Gene enrichment analysis implicated mitochondrial homeostasis, apoptosis, cell cycle and innate and adaptive immunity in the pathogenesis underlying these groups. We developed and validated a 13-gene cluster classifier that predicted mortality in IPF (high-risk clusters vs low-risk cluster: HR 4.25, 95% CI 2.14 to 8.46,  $p=3.7\times 10^{-5}$ ).

**Interpretation** We have identified blood gene expression signatures capable of discerning groups of patients with IPF with significant differences in survival. These clusters could be representative of distinct pathophysiological states, which would support the theory of multiple endotypes of IPF. Although more work must be done to confirm the existence of these endotypes, our classifier could be a useful tool in patient stratification and outcome prediction in IPF.

## INTRODUCTION

Idiopathic pulmonary fibrosis (IPF) is a complex, ultimately fatal disease, characterised by progressive scarring of the lungs, with a median survival of 3–5 years postdiagnosis.<sup>1,2</sup> Currently, there is no cure for IPF and the two drugs approved for treatment (nintedanib and pirfenidone) only slow disease progression, do not work in all patients and are often not well tolerated.<sup>3,4</sup> The clinical course of IPF is highly variable with slow progression in some

## WHAT IS ALREADY KNOWN ON THIS TOPIC

- ⇒ The clinical course of idiopathic pulmonary fibrosis (IPF) is highly heterogeneous, which has prompted speculation that the disease may consist of multiple 'endotypes'.
- ⇒ Gene expression profiles could be used to identify these endotypes but previous studies have been limited by sample size, ability to validate and clinical interpretation.

## WHAT THIS STUDY ADDS

- ⇒ By combining and clustering multiple gene expression datasets, we identified three distinct clusters of patients with IPF with significant clinical differences between groups, as well as differences in gene expression that implicated mitochondrial homeostasis, apoptosis, cell cycle and innate and adaptive immunity.
- ⇒ We went on to develop a 13-gene cluster classifier that was able to predict mortality in two validation cohorts of patients with IPF.

## HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE AND/OR POLICY

- ⇒ Our findings support the hypothesis of multiple endotypes of IPF and highlight distinct underlying biological mechanisms that could inform a precision medicine strategy for IPF.

patients, rapid progression in others, while many experience a slowly progressive course interspersed with periods of rapid lung function deterioration.<sup>1</sup> It is plausible that these clinical phenotypes could reflect different disease endotypes.

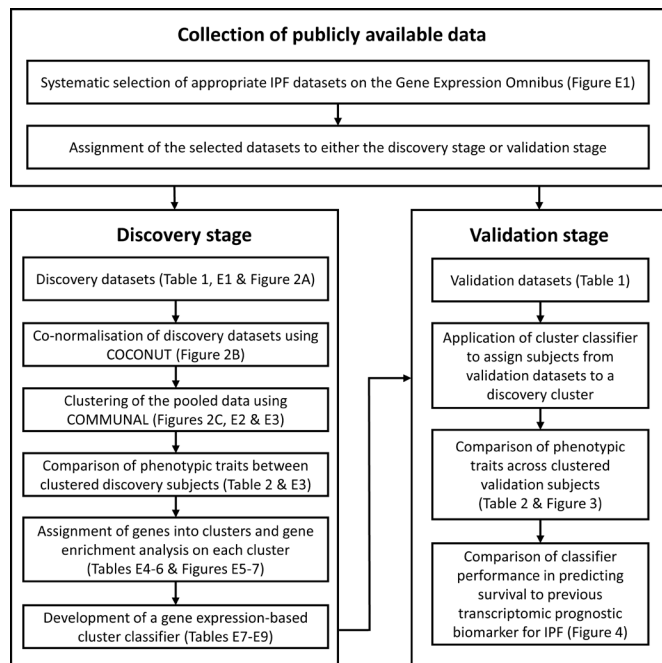
Disease endotypes are subtypes of a disease as defined by a particular pathophysiological mechanism. It has been speculated that distinct endotypes of IPF exist,<sup>5,6</sup> as in asthma and lung cancer,<sup>7,8</sup> although these are not yet well understood. Identification of endotypes would greatly increase our understanding of the behaviour and heterogeneity of the disease, and may allow for the development of biomarkers and more precise, tailored approaches to treatment.

Transcriptomic data can be used to define disease endotypes, as similar transcriptomic profiles in affected individuals may reflect common underlying biological mechanisms. Previous transcriptomic



© Author(s) (or their employer(s)) 2023. Re-use permitted under CC BY. Published by BMJ.

**To cite:** Kraven LM, Taylor AR, Molyneaux PL, et al. *Thorax* 2023;**78**:551–558.



**Figure 1** A flow chart showing the design of our study. COCONUT, Combat CO-Normalisation Using conTrols; COMMUNAL, Combined Mapping of Multiple cUsteriNg ALgorithms; IPF, idiopathic pulmonary fibrosis.

analyses of patients with cancer have been particularly successful in defining clinically significant patient subgroups, which have led to improvements in treatment.<sup>9,10</sup> Previous studies in patients with IPF have used transcriptomic or limited biomarker data with supervised clustering approaches to develop binary signatures predictive of disease progression, measured using mortality or transplant-free survival.<sup>11,12</sup> Studies using unsupervised clustering approaches to discover disease endotypes have been limited by sample size,<sup>13</sup> ability to validate<sup>13,14</sup> and clinical interpretation.<sup>14</sup> However, these studies have consistently reported association of inflammatory genes,<sup>13</sup> in particular those associated with T cell activation<sup>11</sup> and differentiation,<sup>14</sup> with worse outcomes.

In this study, we aimed to conduct the largest unsupervised clustering analysis of available transcriptomic datasets to date, with independent validation, to identify clinically distinct groups of patients with IPF. We hypothesised that these groups could represent individuals with different endotypes of IPF. Rather than undertake single dataset analyses, we co-normalised and pooled multiple datasets together to increase the sample size and enhance statistical power. Additionally, we used classification to develop a method to accurately assign additional individuals with IPF to one of these groups. This classifier displayed the ability to predict survival in IPF and so we then compared the performance of our classifier in independent validation datasets to a previous method of outcome prediction in IPF.

## METHODS

### Collection of publicly available data

The design of our study is shown in [figure 1](#). First, we reviewed the IPF datasets available on the Gene Expression Omnibus<sup>15</sup> and systematically selected several suitable datasets of gene expression data measured from whole blood (see online supplemental file for details). The datasets were then assigned to either the discovery stage or the validation stage (online supplemental file).

Cohorts used in the discovery stage must have included healthy controls to enable the data co-normalisation. The methods used to preprocess the transcriptomic data before the co-normalisation are described in the online supplemental file.

### Discovery stage

As the discovery datasets originated from different studies and the transcriptomic data were collected using varying platforms, there would have been considerable technical (non-biological) differences in gene expression between them. As such, the discovery datasets required adjustment before they could be combined and clustered. We co-normalised the discovery datasets using the Combat CO-Normalisation Using conTrols (COCONUT) method,<sup>16</sup> using R V.4.0.0 and the ‘COCONUT’ package V.1.0.2 (online supplemental file). All healthy control subjects were then removed from further analysis.

We used R V.3.4.0 and the Combined Mapping of Multiple cUsteriNg ALgorithms (COMMUNAL)<sup>17</sup> package V.1.1.0 to identify the optimal number of clusters within the pooled, co-normalised data. COMMUNAL integrates data from multiple clustering algorithms across a range of genes and evaluates the validity of each number of clusters using multiple validity measures. Details on the configuration of COMMUNAL used in this study and the process used to determine the optimal cluster assignment can be found in the online supplemental file. Once an optimal cluster assignment was chosen, principal components analysis and heatmaps were used to visualise the separation of the clusters. Unclustered samples were excluded from further analysis.

Clinical and demographic characteristics of clustered subjects were compared using  $\chi^2$  tests for count data, analysis of variance for non-skewed continuous data, Kruskal-Wallis tests for skewed continuous data and survival analysis methods for time-to-event data (online supplemental file). Gene enrichment analysis was performed in R V.4.0.0 with the in-house ‘metabaser’ package (database V.20.3, package V.4.2.3) to highlight biological mechanisms that were significantly enriched for the subjects in each cluster (online supplemental file).

We developed a gene expression-based classifier to assign new individuals with IPF to one of the clusters using only the most informative differentially expressed genes. This classifier was designed following the approach described by Sweeney *et al* in their study of bacterial sepsis (online supplemental file).<sup>18</sup>

### Validation stage

The classifier was used to assign all subjects with IPF in each validation dataset to a discovery cluster. Phenotypic traits were compared across clusters, as in the discovery stage (online supplemental file).

We compared the classifier’s performance at predicting survival in IPF to a previous transcriptomic prognostic biomarker for IPF by Herazo-Maya *et al*.<sup>19</sup> Each of the validation subjects with survival data available were assigned into a ‘high-risk’ or ‘low-risk’ group (in terms of mortality or requiring a lung transplant) using the method described by Herazo-Maya *et al*, the Scoring Algorithm for Molecular Subphenotypes (SAMS). For this we used as many of the genes in their signature as were present in the validation datasets. Similarly, each subject was assigned into one of our discovery clusters, which were each classed as low risk/high risk based on the discovery stage findings. Survival analysis methods were used to determine which method performed best at predicting survival (online supplemental file).

**Table 1** Summary information on the publicly available datasets that were included in this study, as well as summary statistics for all individuals whose data were included in the analysis.

	Discovery stage						Validation stage		
	GSE38958		GSE33566		GSE93606		GSE132607	GSE27957	GSE28042
GEO accession number	GSE38958		GSE33566		GSE93606		GSE132607	GSE27957	GSE28042
Reference	Huang <i>et al</i> <sup>34</sup>		Yang <i>et al</i> <sup>35</sup>		Molyneaux <i>et al</i> <sup>36</sup>		*	†11	†11
Country	USA		USA		UK		USA	USA	USA
Disease status	IPF	Control	IPF	Control	IPF	Control	IPF	IPF	IPF
Sample size	70	45	93	30	57	20	74	45	75
Age (years, SD)	68.2 (7.2)	69.3 (9.3)	67.2 (11.4)	62.4 (14.3)	67.4 (8.0)	66.0 (10.6)	66.6 (7.6)	67.1 (8.2)	68.9 (8.1)
Sex (% male)	82.6%	60.0%	65.6%	46.7%	66.7%	60.0%	70.3%	88.9%	69.3%
Ancestry (% European)	82.8%	71.1%	Unknown	Unknown	Unknown	Unknown	94.6%	82.2%	97.3%
FVC % predicted (SD)	62.4 (15.0)	Unknown	62.0 (28.8)	Unknown	72.2 (20.3)	Unknown	69.7 (18.4)	60.6 (14.3)	65.4 (16.7)
DL <sub>co</sub> % predicted (SD)	43.3 (18.7)	Unknown	52.1 (27.9)	Unknown	39.2 (14.1)	Unknown	45.6 (15.4)	43.4 (17.7)	48.9 (18.6)
Mortality (%)	Unknown	Unknown	Unknown	Unknown	40.4%	Unknown	Unknown	37.8%	32.0%
MUC5B genotype (% GG)	Unknown	Unknown	28.0%	53.8%	40.0%	Unknown	18.8%	Unknown	Unknown
MUC5B genotype (% GT)	Unknown	Unknown	66.0%	42.3%	50.0%	Unknown	78.1%	Unknown	Unknown
MUC5B genotype (% TT)	Unknown	Unknown	6.0%	3.8%	10.0%	Unknown	3.1%	Unknown	Unknown
Immunosuppressive therapy (%)	Unknown	Unknown	0.0%	Unknown	0.0%	Unknown	Unknown	4.4%	14.7%

\*As of March 2022, the dataset with GEO accession number GSE132607 had not been associated with any published study.

†The datasets with GEO accession numbers GSE27957 and GSE28042 originated from the same study,<sup>11</sup> where the data in GSE27957 were used in discovery and the data in GSE28042 were used as independent validation data.

DL<sub>co</sub>, diffusing capacity of lung for carbon monoxide; GEO, Gene Expression Omnibus; MUC5B genotype, genotype for the MUC5B promoter polymorphism rs35705950.

## RESULTS

### Collection of publicly available data

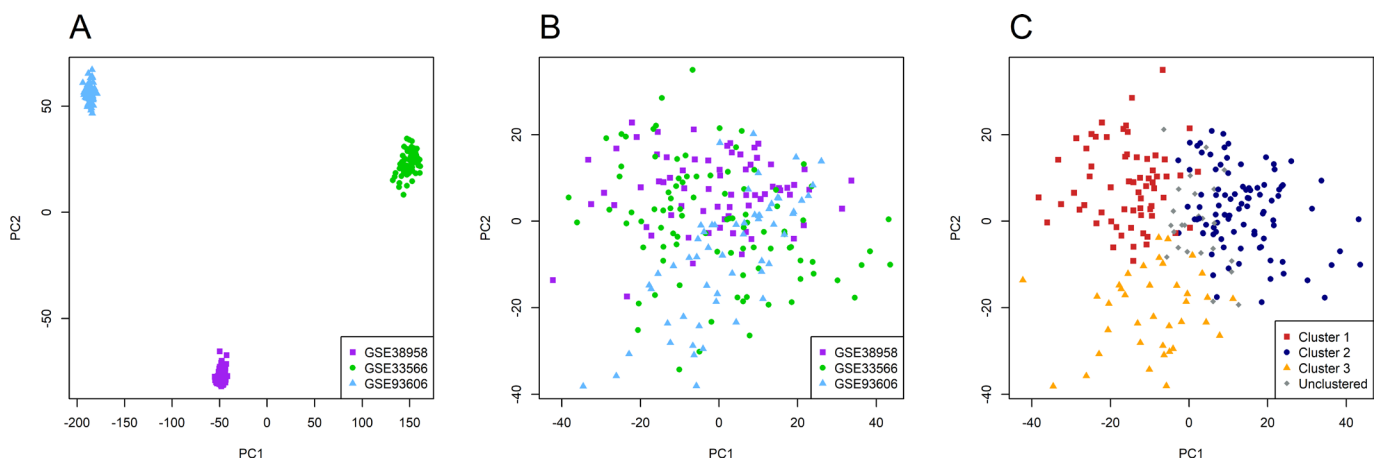
Six independent whole blood gene expression datasets were selected for inclusion in the analysis (online supplemental figure E1). Summary statistics for all subjects are shown in table 1.

### Discovery stage

All three discovery stage datasets were microarray-based (online supplemental table E1). There were expression levels measured for 9371 common genes across the three datasets, which consisted of a total of 220 subjects with IPF and 95 healthy control subjects. There were no significant differences in age or sex between healthy controls across the three studies (online supplemental table E2).

Prior to COCONUT co-normalisation, the data from the three cohorts were entirely separated in high-dimensional space due to technical differences between the studies (figure 2A). Whereas after COCONUT (figure 2B), the data were overlapping in high-dimensional space, indicating that the technical differences between datasets had been reduced and that the co-normalised data were suitable for clustering.

COMMUNAL was run on the co-normalised data and the resulting optimality map is shown in online supplemental figure E2. The clustering assignment with 3 clusters using 2500 genes was chosen as the optimal assignment (online supplemental file), with 64 subjects assigned to cluster 1, 95 assigned to cluster 2, 37 assigned to cluster 3 and 24 (10.4%) that were unclustered (figure 2C and online supplemental figure E3).



**Figure 2** Plots of the first two principal components of the gene expression data for the idiopathic pulmonary fibrosis samples prior to co-normalisation and stratified by original study (A), post co-normalisation and stratified by original study (B) and post co-normalisation stratified by cluster (C). The x-axis represents the first principal component of the data and the y-axis represents the second principal component of the data.

**Table 2** Comparison of clinical and demographic traits of clustered subjects in the discovery and validation stages

	Discovery stage (n=196)					Validation stage (n=194)				
	Cluster 1	Cluster 2	Cluster 3	P value	N used	Cluster 1	Cluster 2	Cluster 3	P value	N used
n subjects in cluster	64	95	37			52	101	41		
Age (years) (mean, SD)	67.8 (8.9)	66.9 (10.2)	68.8 (9.4)	0.592	188	67.1 (8.1)	68.5 (7.6)	66.2 (8.6)	0.239	194
Male (%)	52 (81.3%)	66 (69.5%)	23 (62.2%)	0.091	196	38 (73.1%)	72 (71.3%)	34 (82.9%)	0.347	194
European ancestry (%)	17 (81.0%)	29 (82.9%)	3 (75.0%)	0.883	60	51 (98.1%)	91 (90.1%)	38 (92.7%)	0.196	194
Ever smoker (%)	NA	15 (62.5%)	18 (78.3%)	0.389	47	11 (57.9%)	21 (60.0%)	17 (85.0%)	0.114	74
Death observed during study (%)	NA	6 (25.0%)	16 (66.7%)	<b>0.009</b>	48	16 (48.5%)	13 (19.7%)	12 (57.1%)	<b>0.001</b>	120
FVC % predicted (median, IQR)	63.0 (35.0)	70.5 (30.1)	60.1 (23.4)	0.342	154	64.3 (23.6)	65.0 (24.3)	63.1 (15.3)	0.467	193
DL <sub>CO</sub> % predicted (median, IQR)	35.0 (30.0)	45.0 (29.2)	34.4 (17.3)	<b>0.009</b>	133	42.1 (26.4)	48.2 (21.1)	43.4 (20.3)	0.069	194
FEV <sub>1</sub> % predicted (median, IQR)	NA	74.9 (23.1)	65.4 (22.7)	0.216	48	74.8 (21.7)	75.2 (22.2)	75.4 (17.7)	0.913	75
GAP index (mean, SD)	4.9 (1.4)	3.9 (1.5)	4.4 (1.7)	<b>0.006</b>	132	4.1 (1.6)	4.0 (1.5)	4.3 (1.5)	0.753	193
MUC5B genotype: GG (%)	5 (29.4%)	11 (27.5%)	14 (51.9%)	0.230	84	2 (11.8%)	6 (19.4%)	4 (25.0%)	0.780	64
MUC5B genotype: GT (%)	10 (58.8%)	26 (65.0%)	10 (37.0%)			14 (82.4%)	24 (77.4%)	12 (75.0%)		
MUC5B genotype: TT (%)	2 (11.8%)	3 (7.5%)	3 (11.1%)			1 (5.9%)	1 (3.2%)	0 (0%)		

Data are presented as count (percentage), mean (SD) or median (IQR). GAP index, Gender, age and physiology index for IPF mortality.<sup>20</sup> P value for count data is from a  $\chi^2$  test, test comparing means is analysis of variance and test comparing medians is the Kruskal-Wallis log rank test. Significant p values (p<0.05) are highlighted in bold. For percentages, the denominator was the number of participants in that cluster with non-missing data for that trait.

DL<sub>CO</sub>, diffusing capacity for carbon monoxide; FEV<sub>1</sub>, forced expiratory volume in 1 second; FVC, forced vital capacity; IPF, idiopathic pulmonary fibrosis; MUC5B genotype, genotype for the MUC5B promoter polymorphism rs35705950; NA, data not available.

With all studies combined and unclustered individuals removed (table 2), there was a statistically significant difference in average predicted diffusing capacity of the lung for carbon monoxide (DL<sub>CO</sub>) across clusters (p=0.009). Subjects in cluster 1 had a similar median predicted DL<sub>CO</sub> to those in cluster 3, whilst subjects in cluster 2 had the greatest median predicted DL<sub>CO</sub>, indicating that these individuals had relatively preserved lung function. Additionally, there was a significant difference in average score from the gender, age and physiology (GAP) index for IPF mortality (p=0.006),<sup>20</sup> with those in cluster 1 having the greatest GAP score and those in cluster 2 having the lowest average GAP score. There was a statistically significant difference in mortality between clusters 2 and 3, with death observed for 25% of subjects in cluster 2 and 67% of subjects in cluster 3 (p=0.009). Furthermore, those in cluster 3 had consistently poorer survival over time than those in cluster 2 (online supplemental figure E4). A Cox proportional hazards (PH) model estimated that the HR between clusters 2 and 3 was 3.59 (95% CI 1.40 to 9.19, p=0.008), and so at any follow-up time, subjects in cluster 3 were estimated to be 3.59 times as likely to die as subjects in cluster 2. The clinical and demographic traits of the subjects in each cluster stratified by original study are shown in online supplemental table E3.

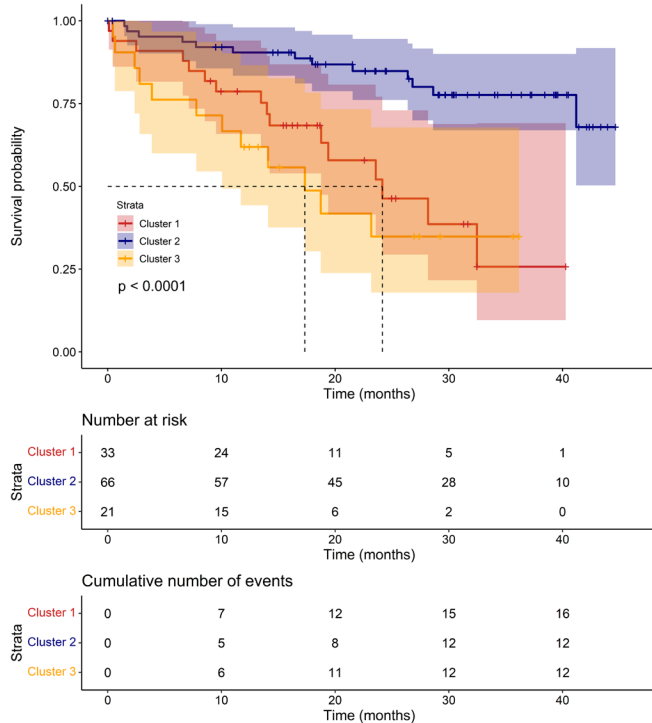
Gene enrichment analysis revealed that cluster 1 was significantly enriched for biological mechanisms relating to metabolic changes, including electron transport and cellular respiration (online supplemental table E4 and figure E5). Cluster 2 was significantly enriched for biological processes and pathways relating to gene regulation, DNA repair, cell cycle and apoptosis (online supplemental table E5 and figure E6), while cluster 3 was significantly enriched for terms relating to the immune response (online supplemental table E6 and figure E7). In addition, the genes assigned to clusters 2 and 3 were each found to be statistically overconnected (in terms of direct gene regulation) to a significant number of genes that have been previously implicated in the development of IPF (see the ‘Gene enrichment analysis’ section in the online supplemental file for more details).

We used the pooled, co-normalised gene expression data for all 196 subjects who were successfully clustered in the discovery analysis to train a gene expression-based cluster classifier (online supplemental file). The classifier (online supplemental tables E7 and E8) used expression data from 13 genes and was able to accurately reassign 99.0% of discovery subjects (online supplemental table E9).

### Validation stage

There were 194 subjects with IPF across the three validation cohorts. Expression levels for all 13 genes used in the classifier were available in all three validation cohorts. We used the classifier to assign each individual to a cluster and compared phenotypic traits across clusters (table 2). As in the discovery stage, there were statistically significant differences in mortality between clusters (p=0.001) and those in cluster 2 had the best survival over time (figure 3). Additionally, individuals in cluster 2 had the highest average DL<sub>CO</sub>, although the difference in DL<sub>CO</sub> between validation clusters was not statistically significant (p=0.069). Cox PH models (online supplemental table E10) estimated that at any follow-up time, an individual in cluster 1 was 3.80 times more likely to die than an individual in cluster 2 (95% CI 1.78 to 8.12, p=0.001), while an individual in cluster 3 was 5.05 times more likely to die than an individual in cluster 2 (95% CI 2.24 to 11.35, p=9.1×10<sup>-5</sup>). However, the difference in survival over time between clusters 1 and 3 was not statistically significant (HR 1.47 (95% CI 0.67 to 3.22, p=0.341).

Finally, we compared the performance of our classifier at predicting survival in IPF with SAMS, a method used by Herazo-Maya *et al* to predict outcome in IPF using a 52-gene signature.<sup>19</sup> There were no common genes between the classifier and the 52-gene signature, although many were highly correlated in the validation subjects (online supplemental figure E8). The subjects in the GSE27957 and GSE28042 validation cohorts (GSE132607 did not report mortality) were each classed as ‘high risk’ or ‘low risk’ using both gene expression-based methods.



**Figure 3** A Kaplan-Meier plot showing survival over time for the clustered validation subjects. The p value shown on the plot is from a log-rank test testing the three curves for equality. Median survival in each cluster is shown by dotted lines, where possible.

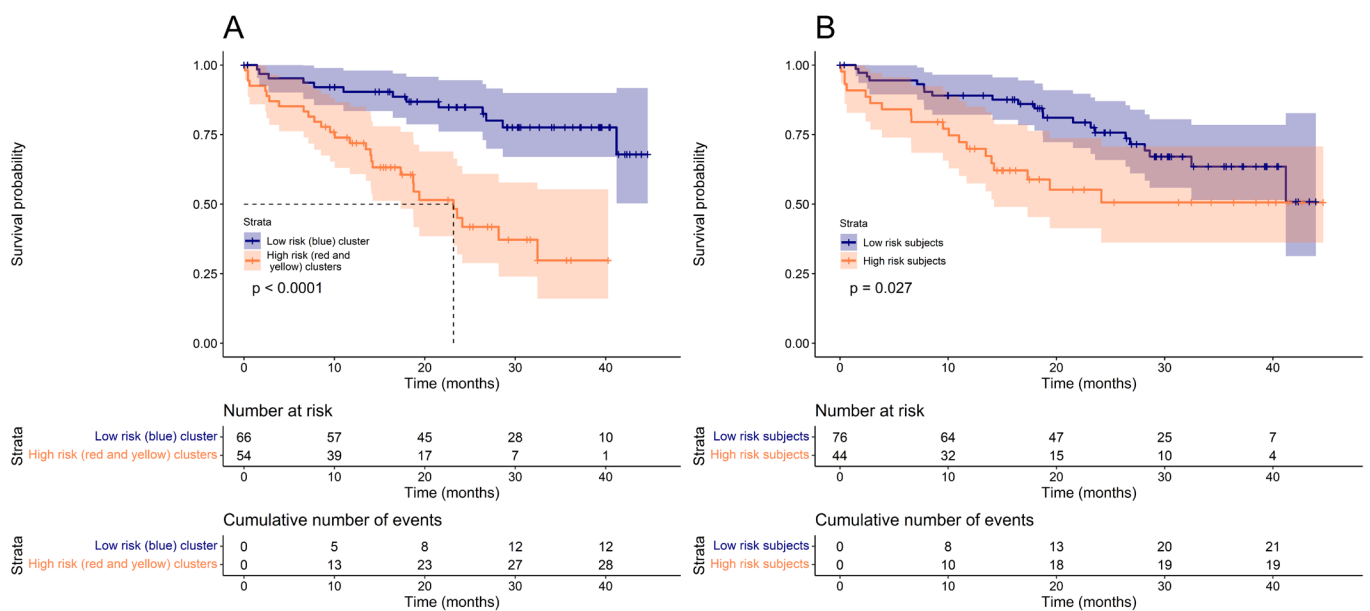
As clusters 1 and 3 were not significantly distinct in terms of survival, both clusters were considered equally ‘high risk’ for the assignments based on the 13-gene classifier. Fifty-one out of 52 (98.1%) genes in the gene signature by Herazo-Maya *et al* were present in the GSE27957 dataset and 50/52 (96.2%) were available in the GSE28042 dataset. Overall, there was 68.3% agreement between the two methods (online supplemental table E11).

Our classifier performed well at predicting survival (figure 4A, E9A and E9C), with the subjects in the ‘high-risk’ clusters having far poorer survival over time than those in the ‘low-risk’ cluster. A univariate Cox PH model estimated that at any follow-up time, an individual in a high-risk cluster was 4.25 times more likely to die than an individual in the low-risk cluster (95% CI 2.14 to 8.46,  $p=3.7 \times 10^{-5}$ ). This model had a C-index (the equivalent of the area under the curve for a receiver operating characteristic curve) of 0.664 (95% CI 0.590 to 0.737). SAMS (figure 4B, E9B and E9D) performed less well, with a Cox PH model estimating that at any time, those in the high-risk group were 1.98 times as likely to die than those in the low-risk group (95% CI 1.07 to 3.68,  $p=0.030$ ) and a C-index of 0.609 (95% CI 0.531 to 0.686).

The risk predictions made using the classifier remained statistically significant ( $p=0.007$ ) after adjusting for age, sex, ancestry, FVC and DL<sub>CO</sub> (online supplemental table E12), with an HR of 2.70 between the high-risk and low-risk clusters (95% CI 1.32 to 5.53). This model had a C-index of 0.773 (95% CI 0.697 to 0.848), which was greater than that of the Cox model containing only age, sex, ancestry, FVC and DL<sub>CO</sub> (C-index=0.747, 95% CI 0.670 to 0.825), suggesting an improvement in predictive ability. A likelihood ratio test between the two models gave a p value of 0.005, suggesting that the improvement in predictive ability when including the classifier’s risk predictions was statistically significant. The multivariate Cox model containing SAMS’ risk predictions had a C-index of 0.760 (95% CI 0.684 to 0.837), which suggested an improvement over the Cox model containing only age, sex, ancestry, FVC and DL<sub>CO</sub>, although the likelihood ratio test p value between these two models was not statistically significant ( $p=0.105$ ).

**DISCUSSION**

By applying new statistical methods for data co-normalisation and machine learning to multiple publicly available datasets, we identified three clusters of patients with IPF with statistically significant differences in lung function and survival. As the



**Figure 4** Survival over time for the subjects with IPF in GSE27957 and GSE28042, stratified by risk group according to our 13 gene classifier (A) and SAMS method by Herazo-Maya *et al* (B). The p value on each plot is from a log-rank test testing the two curves for equality. A dotted line on the plot indicates the median survival time for the risk group if this could be calculated.

clustering in this study was undertaken independently of clinical data, yet significant differences in clinical traits were observed between clusters, this suggests that they may be representative of distinct and clinically relevant endotypes of IPF.

In this study, we used datasets in which the gene expression had been measured from whole blood samples. However, as IPF is a lung disease, characterised by damage to the alveolar epithelium, patterns of gene expression identified in blood may not reflect the underlying pathology of the disease and may instead reflect downstream effects or the presence of confounders, such as secondary infections or treatment effects. Nonetheless, blood is more accessible than a lung-specific tissue/cell type and the expression of a gene in blood is often a significant predictor of the expression of that gene in lung.<sup>21</sup> Furthermore, the blood expression datasets available on GEO provided a larger sample size and more comprehensive accompanying clinical data than lung-specific tissue types, which allowed us to identify statistically significant clinical differences between clusters. In addition, this allowed us to develop a blood-based classifier, which has more clinical utility than one that requires measurements from lung, as this would require more invasive sample collection.

The genes that were most differently expressed in subjects in cluster 1 were significantly enriched for biological mechanisms related to electron transport and cellular respiration. Recent findings appear to suggest that metabolic dysregulation could be a contributing factor to fibrosis, although its role is not yet fully understood.<sup>22, 23</sup> The genes in cluster 1 were also significantly enriched for pathways related to transforming growth factor- $\beta$  signalling, which is a central mediator of fibrosis.<sup>24–26</sup>

Among the biological pathways that were significantly enriched for cluster 2 were pathways related to apoptosis and cell cycle. It has been previously reported that apoptosis is increased in alveolar epithelial cells of patients with IPF but decreased in myofibroblasts,<sup>27</sup> with this imbalance contributing to IPF pathogenesis.<sup>28</sup> Furthermore, the use of therapies that can selectively manipulate apoptosis have been proposed.<sup>29</sup> Additionally, genetic variants within cell cycle genes have been shown to be associated with IPF development and progression.<sup>30</sup> The results for this cluster could further support the idea that apoptosis and cell cycle each play an important role in the pathology of IPF.

Cluster 3 was enriched for terms related to the immune system response. The role of the immune system in IPF has been controversial in the past; failed immunomodulatory therapies in IPF, some of which have led to worse outcomes, have led to speculation that certain immune responses are protective while others are harmful.<sup>31, 32</sup> An improved understanding of immune-driven endotypes could inform novel treatment approaches.

The 13-gene expression-based cluster classifier was able to assign the subjects with IPF from the validation datasets to clusters with statistically significant differences in survival between clusters 2 and 3 ( $p=9.1 \times 10^{-5}$ ), which was consistent with the findings in the discovery stage ( $p=0.008$ ). In addition, while survival information was not directly available for the individuals in cluster 1 in the discovery stage, the significantly low average DL<sub>CO</sub> and high average GAP score for the individuals in that cluster is consistent with the poor survival that was observed for cluster 1 in the validation stage. As the classifier appears to have the ability to assign subjects who are at a lower risk of death into cluster 2 and the subjects who are at a greater risk of death into the other two clusters, it could potentially be used to predict survival in IPF.

The performance of the classifier in predicting survival was compared with SAMS, a similar approach to outcome prediction in IPF.<sup>19</sup> Despite using data from one-quarter of the number

of genes used for SAMS, the differences in survival over time observed between the risk groups in the two validation datasets had greater statistical significance and effect size when predictions were made using the classifier. Additionally, including the classifier's predictions in a survival model that adjusted for important covariate factors led to a statistically significant increase in predictive ability.

One of the main strengths of this study was that the utilisation of a new statistical approach to co-normalisation (COCONUT) allowed for three datasets to be combined,<sup>16</sup> resulting in one of the largest transcriptomic studies in IPF to date with a total of 414 IPF cases across the discovery and validation stages. Another strength of our study was that the application of COMMUNAL, which considered two different clustering algorithms and tested five validity measures over a range of genes, meant that our clustering was more reliable and more likely to be reproducible than the standard approach, which would have been to apply one clustering algorithm and test one validity measure.

There were several limitations to this study. First, as we relied on the use of publicly available data, some clinical variables were relatively underpowered due to missingness within the data or having not been reported in all studies. In particular, survival information was only available in one of the three discovery cohorts and two of the three validation cohorts, which may have limited our ability to clinically distinguish clusters 1 and 3 in terms of survival. In addition, we lacked detailed data for clinically significant traits such as patient reported outcomes, lung function decline over time and the incidence rate of acute exacerbations. Additionally, we did not possess information regarding the background therapy of the subjects with IPF. However, for the three cohorts with survival data available, we were able to glean from the original papers that the patients with IPF were either treatment-naïve populations (GSE93606) or that there were only a small proportion that were receiving immunosuppressive therapy at the time of the blood collection (GSE27957 and GSE28042). In addition, these populations were not given anti-fibrotics and so treatment effects are unlikely to have been driving the large differences in survival that were observed between clusters. These limitations highlight the need for a single large prospective study on this topic with more comprehensive phenotyping.

A further weakness of our study is that each participating cohort of subjects with IPF was subject to survival bias, as only subjects who survived long enough to enrol into each study could have contributed their transcriptomic data to it. This could have restricted the level of heterogeneity of IPF that we were able to capture in the study and limited the generalisability of our findings.

Additionally, COCONUT makes the assumption that the healthy controls across the different studies came from the same statistical distribution and so all differences between healthy controls across studies must have been due to non-biological variation. This means that any large differences in confounding factors between the groups of healthy controls would have restricted the efficacy of the co-normalisation. However, there were no significant differences in age ( $p=0.187$ ) or sex ( $p=0.477$ ) between the healthy controls across the three studies.

If the clusters identified in this study do truly represent endotypes of IPF, it may be worth speculating about the nature of these endotypes. As IPF is a complex disease, with many known common genetic and environmental exposures, it is unlikely that it would behave under a traditional discrete endotype model and instead more likely that it would behave under a more complex model, such as the palette model described by McCarthy.<sup>33</sup>

Our gene enrichment analysis results could implicate metabolic changes and the immune system response as being among the component pathways for IPF.

To conclude, these results could support the hypothesis of multiple endotypes of IPF as there appear to be at least two clinically distinct groups of patients with IPF that can be identified through cluster analysis of transcriptomic data. As these clusters were defined using expression from groups of genes that were significantly enriched for many different biological pathways and processes, they could be representative of distinct pathophysiological states. Additionally, a classifier with the ability to assign additional individuals with IPF to one of the clusters was developed. With further development, this classifier could be a useful tool in outcome prediction in IPF as well as helping us gain a better understanding of the underlying biological processes that may be driving the observed differences in survival.

#### Author affiliations

<sup>1</sup>Department of Health Sciences, University of Leicester, Leicester, UK

<sup>2</sup>Research & Development, GlaxoSmithKline, Stevenage, UK

<sup>3</sup>Guy's and St Thomas' NHS Foundation Trust, Royal Brompton and Harefield Hospitals, London, UK

<sup>4</sup>National Heart and Lung Institute, Imperial College London, London, UK

<sup>5</sup>Keck School of Medicine, University of Southern California, Los Angeles, California, USA

<sup>6</sup>Division of Pulmonary, Critical Care & Sleep Medicine, Yale School of Medicine, New Haven, Connecticut, USA

<sup>7</sup>Division of Respiratory, Western University, London, Ontario, Canada

<sup>8</sup>Department of Medicine, University of Colorado, Denver, Colorado, USA

<sup>9</sup>Division of Pulmonary & Critical Care Medicine, University of Virginia, Charlottesville, Virginia, USA

<sup>10</sup>National Institute for Health Research Respiratory Clinical Research Facility, Royal Brompton Hospital, London, UK

<sup>11</sup>National Institute for Health Research, Glenfield Hospital, Leicester, UK

**Correction notice** This article has been corrected since it was first published. The open access licence has been updated to CC BY.

**Twitter** Luke M Kraven @KravenLuke

**Acknowledgements** We thank the research teams who have made their data publicly available via the Gene Expression Omnibus and to all study participants for contributing their data and samples.

**Contributors** LMK, ART, AJY, WAF, GJ and LVW designed the study and led the writing of the manuscript. LMK collected and analysed the data. ART accessed and verified the data. PLM, TM, JMcD, MM, IVY, DAS, YH, IN and SFM provided additional clinical data for the individuals in the study. All authors contributed to drafting and providing critical feedback on the manuscript. LVW is the guarantor of the paper.

**Funding** LVW holds a GSK/Asthma+Lung UK Chair in Respiratory Research (C17-1). GJ is supported by a National Institute for Health Research (NIHR) Research Professorship (NIHR reference RP-2017-08-ST2-014). LMK was funded by a Medical Research Council PhD studentship (MR/N013913/1). PLM is supported by an Action for Pulmonary Fibrosis Mike Bray fellowship. TM is supported by a National Institute for Health Research Clinician Scientist Fellowship (CS-2013-13-017) and an Asthma+Lung UK Chair in Respiratory Research (C17-3). IN is supported by a National Heart, Lung, and Blood Institute (NHLBI) grant (R01HL145266). DAS is supported by NHLBI grants (UG3HL151865, R01HL097163, P01HL092870, X01HL134585 and UH3HL123442) and a United States Department of Defense grant (W81XWH-17-1-0597). The GSE110147 study was supported by the Roche Multi Organ Transplant Academic Enrichment Fund, Lawson Research Institute Internal Research Fund and Western Strategic Support for CIHR Success, Seed Grant. The research was partially supported by the NIHR Leicester Biomedical Research Centre.

**Disclaimer** The views expressed are those of the author(s) and not necessarily those of the National Health Service (NHS), the National Institute for Health Research or the Department of Health.

**Competing interests** ART, AJY and WAF are employees and shareholders of GlaxoSmithKline. LVW reports recent and current research grant funding from GlaxoSmithKline and Orion and consultancy fees from Galapagos. PLM reports recent and current research grant funding from AstraZeneca, consulting fees from Hoffman-La Roche, Boehringer Ingelheim and AstraZeneca and speaker fees from Boehringer Ingelheim and Hoffman-La Roche. TM reports consulting fees from Boehringer Ingelheim, Roche/Genentech, AstraZeneca, Bayer, Blade Therapeutics,

Bristol-Myers Squibb, Galapagos, Galecto, GlaxoSmithKline, IQVIA, Pliant, Resivant, Theravance and Veracyte and speaker fees from Boehringer Ingelheim and Roche/Genentech. IVY reports consulting fees from Eleven P15 and is co-chair for the ATS Section of Genetics and Genomics. DAS is a consultant for Vertex Pharmaceuticals and is the founder and chief scientific officer of Eleven P15, a company focused on the early diagnosis and treatment of pulmonary fibrosis. IN reports consulting fees from Boehringer Ingelheim, Genentech and Sanofi Aventis and participation on the Yale Data Safety Monitoring Board. GJ reports research grant funding from AstraZeneca, Biogen, Galecto, GlaxoSmithKline, RedX and Pliant, consulting fees from Bristol Myers Squibb, Daewoong, Veracyte, Resolution Therapeutics and Pliant, speaker fees from Chiesi, Roche, PatientMPower and AstraZeneca, participation on Boehringer Ingelheim, Galapagos and Vicore data advisory boards and is a trustee for Action for Pulmonary Fibrosis.

**Patient consent for publication** Not applicable.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** Data are available in a public, open access repository and available on reasonable request. All gene expression data used in this study are freely available on the Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/>). Additional clinical data for some participants were obtained directly from the study authors and are available on reasonable request.

**Supplemental material** This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

**Open access** This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: <https://creativecommons.org/licenses/by/4.0/>.

#### ORCID iDs

Luke M Kraven <http://orcid.org/0000-0003-1908-6281>

Marco Mura <http://orcid.org/0000-0002-2159-7083>

R Gisli Jenkins <http://orcid.org/0000-0002-7929-2119>

#### REFERENCES

- Ley B, Collard HR, King TE. Clinical course and prediction of survival in idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2011;183:431–40.
- Lederer DJ, Martinez FJ. Idiopathic pulmonary fibrosis. *N Engl J Med* 2018;378:1811–23.
- Rodríguez-Portal JA. Efficacy and safety of nintedanib for the treatment of idiopathic pulmonary fibrosis: an update. *Drugs R D* 2018;18:19–25.
- Okuda R, Hagiwara E, Baba T, et al. Safety and efficacy of pirfenidone in idiopathic pulmonary fibrosis in clinical practice. *Respir Med* 2013;107:1431–7.
- Kropski JA, Lawson WE, Blackwell TS. Personalizing therapy in idiopathic pulmonary fibrosis: a glimpse of the future? *Am J Respir Crit Care Med* 2015;192:1409–11.
- Jenkins G. Endotyping idiopathic pulmonary fibrosis should improve outcomes for all patients with progressive fibrotic lung disease. *Thorax* 2015;70:9–10.
- Woodruff PG, Modrek B, Choy DF, et al. T-helper type 2-driven inflammation defines major subphenotypes of asthma. *Am J Respir Crit Care Med* 2009;180:388–95.
- Aggarwal C. Targeted therapy for lung cancer: present and future. *Ann Palliat Med* 2014;3:229–35.
- van 't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002;415:530–6.
- Slodkowska EA, Ross JS. Mammagrin 70-gene signature: another milestone in personalized medical care for breast cancer patients. *Expert Rev Mol Diagn* 2009;9:417–22.
- Herazo-Maya JD, Noth I, Duncan SR, et al. Peripheral blood mononuclear cell gene expression profiles predict poor outcome in idiopathic pulmonary fibrosis. *Sci Transl Med* 2013;5:205ra136.
- Organ LA, Duggan A-MR, Oballa E, et al. Biomarkers of collagen synthesis predict progression in the profile idiopathic pulmonary fibrosis cohort. *Respir Res* 2019;20:148.
- Wang Y, Yella J, Chen J, et al. Unsupervised gene expression analyses identify IPF-severity correlated signatures, associated genes and biomarkers. *BMC Pulm Med* 2017;17:133.
- Zhang N, Guo Y, Wu C, et al. Identification of the molecular subgroups in idiopathic pulmonary fibrosis by gene expression profiles. *Comput Math Methods Med* 2021;2021:7922594.

- 15 Edgar R, Domrachev M, Lash AE. Gene expression Omnibus: NCBI gene expression and hybridization array data Repository. *Nucleic Acids Res* 2002;30:207–10.
- 16 Sweeney TE, Wong HR, Khatri P. Robust classification of bacterial and viral infections via integrated host gene expression diagnostics. *Sci Transl Med* 2016;8:346ra91.
- 17 Sweeney TE, Chen AC, Gevaert O. Combined mapping of multiple clustering algorithms (communal): a robust method for selection of cluster number, K. *Sci Rep* 2015;5:1–10.
- 18 Sweeney TE, Azad TD, Donato M, et al. Unsupervised analysis of transcriptomics in bacterial sepsis across multiple datasets reveals three robust clusters. *Crit Care Med* 2018;46:915.
- 19 Herazo-Maya JD, Sun J, Molyneaux PL, et al. Validation of a 52-gene risk profile for outcome prediction in patients with idiopathic pulmonary fibrosis: an international, multicentre, cohort study. *Lancet Respir Med* 2017;5:857–68.
- 20 Ley B, Ryerson CJ, Vittinghoff E, et al. A multidimensional index and staging system for idiopathic pulmonary fibrosis. *Ann Intern Med* 2012;156:684–91.
- 21 Halloran JW, Zhu D, Qian DC, et al. Prediction of the gene expression in normal lung tissue by the gene expression in blood. *BMC Med Genomics* 2015;8:77.
- 22 Zhao YD, Yin L, Archer S, et al. Metabolic heterogeneity of idiopathic pulmonary fibrosis: a metabolomic study. *BMJ Open Respir Res* 2017;4:e000183.
- 23 Bargagli E, Refini RM, d'Alessandro M, et al. Metabolic dysregulation in idiopathic pulmonary fibrosis. *Int J Mol Sci* 2020;21:5663.
- 24 Biernacka A, Dobaczewski M, Frangogiannis NG. TGF- $\beta$  signaling in fibrosis. *Growth Factors* 2011;29:196–202.
- 25 Meng X-M, Nikolic-Paterson DJ, Lan HY. TGF- $\beta$ : the master regulator of fibrosis. *Nat Rev Nephrol* 2016;12:325.
- 26 Györfi AH, Matei A-E, Distler JHW. Targeting TGF- $\beta$  signaling for the treatment of fibrosis. *Matrix Biol* 2018;68-69:8–27.
- 27 Platakis M, Koutsopoulos AV, Darivianaki K, et al. Expression of apoptotic and antiapoptotic markers in epithelial cells in idiopathic pulmonary fibrosis. *Chest* 2005;127:266–74.
- 28 Wang Q, Xie Z-L, Wu Q, et al. Role of various imbalances centered on alveolar epithelial cell/fibroblast apoptosis imbalance in the pathogenesis of idiopathic pulmonary fibrosis. *Chin Med J* 2021;134:261.
- 29 du Bois RM. Strategies for treating idiopathic pulmonary fibrosis. *Nat Rev Drug Discov* 2010;9:129–40.
- 30 Korthagen NM, van Moersel CHM, Barlo NP, et al. Association between variations in cell cycle genes and idiopathic pulmonary fibrosis. *PLoS One* 2012;7:e30442.
- 31 Adegunsoye A, Hrusch CL, Bonham CA, et al. Skewed Lung CCR4 to CCR6 CD4<sup>+</sup> T Cell Ratio in Idiopathic Pulmonary Fibrosis Is Associated with Pulmonary Function. *Front Immunol* 2016;7:516.
- 32 Desai O, Winkler J, Minasyan M, et al. The role of immune and inflammatory cells in idiopathic pulmonary fibrosis. *Front Med* 2018;5:43.
- 33 McCarthy MI. Painting a new picture of personalised medicine for diabetes. *Diabetologia* 2017;60:793–9.
- 34 Huang LS, Berdyshev EV, Tran JT, et al. Sphingosine-1-Phosphate lyase is an endogenous suppressor of pulmonary fibrosis: role of S1P signalling and autophagy. *Thorax* 2015;70:1138–48.
- 35 Yang IV, Luna LG, Cotter J, et al. The peripheral blood transcriptome identifies the presence and extent of disease in idiopathic pulmonary fibrosis. *PLoS One* 2012;7:e37708.
- 36 Molyneaux PL, Willis-Owen SAG, Cox MJ, et al. Host-Microbial interactions in idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2017;195:1640–50.