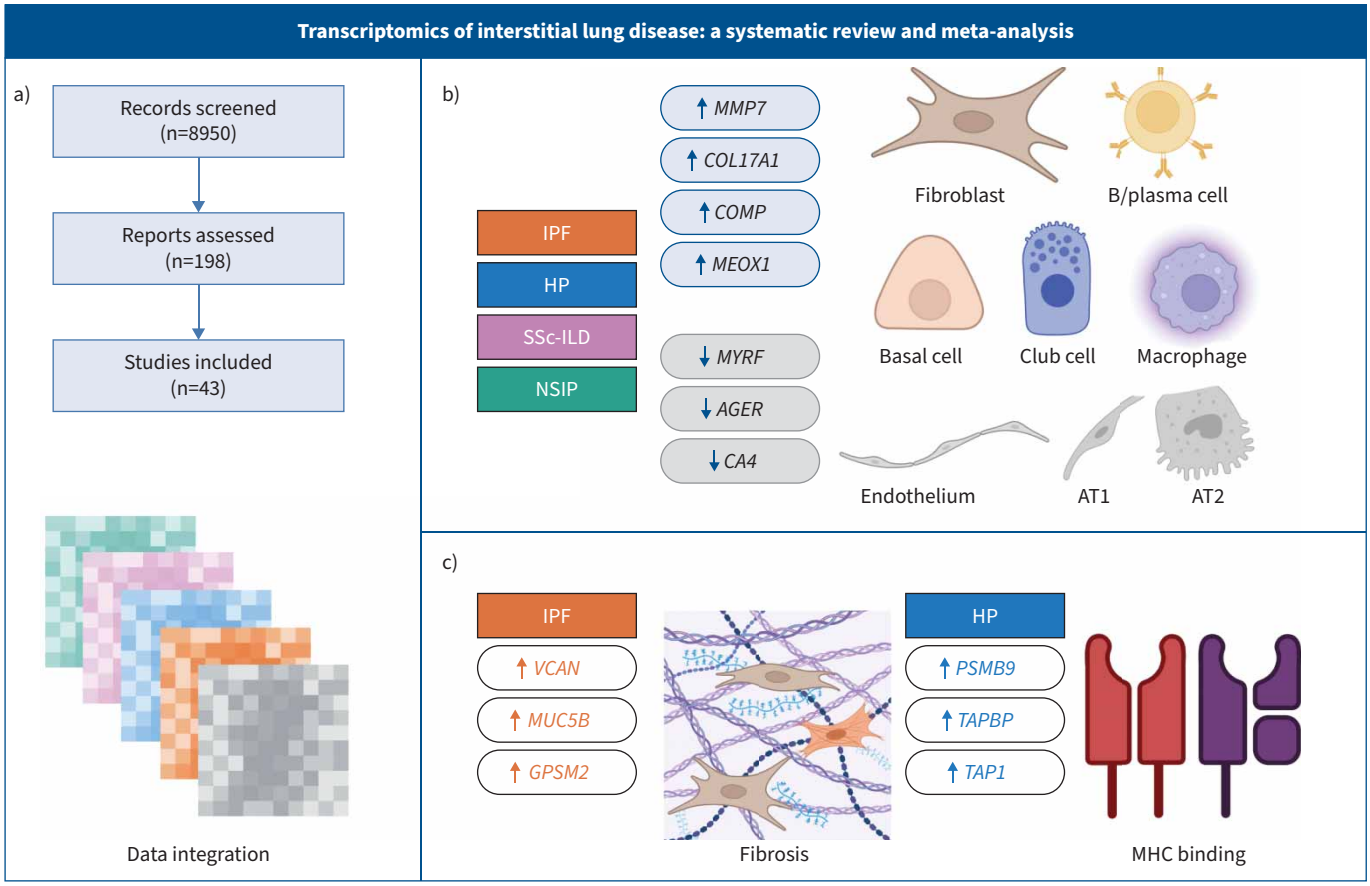
Transcriptomics of interstitial lung disease: a systematic review and meta-analysis

Daniel He, Sabina A. Guler, Casey P. Shannon, Christopher J. Ryerson and Scott J. Tebbutt



GRAPHICAL ABSTRACT Overview of the systematic review and meta-analysis results. a) 8950 records were identified from medical literature and transcriptomics databases, of which 43 were included after full-text screening. b) Datasets were extracted and integrated to develop classification models differentiating between lung transcriptomics samples obtained from patients with interstitial lung disease (ILD) subtypes and healthy controls. Shared differentially expressed genes were identified across multiple ILD subtypes, some of which were due to differences in cell populations as determined by deconvolution analysis. c) Comparison of specific ILD subtypes against all other ILD subtypes identified subtype-specific gene signatures that were validated on external datasets and associated with fibrosis in idiopathic pulmonary fibrosis (IPF) and major histocompatibility complex (MHC) binding in hypersensitivity pneumonitis (HP). AT1: alveolar type 1 cell; AT2: alveolar type 2 cell; SSc-ILD: systemic sclerosis-associated interstitial lung disease; NSIP: nonspecific interstitial pneumonia.

https://doi.org/10.1183/13993003.01070-2024

Transcriptomics of interstitial lung disease: a systematic review and meta-analysis

Daniel He [1,2,3], **Sabina A. Guler** [4,5], **Casey P. Shannon** [2,3], **Christopher J. Ryerson**[1,2,6] **and Scott J. Tebbutt** [1,2,3,6]

[1]Department of Medicine, University of British Columbia, Vancouver, BC, Canada. [2]Centre for Heart Lung Innovation, St Paul's Hospital, Vancouver, BC, Canada. [3]Prevention of Organ Failure (PROOF) Centre of Excellence, Providence Research, Vancouver, BC, Canada. [4]Department for Pulmonary Medicine, Allergology and Clinical Immunology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland. [5]Lung Precision Medicine (LPM), Department for BioMedical Research (DBMR), University of Bern, Bern, Switzerland. [6]C.J. Ryerson and S.J. Tebbutt contributed equally to this article as lead authors and supervised the work.

Corresponding author: Scott J. Tebbutt (scott.tebbutt@hli.ubc.ca)

Shareable abstract (@ERSpublications)
A systematic review and meta-analysis of ILD transcriptomics reveals distinct profiles associated with individual subtypes and putative molecular endotypes linked to decreased lung function
https://bit.ly/40VyZUw

Cite this article as: He D, Guler SA, Shannon CP, *et al*. Transcriptomics of interstitial lung disease: a systematic review and meta-analysis. *Eur Respir J* 2025; 65: 2401070 [DOI: 10.1183/13993003.01070-2024].

This article has an editorial commentary: https://doi.org/10.1183/13993003.02471-2024

Received: 2 June 2024
Accepted: 17 Nov 2024

Abstract

Objective Gene expression (transcriptomics) studies have revealed potential mechanisms of interstitial lung disease, yet sample sizes of studies are often limited and between-subtype comparisons are scarce. The aim of this study was to identify and validate consensus transcriptomic signatures of interstitial lung disease subtypes.

Methods We performed a systematic review and meta-analysis of fibrotic interstitial lung disease transcriptomics studies using an individual participant data approach. We included studies examining bulk transcriptomics of human adult interstitial lung disease samples and excluded those focusing on individual cell populations. Patient-level data and expression matrices were extracted from 43 studies and integrated using a multivariable integrative algorithm to develop interstitial lung disease classification models.

Results Using 1459 samples from 24 studies, we identified transcriptomic signatures for idiopathic pulmonary fibrosis, hypersensitivity pneumonitis, idiopathic nonspecific interstitial pneumonia and systemic sclerosis-associated interstitial lung disease against control samples, which were validated on 308 samples from eight studies (idiopathic pulmonary fibrosis area under receiver operating curve (AUC) 0.99, 95% CI 0.99–1.00; hypersensitivity pneumonitis AUC 0.91, 95% CI 0.84–0.99; nonspecific interstitial pneumonia AUC 0.94, 95% CI 0.88–0.99; systemic sclerosis-associated interstitial lung disease AUC 0.98, 95% CI 0.93–1.00). Significantly, meta-analysis allowed us to identify, for the first time, robust lung transcriptomics signatures to discriminate idiopathic pulmonary fibrosis (AUC 0.71, 95% CI 0.63–0.79) and hypersensitivity pneumonitis (AUC 0.76, 95% CI 0.63–0.89) from other fibrotic interstitial lung disease, and unsupervised learning algorithms identified putative molecular endotypes of interstitial lung disease associated with decreased forced vital capacity and diffusing capacity of the lungs for carbon monoxide % predicted. Transcriptomics signatures were reflective of both cell-specific and disease-specific changes in gene expression.

Conclusion We present the first systematic review and largest meta-analysis of fibrotic interstitial lung disease transcriptomics to date, identifying reproducible transcriptomic signatures with clinical relevance.

Introduction

Interstitial lung disease (ILD) is a heterogeneous group of inflammatory and fibrotic disorders that affect the lung interstitium. Individual ILD subtypes are rare, with an estimated worldwide prevalence of 6.3 to 71 per 100 000 people [1]. Diagnosis and treatment of ILD remains challenging even though the field has advanced significantly. High-performance gene expression analysis (transcriptomics) has been crucial in uncovering the pathogenesis of ILD, including matrix metalloproteinases (MMPs), Wnt signalling and

senescence [2]. Although some biomarkers have been identified, such as MMP-7 as a diagnostic and prognostic biomarker of idiopathic pulmonary fibrosis (IPF) [3], molecular biomarkers are not yet widely used in the clinical management of ILD [4]. A genomic classifier (Envisia) based on transbronchial biopsies has been developed to identify the histological pattern of usual interstitial pneumonia (UIP), which allows for a diagnosis of IPF in matching clinical characteristics [5], but the clinical application of this genomic classifier has not yet been recommended in diagnostic guidelines [6] and it is limited by its availability. The majority of transcriptomics research has been performed in IPF and, while many differentially expressed genes (DEGs) have been identified and reproduced across different studies, few investigations have been performed in other ILD subtypes, such as hypersensitivity pneumonitis (HP). Furthermore, studies with adequate sample sizes are rare owing to the cost of microarray and RNA sequencing (RNA-seq) technologies. Hence, the objective of this systematic review and individual participant data meta-analysis was to determine consensus transcriptomic signatures of patients with fibrotic ILD, and to identify and validate biomarkers for the classification of major fibrotic ILD subtypes.

Methods

Search strategy

In accordance with Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, we performed a systematic review to identify transcriptomics studies focusing on fibrotic ILD samples and we registered the study with PROSPERO (CRD42018085682). Literature searches were performed in MEDLINE, Embase and Gene Expression Omnibus (GEO), and updated on 23 November 2023. Our search terms included variations of "interstitial lung disease", common ILD subtypes and transcriptomics-related key words. Given the recency of transcriptomics technology, the search was limited to include results published from 2000 onwards. Screening of titles, abstracts and full-text studies was performed by two independent reviewers (D. He and S.A. Guler) using Covidence (www.covidence.org). We included studies investigating bulk transcriptomics of fibrotic ILD *via* lung or blood sampling, and excluded those focused on isolated cell populations (supplementary table S1). To assess the risk of bias, we used a modified checklist based on the criteria outlined by Dupuy and Simon [7] (supplementary table S2). A full search strategy can be found in the supplementary material.

Data extraction, integration and analysis

We extracted participant-level metadata from included studies using the data repository sites GEO, ArrayExpress and the Sequence Read Archive, in addition to author-provided normalised matrices. If normalised matrices were unavailable, raw data were processed using recommended protocols from the "limma" package (v3.55.3; www.bioconductor.org) for microarrays (manufacturer-based) and RNA-seq (trimmed mean of M values), and outliers were removed if they were more than 3 SD away from the mean on the first and second principal components (supplementary figure S1). To synthesise transcriptomics data, we used an individual participant data approach to our meta-analysis using the Multivariate INTegrative (MINT) method ("mixOmics" v6.22.0; http://mixomics.org), which is an integrative classification method based on partial least squares regression. MINT does not require batch correction of training with test sets or a particular normalisation method prior to sample classification, thereby resulting in an inductive (not transductive) approach that is less prone to overfitting. A more detailed explanation of the methodology can be found in Rohart *et al*. [8, 9]. We benchmarked the MINT data integration by comparing it with DEG list integration through robust rank aggregation [10], which also allowed for the inclusion of older studies for which raw data were unavailable.

We used 24 of 32 datasets (from 31 studies) as a training set, and integrated them to develop classification models that were tuned *via* leave-one-group-out cross-validation to have the lowest balanced error rate (BER), which is the average error rate in each class (supplementary figure S2). We validated these on the remaining eight held-out test datasets by examining area under the receiver operating characteristic curve (AUC) and BER (supplementary table S4). Eight test sets were selected to cover 15–20% of the total number of samples to create an 80/20 training/test split, in addition to balancing the number of microarray and RNA-seq studies and covering all ILD subtypes of interest (IPF, HP, nonspecific interstitial pneumonia (NSIP) and systemic sclerosis-associated ILD (SSc-ILD)), Lung Tissue Research Consortium (LTRC) and non-LTRC studies, American and non-American cohorts, and recently published datasets [11–18]. All data processing and analysis were performed in R (v4.2.3; www.r-project.org). Additional details on our methodology are provided in the supplementary material.

Biclustering

Transcriptomics datasets were concatenated and genes containing missing values were removed. ComBat ("sva" v3.46.0) was used to correct for study-specific batch effects, and biclusters of the batch-corrected matrix were identified using algorithms implemented in the "MoSBi" package (v1.4.0) [19] and the

unsupervised patient stratification (UnPaSt) algorithm, which is an unconstrained version of the Differentially ExpreSsed gene Modules iN Diseases (DESMOND) algorithm [20]. COPD samples from GSE47460 were used as a disease control. The association between biclusters and clinical variables was assessed using a Chi-squared test (categorical; diagnostic class) or linear regression (continuous; forced vital capacity (FVC) % predicted, diffusing capacity of the lungs for carbon monoxide (D_{LCO}) % predicted).

Cell deconvolution and single-cell RNA-seq analysis

Cell deconvolution analysis was performed using the "BRETIGEA" package (v1.0.3) [21] by using the top 20 gene markers for cell types described in GSE122960 [22] on the ComBat-corrected gene expression data matrix. Comparisons between ILD subtypes and controls were performed using ANOVA and Tukey-adjusted for multiple comparisons. GSE122960 was downloaded from GEO and processed according to recommended protocols using the "Seurat" package (v5.0.3) [23]. Cluster annotations were performed using a combination of manual annotation and reference mapping with the "celldex" package (v1.12.0). Cell type abundance analysis was performed using a binomial generalised linear model implemented in the "emmeans" (v1.10.1) package, while single-cell RNA-seq differential expression analysis was performed using the "MAST" framework in Seurat.

Results

Search results

From our initial search of medical literature databases (3844 from MEDLINE and 6624 from Embase) and a genomic database (215 from GEO), we assessed 198 records for eligibility and identified 43 studies that used transcriptomics to profile ILD samples (figure 1). Four studies investigated blood transcriptomic

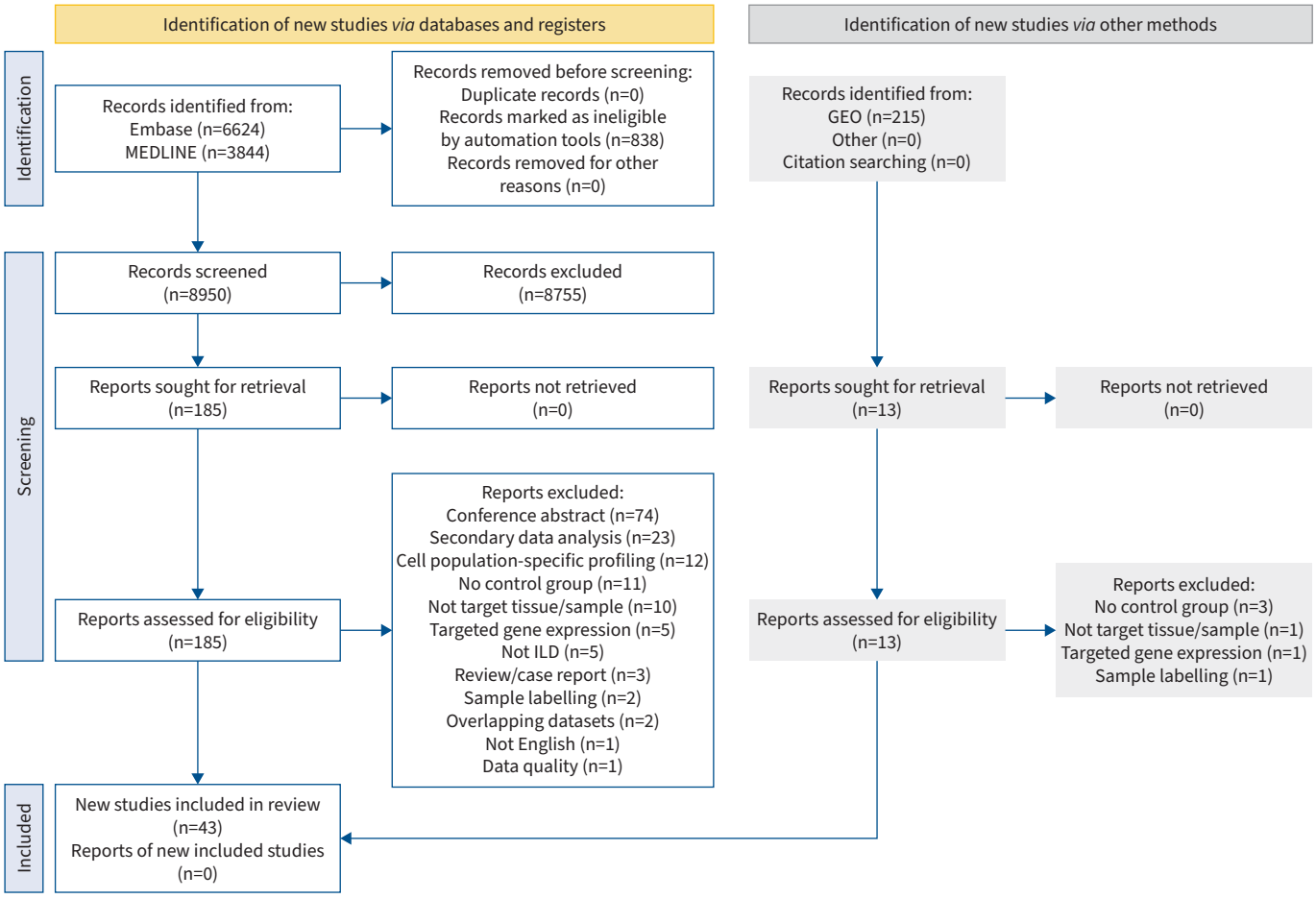


FIGURE 1 Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flowchart of study identification and screening. MEDLINE, Embase and the Gene Expression Omnibus (GEO) were examined for studies using transcriptomics to investigate samples from patients with interstitial lung disease (ILD).

profiles (IPF, HP and SSc-ILD) [24–27] and four looked at transcriptomic profiles of peripheral blood mononuclear cells (PBMCs) from IPF and SSc-ILD [28–31]. Of the remaining 35 studies that used transcriptomics to investigate fibrotic ILD lung samples, 24 studies reported DEGs between ILD and control, 29 studies uploaded publicly available data and 20 studies had both (table 1). Eight datasets were associated with the LTRC, and duplicate LTRC samples across studies were removed where LTRC identifiers were available [11, 13, 39, 42–44, 50, 54–56]. The most commonly profiled fibrotic ILD subtype was IPF, appearing in all datasets except one, followed by idiopathic NSIP, fibrotic HP and SSc-ILD. Classifiers for subtypes such as respiratory bronchiolitis-ILD were not generated because these samples were few in number; classifiers for HP and SSc-ILD blood/PBMC samples were also not created owing to insufficient numbers of datasets.

TABLE 1 Identified lung transcriptomics studies

Study	Gene list (location)	Platform	Accession	Set	Samples (n)				
					Control	IPF	HP	NSIP	SSc
Zuo *et al.* [32]	ILD ↑↓ (online link)	Microarray	–	–	4	3[#]			
Pardo *et al.* [33]		Microarray	GSE2052	Training	11	12[¶]			
Selman *et al.* [34]	IPF ↑ (table E1), HP ↑ (table E2)	Microarray	–	–	4	15	12	8	
Yang *et al.* [35]	ILD ↑↓ (table E2)	Microarray	GSE5774	Training	8	14		2	
Bridges *et al.* [36]	IPF ↑ (table 2)	Microarray	–	–	7	10			
Konishi *et al.* [37]	IPF ↑↓ (table 2)	Microarray	GSE10667	Training	15	23			
Rajkumar *et al.* [38]		Microarray	GSE15197	Training	13	8			
Cho *et al.* [39]		Microarray	GSE21369	Training	5[¶]	10[¶]	2	4[¶]	
Hsu *et al.* [40]	SSc-ILD ↑↓ (table S2)	Microarray	GSE48149	Training	9	13			13
Meltzer *et al.* [41]	IPF ↑↓ (table S5)	Microarray	GSE24206	Training	6	11			
Sanders *et al.* [42]	IPF ↑↓ (table E8)	Microarray	GSE35145	Training	4	4			
Deng *et al.* [43]	IPF ↑↓ (table S1)	RNA-seq	SRA048904	Training	3	3			
Yang *et al.* [44]	IPF ↑↓ (table S2)	Microarray	GSE32537	Training	50	34		4	
Nance *et al.* [45]	IPF ↑↓ (table 1)	RNA-seq	GSE52463	Training	7	8			
Bauer *et al.* [11]	IPF ↑↓ (tables 2, E15, E16)	Microarray	GSE47460–GPL14550	Training	91	122	21	14	
		Microarray	GSE47460–GPL6480	Test	17	38	9	3	
DePianto *et al.* [12]	IPF ↑↓ (table S3)	Microarray	GSE53845	Test	8	39			
Geng *et al.* [46]		Microarray	GSE72073	Training	3	5			
Christmann *et al.* [47]	SSc-ILD ↑↓ (table 1)	Microarray	GSE81292	Training	5				11
Horimasu *et al.* [48]	HP ↑ (table 4)	Microarray	–	–	3		9		
Horimasu *et al.* [49]	ILD ↑↓ (tables 2 and 3)	Microarray	GSE101286	Training	3	7		5	
Schafer *et al.* [13]		RNA-seq	GSE92592	Test	19	20			
Vukmirovic *et al.* [50]	IPF ↑↓ (table S4)	RNA-seq	GSE83717	Training	5	6			
Yu *et al.* [14]		RNA-seq	GSE73189	Test	5	4		3	
Cecchini *et al.* [15]	IPF ↑ (table S1)	Microarray	GSE110147	Test	11	22		10	
Luzina *et al.* [51]	IPF ↑↓ (supplementary dataset 4)	RNA-seq	GSE99621	Training	3	3			
Sivakumar *et al.* [52]	IPF ↑↓ (table S1)	RNA-seq	GSE134692	Training	17	36			
McDonough *et al.* [53]	IPF ↑↓ (table S1)	RNA-seq	GSE124685	Training	6	10			
Furusawa *et al.* [54]	HP ↑↓ (table E2)	RNA-seq	GSE150910	Training	103	102[¶]	81[¶]		
Konigsberg *et al.* [55]	IPF ↑↓ (table S1a)	RNA-seq	GSE173355	Training	14	23			
DePianto *et al.* [16]		RNA-seq	GSE166036	Test	9	20			6
Borie *et al.* [56]		RNA-seq	GSE175457	Training	188	234			
Wang *et al.* [57]		RNA-seq	GSE199152	Training	4	20			
Huang *et al.* [58]	IPF ↑↓ (table E1)	RNA-seq	GSE199949	Training	8	13			
De Sadeleer *et al.* [17]		RNA-seq	GSE184316	Test	6	10	9		
Jaffar *et al.* [18]		RNA-seq	GSE213001	Test	12	19	4	4	
				Training	581	721	104	29	24
				Test	87	173	22	20	6

Upregulated (↑) and/or downregulated (↓) differentially expressed gene lists (compared to control samples) and their location within the citation in parentheses, type of sequencing platform, Gene Expression Omnibus (GEO) or Sequence Read Archive (SRA) accession number, set assignment and number of samples per subtype. For a list of peripheral transcriptomics studies, please see supplementary table S2. [#]: five pulmonary fibrosis patients were profiled; three with idiopathic pulmonary fibrosis/usual interstitial pneumonia, one with rheumatoid arthritis and one with Sjogren's syndrome; [¶]: one sample removed as an outlier. ILD: interstitial lung disease: IPF: interstitial pulmonary fibrosis; HP: hypersensitivity pneumonitis; SSc-ILD: systemic sclerosis-associated interstitial lung disease; RNA-seq: RNA sequencing.

Data integration identified reproducible lung gene expression signatures in fibrotic ILD subtypes

As a proof of concept, we created lung gene signatures for each fibrotic ILD subtype compared to control samples *via* classification models generated from our MINT-integrated datasets (figure 2). In IPF, we identified a 55-gene signature using 23 training datasets, which was then validated on eight test datasets with AUC 0.99 (95% CI 0.99–1.00) (figure 3a). To confirm the validity of MINT in identifying biologically relevant DEGs, we compared our IPF *versus* Control signature against an IPF *versus* Control aggregate DEG list generated by robust rank aggregation and found that 39 of the 55 genes were shared, including *COMP*, *DIO2* and *MMP7* (all IPF upregulated), and *CA4*, *FAM167A* and *MYRF* (all IPF downregulated) (supplementary table S5 and figure S3a). For the peripheral transcriptome, we generated a classification model of IPF *versus* Control for whole blood datasets, which was validated on two PBMC datasets (AUC 0.74, 95% CI 0.67–0.81); likewise, a classification model of IPF *versus* Control for PBMCs was validated on two whole blood datasets (AUC 0.73, 95% CI 0.65–0.81) (supplementary figure S4 and table S3).

We applied similar strategies for the discrimination between controls, HP, SSc-ILD and NSIP. The optimal HP *versus* Control model trained on three datasets had a 235-gene signature, which had an AUC of 0.91 (95% CI 0.84–0.99) on the test datasets (figure 3b). The best-performing NSIP *versus* Control model trained on five datasets had a 378-gene signature, which was validated on four test datasets with an AUC of 0.94 (95% CI 0.88–0.99) (figure 3c). For SSc-ILD, the model trained on two datasets contained 52 genes and was validated on the test dataset with an AUC of 0.98 (95% CI 0.93–1.00) in distinguishing SSc-ILD from control samples (figure 3d). Model performance on individual datasets and cumulative performance can be found in supplementary table S4, and a list of gene signatures and model weights can be found in supplementary table S6.

The same methodology was used to perform sex-specific analyses for IPF, HP and NSIP by inferring sex from sex-specific gene expression (supplementary material). All of our sex-specific models had comparable performance between the test set of the same and opposite sex, with the exception of the male NSIP classification model (male test set: AUC 0.98, 95% CI 0.93–1.00; female test set: AUC 0.79, 95% CI 0.68–0.90) (supplementary table S7). Pathway analysis using genes unique to each sex-specific model revealed an overrepresentation of immune-related pathways in females and fibrosis-related pathways in males (supplementary tables S8 and S9).

Fibrotic ILD subtypes are differentiated at the transcriptional level

To examine identifying features of fibrotic ILD subtypes at the transcriptional level, we generated classification models for IPF, HP and NSIP against all other ILD subtypes. The 57-gene IPF *versus* other ILD model trained on seven datasets had good performance on the six test datasets (AUC 0.71, 95% CI 0.63–0.79), as did the 132-gene HP *versus* other ILD model trained on three datasets on the two test

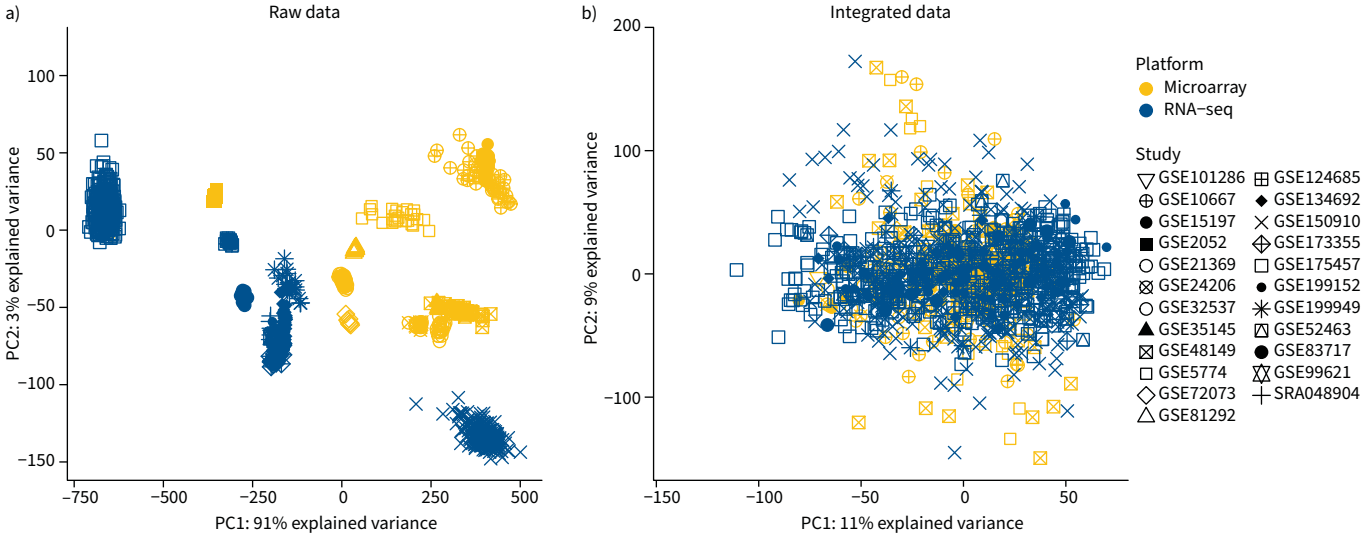


FIGURE 2 Integration of lung transcriptomics datasets. Author-normalised datasets were subsetted with genes shared in >80% of datasets (a) and integrated using the Multivariate INTegrative (MINT) method. MINT-integrated samples projected into a common latent space *via* multi-group principal component (PC) analysis are shown in (b). RNA-seq: RNA sequencing.

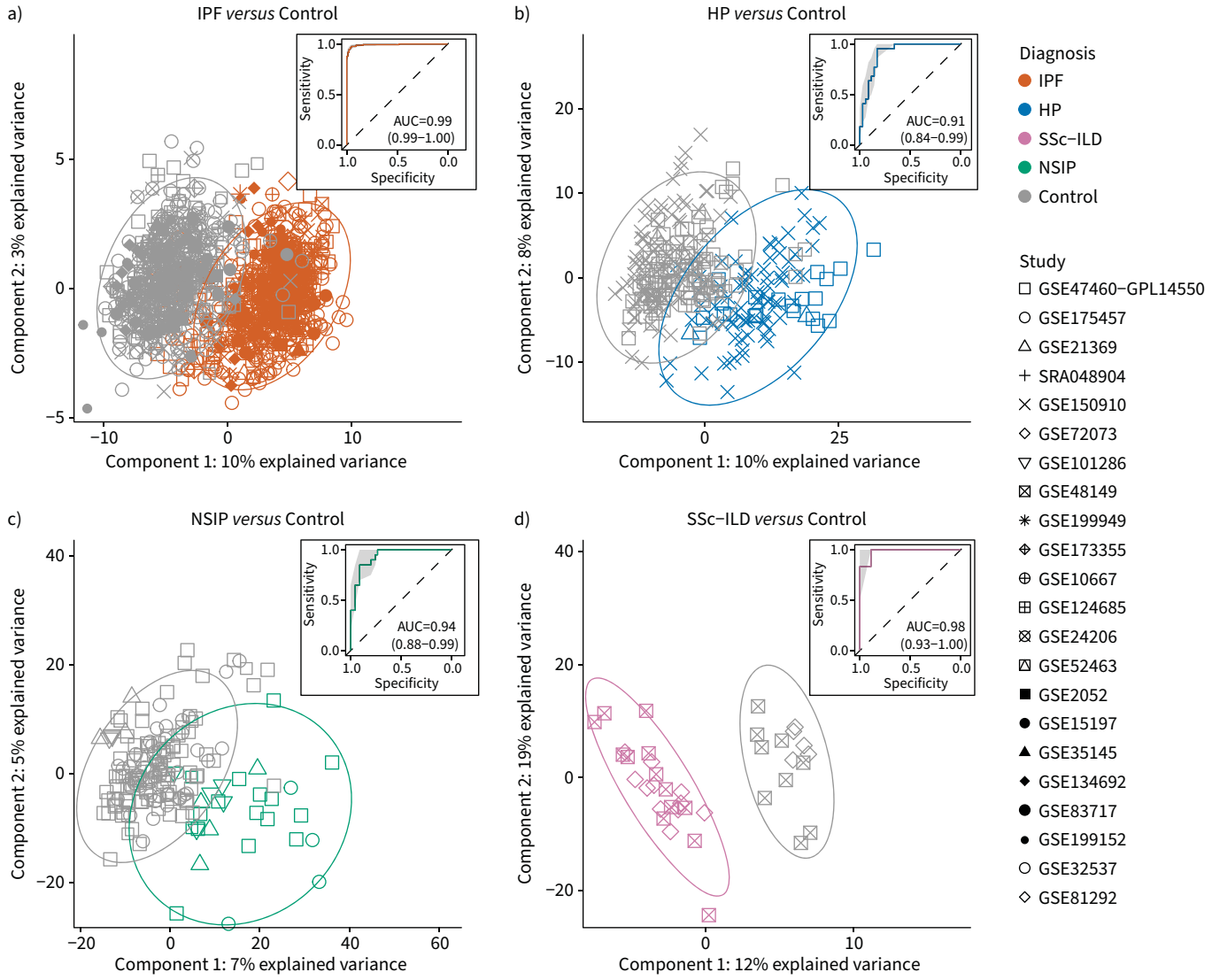


FIGURE 3 Classification models for interstitial lung disease (ILD) subtypes against controls. Training set datasets were integrated and tuned to develop classification models, which were then validated on test set datasets. Projection of Multivariate INTegrative (MINT)-integrated training set samples into the space of the first two partial least squares components for a) idiopathic pulmonary fibrosis (IPF) (n=576 control, n=721 IPF), b) hypersensitivity pneumonitis (HP) (n=199 control, n=104 HP), c) nonspecific interstitial pneumonia (NSIP) (n=157 control, n=29 NSIP) and d) systemic sclerosis-associated ILD (SSc-ILD) (n=14 control, n=24 SSc-ILD) classification models. *Inset*: area under the receiver operating characteristic curve (AUC) performance of each ILD classification model on held-out test datasets (a: n=107 control, n=193 IPF; b: n=35 control, n=22 HP; c: n=45 control, n=20 NSIP; d: n=9 control, n=6 SSc-ILD). Detailed performance metrics for each dataset can be found in supplementary table S2.

datasets (AUC 0.76, 95% CI 0.63–0.89) (figure 4). Our NSIP *versus* other ILD model trained on five datasets did not have a good performance on four test datasets (AUC 0.60, 95% CI 0.49–0.72) (supplementary figure S5). For between-subtype comparisons, a 95-gene signature differentiating IPF from HP samples had an AUC of 0.76 (95% CI 0.64–0.87) on three test datasets, while a 67-gene signature differentiating IPF from NSIP samples had an AUC of 0.76 (95% CI 0.64–0.88) on four test datasets (supplementary figure S6a–d). A classifier differentiating HP from NSIP had variable performance on two test datasets (AUC 0.74, 95% CI 0.51–0.96) (supplementary figure S6e, f).

ILD subtypes have both common and unique disease pathways

Because the ILD classification models had few genes for pathway analysis, we generated "expanded" classification models with an increased number of genes using Bayesian changepoint analysis and verified

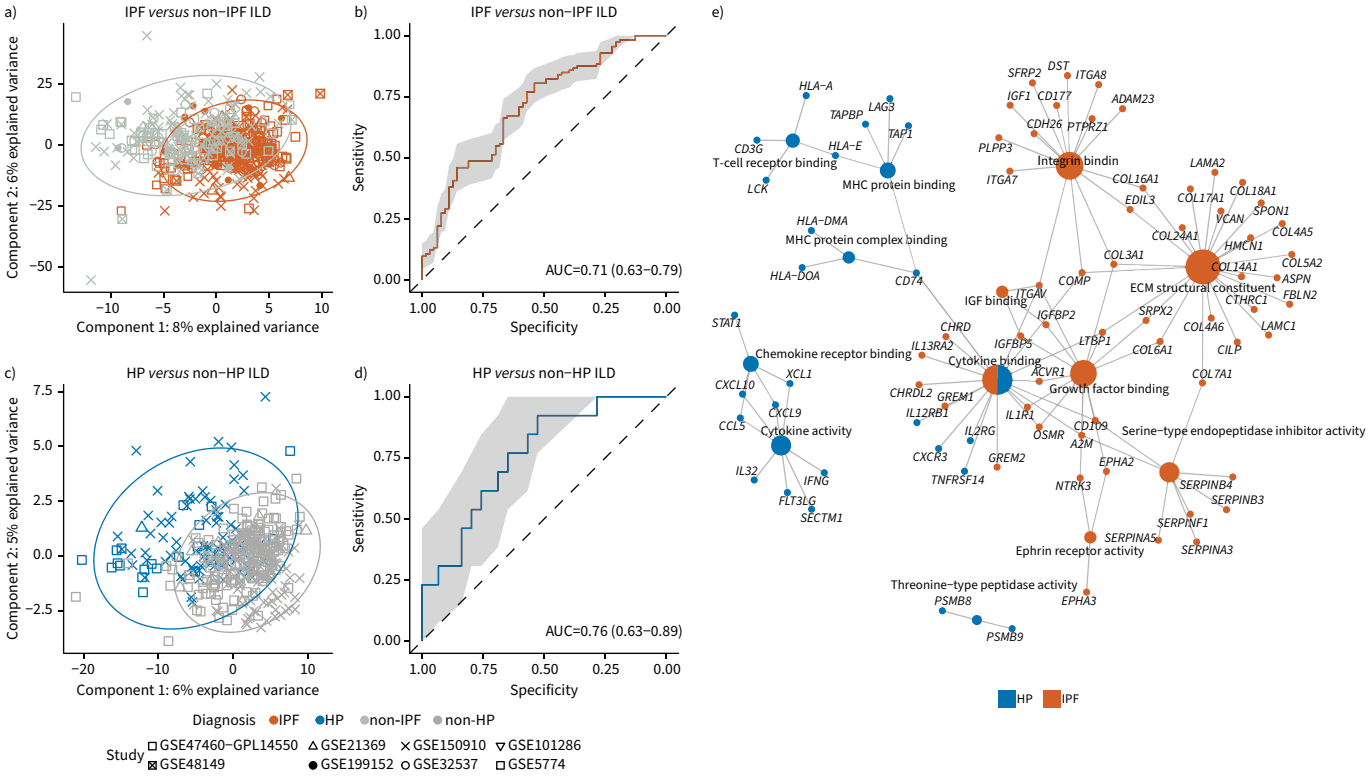


FIGURE 4 Idiopathic pulmonary fibrosis (IPF) and hypersensitivity pneumonitis (HP) have unique transcriptomic signatures compared to other interstitial lung disease (ILD) subtypes. IPF-specific and HP-specific classification models were developed through comparison against all other ILD subtypes using Multivariate INTegrative (MINT)-integrated lung transcriptomics datasets. The training set consisted of 721 IPF, 104 HP, 29 nonspecific interstitial pneumonia (NSIP), 24 systemic sclerosis-associated ILD (SSc-ILD), 14 respiratory bronchiolitis-ILD (RB-ILD), 14 unknown fibrosis, six cryptogenic organising pneumonia (COP), four desquamative interstitial pneumonia (DIP), three rheumatoid arthritis-ILD (RA-ILD) and one connective tissue disease-ILD (CTD-ILD) samples, while the test set consisted of 193 IPF, 22 HP, 20 NSIP, six SSc-ILD, five mixed IPF-NSIP, three unknown fibrosis, two RB-ILD, two DIP, two CTD-ILD, one COP and one combined pulmonary fibrosis and emphysema samples. The figure shows the MINT-partial least squares projection of samples for classification models of IPF *versus* non-IPF ILD (a) and HP *versus* non-HP ILD (c), with corresponding performance on held-out test datasets (b, d). e) Upregulated gene ontology pathways using gene signatures identified in classification models for IPF against non-IPF ILD and HP against non-HP ILD. AUC: area under receiver operating characteristic curve.

their discriminative ability using the same test sets (supplementary figure S7, table S10 and table S11). Using the expanded gene signatures of ILD subtypes against controls, we identified *MMP7*, *COMP*, *DIO2*, *THBS4*, *IL13RA2*, *MEOX1*, *COL17A1* and *SCG5* as shared upregulated genes across all subtypes when compared against controls (supplementary figure S8a, figure S3c and table S6). Shared downregulated genes suggested a decrease in alveolar type 1 (AT1) (*MYRF* and *AGER*) and endothelial (*CA4*, *PRX* and *VIPR1*) cells (supplementary figures S8b and figure S3d). No genes in the signatures for the lung IPF *versus* Control model were shared with peripheral signatures; however, blood and PBMC gene signatures shared 21 genes (supplementary figure S3b).

When comparing IPF against other ILD subtypes, we identified *GPSM2*, *HSPA4L* (previously shown to be involved in the differentiation of AT2 cells [59]) and extracellular matrix-related genes such as *COL16A1*, *ITGA7* and *VCAN* as uniquely upregulated in IPF (supplementary table S6). In HP, we identified antigen presentation and major histocompatibility complex binding as enriched pathways *via PSMB8*, *PSMB9* and *TAPBP* expression (figure 4e), and these genes were also included in the models differentiating HP from IPF and NSIP samples. When compared to controls, SSc-ILD samples were uniquely characterised by expression of *CDH3*, which was not found in signatures of other fibrotic ILD subtypes.

Unsupervised analysis identifies putative molecular endotypes of ILD

We next used biclustering, an unsupervised learning method, to examine lung transcriptomics for specific molecular endotypes. Biclustering algorithms perform simultaneous clustering of rows (genes) and

columns (samples) of a data matrix to identify subsets known as biclusters [19]. We extracted all studies containing lung gene expression data, performed batch correction and used a series of biclustering algorithms to identify 16 biclusters showing enrichment in specific lung disease groups (supplementary figure S9, figure S10 and table S12). Top results from pathway analysis of bicluster genes were used to annotate each bicluster (supplementary table S13 and figure 5a, b). Of the samples with available FVC % and D_{LCO} % predicted data, we identified IPF samples in the "M13-Proliferation" bicluster (n=53) as having lower FVC % compared to non-cluster (n=155) IPF samples (−6.29%, FDR 0.09) (figure 5c), which we confirmed was not due to age (−1.18 years, p=0.15) (supplementary table S14). ILD samples in the "U4-AT1 cells", "U0-Endothelium", "M2-Cytoskeleton", "U1-EMT", "M9-Fibrosis" and "M13-Proliferation" biclusters had significantly lower D_{LCO}% (−7.95% to −13.9%, all FDR<0.05) compared to non-cluster samples, though these samples were all significantly older with the exception of the "M13-Proliferation" samples, which were younger (−2.18 years, FDR 9.17×10^{-3}) (supplementary table S15).

ILD gene signatures and endotypes are associated with cell types

In order to examine the contribution of cell populations in our gene signatures and endotypes, we performed a cellular deconvolution analysis using cell type markers obtained from GSE122960, which is a single-cell lung dataset consisting of eight healthy control, four IPF, two SSc-ILD, one HP and one myositis-ILD samples (supplementary figure S11). IPF, HP, NSIP and SSc-ILD gene signatures were enriched in similar cell types, while fibroblasts were only enriched in IPF gene signatures and the HP signature was uniquely enriched in immune cells when compared to other ILD subtypes (figure 6a and supplementary table S16). Next, we re-analysed GSE122960 to discern whether observed transcriptional

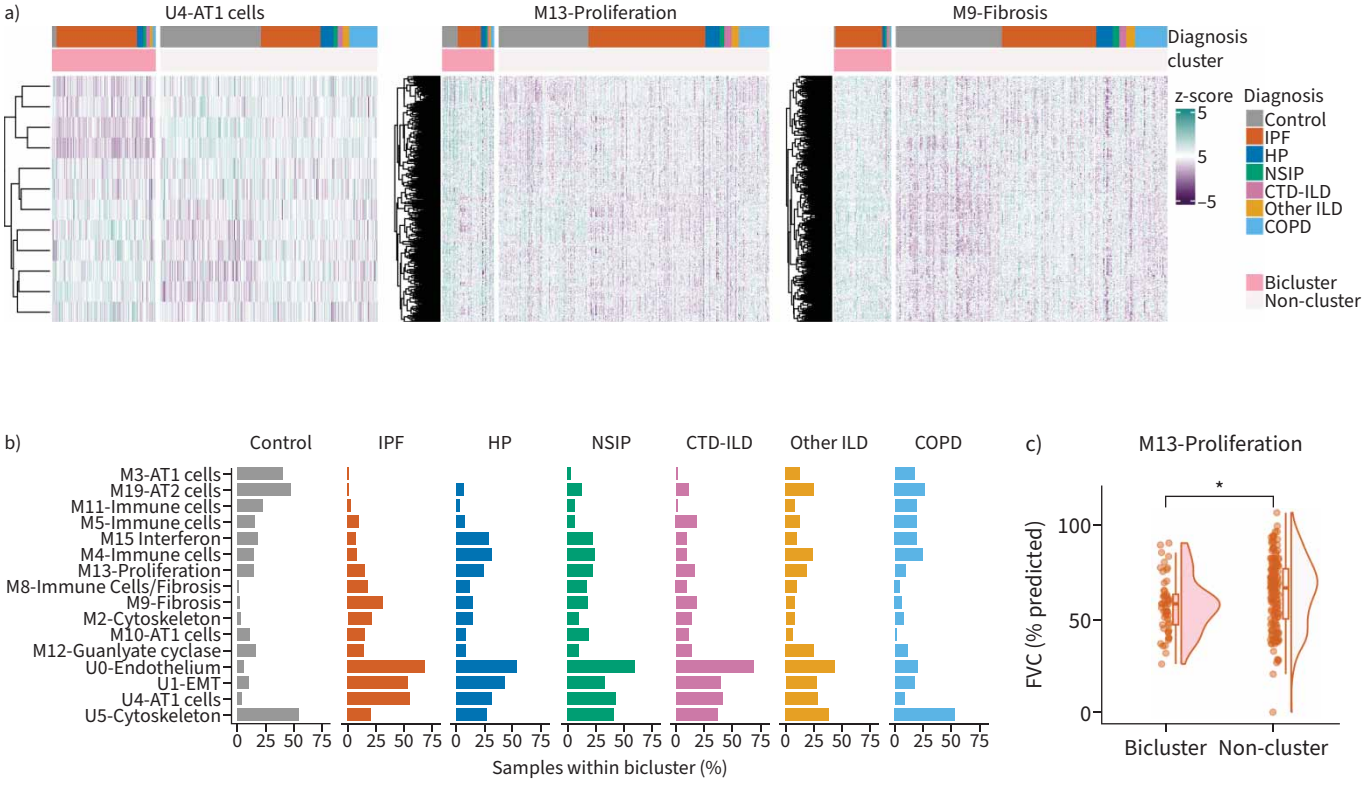


FIGURE 5 Molecular endotypes identified through unsupervised learning are associated with cell types and lung function in interstitial lung disease (ILD). ComBat-corrected lung transcriptomics datasets containing healthy control, ILD and COPD (disease control) samples (from GSE47460) were analysed using biclustering algorithms. a) Heatmaps of genes found in select biclusters showing differences in gene expression between bicluster and non-bicluster samples. b) Proportion of samples within each subtype identified in each bicluster (*i.e.* 32.6% of idiopathic pulmonary fibrosis (IPF) samples are in the "M9-Fibrosis" bicluster). c) Comparison of reported forced vital capacity (FVC % predicted) in IPF samples within and outside of the "M13-Proflieration" bicluster. Significance was determined through linear regression analysis of bicluster membership against FVC %. HP: hypersensitivity pneumonitis; NSIP: nonspecific interstitial pneumonia; CTD-ILD: connective tissue disease interstitial lung disease; AT1: alveolar type 1; AT2: alveolar type 2; EMT: epithelial-to-mesenchymal transition. *: false discovery rate <0.10.

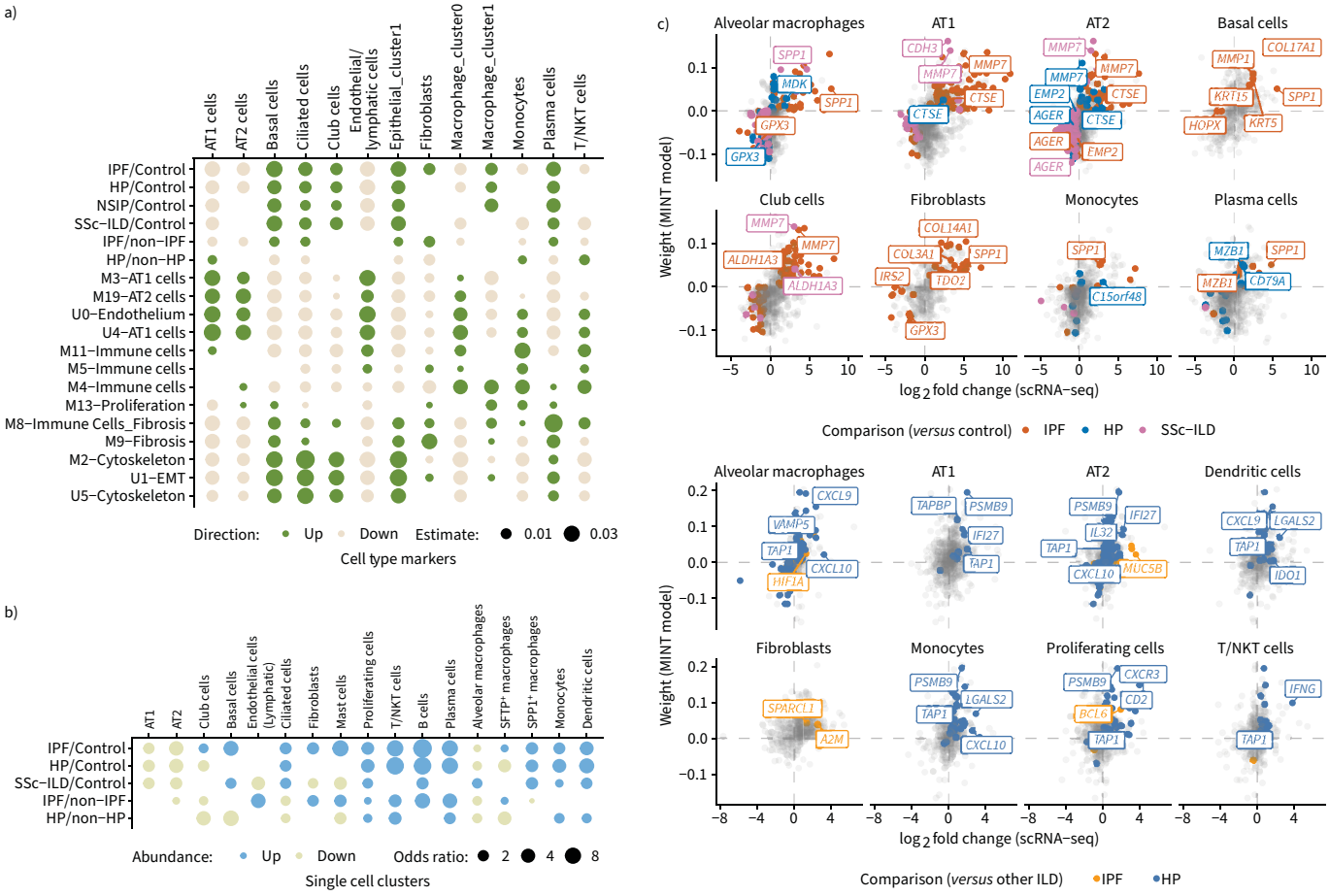


FIGURE 6 Integrated gene signatures are driven by changes in cellular abundance and transcriptional activity. a) Estimate of differences between eigengene values (as determined by cellular deconvolution analysis) of interstitial lung disease (ILD) subtypes using integrated data. Between-subtype comparisons can be read as subtype 1 *versus* subtype 2, *i.e.* "IPF/Control" compares idiopathic pulmonary fibrosis (IPF) against Control samples, with dark green and beige dots representing cell types increased or decreased, respectively, in IPF. Listed bicluster subtypes represent comparisons between bicluster and non-cluster samples. Blank spaces indicate comparisons that were not significant. b) Differences in single-cell RNA sequencing (scRNA-seq) cellular abundance in GSE122960 between indicated subtypes as determined by a binomial generalised linear model. c) Comparison of Multivariate INTegrative (MINT) gene signatures with differentially expressed gene (DEG) analysis of scRNA-seq data. Model weights of each gene identified in classification models were plotted on the y-axis against their $\log_2$ fold change values in scRNA-seq DEG analysis on the x-axis, and coloured if the gene was significantly different in scRNA-seq with the same direction of change as the MINT signature. Selected genes of interest are labelled. A full list of genes can be found in supplementary table S17. HP: hypersensitivity pneumonitis; NSIP: nonspecific interstitial pneumonia; SSc-ILD: systemic sclerosis-associated interstitial lung disease; AT1: alveolar type 1; AT2: alveolar type 2; EMT: epithelial-to-mesenchymal transition; NKT cell: natural killer T-cell.

differences were driven by increases in cellular abundance or transcriptional activity. We compared cell type abundance between ILD subtypes and controls and identified similar results to the deconvolution analysis, with an increase in basal cells, fibroblasts, plasma cells and specific macrophage subsets alongside a decrease in AT1, AT2 and endothelial cells (figure 6b). Within these cell types, we also identified a number of DEGs (*e.g. MMP7*, *SPP1*, *AGER*, *GPX3*, *PSMB9*, *TAPBP* and *SPARCL1*) that were up- or downregulated in the MINT-derived gene signatures as well, which suggests that these molecules are differentially expressed regardless of cellular abundance (figure 6c and supplementary table S17).

Discussion

ILD arises as a result of complex interactions between ageing, genetics and environmental factors that are reflected in the lung transcriptome. Summarising the findings of bulk tissue sequencing is valuable for comparison with future single-cell RNA-seq findings. We identified fibrotic ILD subtype-specific gene signatures in the largest meta-analysis of fibrotic ILD transcriptomics to date, comprising 1767 samples. We also performed the first comprehensive between-subtype transcriptomics analysis of fibrotic ILD to

identify candidate IPF- and HP-specific biomarkers as well as molecular endotypes associated with decreased pulmonary function. The identified fibrotic ILD gene signatures were validated on three recently published datasets, suggesting that they may be applied to future ILD transcriptomics datasets. Although certain genes in the identified signatures, such as *MYRF*, *IL13RA2* and *COMP*, were not differentially expressed at the single-cell level, others, such as *MMP7*, *SFRP2*, *BCL6*, *PSMB9*, *VAMP5* and *TAPBP*, were differentially expressed in specific single-cell clusters, which suggests that these signatures capture changes in cellular abundance and transcriptional activity.

Well-studied genes (*e.g. MMP7*) and disease processes were upregulated across all ILD subtypes, which is consistent with the overlapping radiological and pathological morphology, disease progression and treatment response [60]. In multiple subtypes, upregulated chemokines such as *CCL7*, *CCL11*, *CCL19*, *CCL22*, *CXCL6*, *CXCL12* and *CXCL13* suggest recruitment of inflammatory cell types and fibroblast activation [61–67]. Downregulation of β-catenin binding molecules such as *GSK3B*, which is a component of the destruction complex, as well as *KLF4* [68] and *SMAD7* [69, 70], suggests an increase in β-catenin-mediated activity. Of the Wnt-associated frizzled receptors, *FZD5* was found to be downregulated in IPF, HP and NSIP, and has been shown to prevent epithelial-to-mesenchymal transition (EMT) in cancer cell lines [71, 72].

Unique genes expressed in IPF were primarily associated with the extracellular matrix; in particular, *VCAN*, *SPON1* and *FBLN2* are increased in fibroblastic foci of UIP/IPF lungs [73]. Other genes, such as *COMP* and *COL17A1*, were upregulated when comparing IPF against non-IPF ILD subtypes, which suggests a greater extent of fibrosis in patients with IPF. *SPP1*, which is expressed by pro-fibrotic macrophages [22, 74], was upregulated in IPF and SSc-ILD samples compared to controls, but not in HP or NSIP. The presence of histocompatibility complex and antigen presentation-associated genes (*TAPBP*, *PSMB8* and *PSMB9*) in both HP *versus* Control and HP *versus* non-HP models suggests that they may have diagnostic utility as biomarkers because they are transcriptionally overexpressed in HP and have HP-associated haplotypes [75]. E-cadherin (*CDH3*), which was uniquely identified in the SSc-ILD *versus* Control model but not in other ILD subtypes, forms a key component of the adherens junction that stabilises the airway epithelial barrier and its loss of expression is a hallmark of EMT [76]. Perhaps unsurprisingly, we were unable to identify a model for NSIP *versus* other ILD subtypes, which is consistent with previous studies whose data are included in our systematic review [35, 39, 49]. Given that NSIP is observed in both idiopathic NSIP and many types of connective tissue disease-associated ILD, a consensus gene signature may simply not exist in this heterogenous population.

Sex-stratified analysis suggests there are similarities between females and males in the pathogenesis of ILD, because the majority of our sex-specific models had comparable performance on sex-stratified test sets. Pathway overrepresentation analysis of genes unique to each sex-specific model revealed immune-associated pathways in female classification models. Both female-specific IPF and NSIP *versus* Control models contained *CD79A*, a B-cell marker, which is consistent with females generally having higher B-cell numbers and associated gene expression than males [77]. Previous studies have identified increases in B-cells in patients with ILD [73, 78, 79], thus warranting further investigations into the role of B-cells and sex in ILD.

Through the use of two unsupervised biclustering algorithms, we identified molecular endotypes of ILD representative of disease mechanisms such as immune infiltration and fibrosis. The “M2-Cytoskeleton” bicluster contained genes associated with ciliated cells (*e.g. DYNAH7* and *DYNAH9*), which may be associated with the microscopic honeycombing seen in IPF/UIP [44]. Two biclusters, “U4-AT1 cells” and “U0-Endothelium”, were annotated due to downregulation of genes associated with AT1 cells (*AGER*, *CAV2* and *EMP2*) and endothelial cells (*EPAS1*, *TEK* and *STARD13*), and destruction of these tissues during the pathogenesis of ILD may explain the association with reduced D_{LCO} % in ILD samples. The “M13-Proliferation” bicluster was associated with decreased FVC % and D_{LCO} % in all ILD samples, and in IPF patients it was associated with an ~6% decline in FVC %. Pathway analysis of this bicluster’s genes (*e.g. CDKN2*, *TOP2A* and *CCNB1*) revealed associations with proliferating macrophages, basal cells and natural killer/T-cells [80], which we were able to corroborate in our cellular deconvolution analysis; one of the genes, *CCNA2* is upregulated in acute exacerbations of IPF [37].

Taken together, our supervised and unsupervised analyses suggest that current diagnostic classifications do not fully capture the molecular heterogeneity of ILD subtypes. When comparing genes from our expanded IPF *versus* non-IPF ILD model to the UIP Envisia classifier, only 10 out of 190 genes (*PDLIM5*, *GREM1*, *TUBB2B*, *CHRDL2*, *KIF12*, *SLC4A3*, *PPIC*, *ZNF454*, *SELE* and *MYO3B*) were shared (supplementary figure S12), which might be explained by UIP patterns being observed in other fibrotic ILD subtypes as

well as IPF diagnoses not being limited to UIP patterns. While our classification models perform well in separating IPF or HP from other ILD subtypes, it is evident from the unsupervised analysis that there exists molecular heterogeneity within these subtypes, and these putative endotypes are associated with lung function. Considering the complex drivers and heterogenous presentations of ILD subtypes, future classification based on data-driven molecular endotypes may be clinically beneficial with respect to disease management as well as research and development into novel therapeutics.

This meta-analysis has several limitations. The majority of studies reported clinical and demographic variables (*e.g.* radiological pattern, age, pulmonary function tests and mortality) as summary tables, thereby prohibiting detailed subgroup analyses. However, using available metadata, we were able to examine age (n=852), FVC % predicted (n=309) and D_{LCO} % predicted (n=259) from ILD samples in our unsupervised analysis. Future work in fibrotic ILD transcriptomics should consider the investigation of less profiled subtypes to enrich our understanding of non-IPF pulmonary fibrosis, because the overrepresentation of IPF in our analysis limits the generalisability of our findings, particularly at the single-cell level. We have previously noted compartmental differences between blood and PBMCs using NanoString assays, finding that blood shows greater sensitivity in detecting DEGs between asthma and control samples [81]. Our IPF whole blood and PBMC models were validated on training datasets for the other medium (*i.e.* whole blood model validated on PBMC training data), and 21 genes were shared between both gene signatures (supplementary figure S3), which suggests that the weighting performed by MINT classification can overcome differences in method- and sample-specific RNA detection. Finally, prediction thresholds in MINT are based upon minimum sample distance to class centroids within the partial least squares projection space, thus making direct comparison to existing classifiers difficult. Nevertheless, our models had good AUCs and BERs on held-out datasets, thereby providing more confidence in their discriminatory ability and relevance to ILD pathobiology.

In summary, by using both supervised and unsupervised approaches, our meta-analysis has identified reproducible fibrotic ILD subtype-specific gene signatures and classifiers, and putative molecular endotypes of ILD that reflect the pathogenesis of fibrotic lung disease. Further studies are required to investigate the utility of the herein identified gene signatures and molecular endotypes in a clinical setting.

Acknowledgements: We would like to thank Helen Brown from the UBC library for her assistance in creating a systematic review protocol, as well as Jan Baumbach and Olga Zolotareva from the University of Hamburgh for their guidance in biclustering methods.

Author contributions: D. He, S.A. Guler, S.J. Tebbutt and C.J. Ryerson conceived the study. D. He and S.A. Guler created and performed the systematic review protocol. D. He performed the analyses, with support from C.P. Shannon. D. He wrote the manuscript draft with support from all authors. All authors have reviewed and agreed on the final submitted version of the manuscript.

Conflict of interest: S.A. Guler reports grants from Boehringer Ingelheim, Roche, MSD and Janssen; payment or honoraria for lectures, presentations, manuscript writing or educational events and support for attending meetings from Boehringer Ingelheim; and participation on a data safety monitoring board or advisory board with Boehringer Ingelheim, MSD and Janssen. C.J. Ryerson reports grants from Boehringer Ingelheim; consulting fees from Boehringer Ingelheim, Pliant Therapeutics, AstraZeneca, Trevi Therapeutics and Veracyte; payment or honoraria for lectures, presentations, manuscript writing or educational events from Hoffmann-La Roche, Boehringer Ingelheim and Cipla; payment for expert testimony from Boehringer Ingelheim; and support for attending meetings from Cipla and Boehringer Ingelheim. The remaining authors have no potential conflicts of interest to disclose.

Support statement: Salary support in the form of grants and scholarships from Canadian Institutes for Health Research, BC Lung Foundation, MITACS.

References

1 Kaul B, Cottin V, Collard HR, *et al.* Variability in global prevalence of interstitial lung disease. *Front Med* 2021; 8: 751181.
2 Vukmirovic M, Kaminski N. Impact of transcriptomics on our understanding of pulmonary fibrosis. *Front Med* 2018; 5: 87.
3 Khan FA, Stewart I, Saini G, *et al.* A systematic review of blood biomarkers with individual participant data meta-analysis of matrix metalloproteinase-7 in idiopathic pulmonary fibrosis. *Eur Respir J* 2022; 59: 2101612.

4 Spagnolo P, Ryerson CJ, Putman R, *et al.* Early diagnosis of fibrotic interstitial lung disease: challenges and opportunities. *Lancet Respir Med*; 2021; 9: 1065–1076.
5 Lasky JA, Case A, Unterman A, *et al.* The impact of the Envisia genomic classifier in the diagnosis and management of patients with idiopathic pulmonary fibrosis. *Ann Am Thorac Soc* 2022; 19: 916–924.
6 Raghu G, Remy-Jardin M, Richeldi L, *et al.* Idiopathic pulmonary fibrosis (an update) and progressive pulmonary fibrosis in adults: an official ATS/ERS/JRS/ALAT clinical practice guideline. *Am J Respir Crit Care Med* 2022; 205: e18–e47.
7 Dupuy A, Simon RM. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J Natl Cancer Inst* 2007; 99: 147–157.
8 Rohart F, Eslami A, Matigian N, *et al.* MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. *BMC Bioinformatics* 2017; 18: 128.
9 Rohart F, Gautier B, Singh A, *et al.* mixOmics: an R package for 'omics feature selection and multiple data integration. *PLOS Comput Biol* 2017; 13: e1005752.
10 Kolde R, Laur S, Adler P, *et al.* Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* 2012; 28: 573–580.
11 Bauer Y, Tedrow J, de Bernard S, *et al.* A novel genomic signature with translational significance for human idiopathic pulmonary fibrosis. *Am J Respir Cell Mol Biol* 2015; 52: 217–231.
12 DePianto DJ, Chandriani S, Abbas AR, *et al.* Heterogeneous gene expression signatures correspond to distinct lung pathologies and biomarkers of disease severity in idiopathic pulmonary fibrosis. *Thorax* 2015; 70: 48–56.
13 Schafer MJ, White TA, Iijima K, *et al.* Cellular senescence mediates fibrotic pulmonary disease. *Nat Commun* 2017; 8: 14532.
14 Yu X, Gu P, Huang Z, *et al.* Reduced expression of BMP3 contributes to the development of pulmonary fibrosis and predicts the unfavorable prognosis in IIP patients. *Oncotarget* 2017; 8: 80531–80544.
15 Cecchini MJ, Hosein K, Howlett CJ, *et al.* Comprehensive gene expression profiling identifies distinct and overlapping transcriptional profiles in non-specific interstitial pneumonia and idiopathic pulmonary fibrosis. *Respir Res* 2018; 19: 153.
16 DePianto DJ, Heiden JAV, Morshead KB, *et al.* Molecular mapping of interstitial lung disease reveals a phenotypically distinct senescent basal epithelial cell population. *JCI Insight* 2021; 6: e143626.
17 De Sadeleer LJ, McDonough JE, Schupp JC, *et al.* Lung microenvironments and disease progression in fibrotic hypersensitivity pneumonitis. *Am J Respir Crit Care Med* 2022; 205: 60–74.
18 Jaffar J, Wong M, Fishbein GA, *et al.* Matrix metalloproteinase-7 is increased in lung bases but not apices in idiopathic pulmonary fibrosis. *ERJ Open Res* 2022; 8: 00191-2022.
19 Rose TD, Bechtler T, Ciora O-A, *et al.* MoSBi: automated signature mining for molecular stratification and subtyping. *Proc Natl Acad Sci USA* 2022; 119: e2118210119.
20 Zolotareva O, Khakabimamaghani S, Isaeva OI, *et al.* Identification of differentially expressed gene modules in heterogeneous diseases. *Bioinformatics* 2021; 37: 1691–1698.
21 McKenzie AT, Wang M, Hauberg ME, *et al.* Brain cell type specific gene expression and co-expression network architectures. *Sci Rep* 2018; 8: 8868.
22 Reyfman PA, Walter JM, Joshi N, *et al.* Single-cell transcriptomic analysis of human lung provides insights into the pathobiology of pulmonary fibrosis. *Am J Respir Crit Care Med* 2019; 199: 1517–1536.
23 Hao Y, Stuart T, Kowalski MH, *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* 2024; 42: 293–304.
24 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.
25 Molyneaux PL, Willis-Owen SAG, Cox MJ, *et al.* Host–microbial interactions in idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2017; 195: 1640–1650.
26 Koth LL, Solberg OD, Peng JC, *et al.* Sarcoidosis blood transcriptome reflects lung inflammation and overlaps with tuberculosis. *Am J Respir Crit Care Med* 2011; 184: 1153–1163.
27 Jia G, Ramalingam TR, Heiden JV, *et al.* An interleukin 6 responsive plasma cell signature is associated with disease progression in systemic sclerosis interstitial lung disease. *iScience* 2023; 26: 108133.
28 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: 1138–1148.
29 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.
30 Cheadle C, Berger AE, Mathai SC, *et al.* Erythroid-specific transcriptional changes in PBMCs from pulmonary hypertension patients. *PLoS One* 2012; 7: e34951.
31 Assassi S, Volkmann ER, Zheng WJ, *et al.* Peripheral blood gene expression profiling shows predictive significance for response to mycophenolate in systemic sclerosis-related interstitial lung disease. *Ann Rheum Dis* 2022; 81: 854–860.
32 Zuo F, Kaminski N, Eugui E, *et al.* Gene expression analysis reveals matrilysin as a key regulator of pulmonary fibrosis in mice and humans. *Proc Natl Acad Sci USA* 2002; 99: 6292–6297.

33 Pardo A, Gibson K, Cisneros J, *et al.* Up-regulation and profibrotic role of osteopontin in human idiopathic pulmonary fibrosis. *PLoS Med* 2005; 2: e251.

34 Selman M, Pardo A, Barrera L, *et al.* Gene expression profiles distinguish idiopathic pulmonary fibrosis from hypersensitivity pneumonitis. *Am J Respir Crit Care Med* 2006; 173: 188–198.

35 Yang IV, Burch LH, Steele MP, *et al.* Gene expression profiling of familial and sporadic interstitial pneumonia. *Am J Respir Crit Care Med* 2007; 175: 45–54.

36 Bridges RS, Kass D, Loh K, *et al.* Gene expression profiling of pulmonary fibrosis identifies twist1 as an antiapoptotic molecular "rectifier" of growth factor signaling. *Am J Pathol* 2009; 175: 2351–2361.

37 Konishi K, Gibson KF, Lindell KO, *et al.* Gene expression profiles of acute exacerbations of idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2009; 180: 167–175.

38 Rajkumar R, Konishi K, Richards TJ, *et al.* Genomewide RNA expression profiling in lung identifies distinct signatures in idiopathic pulmonary arterial hypertension and secondary pulmonary hypertension. *Am J Physiol Heart Circ Physiol* 2010; 298: H1235–H1248.

39 Cho J-H, Gelinas R, Wang K, *et al.* Systems biology of interstitial lung diseases: integration of mRNA and microRNA expression changes. *BMC Med Genomics* 2011; 4: 8.

40 Hsu E, Shi H, Jordan RM, *et al.* Lung tissues in patients with systemic sclerosis have gene expression patterns unique to pulmonary fibrosis and pulmonary hypertension. *Arthritis Rheum* 2011; 63: 783–794.

41 Meltzer EB, Barry WT, D'Amico TA, *et al.* Bayesian probit regression model for the diagnosis of pulmonary fibrosis: proof-of-principle. *BMC Med Genomics* 2011; 4: 70.

42 Sanders YY, Ambalavanan N, Halloran B, *et al.* Altered DNA methylation profile in idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2012; 186: 525–535.

43 Deng N, Sanchez CG, Lasky JA, *et al.* Detecting splicing variants in idiopathic pulmonary fibrosis from non-differentially expressed genes. *PLOS One* 2013; 8: e68352.

44 Yang IV, Coldren CD, Leach SM, *et al.* Expression of cilium-associated genes defines novel molecular subtypes of idiopathic pulmonary fibrosis. *Thorax* 2013; 68: 1114–1121.

45 Nance T, Smith KS, Anaya V, *et al.* Transcriptome analysis reveals differential splicing events in IPF lung tissue. *PloS One* 2014; 9: e92111.

46 Geng J, Huang X, Li Y, *et al.* Down-regulation of USP13 mediates phenotype transformation of fibroblasts in idiopathic pulmonary fibrosis. *Respir Res* 2015; 16: 124.

47 Christmann RB, Wooten A, Sampaio-Barros P, *et al.* miR-155 in the progression of lung fibrosis in systemic sclerosis. *Arthritis Res Ther* 2016; 18: 155.

48 Horimasu Y, Ishikawa N, Iwamoto H, *et al.* Clinical and molecular features of rapidly progressive chronic hypersensitivity pneumonitis. *Sarcoidosis Vasc Diffuse Lung Dis* 2017; 34: 48–57.

49 Horimasu Y, Ishikawa N, Taniwaki M, *et al.* Gene expression profiling of idiopathic interstitial pneumonias (IIPs): identification of potential diagnostic markers and therapeutic targets. *BMC Med Genet* 2017; 18: 88.

50 Vukmirovic M, Herazo-Maya JD, Blackmon J, *et al.* Identification and validation of differentially expressed transcripts by RNA-sequencing of formalin-fixed, paraffin-embedded (FFPE) lung tissue from patients with idiopathic pulmonary fibrosis. *BMC Pulm Med* 2017; 17: 15.

51 Luzina IG, Salcedo MV, Rojas-Peña ML, *et al.* Transcriptomic evidence of immune activation in macroscopically normal-appearing and scarred lung tissues in idiopathic pulmonary fibrosis. *Cell Immunol* 2018; 325: 1–13.

52 Sivakumar P, Thompson JR, Ammar R, *et al.* RNA sequencing of transplant-stage idiopathic pulmonary fibrosis lung reveals unique pathway regulation. *ERJ Open Res* 2019; 5: 00117-2019.

53 McDonough JE, Ahangari F, Li Q, *et al.* Transcriptional regulatory model of fibrosis progression in the human lung. *JCI Insight* 2019; 4: e131597.

54 Furusawa H, Cardwell JH, Okamoto T, *et al.* Chronic hypersensitivity pneumonitis, an interstitial lung disease with distinct molecular signatures. *Am J Respir Crit Care Med* 2020; 202: 1430–1444.

55 Konigsberg IR, Borie R, Walts AD, *et al.* Molecular signatures of idiopathic pulmonary fibrosis. *Am J Respir Cell Mol Biol* 2021; 65: 430–441.

56 Borie R, Cardwell J, Konigsberg IR, *et al.* Colocalization of gene expression and DNA methylation with genetic risk variants supports functional roles of MUC5B and DSP in idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2022; 206: 1259–1270.

57 Wang S, Liu M, Li X, *et al.* Canonical and noncanonical regulatory roles for JAK2 in the pathogenesis of rheumatoid arthritis-associated interstitial lung disease and idiopathic pulmonary fibrosis. *FASEB J* 2022; 36: e22336.

58 Huang Y, Guzy R, Ma S-F, *et al.* Central lung gene expression associates with myofibroblast features in idiopathic pulmonary fibrosis. *BMJ Open Respir Res* 2023; 10: e001391.

59 Mohamed BA, Barakat AZ, Held T, *et al.* Respiratory distress and early neonatal lethality in Hspa4l/Hspa4 double-mutant mice. *Am J Respir Cell Mol Biol* 2014; 50: 817–824.

60 Flaherty KR, Wells AU, Cottin V, *et al.* Nintedanib in progressive fibrosing interstitial lung diseases. *N Engl J Med* 2019; 381: 1718–1727.

61 Choi ES, Jakubzick C, Carpenter KJ, *et al.* Enhanced monocyte chemoattractant protein-3/CC chemokine ligand-7 in usual interstitial pneumonia. *Am J Respir Crit Care Med* 2004; 170: 508–515.
62 Puxeddu I, Bader R, Piliponsky AM, *et al.* The CC chemokine eotaxin/CCL11 has a selective profibrogenic effect on human lung fibroblasts. *J Allergy Clin Immunol* 2006; 117: 103–110.
63 Pierce EM, Carpenter K, Jakubzick C, *et al.* Idiopathic pulmonary fibrosis fibroblasts migrate and proliferate to CC chemokine ligand 21. *Eur Respir J* 2007; 29: 1082–1093.
64 Yogo Y, Fujishima S, Inoue T, *et al.* Macrophage derived chemokine (CCL22), thymus and activation-regulated chemokine (CCL17), and CCR4 in idiopathic pulmonary fibrosis. *Respir Res* 2009; 10: 80.
65 Sieber P, Schäfer A, Lieberherr R, *et al.* NF-κB drives epithelial–mesenchymal mechanisms of lung fibrosis in a translational lung cell model. *JCI Insight* 2023; 8: e154719.
66 Li F, Xu X, Geng J, *et al.* The autocrine CXCR4/CXCL12 axis contributes to lung fibrosis through modulation of lung fibroblast activity. *Exp Ther Med* 2020; 19: 1844–1854.
67 Vuga LJ, Tedrow JR, Pandit KV, *et al.* C-X-C motif chemokine 13 (CXCL13) is a prognostic biomarker of idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2014; 189: 966–974.
68 Evans PM, Chen X, Zhang W, *et al.* KLF4 interacts with β-catenin/TCF4 and blocks p300/CBP recruitment by β-catenin. *Mol Cell Biol* 2010; 30: 372–381.
69 Edlund S, Lee SY, Grimsby S, *et al.* Interaction between Smad7 and β-catenin: importance for transforming growth factor β-induced apoptosis. *Mol Cell Biol* 2005; 25: 1475–1488.
70 Tang Y, Liu Z, Zhao L, *et al.* Smad7 stabilizes β-catenin binding to E-cadherin complex and promotes cell–cell adhesion. *J Biol Chem* 2008; 283: 23956–23963.
71 Dong D, Na L, Zhou K, *et al.* FZD5 prevents epithelial–mesenchymal transition in gastric cancer. *Cell Commun Signal* 2021; 19: 21.
72 Na L, Wang Z, Bai Y, *et al.* WNT7B represses epithelial–mesenchymal transition and stem-like properties in bladder urothelial carcinoma. *Biochim Biophys Acta Mol Basis Dis* 2022; 1868: 166271.
73 Herrera JA, Dingle L, Montero MA, *et al.* The UIP/IPF fibroblastic focus is a collagen biosynthesis factory embedded in a distinct extracellular matrix. *JCI Insight* 2022; 7: e156115.
74 Morse C, Tabib T, Sembrat J, *et al.* Proliferating SPP1/MERTK-expressing macrophages in idiopathic pulmonary fibrosis. *Eur Respir J* 2019; 54: 1802441.
75 Camarena A, Aquino-Galvez A, Falfán-Valencia R, *et al.* PSMB8 (LMP7) but not PSMB9 (LMP2) gene polymorphisms are associated to pigeon breeder's hypersensitivity pneumonitis. *Respir Med* 2010; 104: 889–894.
76 Bartis D, Mise N, Mahida RY, *et al.* Epithelial–mesenchymal transition in lung development and disease: does it exist and is it important? *Thorax* 2014; 69: 760–765.
77 Klein SL, Flanagan KL. Sex differences in immune responses. *Nat Rev Immunol* 2016; 16: 626–638.
78 Schiller HB, Mayr CH, Leuschner G, *et al.* Deep proteome profiling reveals common prevalence of MZB1-positive plasma B cells in human lung and skin fibrosis. *Am J Respir Crit Care Med* 2017; 196: 1298–1310.
79 Ali MF, Egan AM, Shaughnessy GF, *et al.* Antifibrotics modify B-cell–induced fibroblast migration and activation in patients with idiopathic pulmonary fibrosis. *Am J Respir Cell Mol Biol* 2021; 64: 722–733.
80 Travaglini KJ, Nabhan AN, Penland L, *et al.* A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* 2020; 587: 619–625.
81 He D, Yang CX, Sahin B, *et al.* Whole blood vs PBMC: compartmental differences in gene expression profiling exemplified in asthma. *Allergy Asthma Clin Immunol* 2019; 15: 67.