Original research

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

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

► 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

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

Received 8 December 2021
Accepted 14 April 2022

To cite: Kraven LM, Taylor AR, Molyneaux PL, *et al*. *Thorax* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/thoraxjnl-2021-218563

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, $p=3.7\times10^{-5}$).

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.

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.

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.[1 2] 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.[3 4] The clinical course of IPF is highly variable with slow progression in some patients, rapid progression in others, while many experience a slowly progressive course interspersed with periods of rapid lung function deterioration.[1] 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,[5 6] as in asthma and lung cancer,[7 8] 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

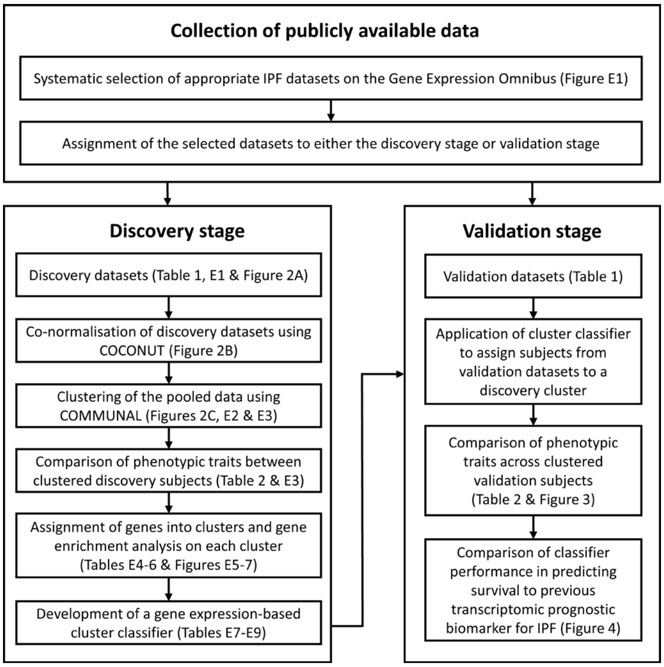

Collection of publicly available data
Systematic selection of appropriate IPF datasets on the Gene Expression Omnibus (Figure E1)
Assignment of the selected datasets to either the discovery stage or validation stage
Discovery stage
Discovery datasets (Table 1, E1 & Figure 2A)
Co-normalisation of discovery datasets using COCONUT (Figure 2B)
Clustering of the pooled data using COMMUNAL (Figures 2C, E2 & E3)
Comparison of phenotypic traits between clustered discovery subjects (Table 2 & E3)
Assignment of genes into clusters and gene enrichment analysis on each cluster (Tables E4-6 & Figures E5-7)
Development of a gene expression-based cluster classifier (Tables E7-E9)
Validation stage
Validation datasets (Table 1)
Application of cluster classifier to assign subjects from validation datasets to a discovery cluster
Comparison of phenotypic traits across clustered validation subjects (Table 2 & Figure 3)
Comparison of classifier performance in predicting survival to previous transcriptomic prognostic biomarker for IPF (Figure 4)

Figure 1 A flow chart showing the design of our study. COCONUT, COmbat CO-Normalisation Using conTrols; COMMUNAL, Combined Mapping of Multiple clUsteriNg 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.[9 10] 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.[11 12] Studies using unsupervised clustering approaches to discover disease endotypes have been limited by sample size,[13] ability to validate[13 14] and clinical interpretation.[14] However, these studies have consistently reported association of inflammatory genes,[13] in particular those associated with T cell activation[11] and differentiation,[14] 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[15] 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,[16] 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 clUsteriNg ALgorithms (COMMUNAL)[17] 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 χ^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).[18]

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*.[19] 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		
GEO accession number	GSE38958		GSE33566		GSE93606		GSE132607	GSE27957	GSE28042
Reference	Huang *et al*[34]		Yang *et al*[35]		Molyneaux *et al*[36]		*	†[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_{CO} % 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,[11] where the data in GSE27957 were used in discovery and the data in GSE28042 were used as independent validation data.

DL_{CO}, 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).

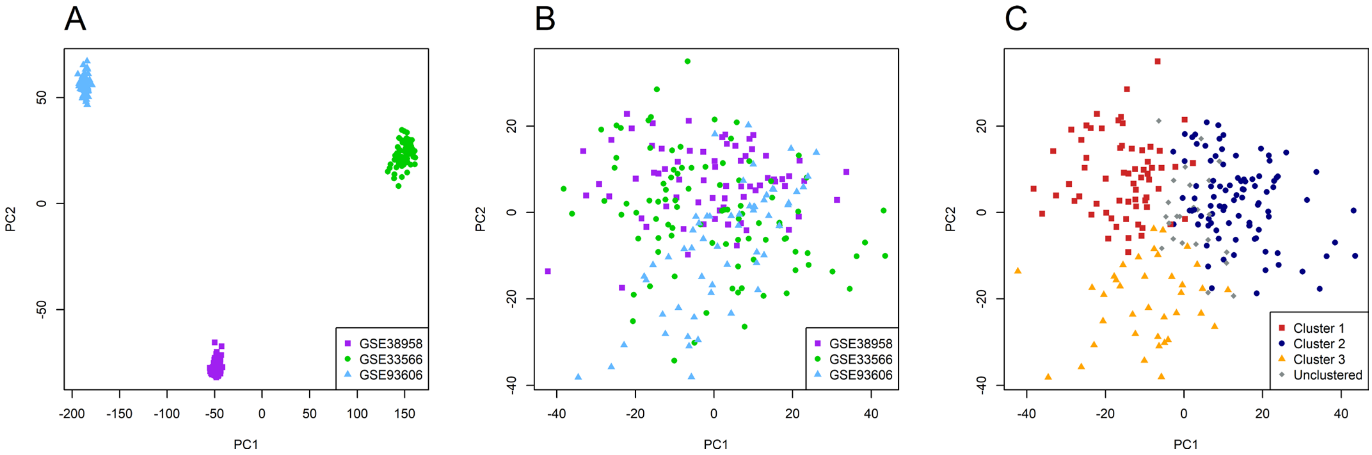


Figure 2 Plots of the first two principal components of the gene expression data for the idiopathic pulmonary fibrosis samples prior to co-normalisation and stratified by original study (A), post co-normalisation and stratified by original study (B) and post co-normalisation stratified by cluster (C). 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%)	**0.009**	48	16 (48.5%)	13 (19.7%)	12 (57.1%)	**0.001**	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_{CO} % predicted (median, IQR)	35.0 (30.0)	45.0 (29.2)	34.4 (17.3)	**0.009**	133	42.1 (26.4)	48.2 (21.1)	43.4 (20.3)	0.069	194
FEV_1 % 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)	**0.006**	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.[20] P value for count data is from a χ^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_{CO}, diffusing capacity for carbon monoxide; FEV1, 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_{CO}) across clusters (p=0.009). Subjects in cluster 1 had a similar median predicted DL_{CO} to those in cluster 3, whilst subjects in cluster 2 had the greatest median predicted DL_{CO}, 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),[20] 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_{CO}, although the difference in DL_{CO} 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, $p=9.1\times10^{-5}$). 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.[19] 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.

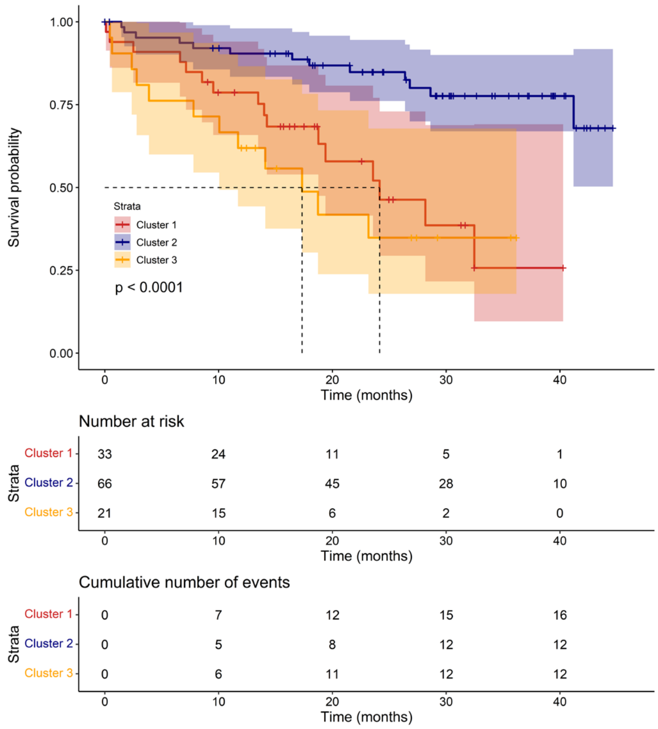


Figure 3 A Kaplan-Meier plot showing survival over time for the clustered validation subjects. The p value shown on the plot is from a log-rank test testing the three curves for equality. Median survival in each cluster is shown by dotted lines, where possible.

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, $p=3.7\times10^{-5}$). 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_{CO} (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_{CO} (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_{CO}, 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

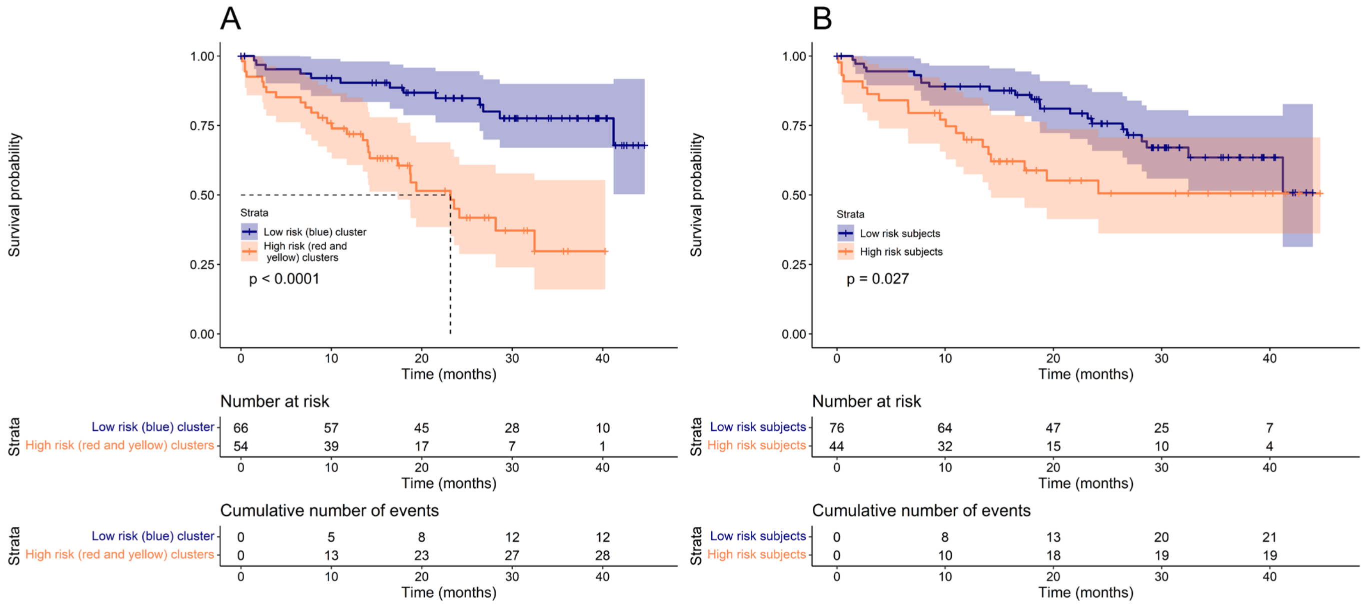


Figure 4 Survival over time for the subjects with IPF in GSE27957 and GSE28042, stratified by risk group according to our 13 gene classifier (A) and SAMS method by Herazo-Maya *et al* (B) . The p value on each plot is from a log-rank test testing the two curves for equality. A dotted line on the plot 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.[21] 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.[22 23] The genes in cluster 1 were also significantly enriched for pathways related to transforming growth factor-β signalling, which is a central mediator of fibrosis.[24–26]

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,[27] with this imbalance contributing to IPF pathogenesis.[28] Furthermore, the use of therapies that can selectively manipulate apoptosis have been proposed.[29] Additionally, genetic variants within cell cycle genes have been shown to be associated with IPF development and progression.[30] 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.[31 32] 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 ($p=9.1\times10^{-5}$), 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_{CO} 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.[19] 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,[16] 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.[33]

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
[1]Department of Health Sciences, University of Leicester, Leicester, UK
[2]Research & Development, GlaxoSmithKline, Stevenage, UK
[3]Guy's and St Thomas' NHS Foundation Trust, Royal Brompton and Harefield Hospitals, London, UK
[4]National Heart and Lung Institute, Imperial College London, London, UK
[5]Keck School of Medicine, University of Southern California, Los Angeles, California, USA
[6]Division of Pulmonary, Critical Care & Sleep Medicine, Yale School of Medicine, New Haven, Connecticut, USA
[7]Division of Respirology, Western University, London, Ontario, Canada
[8]Department of Medicine, University of Colorado, Denver, Colorado, USA
[9]Division of Pulmonary & Critical Care Medicine, University of Virginia, Charlottesville, Virginia, USA
[10]National Institute for Health Research Respiratory Clinical Research Facility, Royal Brompton Hospital, London, UK
[11]National Institute for Health Research, Glenfield Hospital, Leicester, UK

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, Respivant, 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.

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

1 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.
2 Lederer DJ, Martinez FJ. Idiopathic pulmonary fibrosis. *N Engl J Med* 2018;378:1811–23.
3 Rodríguez-Portal JA. Efficacy and safety of nintedanib for the treatment of idiopathic pulmonary fibrosis: an update. *Drugs R D* 2018;18:19–25.
4 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.
5 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.
6 Jenkins G. Endotyping idiopathic pulmonary fibrosis should improve outcomes for all patients with progressive fibrotic lung disease. *Thorax* 2015;70:9–10.
7 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.
8 Aggarwal C. Targeted therapy for lung cancer: present and future. *Ann Palliat Med* 2014;3:229–35.
9 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.
10 Slodkowska EA, Ross JS. Mammaprint 70-gene signature: another milestone in personalized medical care for breast cancer patients. *Expert Rev Mol Diagn* 2009;9:417–22.
11 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.
12 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.
13 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.
14 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-β signaling in fibrosis. *Growth Factors* 2011;29:196–202.

25 Meng X-M, Nikolic-Paterson DJ, Lan HY. TGF-β: the master regulator of fibrosis. *Nat Rev Nephrol* 2016;12:325.

26 Györfi AH, Matei A-E, Distler JHW. Targeting TGF-β signaling for the treatment of fibrosis. *Matrix Biol* 2018;68-69:8–27.

27 Plataki 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 Moorsel 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 $CD4^+$ 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.

Cluster analysis of transcriptomic datasets to identify endotypes of Idiopathic Pulmonary Fibrosis – online data supplement

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

Contents

Additional text

Systematic selection of publicly available datasets

We performed our systematic search in March 2020 to select the datasets that were suitable for inclusion in the study (Figure E1). We required multiple sets of transcriptomic data from independent cohorts. We searched the Gene Expression Omnibus (GEO) (1) for all collections that contained the term ‘IPF’, excluding any that did not contain human samples. We restricted the search to collections with at least 30 samples as this allowed for inclusion of the largest datasets with the most IPF cases and healthy control subjects, which are the datasets that were the most likely to successfully co-normalise due to the higher counts of healthy control subjects. We did not restrict the search by platform. Each of the remaining collections were then reviewed to assess whether they contained data for IPF cases. All collections that did not contain data for IPF subjects were excluded.

For a successful co-normalisation and meaningful clustering results, we were required to choose an optimal tissue/cell type to use for the analysis. After reviewing the IPF datasets on GEO, we chose whole blood as our optimal tissue/cell type. There were three main reasons for this. Firstly, there were several relatively large whole blood datasets available on GEO and these would have provided the largest sample size and greatest statistical power for the study compared to other tissue types. Secondly, we required multiple datasets that contained data for healthy controls in addition to the IPF patients (so that the data could be co-normalised using COCONUT) and the whole blood datasets fulfilled this requirement. Thirdly, the accompanying clinical data for the whole blood datasets was far more comprehensive than for other tissue types, such as whole lung. This clinical data was vital to the study as it was required for the characterisation of the clusters in both the discovery and validation stages. So, all GEO collections containing expression data measured from a non-blood tissue/cell type were excluded.

As multiple transcriptomic datasets were to be combined, it was important to check for the presence of common individuals across cohorts, which would have meant that the cohorts were not independent and could have biased the results of the study. To this end, the subjects in each collection were checked for unique study identification codes. Using these, we found that two of the blood collections, GSE132607 (n=74) and GSE85268 (n=68), both contained subjects from the Correlating Outcomes With Biochemical Markers to Estimate Time-progression in Idiopathic Pulmonary Fibrosis (COMET) study (ClinicalTrials.gov identifier: NCT01071707). There were a large number of IPF subjects in common between the two cohorts (n=58) and so we excluded the GSE85268 dataset as it was the collection with fewer IPF subjects.

The seven remaining collections of data were uploaded by research groups from across the USA (including the University of Virginia, Yale University, the University of Nevada and the University of Colorado) and the UK (Imperial College London). GSE27957 and GSE28042 were uploaded by the Kaminski Lab in Yale. These two collections were both used in the same study (2), where GSE27957 was used as discovery data and GSE28042 was used as independent replication data. Similarly, the data found in GSE133298 and GSE132607 were uploaded by researchers at the University of Virginia and were used as independent cohorts in the same study (unpublished as of October 2020, both collections uploaded to GEO in September 2019). All remaining collections were uploaded by separate research groups and no additional evidence of common subjects across cohorts was found so the seven cohorts of IPF subjects were deemed independent. However, the possibility that subjects could be common in two or more studies cannot be ruled out.

The human biological samples were sourced ethically and their research use was in accord with the terms of the informed consents under an institutional review board/ethical committee (IRB/EC)-approved protocol.

Assignment of datasets to discovery and validation stages

All cohorts included in the discovery stage must have contained healthy controls in order to enable the data co-normalization step. Four of the seven selected blood datasets contained data for healthy controls. We used the three with the greatest number of controls in discovery as these were the most likely to successfully co-normalize. The four remaining datasets were reserved for use in the validation stage. One dataset (GSE133298) was excluded during the validation stage as not all of the genes that were required to fully apply the classifier were present in the dataset.

Discovery stage studies

GSE38958: This dataset originates from an American observational study (3) that was investigating the relationship between sphingosine-1-phosphate lyase and pulmonary fibrosis. IPF cases were recruited from the University of Chicago. The authors studied gene expression data from peripheral blood mononuclear cells of IPF subjects (n=70) and compared this to gene expression from healthy controls (n=45).

GSE33566: This dataset contained data for 123 IPF subjects and 30 healthy controls. A subset of this data was used in an American observational study (4), where the authors hypothesised that a peripheral blood biomarker for IPF would be able to identify the disease in its early stages and allow for disease progression to be monitored. The IPF cases were recruited through the Interstitial Lung Disease or the Familial Pulmonary Fibrosis Programs conducted at National Jewish Health and Duke University. In the study, 40 IPF subjects were split into groups based on their predicted FVC and D_{LCO}, then the authors looked for differentially expressed genes between groups.

GSE93606: This dataset contained data from a British prospective cohort study (5) (n=57 IPF subjects and n=20 healthy age, sex and smoking history matched controls) which had the objective of examining host-microbial interactions in IPF subjects over time. IPF cases were prospectively recruited from the Interstitial Lung Disease Unit at the Royal Brompton Hospital, London, within six months of their initial diagnosis. The study was approved by the local research ethics committee (reference numbers 10/H0720/12 and 12/LO/1034). In this study, gene expression data from peripheral blood and lung function measurements were collected at multiple time points. However, only baseline gene expression and lung function data was used in our study. IPF patient survival was also recorded up to a maximum follow-up time of 34 months.

Validation stage studies

GSE132607: This dataset originates from a study (unpublished as of March 2022) which aimed to develop a predictor of FVC progression by studying gene expression differences in 74 IPF subjects over time. The subjects included in this analysis were participants in COMET-IPF (Correlating Outcomes with biochemical Markers to Estimate Time-progression in Idiopathic Pulmonary Fibrosis), a prospective, observational study correlating biomarkers with disease progression. All IPF cases had been recruited in to this study within four years of their initial IPF diagnosis.

GSE27957 and **GSE28042**: both datasets originate from the same study (6), where the data in GSE27957 (n=45 IPF subjects) was used in discovery and the data in GSE28042 (n=75 IPF subjects) was used as independent validation data. Individuals with IPF from the GSE27957 dataset were recruited from the University of Chicago and the individuals with IPF from the GSE28042 dataset were recruited from the University of Pittsburgh. In brief, the authors used these cohorts to develop a 52-gene signature that had the ability to predict transplant-free survival in IPF subjects.

Data pre-processing

In each discovery dataset, probes that did not map to a gene were removed. In the instance where multiple probes mapped to the same gene, only the probe with the greatest mean expression was included in the analysis. Each dataset was then quantile normalised to reduce any technical differences between the gene probes within a study. Following this, each dataset was scaled so that all expression data was on the $\log_2$ scale and thus in a consistent form prior to co-normalisation. Genes were matched across studies based on their gene symbols.

Data co-normalisation using COCONUT

We used COmbat CO-Normalization Using conTrols (COCONUT) (7) (in R v4.0.0 and the 'COCONUT' package) to reduce the technical differences between the three discovery transcriptomic datasets, therefore enabling a cluster analysis to be performed on the pooled, co-normalized data. COCONUT is an unbiased co-normalisation method which assumes that all healthy controls across studies come from the same statistical distribution. It uses the healthy controls in each study to calculate correction factors that remove the technical differences in the data for the diseased subjects, without bias to the number of disease cases present. The method is adapted from the ComBat empiric Bayes normalization method (8), which is often used to adjust for batch effects within a study.

As COCONUT makes the assumption that all healthy controls come from the same background statistical distribution, we tested for significant differences in clinical and demographic traits between the healthy controls in each study, where possible. Clinical and demographic characteristics of the healthy controls were compared using chi-square tests for count data and analysis of variance for non-skewed continuous data.

Data for each study was input into COCONUT by providing a gene expression matrix (on the $\log_2$ scale) of common genes against subjects. These were accompanied by an indicator variable that showed which subjects were cases and which were controls. Following the co-normalisation, we removed all healthy control subjects from further analysis. Plots of the first two principal components of the transcriptomic data before and after COCONUT were used to evaluate the efficacy of the co-normalisation.

Clustering using COMMUNAL

In this study, we ran COMMUNAL using consensus clustering versions of two algorithms, K-means clustering and partitioning around medoids (PAM). Five different metrics were used to assess the validity of the clustering for different numbers of clusters and genes. These were: the gap statistic, connectivity, average silhouette width, the G3 metric, and Pearson's gamma coefficient. We ranked the genes in order of variance, with the 'top' 100 genes referring to the 100 genes with the greatest variance. We then applied the COMMUNAL algorithm using a range of input genes from the top 100 to the top 5,000. The genes with the greatest variance were used as these were the most likely to be informative, so as to minimise the number of non-informative genes and increase the signal-to-noise ratio.

The samples that were not assigned into the same cluster by the COMMUNAL clustering algorithms were labelled 'unclustered'. Since the intention was to use the clustered data to create a classifier and classifiers trained on data with fewer errors are more robust, these uncertain samples were removed from further analysis to improve the accuracy of the classifier.

The results were visualised in the form of a 3-dimensional (3D) map (Figure E2), which we used to select the optimal number of clusters in the data, as well as the optimal number of genes to use in the clustering. The map shows the mean of standardized values of each validity measure across the entire tested space. On the 3D map, blue squares indicate a potentially optimal clustering at a certain number of genes by finding the assignment where the mean combined validation metric is greatest. The absolute maximum number of clusters for any consensus subset is marked with a red square. The points where the blue and red squares overlap indicate stable optima. If stable optima at a particular number of clusters are observed over most of the tested space, this indicates the presence of a strong, consistent biological signal at this number of clusters.

In Figure E2 there are stable optima at K=4 from 250 genes to 1,000 genes, and at K=3 from 2,500 genes to 5,000 genes, as shown by the red and blue squares meeting. Despite the K=4 clustering assignment at 1,000 genes showing the highest mean standardized validity score of all tested clustering assignments, there were stable optima at K=3 clusters over a larger range of tested space, indicating a stronger biological signal. As such, K=3 was chosen as the optimal number of clusters in the pooled IPF dataset. The clustering at 2,500 genes and 3 clusters was chosen as the optimal clustering assignment, under the assumption that the assignment with the fewest number of genes (out of those with stable optima at K=3) has the least amount of redundant signal.

Comparison of phenotypic traits across clusters

We characterised the clusters by comparing the clinical and demographic traits of the subjects that were assigned to each cluster. This was done for each phenotypic trait that was reported in at least one discovery cohort and one validation cohort. The statistical significance of the phenotypic differences across clusters was evaluated for all studies combined using a chi-square test for count data, an analysis of variance to compare means for non-skewed continuous data and a Kruskal-Wallis rank sum test to compare medians for skewed continuous data. For traits in the form of time-to-event data, Kaplan-Meier plots were used to approximate and visualise the survival function for these variables. Further, Cox proportional-hazards (PH) models were fit with cluster as the sole independent variable and the time to the event as the response variable.

Gene enrichment analysis

First, we assigned each of the 2,500 genes used in the optimal COMMUNAL clustering assignment to the cluster in which its expression was most different to its expression in the other two clusters, as this suggests that that gene was contributing to the identity of that cluster. 814 genes were assigned to Cluster 1, 866 were assigned to Cluster 2 and 820 were assigned to Cluster 3.

We then performed multiple ANOVA tests (one for each cluster) for each gene, each comparing the expression of that gene in subjects within a given cluster against the expression of subjects in both other clusters. Each gene was then assigned to the cluster in which it had the lowest ANOVA p-value. One benefit of this approach is that the ANOVA tests allowed for filtering based on statistical significance; a nominal p-value significance threshold of 0.05 was introduced and genes whose lowest ANOVA p-value was greater than this threshold were removed. The rationale for the introduction of this filtering step was that removing genes that were not associated with any cluster would reduce noise and strengthen the gene enrichment analysis for each cluster. The threshold for statistical significance was kept at a nominal level as a correction for all 7,500 ANOVA tests would have likely left too few genes assigned to each cluster to successfully perform the enrichment analysis. After the removal of the genes that were not at least nominally associated to any cluster, there were 769 genes assigned to Cluster 1, 839 assigned to Cluster 2 and 784 assigned to Cluster 3.

Then, gene enrichment analysis was performed separately on the three resulting gene lists using R v.4.0.0 and the in-house package ‘metabaser’ (database v20.3, package v4.2.3). This was used to search databases of gene ontology terms for statistically overrepresented *biological processes* and *biological pathways*. At the time that the analysis was performed, there were 17,552 *biological processes* and 12,222 *biological pathways* in the database accessed by metabaser. metabaser reports ‘q-values’, which are p-values that have been adjusted for multiple tests using the false-discovery rate. Gene ontology terms with q-value < 0.05 were deemed statistically significant. Sankey plots were used to show which of the genes that were assigned to each cluster corresponded to the 20 most significantly enriched *biological pathways* (see Figure 3).

Additionally, the gene lists of each cluster were searched for the presence of the nearest gene for any of the 14 variants that were genome-wide significant in Allen et al. (9), the largest genome-wide association study meta-analysis of IPF susceptibility to-date. The 14 genes were as follows: *AKAP13*, *ATP11A*, *DEPTOR*, *DPP9*, *DSP*, *FAM13A*, *LRRC34*, *IVD*, *KIF15*, *MAD1L1*, *MAPT*, *MUC5B*, *TERC* and *TERT*. Following this, enrichment analysis was performed on the genes of each cluster to investigate whether those genes were statistically overconnected (in terms of direct gene regulation) to any of the IPF-associated genes from Allen et al. (2020). If the genes that were assigned to a particular cluster were found to be overconnected to one or more of the IPF-associated genes listed above (say the exact number of overconnected IPF-associated genes is N), then a hypergeometric test was performed to approximate the statistical significance of the finding that N out of the 14 IPF-associated genes were present within the list of overconnected genes for that cluster.

None of the 14 suspected IPF susceptibility genes from Allen et al. were assigned to Cluster 1, nor were they statistically overconnected to the genes that were assigned to this cluster. *FAM13A* was one of the genes that was assigned to Cluster 2, though it did not belong to any of the top 20 significantly enriched biological pathways. Additionally, the genes in Cluster 2 were statistically overconnected to five other IPF-associated genes. These were: *AKAP13*, *DSP*, *LRRC34*, *MAPT* and *TERT*. The hypergeometric p-value was calculated to be 0.020, indicating that it is significant that five IPF-associated genes were overconnected to the genes Cluster 2 and this is more than would be expected due to random chance. None of the IPF-associated genes from Allen et al. were found in the gene list for Cluster 3, although four were found to be statistically overconnected to the genes in this cluster. These were as follows: DSP, MAD1L1, MAPT and TERT. The statistical significance of this was approximated to be P=0.008 using a hypergeometric test, again indicating that this was significantly more than would be expected under random chance.

Developing the gene expression-based cluster classifier

Classification is a method of supervised machine learning that uses a correctly labelled training dataset to predict which category new observations belong in.

To determine the optimal genes to include in the classifier for the IPF data, we used an iterative algorithm which performed a greedy forward search for each cluster separately to determine the optimal combination of genes to differentiate between subjects in that cluster vs all other clusters. This was done by calculating receiver operating characteristic curves for each combination of genes and selecting the combination of genes which maximised the area under the curve (AUC). In an effort to prevent the classifier from being overfit to the discovery data, a threshold was implemented to stop the algorithm once an AUC of 0.99 had been reached. Each gene was labelled as either overexpressed or underexpressed based on whether the average expression of that gene was greater in the subjects from that particular cluster compared to the average expression across all subjects.

Making predictions with the classifier was a two-stage process. First, each subject was given a classification score for each cluster. This score was calculated as the geometric mean of the overexpressed genes for that cluster minus the geometric mean of the underexpressed genes. These scores were mean centred around zero and scaled to reflect a Z-score (i.e. standard deviation equal to 1). Ideally, subjects that belonged to a certain cluster should have had a high classification Z-score for that cluster and low classification Z-scores for the other clusters.

Then, we used the classification Z-scores to fit a multinomial logistic regression model, with cluster as the independent categorical variable and the Z-scores from each cluster as the dependent variables. This model had the ability to take data from new IPF subjects and predict which cluster they were each most likely to belong in, using only expression data from the optimal genes in the classifier. Importantly, the classifier does not use absolute levels of gene expression in order to make predictions, but instead utilizes relative gene expression between subjects. This meant that the classifier could be applied to a cohort of IPF cases (from the same study) without first requiring the removal of technical effects, which allowed for the use of validation datasets that did not contain data for healthy controls.

We tested the prediction accuracy of the classifier by using it to reassign all of the IPF subjects in the discovery datasets.

Risk classification using the classifier

Each of the IPF subjects in the two validation studies for which survival data was available, GSE27957 (n=45) and GSE28042 (n=75), were assigned into one of the three clusters using the 13 gene classifier. As significant differences in survival were observed between clusters 1 and 2 and 2 and 3, but not between clusters 1 and 3 (Table E9), we used assignment to clusters 1 and 3 to define high risk individuals and assignment to cluster 2 as low risk.

Risk classification using SAMS

Each of these individuals were also classed as high-risk or low-risk using SAMS (2). 7 of the 52 genes used by SAMS were expected to be more highly expressed in high risk cases than low risk cases (‘up genes’). Likewise, the remaining 45 genes were expected to be less highly expressed in high risk cases than low risk cases (‘down’ genes). The method that SAMS used to predict risk is as follows:

1. For each gene, the geometric mean of the expression for that gene across all subjects was calculated. This value represents the average level of expression for that gene across the whole cohort. It was then subtracted from the gene expression of that gene for each subject so that positive values represented subjects that had increased expression of that gene compared to the average and negative values represented subjects that had decreased expression compared to the average.

2. For each subject, the proportion of the 7 ‘up genes’ that were overexpressed was calculated. Similarly, the proportion of the 45 ‘down genes’ that were less highly expressed than average was calculated. So, if a subject had 4 ‘up genes’ that were greater than the average and 30 ‘down genes’ that were lower than the average, these proportions would have been 0.571 and 0.667 respectively.

3. For each subject, the sum of the geometric mean normalised expression data was summed up for the ‘up genes’ that were more highly expressed than average. Then the sum of the geometric mean normalised expression data was summed up for the ‘down genes’ that were less highly expressed than average. So, for example, for the subject above who had 4 of the 7 ‘up genes’ that were more highly expressed than the average, say with expression values 0.185, 0.553, 0.123 and 1.003 for these four genes, the sum would have been 1.864. The sum for the ‘down genes’ must always be negative, for example say that this sum for the subject above was -7.645.

4. The proportion of the ‘up genes’ calculated in step 2 was multiplied by the sum for the ‘up genes’ calculated in step 3 to produce the ‘up score’ for each subject. So, for the example subject above, their up score would have been 0.571×1.864 = 1.064. A ‘down score’ for each subject was also calculated by multiplying their proportion of down genes by their down sum from step 3. For our example subject, this would have been 0.667× -7.645=-5.099.

5. Subjects with up scores greater than the median value and down scores lower than the median value were classed as ‘high risk’, while all other subjects were classed as ‘low risk’.

This was done separately for each cohort and by using data from as many of the 52 genes as were measured in the datasets; 51/52 (98·1%) genes in the SAMS signature were present in GSE27957 and 50/52 (96·2%) were present in GSE28042. Two-way tables were used to compare agreement between the two methods.

Comparing prognostic methods using survival analysis

Kaplan-Meier plots were used to visualise the survival over time for the validation subjects in each risk group under each method. In both cases, the log-rank test was used to test the survival curves of each risk group for equality. Univariate Cox proportional-hazards models were fit to the data with risk group as the sole covariate and time-to-death as the outcome of interest. In both cases, the low-risk group was used as the reference group. The Concordance index (C-index), the equivalent of the area under the curve (AUC) for a receiver operating characteristic (ROC) curve, and the p-values from the log-rank test were used to assess which method performed best at assigning the IPF subjects to the correct risk group and therefore predicting survival.

Following this, multivariate Cox proportional-hazards models were used to assess whether the predictions made by each method were significant predictors of mortality in the validation datasets whilst adjusting for age, sex, ancestry, FVC and DL_{CO}. We used the likelihood ratio test and C-index to assess whether either of the two methods of risk prediction led to a significant increase in predictive ability over a Cox PH model containing only age, sex, ancestry, FVC and DL_{CO}.

References

(1) Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002;30:207-210.

(2) Herazo-Maya JD, Sun J, Molyneaux PL, Li Q, Villalba JA, Tzouvelekis A, Lynn H, Juan-Guardela BM, Risquez C, Osorio JC. Validation of a 52-gene risk profile for outcome prediction in patients with idiopathic pulmonary fibrosis: an international, multicentre, cohort study. *The Lancet Respiratory Medicine* 2017;5:857-868.

(3) Huang LS, Berdyshev EV, Tran JT, Xie L, Chen J, Ebenezer DL, Mathew B, Gorshkova I, Zhang W, Reddy SP. Sphingosine-1-phosphate lyase is an endogenous suppressor of pulmonary fibrosis: role of S1P signalling and autophagy. *Thorax* 2015;70:1138-1148.

(4) Yang IV, Luna LG, Cotter J, Talbert J, Leach SM, Kidd R, Turner J, Kummer N, Kervitsky D, Brown KK. The peripheral blood transcriptome identifies the presence and extent of disease in idiopathic pulmonary fibrosis. *PloS one* 2012;7:e37708.

(5) Molyneaux PL, Willis-Owen SA, Cox MJ, James P, Cowman S, Loebinger M, Blanchard A, Edwards LM, Stock C, Daccord C. Host–microbial interactions in idiopathic pulmonary fibrosis. *American journal of respiratory and critical care medicine* 2017;195:1640-1650.

(6) Herazo-Maya JD, Noth I, Duncan SR, Kim S, Ma S, Tseng GC, Feingold E, Juan-Guardela BM, Richards TJ, Lussier Y. Peripheral blood mononuclear cell gene expression profiles predict poor outcome in idiopathic pulmonary fibrosis. *Science translational medicine* 2013;5:205ra136.

(7) Sweeney TE, Wong HR, Khatri P. Robust classification of bacterial and viral infections via integrated host gene expression diagnostics. *Science translational medicine* 2016;8:346ra91.

(8) Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007;8:118-127.

(9) Allen RJ, Guillen-Guio B, Oldham JM, Ma S, Dressen A, Paynton ML, Kraven LM, Obeidat M, Li X, Ng M. Genome-wide association study of susceptibility to idiopathic pulmonary fibrosis. *American journal of respiratory and critical care medicine* 2020;201:564-574.

Additional Tables

TABLE E1: Information about the transcriptomic data in the discovery datasets and the platform used in each study.

GEO accession	GSE38958	GSE33566	GSE93606
Microarray platform	Affymetrix Human Exon 1.0 ST Array	Agilent-014850 Whole Human Genome Microarray	Affymetrix Human Gene 1.1 ST Array
Number of gene probes	44,280	32,850	33,297
Number of unique genes	17,256	12,171	20,254

TABLE E2: Comparison of the age and sex of the healthy controls in each discovery stage study. Data are presented as count (percentage) or mean (standard deviation, SD). P-value for count data is from a chi-square test and the test comparing means is analysis of variance.

	GSE38958	GSE33566	GSE93606	P-value	n used
Number of healthy controls	45	30	20		
Age (years, SD)	69·3 (9·3)	62·4 (14·3)	66·0 (10·6)	0.187	83
Sex (% male)	27 (60·0%)	14 (46·7%)	12 (60·0%)	0.477	95

TABLE E3: Comparison of clinical and demographic traits of clustered discovery subjects by study and for all studies combined. Data are presented as count (percentage), mean (standard deviation, SD) or median (interquartile range, IQR). NA = data not available, FVC=Forced vital capacity, D_{LCO} = Diffusing capacity for carbon monoxide, FEV_1 = Forced expiratory volume in one second, CPI = composite physiologic index, MUC5B genotype = genotype for the MUC5B promoter polymorphism rs35705950. - indicates that the calculation was not applicable as there were zero subjects in that cluster. P-value for count data is from a chi-square 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.

	GSE38958 (n=65)			**GSE33566 (n=83)**			**GSE93606 (n=48)**			**All studies combined (n=196)**				
	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**P-value**	**Total n used**
n subjects in cluster	22	39	4	42	32	9	0	24	24	64	95	37		
Age (years) (mean, SD)	70·0 (6·3)	68·3 (7·9)	64·0 (2·7)	66·7 (9·8)	67·0 (14·1)	67·0 (12·1)	-	64·8 (5·9)	70·3 (8·8)	67·8 (8·9)	66·9 (10·2)	68·8 (9·4)	0·592	188
Male (%)	20 (91·0%)	30 (77·0%)	4 (100%)	32 (76·2%)	21 (65·6%)	3 (33·3%)	-	15 (62·5%)	16 (66·7%)	52 (81·3%)	66 (69·5%)	23 (62·2%)	0·091	196
European ancestry (%)	17 (81·0%)	29 (82·9%)	3 (75·0%)	NA	NA	NA	-	NA	NA	17 (81·0%)	29 (82·9%)	3 (75·0%)	0·883	60
Ever smoker (%)	NA	NA	NA	NA	NA	NA	-	15 (62·5%)	18 (78·3%)	NA	15 (62·5%)	18 (78·3%)	0·389	47
Death observed during study (%)	NA	NA	NA	NA	NA	NA	-	6 (25·0%)	16 (66·7%)	NA	6 (25·0%)	16 (66.7%)	**0·009**	48
FVC % predicted (median, IQR)	59·5 (19·5)	65·0 (24·0)	51·5 (7·8)	77·0 (36·0)	66·0 (46·0)	73·0 (17·5)	-	71·5 (27·7)	60·8 (24·1)	63.0 (35·0)	70·5 (30·1)	60·1 (23·4)	0·342	154
D_{LCO} % predicted (median, IQR)	34·5 (17·5)	49·0 (21·0)	28·5 (21·0)	65·0 (37·0)	66·0 (40·0)	30·0 (30·0)	-	38·1 (17·1)	36·6 (15·9)	35·0 (30·0)	45·0 (29·2)	34·4 (17·3)	**0·009**	133
FEV_1 % predicted (median, IQR)	NA	NA	NA	NA	NA	NA	-	74·9 (23·1)	65·4 (22·7)	NA	74·9 (23·1)	65·4 (22·7)	0·216	48
GAP index (mean, SD)	5·3 (1·3)	3·9 (1·3)	4·5 (1·3)	4·3 (1·5)	4·1 (1·6)	4·3 (3·1)	-	3·7 (1·8)	4·4 (1·6)	4·9 (1·4)	3·9 (1·5)	4·4 (1·7)	**0·006**	132
MUC5B genotype: GG (%)	NA	NA	NA	5 (29·4%)	6 (28·6%)	3 (60·0%)	-	5 (26·3%)	11 (50·0%)	5 (29·4%)	11 (27·5%)	14 (51·9%)	0·230	84
MUC5B genotype: GT (%)	NA	NA	NA	10 (58·8%)	14 (66·7%)	2 (40·0%)	-	12 (63·2%)	8 (36·4%)	10 (58·8%)	26 (65·0%)	10 (37·0%)		
MUC5B genotype: TT (%)	NA	NA	NA	2 (11·8%)	1 (4·8%)	0 (0%)	-	2 (10·5%)	3 (13·6%)	2 (11·8%)	3 (7·5%)	3 (11·1%)		

TABLE E4: The significantly enriched (q-value <0.05) biological processes for the 769 genes assigned to Cluster 1.

Biological process	Enrichment score	p-value	q-value
Mitochondrial ATP synthesis coupled electron transport	7.18	1.0×10^{-7}	7.8×10^{-4}
ATP synthesis coupled electron transport	7.12	1.2×10^{-7}	7.8×10^{-4}
Respiratory electron transport chain	6.88	1.4×10^{-7}	7.8×10^{-4}
Cellular respiration	5.95	1.3×10^{-6}	0.005
Oxidative phosphorylation	5.84	4.0×10^{-6}	0.012
Electron transport chain	5.56	4.3×10^{-6}	0.012
Homeostasis of number of cells	5.12	1.1×10^{-5}	0.024
Homeostatic process	4.54	1.7×10^{-5}	0.032

TABLE E5: The 20 most significantly enriched (q-value <0.05) biological processes for the 839 genes assigned to Cluster 2.

Biological process	Enrichment score	p-value	q-value
Cell activation	12.78	2.2×10^{-27}	3.7×10^{-24}
Immune system process	11.33	1.7×10^{-25}	1.4×10^{-21}
Leukocyte activation	11.76	2.4×10^{-23}	1.2×10^{-19}
Immune response	9.83	6.0×10^{-19}	2.5×10^{-15}
Regulation of immune system process	9.75	1.5×10^{-18}	4.9×10^{-15}
Regulated exocytosis	8.90	2.5×10^{-14}	6.9×10^{-11}
Response to stimulus	7.30	1.3×10^{-13}	3.1×10^{-10}
Defence response	8.16	1.6×10^{-13}	3.2×10^{-10}
Multi-organism process	7.74	1.9×10^{-13}	3.5×10^{-10}
Lymphocyte activation	8.73	4.5×10^{-13}	7.5×10^{-10}
Translational initiation	9.72	6.4×10^{-13}	9.1×10^{-10}
Symbiotic process	8.24	6.6×10^{-13}	9.1×10^{-10}
Interspecies interaction between organisms	8.02	1.6×10^{-12}	2.1×10^{-9}
Peptide metabolic process	8.31	1.9×10^{-12}	2.1×10^{-9}
Exocytosis	8.06	1.9×10^{-12}	2.1×10^{-9}
Peptide biosynthetic process	8.43	2.9×10^{-12}	2.9×10^{-9}
Translation	8.46	3.2×10^{-12}	3.1×10^{-9}
Regulation of biological quality	7.14	3.8×10^{-12}	3.5×10^{-9}
Myeloid leukocyte activation	8.09	4.1×10^{-12}	3.6×10^{-9}
Regulation of multicellular organismal process	7.20	5.0×10^{-12}	4.0×10^{-9}

TABLE E6: The 20 most significantly enriched (q-value <0.05) biological processes for the 784 genes assigned to Cluster 3.

Biological process	Enrichment score	p-value	q-value
Cell activation	20.78	1.3×10^{-60}	1.5×10^{-56}
Immune response	19.53	1.8×10^{-60}	1.5×10^{-56}
Leukocyte activation	20.87	3.3×10^{-59}	1.8×10^{-55}
Immune system process	18.04	1.6×10^{-57}	6.6×10^{-54}
Immune effector process	19.19	1.2×10^{-52}	4.0×10^{-49}
Myeloid leukocyte activation	20.63	1.7×10^{-52}	4.7×10^{-49}
Leukocyte activation involved in immune response	20.07	9.2×10^{-51}	2.2×10^{-47}
Cell activation involved in immune response	19.98	1.9×10^{-50}	3.9×10^{-47}
Neutrophil activation	20.19	1.0×10^{-48}	1.9×10^{-45}
Granulocyte activation	20.02	3.5×10^{-48}	5.7×10^{-45}
Neutrophil activation involved in immune response	19.55	4.0×10^{-46}	6.1×10^{-43}
Leukocyte degranulation	19.42	5.0×10^{-46}	6.8×10^{-43}
Neutrophil degranulation	19.43	1.3×10^{-45}	1.7×10^{-42}
Myeloid cell activation involved in immune response	19.21	1.5×10^{-45}	1.8×10^{-42}
Neutrophil mediated immunity	19.23	3.6×10^{-45}	3.9×10^{-42}
Myeloid leukocyte mediated immunity	18.99	1.1×10^{-44}	1.1×10^{-41}
Leukocyte mediated immunity	17.11	4.3×10^{-43}	4.2×10^{-40}
Secretion by cell	16.63	3.9×10^{-41}	3.5×10^{-38}
Export from cell	16.50	5.9×10^{-41}	5.2×10^{-38}
Defence response	15.95	1.2×10^{-40}	1.0×10^{-37}

TABLE E7: The 13 genes in the classifier. 'Up genes' refer to genes that were more highly expressed in the subjects for that cluster compared to the mean expression across all subjects, and 'down genes' refer to genes that were less highly expressed in the subjects in that cluster.

Cluster 1		Cluster 2		Cluster 3	
Up genes	Down genes	Up genes	Down genes	Up genes	Down genes
KCNK15	*RPF1*	*NOP58*		*CA4*	
SORBS1		*PSMA5*		*BCL2A1*	
HBB		*RASGRP1*		*UGCG*	
		IFI30			
		HLA-DRA			
		ATM			

TABLE E8: Coefficients of the multinomial logistic regression model fit using classification scores from the genes in the classifier. Note that Cluster 1 is the reference cluster and so the coefficients for this cluster are all zero and have been omitted.

Cluster	Intercept	Cluster 1 score	Cluster 2 score	Cluster 3 score
2	3.12	-9.75	8.87	1.66
3	-16.6	-11.92	-3.15	29.42

TABLE E9: Two-way tables comparing 'true' assignment of subjects from the discovery analysis (determined using COMMUNAL with 2,500 genes) to the reassignment of these subjects using the 13-gene cluster classifier.

		True cluster		
		Cluster 1	Cluster 2	Cluster 3
Classifier predicted cluster	Cluster 1	63	1	0
	Cluster 2	1	94	0
	Cluster3	0	0	37

TABLE E10: Pairwise comparisons showing the differences in survival over time between any two validation clusters, estimated using Cox proportional hazards models.

Reference cluster	**Alternate cluster**	**Hazard Ratio**	**95% CI**	**P-value**
Cluster 2	Cluster 1	3.80	1.78, 8.12	0.001
Cluster 2	Cluster 3	5.05	2.24, 11.35	9.1×10^{-5}
Cluster 1	Cluster 3	1.47	0.67, 3.22	0.341

TABLE E11: The agreement between the cluster classifier and SAMS when validation subjects were assigned to risk groups using each method.

GSE27957 (n=45)		Cluster classifier	
		High risk	Low risk
SAMS	High risk	13	2
	Low risk	5	25
GSE28042 (n=75)		Cluster classifier	
		High risk	Low risk
SAMS	High risk	17	12
	Low risk	19	27
Both datasets combined (n=120)		Cluster classifier	
		High risk	Low risk
SAMS	High risk	30	14
	Low risk	24	52

TABLE E12: Summary statistics from the Cox proportional hazards model adjusting for cluster, age, sex, ancestry, predicted forced vital capacity (FVC) and predicted diffusing capacity of the lung for carbon monoxide (DL_{CO}). OR = odds ratio, SE = standard error and CI = confidence interval.

Variable	OR	SE	P-value	95% CI
Cluster (high-risk cluster)	2.697	0.367	0.007	(1.315, 5.534)
Age (years)	1.006	0.020	0.748	(0.968, 1.046)
Sex (male)	5.720	0.752	0.020	(1.310, 24.969)
Ancestry (non-European)	1.099	0.608	0.876	(0.334, 3.619)
Predicted FVC	0.996	0.013	0.745	(0.971, 1.022)
Predicted DL_{CO}	0.967	0.013	0.008	(0.944, 0.991)

Additional Figures

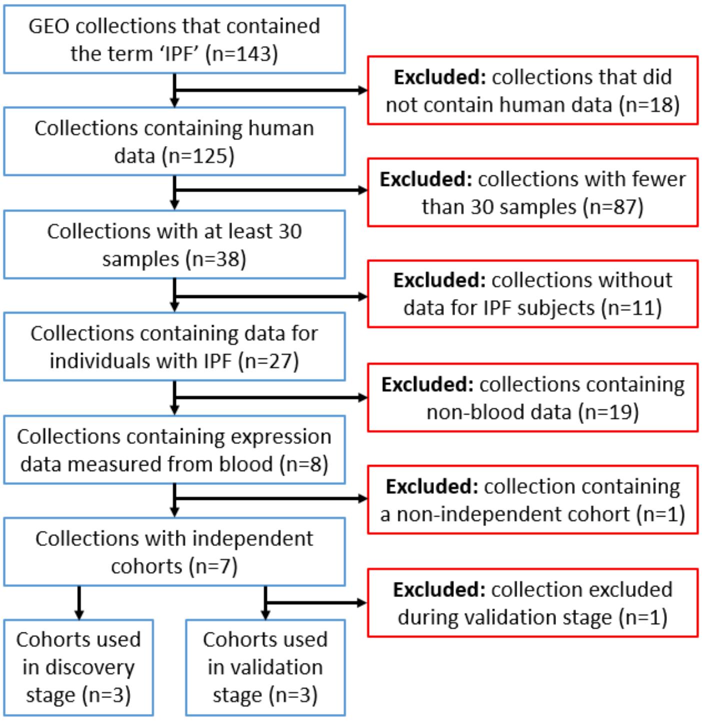

GEO collections that contained the term 'IPF' (n=143)
Excluded: collections that did not contain human data (n=18)
Collections containing human data (n=125)
Excluded: collections with fewer than 30 samples (n=87)
Collections with at least 30 samples (n=38)
Excluded: collections without data for IPF subjects (n=11)
Collections containing data for individuals with IPF (n=27)
Excluded: collections containing non-blood data (n=19)
Collections containing expression data measured from blood (n=8)
Excluded: collection containing a non-independent cohort (n=1)
Collections with independent cohorts (n=7)
Excluded: collection excluded during validation stage (n=1)
Cohorts used in discovery stage (n=3)
Cohorts used in validation stage (n=3)

FIGURE E1: Flow diagram showing the process used to systematically select publicly available IPF gene expression datasets from the Gene Expression Omnibus for use in this study.

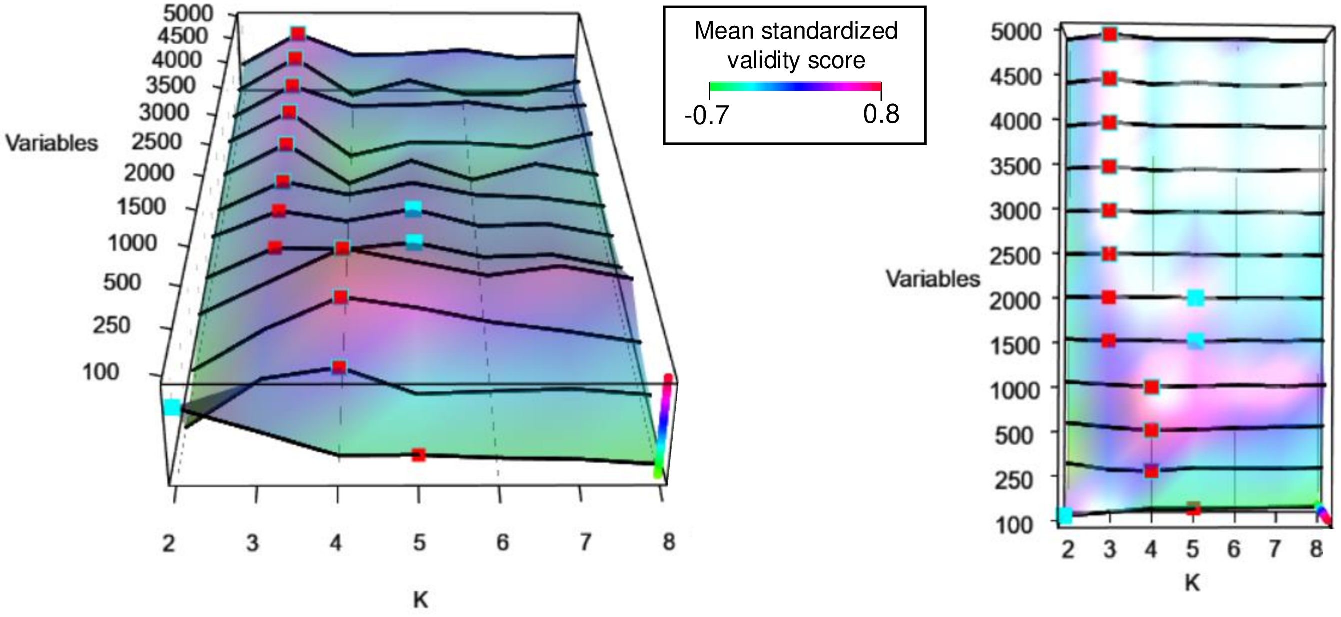
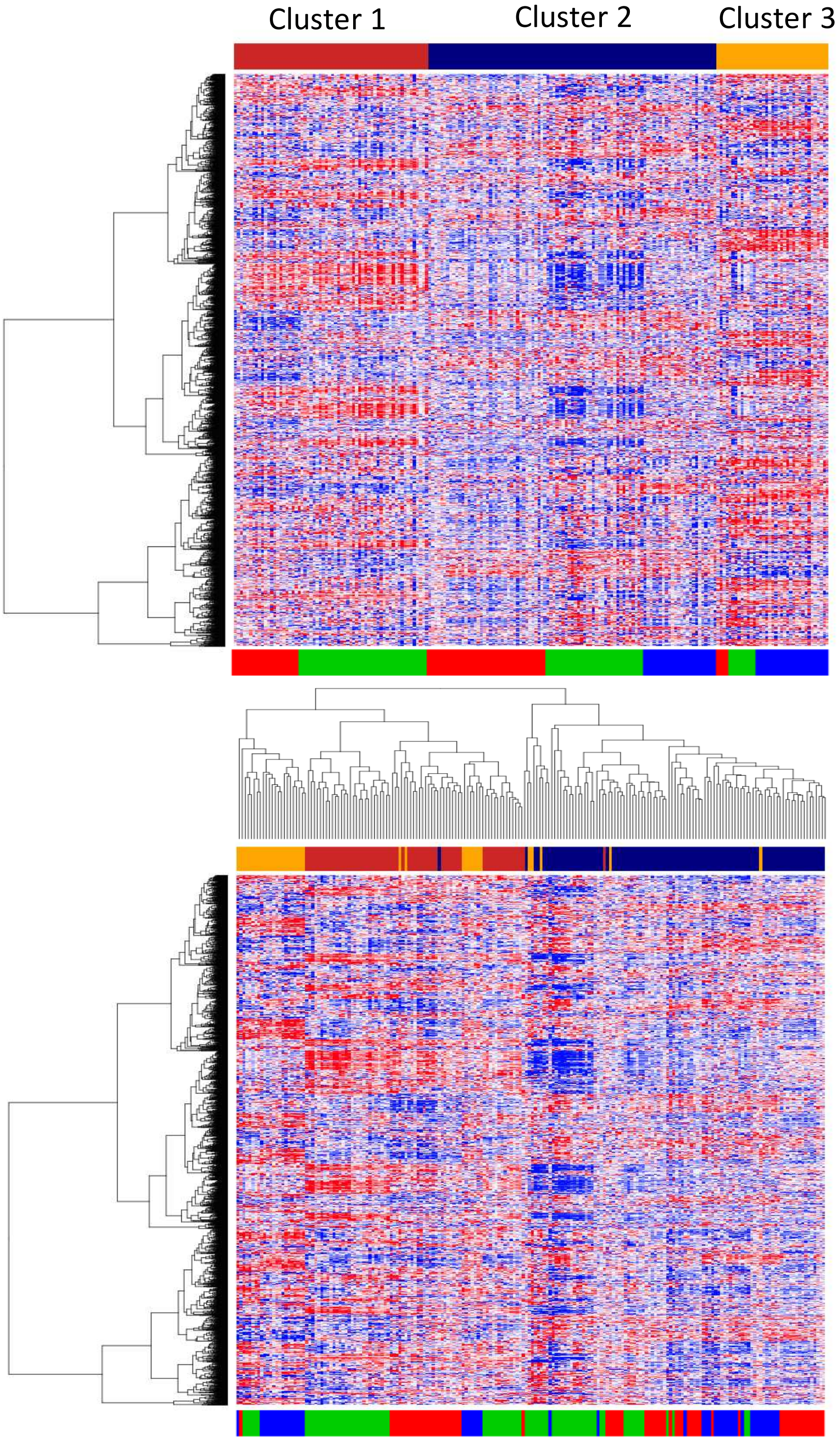


FIGURE E2: The 3D optimality map produced by COMMUNAL to identify the most robust number of clusters in the co-normalised data. A higher validity score indicates a better clustering assignment and stable optima are the points where the blue and red squares meet. In this map there are stable optima at K=4 from 250 genes to 1,000 genes, and at K=3 from 2,500 genes to 5,000 genes, as shown by the red and blue squares meeting. Despite the K=4 clustering assignment at 1,000 genes showing the highest mean standardized validity score of all tested clustering assignments, there were stable optima at K=3 clusters over a larger range of tested space, indicating a stronger biological signal. As such, K=3 was chosen as the optimal number of clusters in the pooled IPF dataset. The clustering at 2,500 genes and 3 clusters was chosen as the optimal clustering assignment, under the assumption that using the fewest number of genes has the least amount of redundant signal.



1 **FIGURE E3:** Heatmaps of gene expression for the clustered samples (x-axis) across the top 2,500 genes (y-axis), without hierarchical clustering of the samples (A) and with hierarchical clustering of the samples (B). Blue inside the heatmap indicates low expression and red indicates high expression. In both plots, the genes have been hierarchically clustered for presentation purposes, the bar above the plot shows the cluster that subject was assigned in to (red = cluster 1, blue = cluster 2 and yellow = cluster 3) and the bar below the plot indicates which original study the subject was in (red = GSE38958, green = GSE33566 and blue = GSE93606).

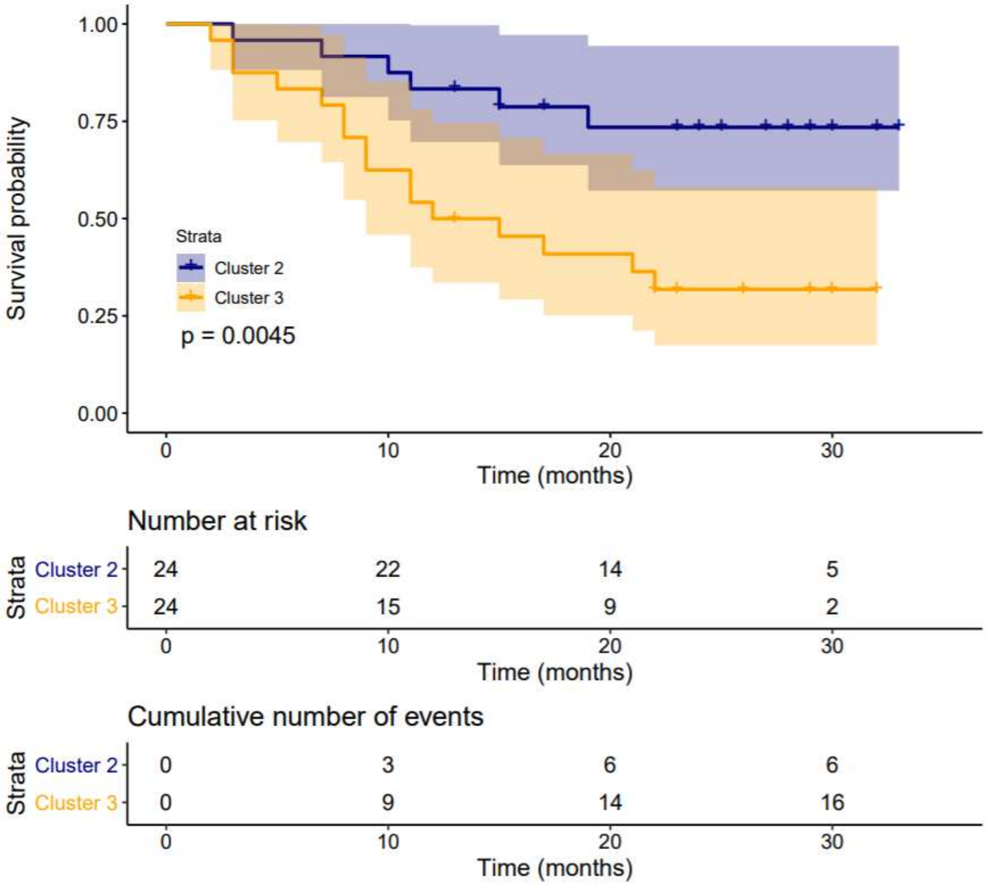


FIGURE E4: Kaplan-Meier curves and corresponding 95% confidence intervals showing survival over time for the subjects from study GSE93606, stratified by the cluster which they were assigned to in this study. The p-value shown on the plot is from a log-rank test testing the two curves for equality.

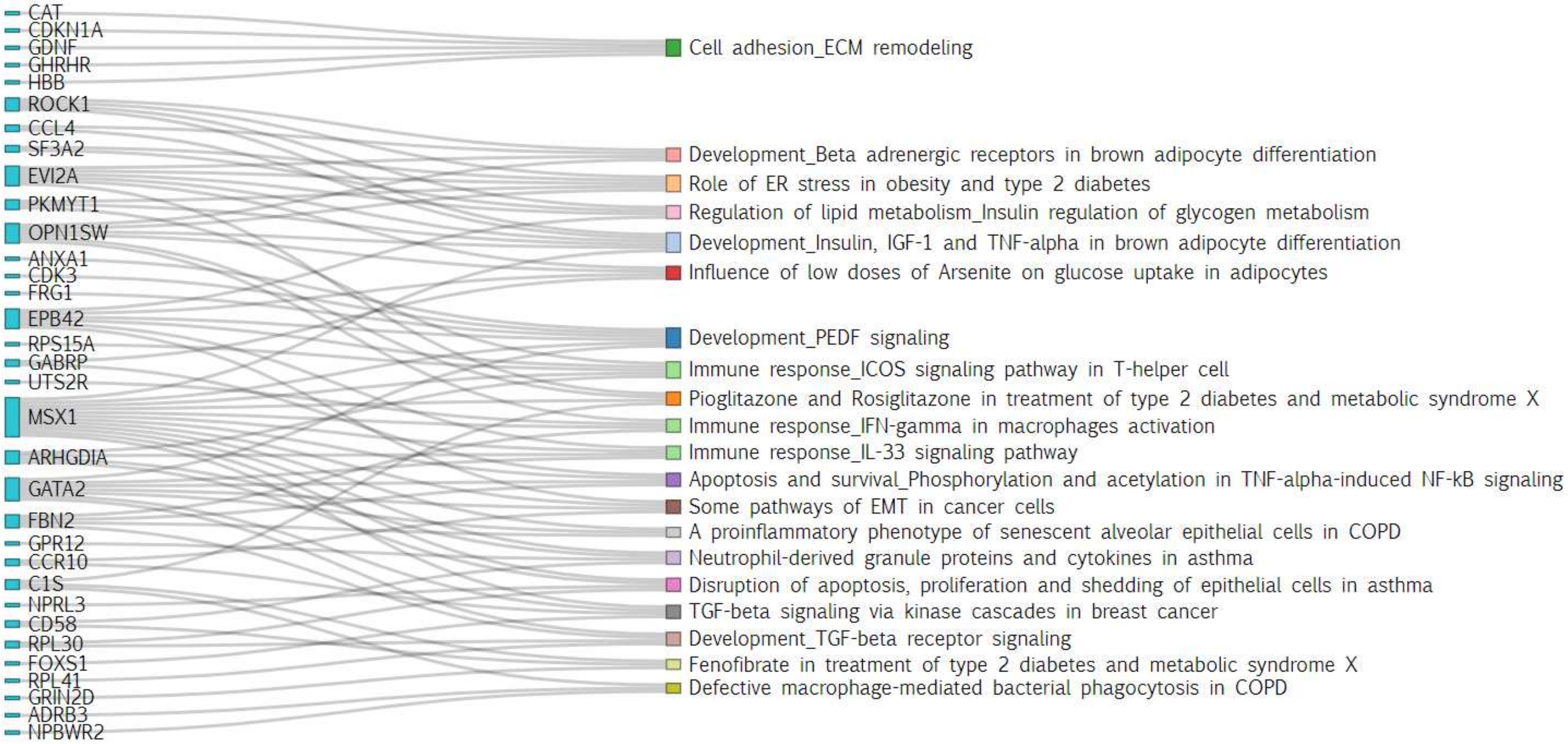

CAT
CDKN1A
GDNF
GHRHR
HBB
ROCK1
CCL4
SF3A2
EVI2A
PKMYT1
OPN1SW
ANXA1
CDK3
FRG1
EPB42
RPS15A
GABRP
UTS2R
MSX1
ARHGDIA
GATA2
FBN2
GPR12
CCR10
C1S
NPRL3
CD58
RPL30
FOXS1
RPL41
GRIN2D
ADRB3
NPBWR2
Cell adhesion_ECM remodeling
Development_Beta adrenergic receptors in brown adipocyte differentiation
Role of ER stress in obesity and type 2 diabetes
Regulation of lipid metabolism_Insulin regulation of glycogen metabolism
Development_Insulin, IGF-1 and TNF-alpha in brown adipocyte differentiation
Influence of low doses of Arsenite on glucose uptake in adipocytes
Development_PEDF signaling
Immune response_ICOS signaling pathway in T-helper cell
Pioglitazone and Rosiglitazone in treatment of type 2 diabetes and metabolic syndrome X
Immune response_IFN-gamma in macrophages activation
Immune response_IL-33 signaling pathway
Apoptosis and survival_Phosphorylation and acetylation in TNF-alpha-induced NF-kB signaling
Some pathways of EMT in cancer cells
A proinflammatory phenotype of senescent alveolar epithelial cells in COPD
Neutrophil-derived granule proteins and cytokines in asthma
Disruption of apoptosis, proliferation and shedding of epithelial cells in asthma
TGF-beta signaling via kinase cascades in breast cancer
Development_TGF-beta receptor signaling
Fenofibrate in treatment of type 2 diabetes and metabolic syndrome X
Defective macrophage-mediated bacterial phagocytosis in COPD

FIGURE E5: A Sankey diagram for Cluster 1 showing the genes that correspond to the 20 most significantly enriched biological pathways. The colour on the right hand side of the plot indicates the category of a particular pathway.

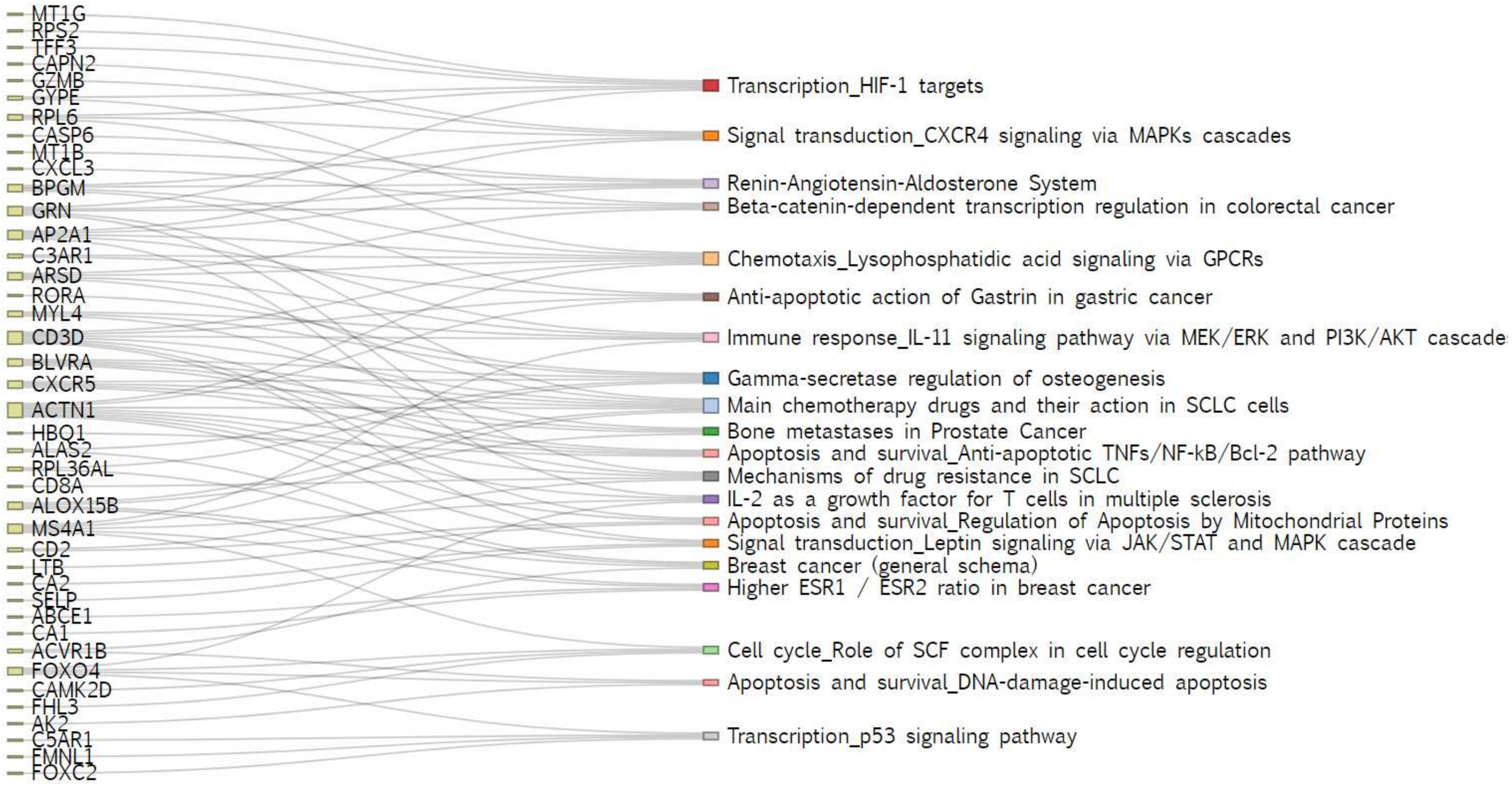

MT1G
RPS2
TFF3
CAPN2
GZMB
GYPE
RPL6
CASP6
MT1B
CXCL3
BPGM
GRN
AP2A1
C3AR1
ARSD
RORA
MYL4
CD3D
BLVRA
CXCR5
ACTN1
HBQ1
ALAS2
RPL36AL
CD8A
ALOX15B
MS4A1
CD2
LTB
CA2
SELP
ABCE1
CA1
ACVR1B
FOXO4
CAMK2D
FHL3
AK2
C5AR1
FMNL1
FOXC2
Transcription_HIF-1 targets
Signal transduction_CXCR4 signaling via MAPKs cascades
Renin-Angiotensin-Aldosterone System
Beta-catenin-dependent transcription regulation in colorectal cancer
Chemotaxis_Lysophosphatidic acid signaling via GPCRs
Anti-apoptotic action of Gastrin in gastric cancer
Immune response_IL-11 signaling pathway via MEK/ERK and PI3K/AKT cascade
Gamma-secretase regulation of osteogenesis
Main chemotherapy drugs and their action in SCLC cells
Bone metastases in Prostate Cancer
Apoptosis and survival_Anti-apoptotic TNFs/NF-kB/Bcl-2 pathway
Mechanisms of drug resistance in SCLC
IL-2 as a growth factor for T cells in multiple sclerosis
Apoptosis and survival_Regulation of Apoptosis by Mitochondrial Proteins
Signal transduction_Leptin signaling via JAK/STAT and MAPK cascade
Breast cancer (general schema)
Higher ESR1 / ESR2 ratio in breast cancer
Cell cycle_Role of SCF complex in cell cycle regulation
Apoptosis and survival_DNA-damage-induced apoptosis
Transcription_p53 signaling pathway

FIGURE E6: A Sankey diagram for Cluster 2 showing the genes that correspond to the 20 most significantly enriched biological pathways. The colour on the right hand side of the plot indicates the category of a particular pathway.

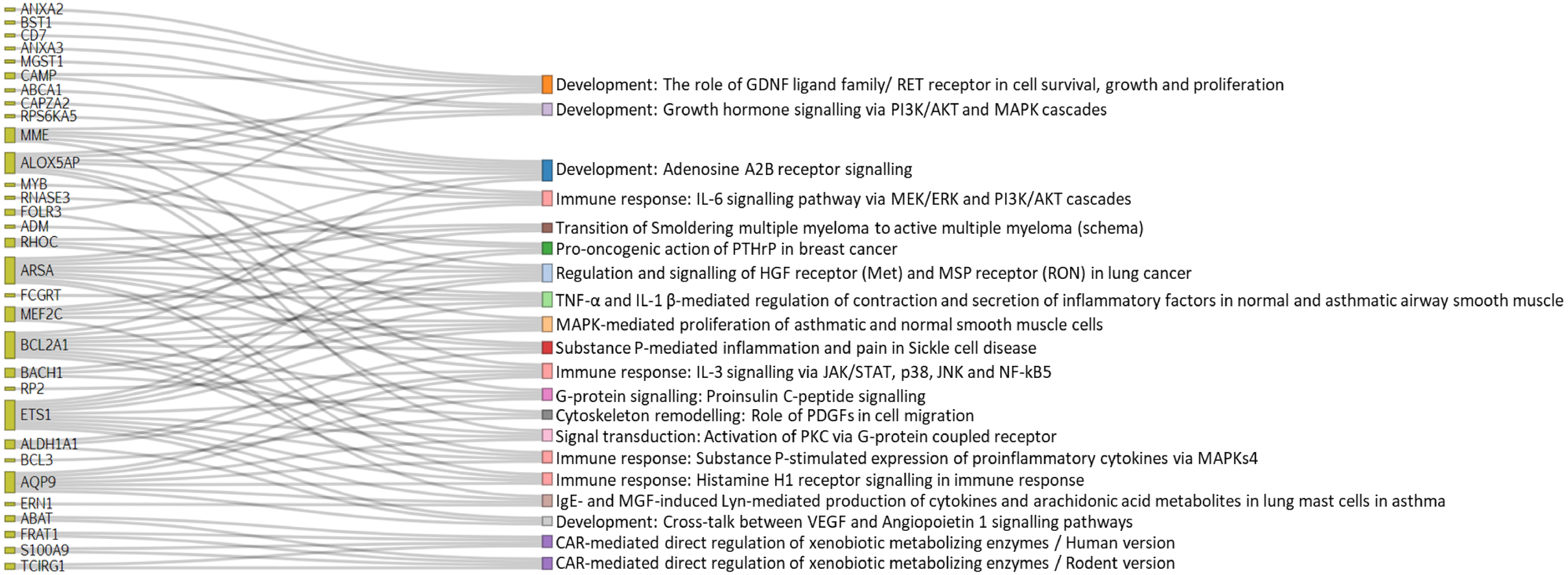

ANXA2
BST1
CD7
ANXA3
MGST1
CAMP
ABCA1
CAPZA2
RPS6KA5
MME
ALOX5AP
MYB
RNASE3
FOLR3
ADM
RHOC
ARSA
FCGRT
MEF2C
BCL2A1
BACH1
RP2
ETS1
ALDH1A1
BCL3
AQP9
ERN1
ABAT
FRAT1
S100A9
TCIRG1
Development: The role of GDNF ligand family/ RET receptor in cell survival, growth and proliferation
Development: Growth hormone signalling via PI3K/AKT and MAPK cascades
Development: Adenosine A2B receptor signalling
Immune response: IL-6 signalling pathway via MEK/ERK and PI3K/AKT cascades
Transition of Smoldering multiple myeloma to active multiple myeloma (schema)
Pro-oncogenic action of PTHrP in breast cancer
Regulation and signalling of HGF receptor (Met) and MSP receptor (RON) in lung cancer
TNF-α and IL-1 β-mediated regulation of contraction and secretion of inflammatory factors in normal and asthmatic airway smooth muscle
MAPK-mediated proliferation of asthmatic and normal smooth muscle cells
Substance P-mediated inflammation and pain in Sickle cell disease
Immune response: IL-3 signalling via JAK/STAT, p38, JNK and NF-kB5
G-protein signalling: Proinsulin C-peptide signalling
Cytoskeleton remodelling: Role of PDGFs in cell migration
Signal transduction: Activation of PKC via G-protein coupled receptor
Immune response: Substance P-stimulated expression of proinflammatory cytokines via MAPKs4
Immune response: Histamine H1 receptor signalling in immune response
IgE- and MGF-induced Lyn-mediated production of cytokines and arachidonic acid metabolites in lung mast cells in asthma
Development: Cross-talk between VEGF and Angiopoietin 1 signalling pathways
CAR-mediated direct regulation of xenobiotic metabolizing enzymes / Human version
CAR-mediated direct regulation of xenobiotic metabolizing enzymes / Rodent version

FIGURE E7: A Sankey diagram for Cluster 3 showing the genes that correspond to the 20 most significantly enriched biological pathways. The colour on the right hand side of the plot indicates the category of a particular pathway.

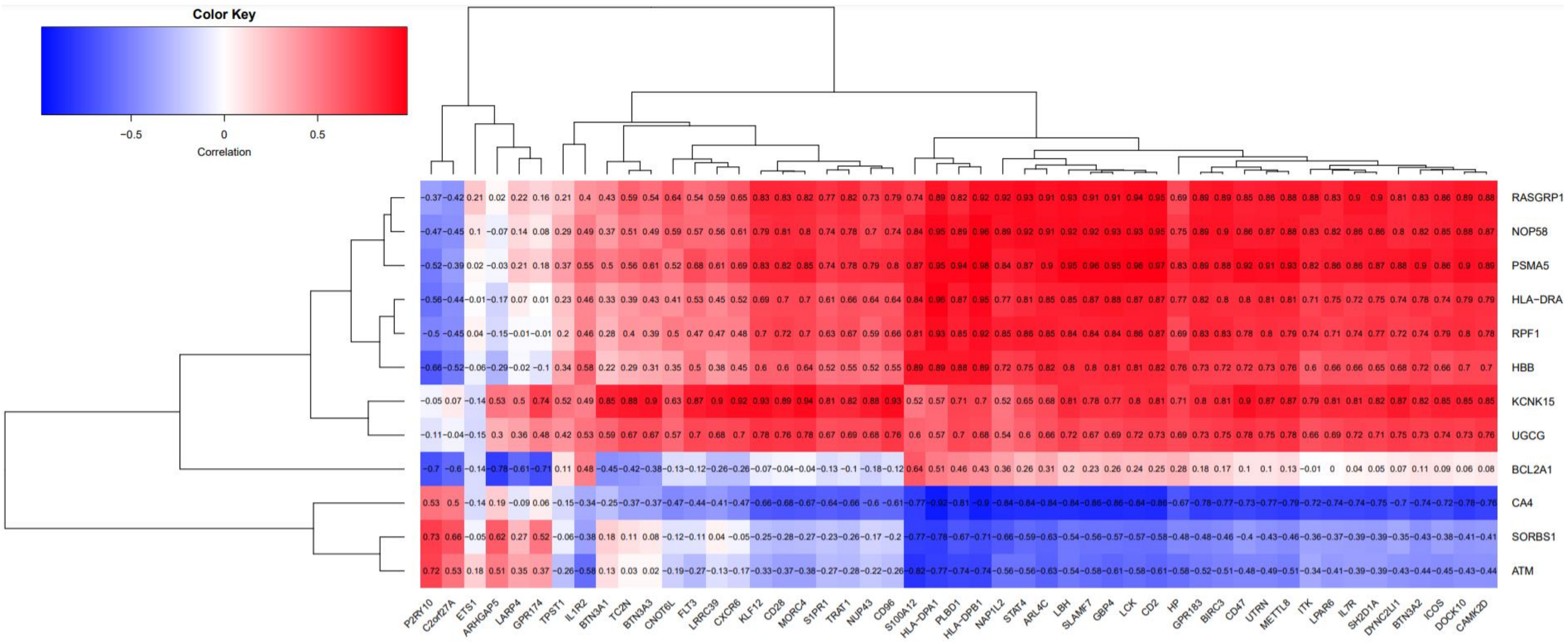


FIGURE E8: A heatmap showing the Pearson correlation between the genes in the classifier (y-axis) and the genes used by SAMS (x-axis). The correlation was calculated using the data from the IPF patients in the three validation cohorts (total n=194) for all genes that had complete data (12/13 genes for the classifier and 49/52 genes for SAMS). Both sets of genes were clustered using hierarchical clustering for presentation purposes.

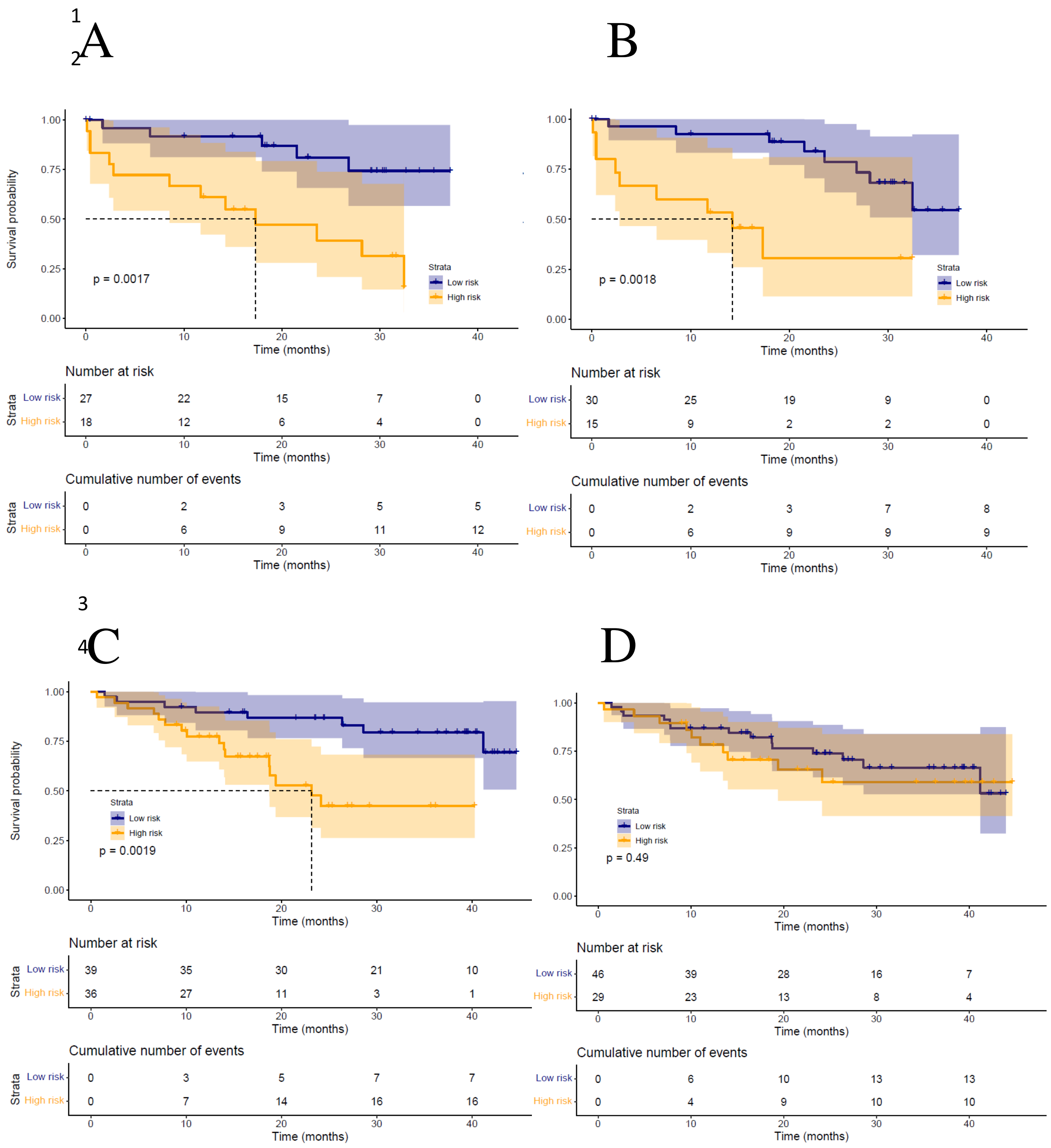


FIGURE E9: Survival over time for the IPF subjects in the validation datasets GSE27957 and GSE28042, stratified by predicted risk group. A) Survival of IPF cases from GSE27957 with risk predicted by our 13 gene classifier. B) Survival of IPF cases from GSE27957 with risk predicted by SAMS. C) Survival of IPF cases from GSE28042 with risk predicted by our 13 gene classifier. D) Survival of IPF cases from GSE28042 with risk predicted by SAMS. 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.

Cluster analysis of transcriptomic datasets to identify endotypes of Idiopathic Pulmonary Fibrosis – online data supplement

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

Contents

Additional text

Systematic selection of publicly available datasets

We performed our systematic search in March 2020 to select the datasets that were suitable for inclusion in the study (Figure E1). We required multiple sets of transcriptomic data from independent cohorts. We searched the Gene Expression Omnibus (GEO) (1) for all collections that contained the term 'IPF', excluding any that did not contain human samples. We restricted the search to collections with at least 30 samples as this allowed for inclusion of the largest datasets with the most IPF cases and healthy control subjects, which are the datasets that were the most likely to successfully co-normalise due to the higher counts of healthy control subjects. We did not restrict the search by platform. Each of the remaining collections were then reviewed to assess whether they contained data for IPF cases. All collections that did not contain data for IPF subjects were excluded.

For a successful co-normalisation and meaningful clustering results, we were required to choose an optimal tissue/cell type to use for the analysis. After reviewing the IPF datasets on GEO, we chose whole blood as our optimal tissue/cell type. There were three main reasons for this. Firstly, there were several relatively large whole blood datasets available on GEO and these would have provided the largest sample size and greatest statistical power for the study compared to other tissue types. Secondly, we required multiple datasets that contained data for healthy controls in addition to the IPF patients (so that the data could be co-normalised using COCONUT) and the whole blood datasets fulfilled this requirement. Thirdly, the accompanying clinical data for the whole blood datasets was far more comprehensive than for other tissue types, such as whole lung. This clinical data was vital to the study as it was required for the characterisation of the clusters in both the discovery and validation stages. So, all GEO collections containing expression data measured from a non-blood tissue/cell type were excluded.

As multiple transcriptomic datasets were to be combined, it was important to check for the presence of common individuals across cohorts, which would have meant that the cohorts were not independent and could have biased the results of the study. To this end, the subjects in each collection were checked for unique study identification codes. Using these, we found that two of the blood collections, GSE132607 (n=74) and GSE85268 (n=68), both contained subjects from the Correlating Outcomes With Biochemical Markers to Estimate Time-progression in Idiopathic Pulmonary Fibrosis (COMET) study (ClinicalTrials.gov identifier: NCT01071707). There were a large number of IPF subjects in common between the two cohorts (n=58) and so we excluded the GSE85268 dataset as it was the collection with fewer IPF subjects.

The seven remaining collections of data were uploaded by research groups from across the USA (including the University of Virginia, Yale University, the University of Nevada and the University of Colorado) and the UK (Imperial College London). GSE27957 and GSE28042 were uploaded by the Kaminski Lab in Yale. These two collections were both used in the same study (2), where GSE27957 was used as discovery data and GSE28042 was used as independent replication data. Similarly, the data found in GSE133298 and GSE132607 were uploaded by researchers at the University of Virginia and were used as independent cohorts in the same study (unpublished as of October 2020, both collections uploaded to GEO in September 2019). All remaining collections were uploaded by separate research groups and no additional evidence of common subjects across cohorts was found so the seven cohorts of IPF subjects were deemed independent. However, the possibility that subjects could be common in two or more studies cannot be ruled out.

The human biological samples were sourced ethically and their research use was in accord with the terms of the informed consents under an institutional review board/ethical committee (IRB/EC)-approved protocol.

Assignment of datasets to discovery and validation stages

All cohorts included in the discovery stage must have contained healthy controls in order to enable the data co-normalization step. Four of the seven selected blood datasets contained data for healthy controls. We used the three with the greatest number of controls in discovery as these were the most likely to successfully co-normalize. The four remaining datasets were reserved for use in the validation stage. One dataset (GSE133298) was excluded during the validation stage as not all of the genes that were required to fully apply the classifier were present in the dataset.

Discovery stage studies

GSE38958: This dataset originates from an American observational study (3) that was investigating the relationship between sphingosine-1-phosphate lyase and pulmonary fibrosis. IPF cases were recruited from the University of Chicago. The authors studied gene expression data from peripheral blood mononuclear cells of IPF subjects (n=70) and compared this to gene expression from healthy controls (n=45).

GSE33566: This dataset contained data for 123 IPF subjects and 30 healthy controls. A subset of this data was used in an American observational study (4), where the authors hypothesised that a peripheral blood biomarker for IPF would be able to identify the disease in its early stages and allow for disease progression to be monitored. The IPF cases were recruited through the Interstitial Lung Disease or the Familial Pulmonary Fibrosis Programs conducted at National Jewish Health and Duke University. In the study, 40 IPF subjects were split into groups based on their predicted FVC and D_{LCO}, then the authors looked for differentially expressed genes between groups.

GSE93606: This dataset contained data from a British prospective cohort study (5) (n=57 IPF subjects and n=20 healthy age, sex and smoking history matched controls) which had the objective of examining host-microbial interactions in IPF subjects over time. IPF cases were prospectively recruited from the Interstitial Lung Disease Unit at the Royal Brompton Hospital, London, within six months of their initial diagnosis. The study was approved by the local research ethics committee (reference numbers 10/H0720/12 and 12/LO/1034). In this study, gene expression data from peripheral blood and lung function measurements were collected at multiple time points. However, only baseline gene expression and lung function data was used in our study. IPF patient survival was also recorded up to a maximum follow-up time of 34 months.

Validation stage studies

GSE132607: This dataset originates from a study (unpublished as of March 2022) which aimed to develop a predictor of FVC progression by studying gene expression differences in 74 IPF subjects over time. The subjects included in this analysis were participants in COMET-IPF (Correlating Outcomes with biochemical Markers to Estimate Time-progression in Idiopathic Pulmonary Fibrosis), a prospective, observational study correlating biomarkers with disease progression. All IPF cases had been recruited in to this study within four years of their initial IPF diagnosis.

GSE27957 and **GSE28042**: both datasets originate from the same study (6), where the data in GSE27957 (n=45 IPF subjects) was used in discovery and the data in GSE28042 (n=75 IPF subjects) was used as independent validation data. Individuals with IPF from the GSE27957 dataset were recruited from the University of Chicago and the individuals with IPF from the GSE28042 dataset were recruited from the University of Pittsburgh. In brief, the authors used these cohorts to develop a 52-gene signature that had the ability to predict transplant-free survival in IPF subjects.

Data pre-processing

In each discovery dataset, probes that did not map to a gene were removed. In the instance where multiple probes mapped to the same gene, only the probe with the greatest mean expression was included in the analysis. Each dataset was then quantile normalised to reduce any technical differences between the gene probes within a study. Following this, each dataset was scaled so that all expression data was on the $\log_2$ scale and thus in a consistent form prior to co-normalisation. Genes were matched across studies based on their gene symbols.

Data co-normalisation using COCONUT

We used COmbat CO-Normalization Using conTrols (COCONUT) (7) (in R v4.0.0 and the 'COCONUT' package) to reduce the technical differences between the three discovery transcriptomic datasets, therefore enabling a cluster analysis to be performed on the pooled, co-normalized data. COCONUT is an unbiased co-normalisation method which assumes that all healthy controls across studies come from the same statistical distribution. It uses the healthy controls in each study to calculate correction factors that remove the technical differences in the data for the diseased subjects, without bias to the number of disease cases present. The method is adapted from the ComBat empiric Bayes normalization method (8), which is often used to adjust for batch effects within a study.

As COCONUT makes the assumption that all healthy controls come from the same background statistical distribution, we tested for significant differences in clinical and demographic traits between the healthy controls in each study, where possible. Clinical and demographic characteristics of the healthy controls were compared using chi-square tests for count data and analysis of variance for non-skewed continuous data.

Data for each study was input into COCONUT by providing a gene expression matrix (on the $\log_2$ scale) of common genes against subjects. These were accompanied by an indicator variable that showed which subjects were cases and which were controls. Following the co-normalisation, we removed all healthy control subjects from further analysis. Plots of the first two principal components of the transcriptomic data before and after COCONUT were used to evaluate the efficacy of the co-normalisation.

Clustering using COMMUNAL

In this study, we ran COMMUNAL using consensus clustering versions of two algorithms, K-means clustering and partitioning around medoids (PAM). Five different metrics were used to assess the validity of the clustering for different numbers of clusters and genes. These were: the gap statistic, connectivity, average silhouette width, the G3 metric, and Pearson's gamma coefficient. We ranked the genes in order of variance, with the 'top' 100 genes referring to the 100 genes with the greatest variance. We then applied the COMMUNAL algorithm using a range of input genes from the top 100 to the top 5,000. The genes with the greatest variance were used as these were the most likely to be informative, so as to minimise the number of non-informative genes and increase the signal-to-noise ratio.

The samples that were not assigned into the same cluster by the COMMUNAL clustering algorithms were labelled 'unclustered'. Since the intention was to use the clustered data to create a classifier and classifiers trained on data with fewer errors are more robust, these uncertain samples were removed from further analysis to improve the accuracy of the classifier.

The results were visualised in the form of a 3-dimensional (3D) map (Figure E2), which we used to select the optimal number of clusters in the data, as well as the optimal number of genes to use in the clustering. The map shows the mean of standardized values of each validity measure across the entire tested space. On the 3D map, blue squares indicate a potentially optimal clustering at a certain number of genes by finding the assignment where the mean combined validation metric is greatest. The absolute maximum number of clusters for any consensus subset is marked with a red square. The points where the blue and red squares overlap indicate stable optima. If stable optima at a particular number of clusters are observed over most of the tested space, this indicates the presence of a strong, consistent biological signal at this number of clusters.

In Figure E2 there are stable optima at K=4 from 250 genes to 1,000 genes, and at K=3 from 2,500 genes to 5,000 genes, as shown by the red and blue squares meeting. Despite the K=4 clustering assignment at 1,000 genes showing the highest mean standardized validity score of all tested clustering assignments, there were stable optima at K=3 clusters over a larger range of tested space, indicating a stronger biological signal. As such, K=3 was chosen as the optimal number of clusters in the pooled IPF dataset. The clustering at 2,500 genes and 3 clusters was chosen as the optimal clustering assignment, under the assumption that the assignment with the fewest number of genes (out of those with stable optima at K=3) has the least amount of redundant signal.

Comparison of phenotypic traits across clusters

We characterised the clusters by comparing the clinical and demographic traits of the subjects that were assigned to each cluster. This was done for each phenotypic trait that was reported in at least one discovery cohort and one validation cohort. The statistical significance of the phenotypic differences across clusters was evaluated for all studies combined using a chi-square test for count data, an analysis of variance to compare means for non-skewed continuous data and a Kruskal-Wallis rank sum test to compare medians for skewed continuous data. For traits in the form of time-to-event data, Kaplan-Meier plots were used to approximate and visualise the survival function for these variables. Further, Cox proportional-hazards (PH) models were fit with cluster as the sole independent variable and the time to the event as the response variable.

Gene enrichment analysis

First, we assigned each of the 2,500 genes used in the optimal COMMUNAL clustering assignment to the cluster in which its expression was most different to its expression in the other two clusters, as this suggests that that gene was contributing to the identity of that cluster. 814 genes were assigned to Cluster 1, 866 were assigned to Cluster 2 and 820 were assigned to Cluster 3.

We then performed multiple ANOVA tests (one for each cluster) for each gene, each comparing the expression of that gene in subjects within a given cluster against the expression of subjects in both other clusters. Each gene was then assigned to the cluster in which it had the lowest ANOVA p-value. One benefit of this approach is that the ANOVA tests allowed for filtering based on statistical significance; a nominal p-value significance threshold of 0.05 was introduced and genes whose lowest ANOVA p-value was greater than this threshold were removed. The rationale for the introduction of this filtering step was that removing genes that were not associated with any cluster would reduce noise and strengthen the gene enrichment analysis for each cluster. The threshold for statistical significance was kept at a nominal level as a correction for all 7,500 ANOVA tests would have likely left too few genes assigned to each cluster to successfully perform the enrichment analysis. After the removal of the genes that were not at least nominally associated to any cluster, there were 769 genes assigned to Cluster 1, 839 assigned to Cluster 2 and 784 assigned to Cluster 3.

Then, gene enrichment analysis was performed separately on the three resulting gene lists using R v.4.0.0 and the in-house package ‘metabaser’ (database v20.3, package v4.2.3). This was used to search databases of gene ontology terms for statistically overrepresented *biological processes* and *biological pathways*. At the time that the analysis was performed, there were 17,552 *biological processes* and 12,222 *biological pathways* in the database accessed by metabaser. metabaser reports ‘q-values’, which are p-values that have been adjusted for multiple tests using the false-discovery rate. Gene ontology terms with q-value < 0.05 were deemed statistically significant. Sankey plots were used to show which of the genes that were assigned to each cluster corresponded to the 20 most significantly enriched *biological pathways* (see Figure 3).

Additionally, the gene lists of each cluster were searched for the presence of the nearest gene for any of the 14 variants that were genome-wide significant in Allen et al. (9), the largest genome-wide association study meta-analysis of IPF susceptibility to-date. The 14 genes were as follows: *AKAP13*, *ATP11A*, *DEPTOR*, *DPP9*, *DSP*, *FAM13A*, *LRRC34*, *IVD*, *KIF15*, *MAD1L1*, *MAPT*, *MUC5B*, *TERC* and *TERT*. Following this, enrichment analysis was performed on the genes of each cluster to investigate whether those genes were statistically overconnected (in terms of direct gene regulation) to any of the IPF-associated genes from Allen et al. (2020). If the genes that were assigned to a particular cluster were found to be overconnected to one or more of the IPF-associated genes listed above (say the exact number of overconnected IPF-associated genes is N), then a hypergeometric test was performed to approximate the statistical significance of the finding that N out of the 14 IPF-associated genes were present within the list of overconnected genes for that cluster.

None of the 14 suspected IPF susceptibility genes from Allen et al. were assigned to Cluster 1, nor were they statistically overconnected to the genes that were assigned to this cluster. *FAM13A* was one of the genes that was assigned to Cluster 2, though it did not belong to any of the top 20 significantly enriched biological pathways. Additionally, the genes in Cluster 2 were statistically overconnected to five other IPF-associated genes. These were: *AKAP13*, *DSP*, *LRRC34*, *MAPT* and *TERT*. The hypergeometric p-value was calculated to be 0.020, indicating that it is significant that five IPF-associated genes were overconnected to the genes Cluster 2 and this is more than would be expected due to random chance. None of the IPF-associated genes from Allen et al. were found in the gene list for Cluster 3, although four were found to be statistically overconnected to the genes in this cluster. These were as follows: DSP, MAD1L1, MAPT and TERT. The statistical significance of this was approximated to be P=0.008 using a hypergeometric test, again indicating that this was significantly more than would be expected under random chance.

Developing the gene expression-based cluster classifier

Classification is a method of supervised machine learning that uses a correctly labelled training dataset to predict which category new observations belong in.

To determine the optimal genes to include in the classifier for the IPF data, we used an iterative algorithm which performed a greedy forward search for each cluster separately to determine the optimal combination of genes to differentiate between subjects in that cluster vs all other clusters. This was done by calculating receiver operating characteristic curves for each combination of genes and selecting the combination of genes which maximised the area under the curve (AUC). In an effort to prevent the classifier from being overfit to the discovery data, a threshold was implemented to stop the algorithm once an AUC of 0.99 had been reached. Each gene was labelled as either overexpressed or underexpressed based on whether the average expression of that gene was greater in the subjects from that particular cluster compared to the average expression across all subjects.

Making predictions with the classifier was a two-stage process. First, each subject was given a classification score for each cluster. This score was calculated as the geometric mean of the overexpressed genes for that cluster minus the geometric mean of the underexpressed genes. These scores were mean centred around zero and scaled to reflect a Z-score (i.e. standard deviation equal to 1). Ideally, subjects that belonged to a certain cluster should have had a high classification Z-score for that cluster and low classification Z-scores for the other clusters.

Then, we used the classification Z-scores to fit a multinomial logistic regression model, with cluster as the independent categorical variable and the Z-scores from each cluster as the dependent variables. This model had the ability to take data from new IPF subjects and predict which cluster they were each most likely to belong in, using only expression data from the optimal genes in the classifier. Importantly, the classifier does not use absolute levels of gene expression in order to make predictions, but instead utilizes relative gene expression between subjects. This meant that the classifier could be applied to a cohort of IPF cases (from the same study) without first requiring the removal of technical effects, which allowed for the use of validation datasets that did not contain data for healthy controls.

We tested the prediction accuracy of the classifier by using it to reassign all of the IPF subjects in the discovery datasets.

Risk classification using the classifier

Each of the IPF subjects in the two validation studies for which survival data was available, GSE27957 (n=45) and GSE28042 (n=75), were assigned into one of the three clusters using the 13 gene classifier. As significant differences in survival were observed between clusters 1 and 2 and 2 and 3, but not between clusters 1 and 3 (Table E9), we used assignment to clusters 1 and 3 to define high risk individuals and assignment to cluster 2 as low risk.

Risk classification using SAMS

Each of these individuals were also classed as high-risk or low-risk using SAMS (2). 7 of the 52 genes used by SAMS were expected to be more highly expressed in high risk cases than low risk cases (‘up genes’). Likewise, the remaining 45 genes were expected to be less highly expressed in high risk cases than low risk cases (‘down’ genes). The method that SAMS used to predict risk is as follows:

1. For each gene, the geometric mean of the expression for that gene across all subjects was calculated. This value represents the average level of expression for that gene across the whole cohort. It was then subtracted from the gene expression of that gene for each subject so that positive values represented subjects that had increased expression of that gene compared to the average and negative values represented subjects that had decreased expression compared to the average.

2. For each subject, the proportion of the 7 'up genes' that were overexpressed was calculated. Similarly, the proportion of the 45 'down genes' that were less highly expressed than average was calculated. So, if a subject had 4 'up genes' that were greater than the average and 30 'down genes' that were lower than the average, these proportions would have been 0.571 and 0.667 respectively.

3. For each subject, the sum of the geometric mean normalised expression data was summed up for the 'up genes' that were more highly expressed than average. Then the sum of the geometric mean normalised expression data was summed up for the 'down genes' that were less highly expressed than average. So, for example, for the subject above who had 4 of the 7 'up genes' that were more highly expressed than the average, say with expression values 0.185, 0.553, 0.123 and 1.003 for these four genes, the sum would have been 1.864. The sum for the 'down genes' must always be negative, for example say that this sum for the subject above was -7.645.

4. The proportion of the 'up genes' calculated in step 2 was multiplied by the sum for the 'up genes' calculated in step 3 to produce the 'up score' for each subject. So, for the example subject above, their up score would have been 0.571×1.864 = 1.064. A 'down score' for each subject was also calculated by multiplying their proportion of down genes by their down sum from step 3. For our example subject, this would have been 0.667× -7.645=-5.099.

5. Subjects with up scores greater than the median value and down scores lower than the median value were classed as 'high risk', while all other subjects were classed as 'low risk'.

This was done separately for each cohort and by using data from as many of the 52 genes as were measured in the datasets; 51/52 (98·1%) genes in the SAMS signature were present in GSE27957 and 50/52 (96·2%) were present in GSE28042. Two-way tables were used to compare agreement between the two methods.

Comparing prognostic methods using survival analysis

Kaplan-Meier plots were used to visualise the survival over time for the validation subjects in each risk group under each method. In both cases, the log-rank test was used to test the survival curves of each risk group for equality. Univariate Cox proportional-hazards models were fit to the data with risk group as the sole covariate and time-to-death as the outcome of interest. In both cases, the low-risk group was used as the reference group. The Concordance index (C-index), the equivalent of the area under the curve (AUC) for a receiver operating characteristic (ROC) curve, and the p-values from the log-rank test were used to assess which method performed best at assigning the IPF subjects to the correct risk group and therefore predicting survival.

Following this, multivariate Cox proportional-hazards models were used to assess whether the predictions made by each method were significant predictors of mortality in the validation datasets whilst adjusting for age, sex, ancestry, FVC and DL_{CO}. We used the likelihood ratio test and C-index to assess whether either of the two methods of risk prediction led to a significant increase in predictive ability over a Cox PH model containing only age, sex, ancestry, FVC and DL_{CO}.

References

(1) Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002;30:207-210.

(2) Herazo-Maya JD, Sun J, Molyneaux PL, Li Q, Villalba JA, Tzouvelekis A, Lynn H, Juan-Guardela BM, Risquez C, Osorio JC. Validation of a 52-gene risk profile for outcome prediction in patients with idiopathic pulmonary fibrosis: an international, multicentre, cohort study. *The Lancet Respiratory Medicine* 2017;5:857-868.

(3) Huang LS, Berdyshev EV, Tran JT, Xie L, Chen J, Ebenezer DL, Mathew B, Gorshkova I, Zhang W, Reddy SP. Sphingosine-1-phosphate lyase is an endogenous suppressor of pulmonary fibrosis: role of S1P signalling and autophagy. *Thorax* 2015;70:1138-1148.

(4) Yang IV, Luna LG, Cotter J, Talbert J, Leach SM, Kidd R, Turner J, Kummer N, Kervitsky D, Brown KK. The peripheral blood transcriptome identifies the presence and extent of disease in idiopathic pulmonary fibrosis. *PloS one* 2012;7:e37708.

(5) Molyneaux PL, Willis-Owen SA, Cox MJ, James P, Cowman S, Loebinger M, Blanchard A, Edwards LM, Stock C, Daccord C. Host–microbial interactions in idiopathic pulmonary fibrosis. *American journal of respiratory and critical care medicine* 2017;195:1640-1650.

(6) Herazo-Maya JD, Noth I, Duncan SR, Kim S, Ma S, Tseng GC, Feingold E, Juan-Guardela BM, Richards TJ, Lussier Y. Peripheral blood mononuclear cell gene expression profiles predict poor outcome in idiopathic pulmonary fibrosis. *Science translational medicine* 2013;5:205ra136.

(7) Sweeney TE, Wong HR, Khatri P. Robust classification of bacterial and viral infections via integrated host gene expression diagnostics. *Science translational medicine* 2016;8:346ra91.

(8) Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007;8:118-127.

(9) Allen RJ, Guillen-Guio B, Oldham JM, Ma S, Dressen A, Paynton ML, Kraven LM, Obeidat M, Li X, Ng M. Genome-wide association study of susceptibility to idiopathic pulmonary fibrosis. *American journal of respiratory and critical care medicine* 2020;201:564-574.

Additional Tables

TABLE E1: Information about the transcriptomic data in the discovery datasets and the platform used in each study.

GEO accession	**GSE38958**	**GSE33566**	**GSE93606**
Microarray platform	Affymetrix Human Exon 1.0 ST Array	Agilent-014850 Whole Human Genome Microarray	Affymetrix Human Gene 1.1 ST Array
Number of gene probes	44,280	32,850	33,297
Number of unique genes	17,256	12,171	20,254

TABLE E2: Comparison of the age and sex of the healthy controls in each discovery stage study. Data are presented as count (percentage) or mean (standard deviation, SD). P-value for count data is from a chi-square test and the test comparing means is analysis of variance.

	GSE38958	**GSE33566**	**GSE93606**	**P-value**	**n used**
Number of healthy controls	45	30	20		
Age (years, SD)	69·3 (9·3)	62·4 (14·3)	66·0 (10·6)	0.187	83
Sex (% male)	27 (60·0%)	14 (46·7%)	12 (60·0%)	0.477	95

TABLE E3: Comparison of clinical and demographic traits of clustered discovery subjects by study and for all studies combined. Data are presented as count (percentage), mean (standard deviation, SD) or median (interquartile range, IQR). NA = data not available, FVC=Forced vital capacity, D_{LCO} = Diffusing capacity for carbon monoxide, FEV_1 = Forced expiratory volume in one second, CPI = composite physiologic index, MUC5B genotype = genotype for the MUC5B promoter polymorphism rs35705950. - indicates that the calculation was not applicable as there were zero subjects in that cluster. P-value for count data is from a chi-square 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.

	GSE38958 (n=65)			**GSE33566 (n=83)**			**GSE93606 (n=48)**			**All studies combined (n=196)**				
	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**P-value**	**Total n used**
n subjects in cluster	22	39	4	42	32	9	0	24	24	64	95	37		
Age (years) (mean, SD)	70·0 (6·3)	68·3 (7·9)	64·0 (2·7)	66·7 (9·8)	67·0 (14·1)	67·0 (12·1)	-	64·8 (5·9)	70·3 (8·8)	67·8 (8·9)	66·9 (10·2)	68·8 (9·4)	0·592	188
Male (%)	20 (91·0%)	30 (77·0%)	4 (100%)	32 (76·2%)	21 (65·6%)	3 (33·3%)	-	15 (62·5%)	16 (66·7%)	52 (81·3%)	66 (69·5%)	23 (62·2%)	0·091	196
European ancestry (%)	17 (81·0%)	29 (82·9%)	3 (75·0%)	NA	NA	NA	-	NA	NA	17 (81·0%)	29 (82·9%)	3 (75·0%)	0·883	60
Ever smoker (%)	NA	NA	NA	NA	NA	NA	-	15 (62·5%)	18 (78·3%)	NA	15 (62·5%)	18 (78·3%)	0·389	47
Death observed during study (%)	NA	NA	NA	NA	NA	NA	-	6 (25·0%)	16 (66·7%)	NA	6 (25·0%)	16 (66.7%)	**0·009**	48
FVC % predicted (median, IQR)	59·5 (19·5)	65·0 (24·0)	51·5 (7·8)	77·0 (36·0)	66·0 (46·0)	73·0 (17·5)	-	71·5 (27·7)	60·8 (24·1)	63.0 (35·0)	70·5 (30·1)	60·1 (23·4)	0·342	154
D_{LCO} % predicted (median, IQR)	34·5 (17·5)	49·0 (21·0)	28·5 (21·0)	65·0 (37·0)	66·0 (40·0)	30·0 (30·0)	-	38·1 (17·1)	36·6 (15·9)	35·0 (30·0)	45·0 (29·2)	34·4 (17·3)	**0·009**	133
FEV_1 % predicted (median, IQR)	NA	NA	NA	NA	NA	NA	-	74·9 (23·1)	65·4 (22·7)	NA	74·9 (23·1)	65·4 (22·7)	0·216	48
GAP index (mean, SD)	5·3 (1·3)	3·9 (1·3)	4·5 (1·3)	4·3 (1·5)	4·1 (1·6)	4·3 (3·1)	-	3·7 (1·8)	4·4 (1·6)	4·9 (1·4)	3·9 (1·5)	4·4 (1·7)	**0·006**	132
MUC5B genotype: GG (%)	NA	NA	NA	5 (29·4%)	6 (28·6%)	3 (60·0%)	-	5 (26·3%)	11 (50·0%)	5 (29·4%)	11 (27·5%)	14 (51·9%)	0·230	84
MUC5B genotype: GT (%)	NA	NA	NA	10 (58·8%)	14 (66·7%)	2 (40·0%)	-	12 (63·2%)	8 (36·4%)	10 (58·8%)	26 (65·0%)	10 (37·0%)		
MUC5B genotype: TT (%)	NA	NA	NA	2 (11·8%)	1 (4·8%)	0 (0%)	-	2 (10·5%)	3 (13·6%)	2 (11·8%)	3 (7·5%)	3 (11·1%)		

TABLE E4: The significantly enriched (q-value <0.05) biological processes for the 769 genes assigned to Cluster 1.

Biological process	**Enrichment score**	**p-value**	**q-value**
Mitochondrial ATP synthesis coupled electron transport	7.18	1.0×10^{-7}	7.8×10^{-4}
ATP synthesis coupled electron transport	7.12	1.2×10^{-7}	7.8×10^{-4}
Respiratory electron transport chain	6.88	1.4×10^{-7}	7.8×10^{-4}
Cellular respiration	5.95	1.3×10^{-6}	0.005
Oxidative phosphorylation	5.84	4.0×10^{-6}	0.012
Electron transport chain	5.56	4.3×10^{-6}	0.012
Homeostasis of number of cells	5.12	1.1×10^{-5}	0.024
Homeostatic process	4.54	1.7×10^{-5}	0.032

TABLE E5: The 20 most significantly enriched (q-value <0.05) biological processes for the 839 genes assigned to Cluster 2.

Biological process	**Enrichment score**	**p-value**	**q-value**
Cell activation	12.78	2.2×10^{-27}	3.7×10^{-24}
Immune system process	11.33	1.7×10^{-25}	1.4×10^{-21}
Leukocyte activation	11.76	2.4×10^{-23}	1.2×10^{-19}
Immune response	9.83	6.0×10^{-19}	2.5×10^{-15}
Regulation of immune system process	9.75	1.5×10^{-18}	4.9×10^{-15}
Regulated exocytosis	8.90	2.5×10^{-14}	6.9×10^{-11}
Response to stimulus	7.30	1.3×10^{-13}	3.1×10^{-10}
Defence response	8.16	1.6×10^{-13}	3.2×10^{-10}
Multi-organism process	7.74	1.9×10^{-13}	3.5×10^{-10}
Lymphocyte activation	8.73	4.5×10^{-13}	7.5×10^{-10}
Translational initiation	9.72	6.4×10^{-13}	9.1×10^{-10}
Symbiotic process	8.24	6.6×10^{-13}	9.1×10^{-10}
Interspecies interaction between organisms	8.02	1.6×10^{-12}	2.1×10^{-9}
Peptide metabolic process	8.31	1.9×10^{-12}	2.1×10^{-9}
Exocytosis	8.06	1.9×10^{-12}	2.1×10^{-9}
Peptide biosynthetic process	8.43	2.9×10^{-12}	2.9×10^{-9}
Translation	8.46	3.2×10^{-12}	3.1×10^{-9}
Regulation of biological quality	7.14	3.8×10^{-12}	3.5×10^{-9}
Myeloid leukocyte activation	8.09	4.1×10^{-12}	3.6×10^{-9}
Regulation of multicellular organismal process	7.20	5.0×10^{-12}	4.0×10^{-9}

TABLE E6: The 20 most significantly enriched (q-value <0.05) biological processes for the 784 genes assigned to Cluster 3.

Biological process	**Enrichment score**	**p-value**	**q-value**
Cell activation	20.78	1.3×10^{-60}	1.5×10^{-56}
Immune response	19.53	1.8×10^{-60}	1.5×10^{-56}
Leukocyte activation	20.87	3.3×10^{-59}	1.8×10^{-55}
Immune system process	18.04	1.6×10^{-57}	6.6×10^{-54}
Immune effector process	19.19	1.2×10^{-52}	4.0×10^{-49}
Myeloid leukocyte activation	20.63	1.7×10^{-52}	4.7×10^{-49}
Leukocyte activation involved in immune response	20.07	9.2×10^{-51}	2.2×10^{-47}
Cell activation involved in immune response	19.98	1.9×10^{-50}	3.9×10^{-47}
Neutrophil activation	20.19	1.0×10^{-48}	1.9×10^{-45}
Granulocyte activation	20.02	3.5×10^{-48}	5.7×10^{-45}
Neutrophil activation involved in immune response	19.55	4.0×10^{-46}	6.1×10^{-43}
Leukocyte degranulation	19.42	5.0×10^{-46}	6.8×10^{-43}
Neutrophil degranulation	19.43	1.3×10^{-45}	1.7×10^{-42}
Myeloid cell activation involved in immune response	19.21	1.5×10^{-45}	1.8×10^{-42}
Neutrophil mediated immunity	19.23	3.6×10^{-45}	3.9×10^{-42}
Myeloid leukocyte mediated immunity	18.99	1.1×10^{-44}	1.1×10^{-41}
Leukocyte mediated immunity	17.11	4.3×10^{-43}	4.2×10^{-40}
Secretion by cell	16.63	3.9×10^{-41}	3.5×10^{-38}
Export from cell	16.50	5.9×10^{-41}	5.2×10^{-38}
Defence response	15.95	1.2×10^{-40}	1.0×10^{-37}

TABLE E7: The 13 genes in the classifier. 'Up genes' refer to genes that were more highly expressed in the subjects for that cluster compared to the mean expression across all subjects, and 'down genes' refer to genes that were less highly expressed in the subjects in that cluster.

Cluster 1		**Cluster 2**		**Cluster 3**	
Up genes	**Down genes**	**Up genes**	**Down genes**	**Up genes**	**Down genes**
KCNK15	*RPF1*	*NOP58*		*CA4*	
SORBS1		*PSMA5*		*BCL2A1*	
HBB		*RASGRP1*		*UGCG*	
		IFI30			
		HLA-DRA			
		ATM			

TABLE E8: Coefficients of the multinomial logistic regression model fit using classification scores from the genes in the classifier. Note that Cluster 1 is the reference cluster and so the coefficients for this cluster are all zero and have been omitted.

Cluster	**Intercept**	**Cluster 1 score**	**Cluster 2 score**	**Cluster 3 score**
2	3.12	-9.75	8.87	1.66
3	-16.6	-11.92	-3.15	29.42

TABLE E9: Two-way tables comparing 'true' assignment of subjects from the discovery analysis (determined using COMMUNAL with 2,500 genes) to the reassignment of these subjects using the 13-gene cluster classifier.

		True cluster		
		Cluster 1	Cluster 2	Cluster 3
Classifier predicted cluster	Cluster 1	63	1	0
	Cluster 2	1	94	0
	Cluster3	0	0	37

TABLE E10: Pairwise comparisons showing the differences in survival over time between any two validation clusters, estimated using Cox proportional hazards models.

Reference cluster	**Alternate cluster**	**Hazard Ratio**	**95% CI**	**P-value**
Cluster 2	Cluster 1	3.80	1.78, 8.12	0.001
Cluster 2	Cluster 3	5.05	2.24, 11.35	9.1×10^{-5}
Cluster 1	Cluster 3	1.47	0.67, 3.22	0.341

TABLE E11: The agreement between the cluster classifier and SAMS when validation subjects were assigned to risk groups using each method.

GSE27957 (n=45)		Cluster classifier	
		High risk	Low risk
SAMS	High risk	13	2
	Low risk	5	25
GSE28042 (n=75)		Cluster classifier	
		High risk	Low risk
SAMS	High risk	17	12
	Low risk	19	27
Both datasets combined (n=120)		Cluster classifier	
		High risk	Low risk
SAMS	High risk	30	14
	Low risk	24	52

TABLE E12: Summary statistics from the Cox proportional hazards model adjusting for cluster, age, sex, ancestry, predicted forced vital capacity (FVC) and predicted diffusing capacity of the lung for carbon monoxide (DL_{CO}). OR = odds ratio, SE = standard error and CI = confidence interval.

Variable	OR	SE	P-value	95% CI
Cluster (high-risk cluster)	2.697	0.367	0.007	(1.315, 5.534)
Age (years)	1.006	0.020	0.748	(0.968, 1.046)
Sex (male)	5.720	0.752	0.020	(1.310, 24.969)
Ancestry (non-European)	1.099	0.608	0.876	(0.334, 3.619)
Predicted FVC	0.996	0.013	0.745	(0.971, 1.022)
Predicted DL_{CO}	0.967	0.013	0.008	(0.944, 0.991)

Additional Figures

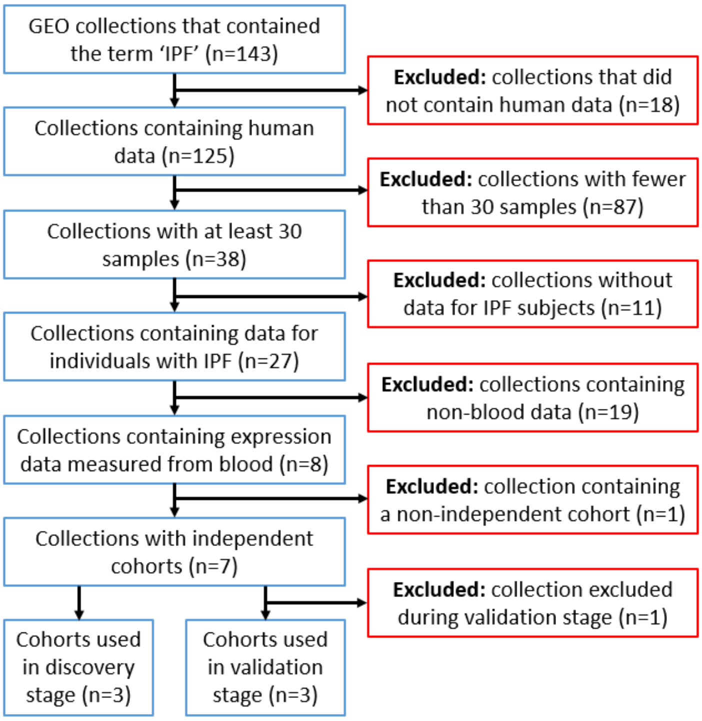

GEO collections that contained the term 'IPF' (n=143)
Excluded: collections that did not contain human data (n=18)
Collections containing human data (n=125)
Excluded: collections with fewer than 30 samples (n=87)
Collections with at least 30 samples (n=38)
Excluded: collections without data for IPF subjects (n=11)
Collections containing data for individuals with IPF (n=27)
Excluded: collections containing non-blood data (n=19)
Collections containing expression data measured from blood (n=8)
Excluded: collection containing a non-independent cohort (n=1)
Collections with independent cohorts (n=7)
Excluded: collection excluded during validation stage (n=1)
Cohorts used in discovery stage (n=3)
Cohorts used in validation stage (n=3)

FIGURE E1: Flow diagram showing the process used to systematically select publicly available IPF gene expression datasets from the Gene Expression Omnibus for use in this study.

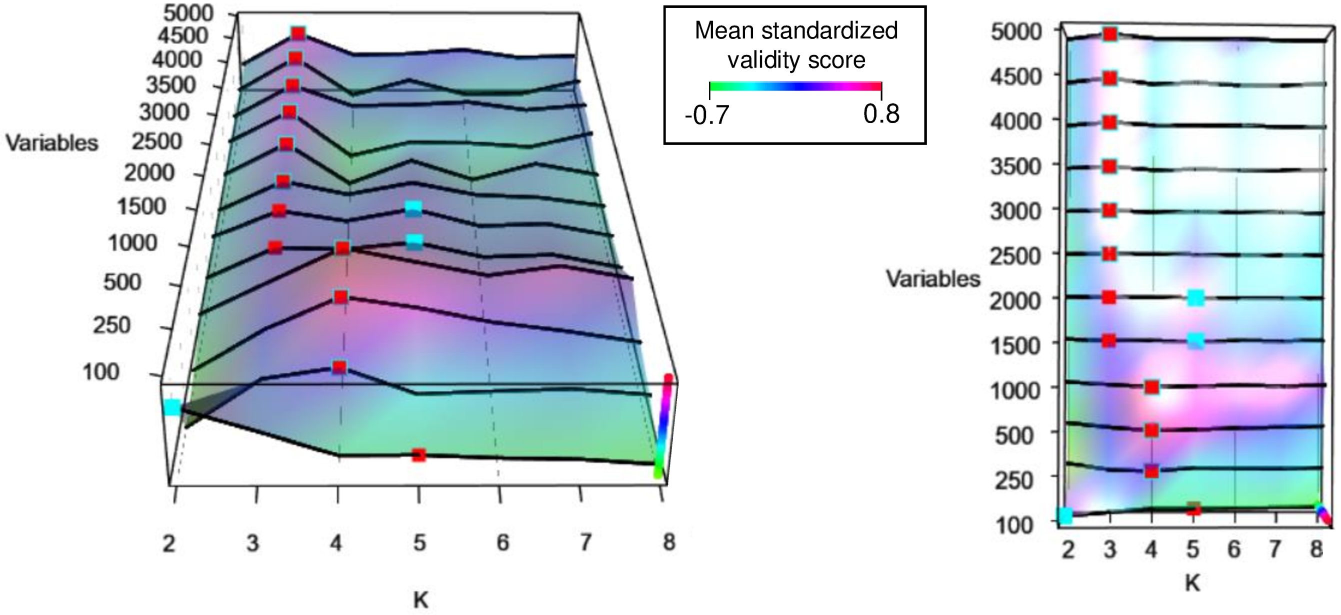
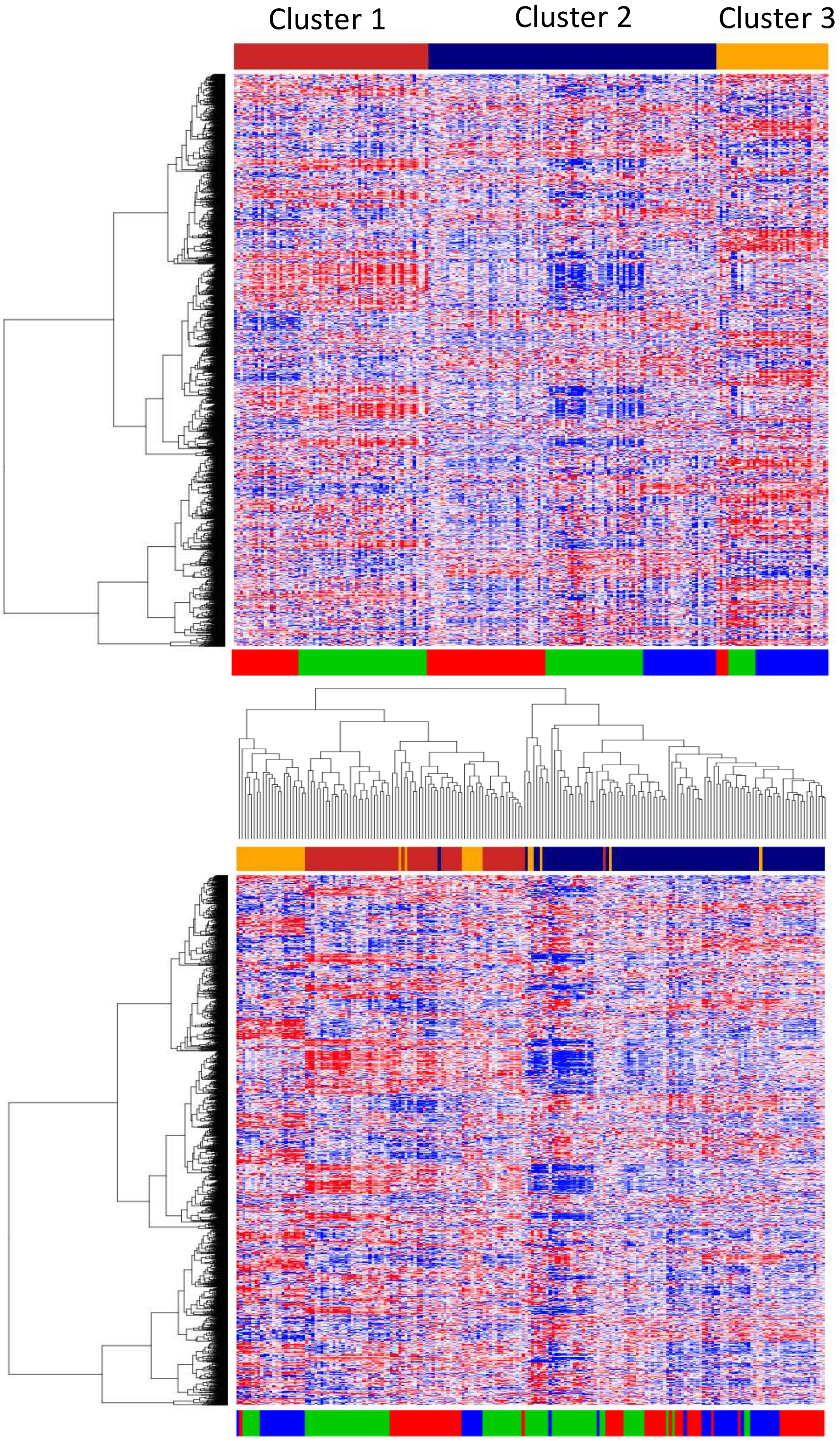


FIGURE E2: The 3D optimality map produced by COMMUNAL to identify the most robust number of clusters in the co-normalised data. A higher validity score indicates a better clustering assignment and stable optima are the points where the blue and red squares meet. In this map there are stable optima at K=4 from 250 genes to 1,000 genes, and at K=3 from 2,500 genes to 5,000 genes, as shown by the red and blue squares meeting. Despite the K=4 clustering assignment at 1,000 genes showing the highest mean standardized validity score of all tested clustering assignments, there were stable optima at K=3 clusters over a larger range of tested space, indicating a stronger biological signal. As such, K=3 was chosen as the optimal number of clusters in the pooled IPF dataset. The clustering at 2,500 genes and 3 clusters was chosen as the optimal clustering assignment, under the assumption that using the fewest number of genes has the least amount of redundant signal.



1 **FIGURE E3:** Heatmaps of gene expression for the clustered samples (x-axis) across the top 2,500 genes (y-axis), without hierarchical clustering of the samples (A) and with hierarchical clustering of the samples (B). Blue inside the heatmap indicates low expression and red indicates high expression. In both plots, the genes have been hierarchically clustered for presentation purposes, the bar above the plot shows the cluster that subject was assigned in to (red = cluster 1, blue = cluster 2 and yellow = cluster 3) and the bar below the plot indicates which original study the subject was in (red = GSE38958, green = GSE33566 and blue = GSE93606).

16

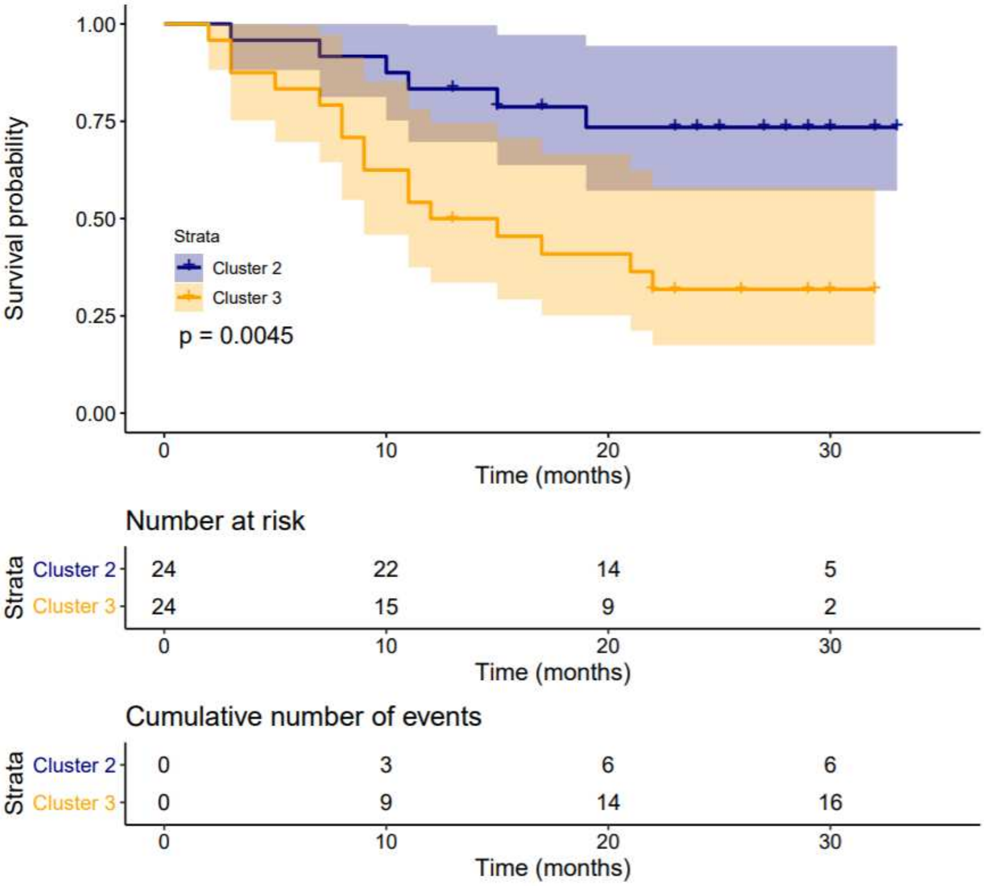


FIGURE E4: Kaplan-Meier curves and corresponding 95% confidence intervals showing survival over time for the subjects from study GSE93606, stratified by the cluster which they were assigned to in this study. The p-value shown on the plot is from a log-rank test testing the two curves for equality.

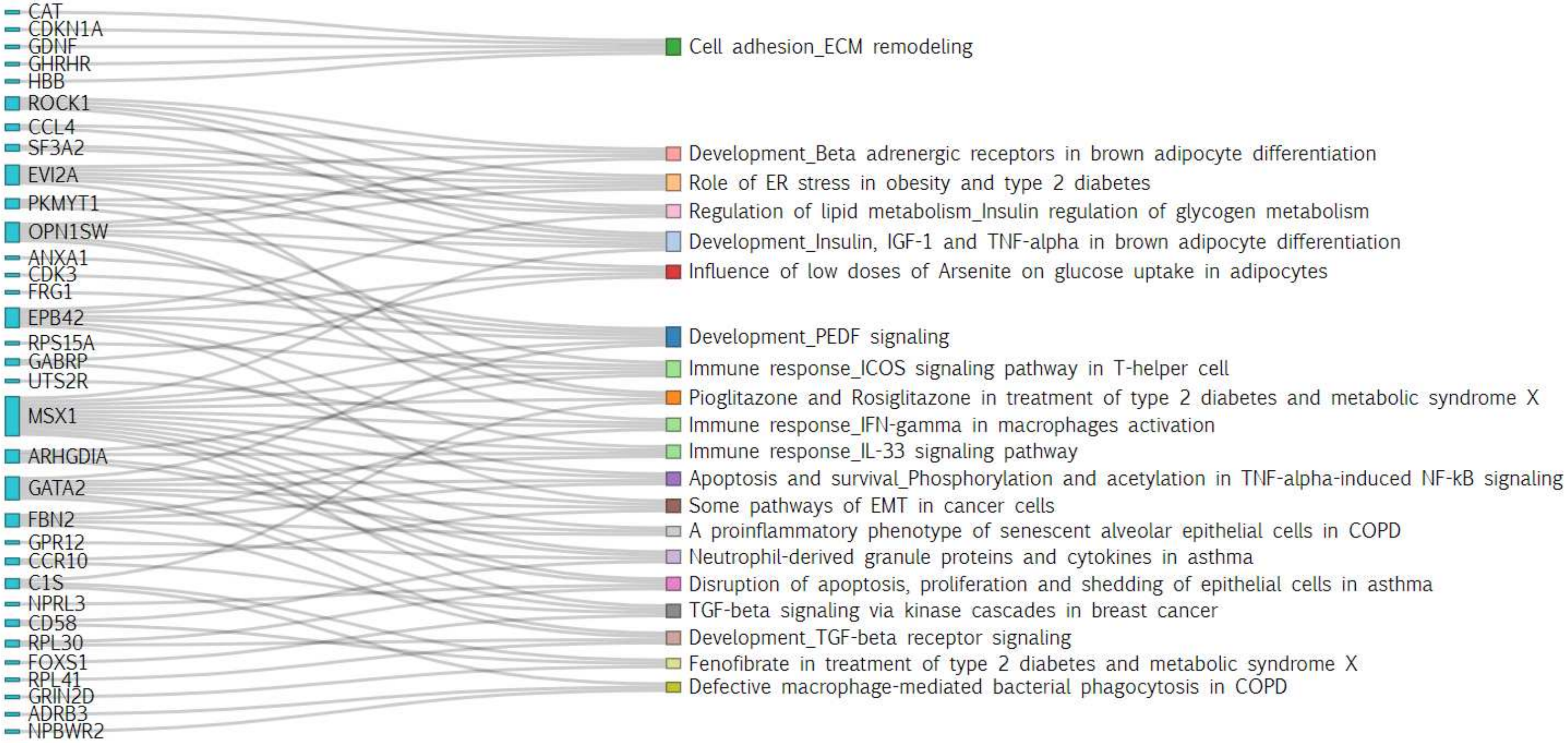

CAT
CDKN1A
GDNF
GHRHR
HBB
ROCK1
CCL4
SF3A2
EVI2A
PKMYT1
OPN1SW
ANXA1
CDK3
FRG1
EPB42
RPS15A
GABRP
UTS2R
MSX1
ARHGDIA
GATA2
FBN2
GPR12
CCR10
C1S
NPRL3
CD58
RPL30
FOXS1
RPL41
GRIN2D
ADRB3
NPBWR2
Cell adhesion_ECM remodeling
Development_Beta adrenergic receptors in brown adipocyte differentiation
Role of ER stress in obesity and type 2 diabetes
Regulation of lipid metabolism_Insulin regulation of glycogen metabolism
Development_Insulin, IGF-1 and TNF-alpha in brown adipocyte differentiation
Influence of low doses of Arsenite on glucose uptake in adipocytes
Development_PEDF signaling
Immune response_ICOS signaling pathway in T-helper cell
Pioglitazone and Rosiglitazone in treatment of type 2 diabetes and metabolic syndrome X
Immune response_IFN-gamma in macrophages activation
Immune response_IL-33 signaling pathway
Apoptosis and survival_Phosphorylation and acetylation in TNF-alpha-induced NF-kB signaling
Some pathways of EMT in cancer cells
A proinflammatory phenotype of senescent alveolar epithelial cells in COPD
Neutrophil-derived granule proteins and cytokines in asthma
Disruption of apoptosis, proliferation and shedding of epithelial cells in asthma
TGF-beta signaling via kinase cascades in breast cancer
Development_TGF-beta receptor signaling
Fenofibrate in treatment of type 2 diabetes and metabolic syndrome X
Defective macrophage-mediated bacterial phagocytosis in COPD

FIGURE E5: A Sankey diagram for Cluster 1 showing the genes that correspond to the 20 most significantly enriched biological pathways. The colour on the right hand side of the plot indicates the category of a particular pathway.

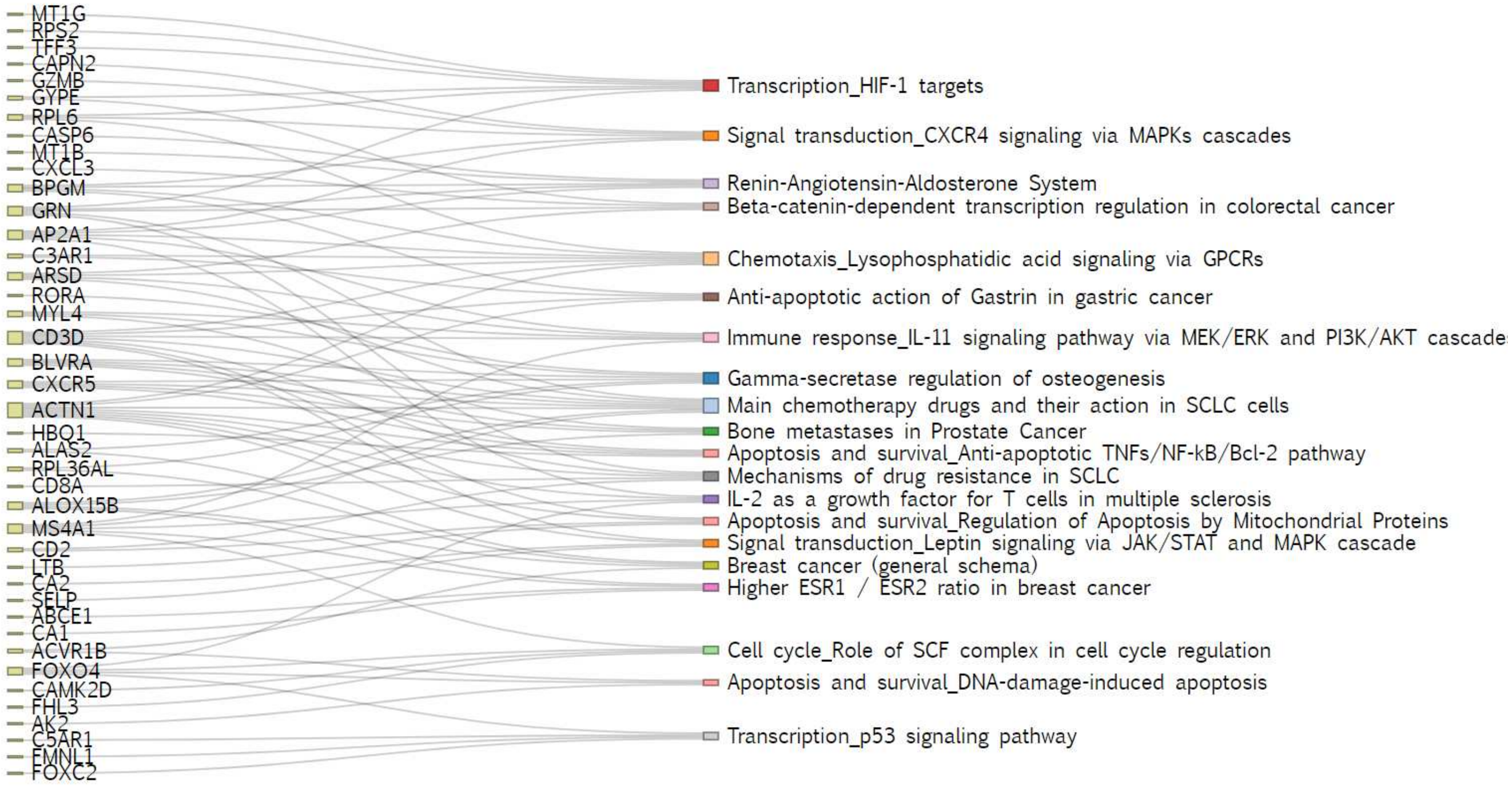

MT1G
RPS2
TFF3
CAPN2
GZMB
GYPE
RPL6
CASP6
MT1B
CXCL3
BPGM
GRN
AP2A1
C3AR1
ARSD
RORA
MYL4
CD3D
BLVRA
CXCR5
ACTN1
HBQ1
ALAS2
RPL36AL
CD8A
ALOX15B
MS4A1
CD2
LTB
CA2
SELP
ABCE1
CA1
ACVR1B
FOXO4
CAMK2D
FHL3
AK2
C5AR1
FMNL1
FOXC2
Transcription_HIF-1 targets
Signal transduction_CXCR4 signaling via MAPKs cascades
Renin-Angiotensin-Aldosterone System
Beta-catenin-dependent transcription regulation in colorectal cancer
Chemotaxis_Lysophosphatidic acid signaling via GPCRs
Anti-apoptotic action of Gastrin in gastric cancer
Immune response_IL-11 signaling pathway via MEK/ERK and PI3K/AKT cascade
Gamma-secretase regulation of osteogenesis
Main chemotherapy drugs and their action in SCLC cells
Bone metastases in Prostate Cancer
Apoptosis and survival_Anti-apoptotic TNFs/NF-kB/Bcl-2 pathway
Mechanisms of drug resistance in SCLC
IL-2 as a growth factor for T cells in multiple sclerosis
Apoptosis and survival_Regulation of Apoptosis by Mitochondrial Proteins
Signal transduction_Leptin signaling via JAK/STAT and MAPK cascade
Breast cancer (general schema)
Higher ESR1 / ESR2 ratio in breast cancer
Cell cycle_Role of SCF complex in cell cycle regulation
Apoptosis and survival_DNA-damage-induced apoptosis
Transcription_p53 signaling pathway

FIGURE E6: A Sankey diagram for Cluster 2 showing the genes that correspond to the 20 most significantly enriched biological pathways. The colour on the right hand side of the plot indicates the category of a particular pathway.

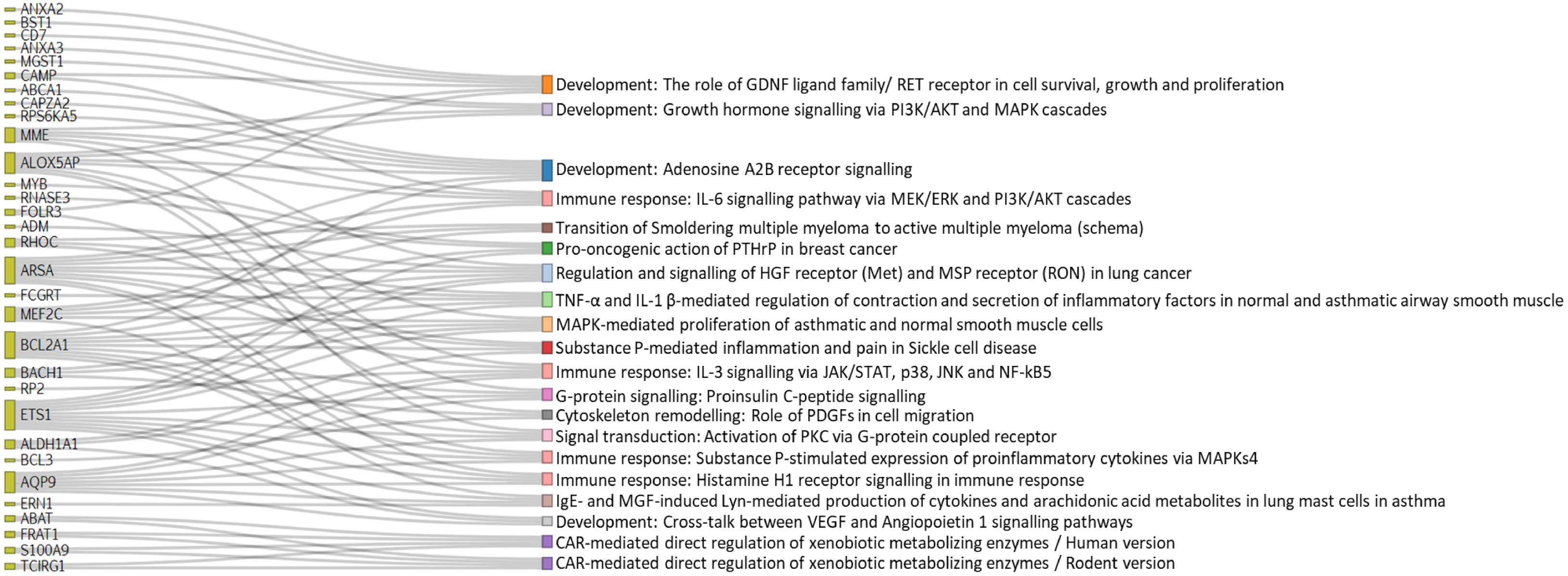

ANXA2
BST1
CD7
ANXA3
MGST1
CAMP
ABCA1
CAPZA2
RPS6KA5
MME
ALOX5AP
MYB
RNASE3
FOLR3
ADM
RHOC
ARSA
FCGRT
MEF2C
BCL2A1
BACH1
RP2
ETS1
ALDH1A1
BCL3
AQP9
ERN1
ABAT
FRAT1
S100A9
TCIRG1
Development: The role of GDNF ligand family/ RET receptor in cell survival, growth and proliferation
Development: Growth hormone signalling via PI3K/AKT and MAPK cascades
Development: Adenosine A2B receptor signalling
Immune response: IL-6 signalling pathway via MEK/ERK and PI3K/AKT cascades
Transition of Smoldering multiple myeloma to active multiple myeloma (schema)
Pro-oncogenic action of PTHrP in breast cancer
Regulation and signalling of HGF receptor (Met) and MSP receptor (RON) in lung cancer
TNF-α and IL-1 β-mediated regulation of contraction and secretion of inflammatory factors in normal and asthmatic airway smooth muscle
MAPK-mediated proliferation of asthmatic and normal smooth muscle cells
Substance P-mediated inflammation and pain in Sickle cell disease
Immune response: IL-3 signalling via JAK/STAT, p38, JNK and NF-kB5
G-protein signalling: Proinsulin C-peptide signalling
Cytoskeleton remodelling: Role of PDGFs in cell migration
Signal transduction: Activation of PKC via G-protein coupled receptor
Immune response: Substance P-stimulated expression of proinflammatory cytokines via MAPKs4
Immune response: Histamine H1 receptor signalling in immune response
IgE- and MGF-induced Lyn-mediated production of cytokines and arachidonic acid metabolites in lung mast cells in asthma
Development: Cross-talk between VEGF and Angiopoietin 1 signalling pathways
CAR-mediated direct regulation of xenobiotic metabolizing enzymes / Human version
CAR-mediated direct regulation of xenobiotic metabolizing enzymes / Rodent version

FIGURE E7: A Sankey diagram for Cluster 3 showing the genes that correspond to the 20 most significantly enriched biological pathways. The colour on the right hand side of the plot indicates the category of a particular pathway.

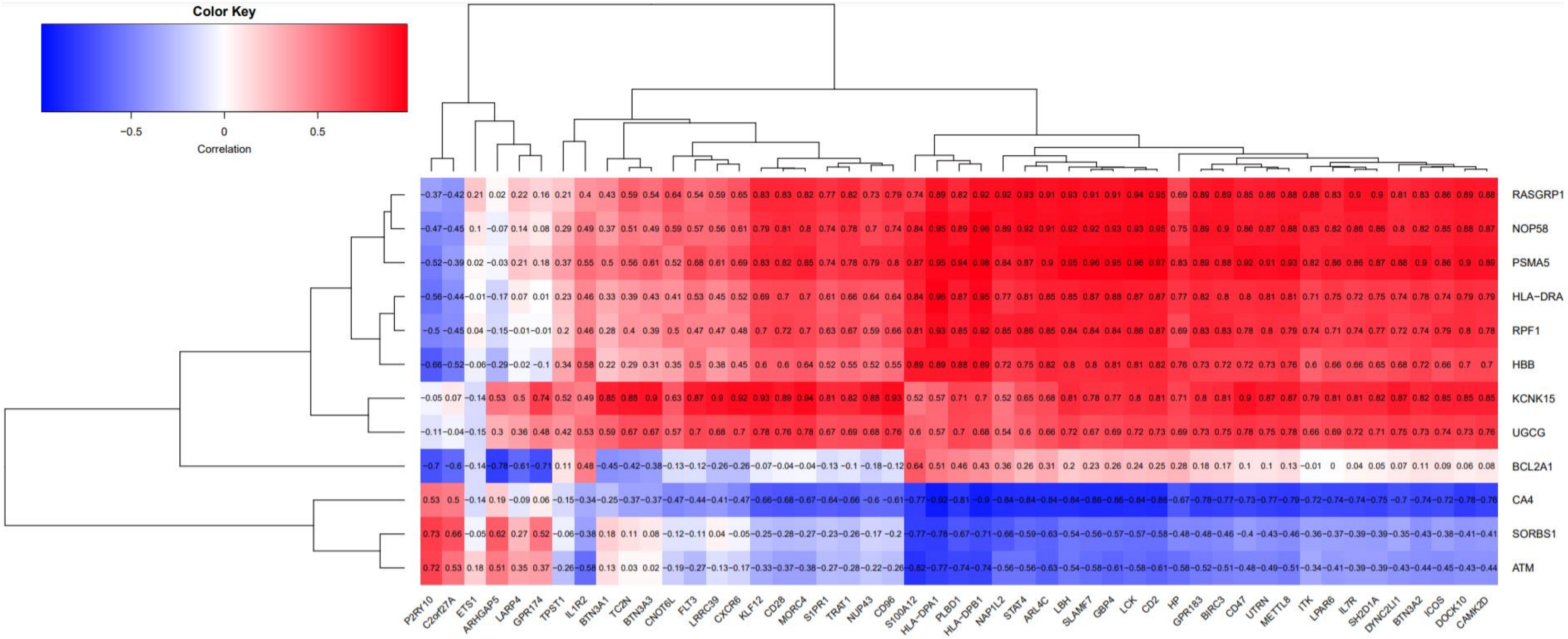


FIGURE E8: A heatmap showing the Pearson correlation between the genes in the classifier (y-axis) and the genes used by SAMS (x-axis). The correlation was calculated using the data from the IPF patients in the three validation cohorts (total n=194) for all genes that had complete data (12/13 genes for the classifier and 49/52 genes for SAMS). Both sets of genes were clustered using hierarchical clustering for presentation purposes.

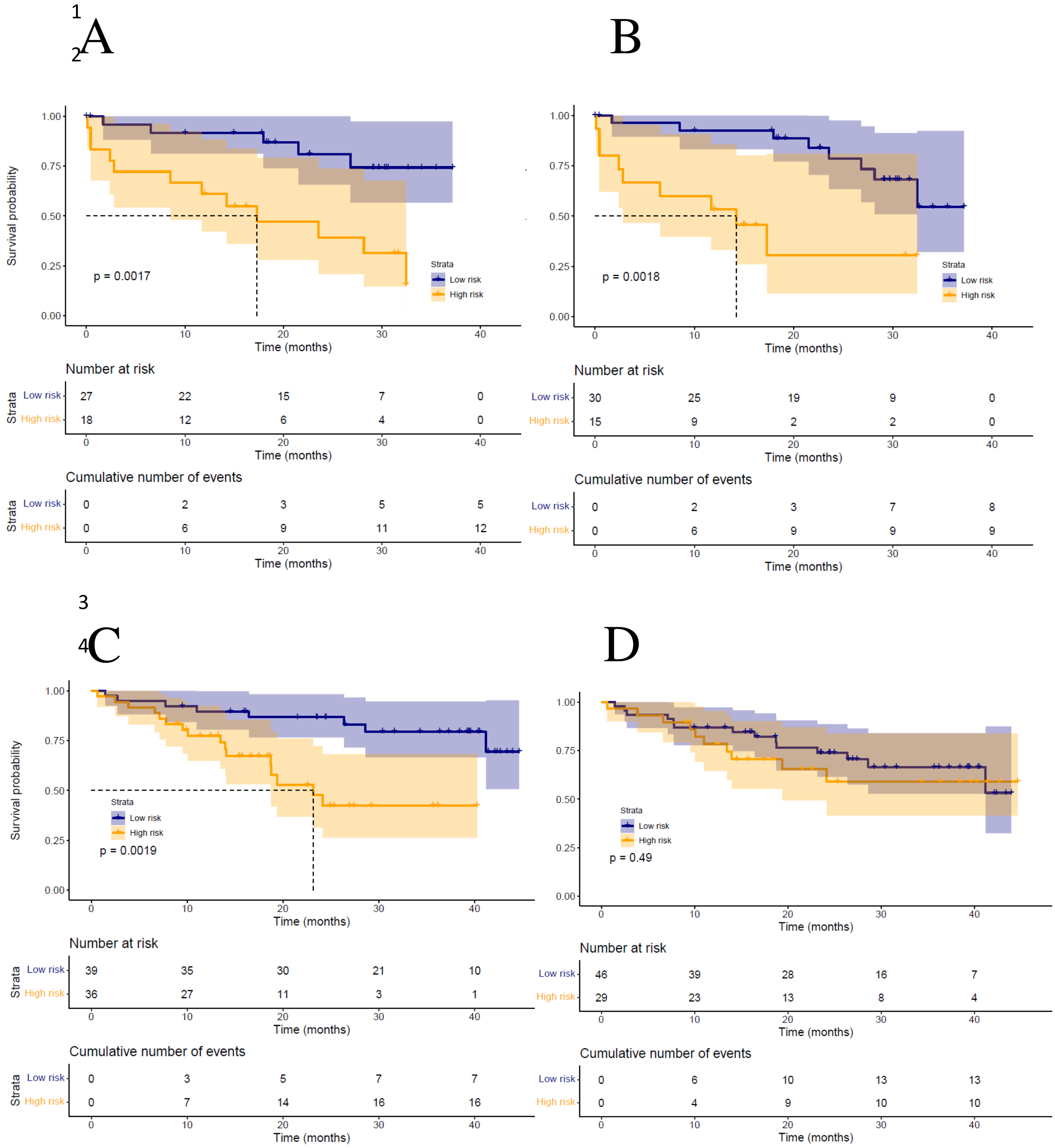


FIGURE E9: Survival over time for the IPF subjects in the validation datasets GSE27957 and GSE28042, stratified by predicted risk group. A) Survival of IPF cases from GSE27957 with risk predicted by our 13 gene classifier. B) Survival of IPF cases from GSE27957 with risk predicted by SAMS. C) Survival of IPF cases from GSE28042 with risk predicted by our 13 gene classifier. D) Survival of IPF cases from GSE28042 with risk predicted by SAMS. 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.