Burden of rare pathogenic variants suggests disrupted cytoskeletal organisation in the pathogenesis of pulmonary fibrosis

Konain Bhatti* [1], Dapeng Wang* [1,2], Hernan Fainberg[1], Andressa Alves Bordignon[1],Yujie Ni[1], Bin Liu[1], Alison John[1,3], Richard Allen[4,5], Louise V. Wain[4,5], Simon R. Johnson[6], Toby M Maher[1,7], Philip L Molyneaux[1,3], Elisabetta Renzoni[1,8], Gauri Saini[6], Deborah Morris-Rosendahl[1,9], R Gisli Jenkins[1,3], and Iain Stewart[1,3]

[1]National Heart and Lung Institute, Imperial College London, London, UK

[2]Shandong Key Laboratory of Intelligent Oil & Gas Industrial Software, Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), Qingdao, China

[3]NIHR Imperial Biomedical Research Centre, Imperial College London, London, UK

[4]Division of Public Health and Epidemiology, School of Medicine, University of Leicester, Leicester, UK

[5]University Hospitals of Leicester NHS Trust, Leicester, UK

[6]Centre for Respiratory Research, NIHR Biomedical Research Centre and Biodiscovery Institute, Translational Medical Sciences, University of Nottingham, Nottingham, UK

[7]Keck Medicine of University of Southern California, Los Angeles, California, USA

[8]The Interstitial Lung Disease Unit, Royal Brompton Hospital, London, United Kingdom

[9]Clinical Genetics and Genomics Laboratory, Royal Brompton Hospital, Guy's and St Thomas' NHS Foundation Trust, London, UK

*These authors contributed equally

Corresponding author:

Iain Stewart, National Heart and Lung Institute, Imperial College London, London, SW3 6LY, UK

Email: iain.stewart@imperial.ac.uk

Abstract

Background

Rare genetic variants contribute to pulmonary fibrosis (PF) risk and outcome, with known variants highlighting the importance of impaired telomere maintenance and surfactant biology. However, much of the disrupted genetic architecture of PF remains unexplained. This study aimed to identify genes with rare pathogenic coding variants that represented a burden at the exon level associated with the pathogenesis of PF.

Methods

PF case cohorts included the PROFILE study of incident idiopathic pulmonary fibrosis (IPF) and PF-classified participants from the Genomics England 100K (GE100KGP) study. Whole genome sequencing data were used to test the burden of rare protein altering variants (PAVs; CADD score >20) defined by minor allele frequency <0.1% in the summary level gnomAD reference database (v3.1.2). A summary exon-level pathogenicity score was derived by standardising PAV predictions from functional annotation tools (AlphaMissense, REVEL, ClinPred, CADD, PolyPhen-2, SIFT) averaged across exons. SKAT-O based kernel regression associated exon-level pathogenicity features with disease progression and survival. Single cell lung transcriptomic datasets were harmonised to evaluate expression patterns and gene ontology characterised enriched molecular functions.

Results

Across 507 PROFILE cases and 451 GE100KGP cases, each compared against 76,156 reference participants, 77 genes showed overlapping significant rare PAV burden, comprising 206 concordant PAVs. Of these, 15 genes had exons with pathogenicity scores that were associated with worse clinical outcomes in IPF, including *FAT4* exon-10 and *DNAH7* exon-43. Dysregulated gene expression of *COL6A3* and *FAT4* was observed in IPF lung fibroblasts, while *DNAH7*, *DNAH12* and *PCM1* showed high expression in IPF epithelial cells. The top enriched molecular functions were related to cytoskeletal motor activity.

Conclusions

Rare PAV burden highlighted that pathogenicity within specific exons was associated with PF risk and worse clinical outcomes, identifying genes with dysregulated expression in disease lung cells. The genetic architecture implicates disrupted cytoskeletal organisation in the pathogenesis of PF.

Introduction

Pulmonary fibrosis (PF) is a chronic, progressive process characteristic of many interstitial lung diseases (ILD) including idiopathic pulmonary fibrosis (IPF) and familial pulmonary fibrosis (FPF)[1]. Although antifibrotic treatments such as pirfenidone and nintedanib can slow disease progression, people with PF continue to experience poor prognosis and limited survival after diagnosis[2]. The considerable overlap in clinical characteristics, genetic variants and environmental risk factors between IPF and FPF[3-5], suggests common molecular mechanisms underlie PF pathogenesis.

Genome-wide association studies (GWAS) have identified common variants in coding and non-coding regulatory regions linked to IPF susceptibility, with over 30 signals associated with increased risk[6]. Additional analyses have identified common variants associated with prognosis, including *PCSK6* which was associated with IPF survival[7], and *PKN2* with longitudinal lung function decline of IPF[8].

Whole genome sequencing (WGS) supports identification of rare pathogenic variants across the genome, including coding regions in which variants may lead to truncation and loss of function that alters protein function as well as changes in gene expression. Whole exome sequencing (WES) has highlighted rare missense variants in genes implicating mitotic checkpoint signalling as a dysregulated pathway in fibrotic development[9, 10], whilst rare variants in surfactant genes identified in sporadic PF have also been identified in FPF[11]. Telomere-related rare variants have been associated with worse survival in two independent IPF cohorts, increasing estimated risk of mortality by 57% in meta-analysis[12], and explain up to 23% of FPF risk[13]. Despite advances in understanding the genetic architecture of PF, a large proportion of genetic risk and genetically dysregulated molecular mechanisms are yet to be elucidated.

This study aims to prioritise biologically plausible genes and pathways that contribute to PF susceptibility and progression by evaluating the burden of rare pathogenic protein-altering variants (PAVs), consolidating the predicted pathogenicity at exon-level, and interpreting these findings in the context of transcriptomic signals and molecular pathways.

Methods

Populations

WGS data were obtained from the PROFILE cohort, a UK based prospective study of incident IPF diagnoses[14], and the Genomics England 100K Genome Project (GE100KGP), a UK initiative aiming to produce approximately 100,000 human genomes using WGS to study the causes of cancer and rare diseases. Analysis of GE100KGP data was performed within the Research Environment of Genomics England. The case samples in GE100KGP cohort were defined based on data release v14 as those recruited to the rare disease FPF programme or patients with Human Phenotype Ontology (HPO) codes HP:0002206, HP:0006530, or International Classification of Diseases (ICD-10) codes J84.1, J84.9, J84.8, J67.9 (Supplementary Table S1). Quality control procedures applied to PROFILE and GE100KGP case cohorts included sex concordance, confirmation of European ancestry and exclusion of related individuals[9] (Supplementary Methods). Control data were derived from summary level variant call format (VCF) files from the Genome Aggregation Database (gnomAD v3.1.2)[15], comprising 76,156 individuals with WGS aligned to the GRCh38 reference genome without enrichment of any specific diseases or restriction on ancestry.

Rare variant burden testing

Rare variants were defined as variants with an allele frequency < 0.1%. Additional filtering for internal allele frequency cut-off was also applied (Supplementary Table S2). Variants in coding regions were defined as protein-altering or synonymous based on the consequence predicted by the Ensembl Variant Effect Predictor (VEP) and Combined Annotation–Dependent Depletion (CADD) score provided within the gnomAD VCF files[16]. The Testing Rare vAriants using Public Data (TRAPD) pipeline was used for the burden testing between the individual level case data and summary level control data[17]. Potentially pathogenic rare variants were identified, the relationships between the variants and genes were constructed, and the number of individuals who carried the specific alleles was calculated for both case and control cohorts. Burden testing was performed by using the Fisher exact test at P-value $< 1 \times 10^{-10}$ under the dominant model. Genes that were significant for synonymous variants were excluded from subsequent prioritisation analysis to reduce false discovery (Supplementary Table S2).

Exon-level burden assessment

Genes showing significant rare PAV burden in both cohorts relative to the gnomAD reference in TRAPD analyses were taken forward. Variants were filtered using gnomAD sequencing coverage metrics, excluding loci with insufficient coverage (depth ≥20× in <50% of gnomAD samples). To support exon-level aggregation, genes were retained only if they contained at least one QC-passing rare PAV within the same exon in both cohorts. Predicted pathogenicity for individual rare variants was summarised using established in-silico predictors, Polyphen-2[18], SIFT[19], CADD[20], alphaMissense[21], ClinPred[22], REVEL[23]. Outputs from each tool were normalised to a common scale 0-1 (higher values indicating greater predicted pathogenicity) and summarised as a variant-level pathogenicity score (range 0-6). Agreement between the summary score and predictor outputs was assessed with Pearson correlation, r=0.7 with AlphaMissense, REVEL, ClinPred and PolyPhen-2 (Supplementary Figure S1). Exon-level burden was defined as the mean summary pathogenicity score of rare PAVs within each exon, with exons retained in downstream analysis if mean score ≥1.5 and containing at least one concordant rare PAV across case cohorts.

Functional association of exon-burden with prognosis

The PROFILE prospective observational cohort of incident IPF participants includes disease progression, defined as 10% relative decline of percent predicted forced vital capacity (ppFVC) or mortality within 12 months, and overall survival. In downstream exon-level association analyses, rare PAVs within retained exons were collapsed into a per-participant exon burden score summarising the cumulative predicted pathogenicity of carried variants. SKAT-O was used as a kernel regression framework applied to exon burden features to evaluate whether variation in exon burden scores was associated with prognosis. The individual-by-exon burden matrix was used as the input feature matrix for SKAT-O to test exon-level associations with dichotomous endpoints of disease progression and survival below median, adjusting for age. Region-based associations have been similarly applied in other diseases[24]. Standardised effect sizes were derived using the squared Wald statistic (F-statistic), and hierarchical clustering was performed using Ward linkage on Euclidean distance (SciPy) to visualise functional relationships. An f-statistic ≥1.5 was interpreted to represent a model in which the variance explained exceeded the residual error by at least 50% and used to select associations. Disease outcomes were unavailable in GE100KGP.

Differential gene expression and pathway analyses

Three publicly available single cell RNA sequencing (scRNA-seq) lung datasets were retrieved and integrated to create a representative dataset; GSE136831[25], GSE135893[26], and GSE128033[27] were obtained from the Gene Expression (GEO) Omnibus (GEO) . To support consistency with the PROFILE cases in functional associations, only IPF and control lung samples were included (Supplementary Methods). Dataset integration was performed using Harmony to align multi-dataset cells while preserving biologically meaningful variability (Supplementary Figure S2-4)[28]. scRNA-seq preprocessing, clustering, and data handling were performed in Seurat, with cell-type annotation guided by the Human Lung Cell Atlas. Differential gene expression analysis was performed with DESeq2 pseudobulk aggregation per donor per cell type. To compare expression of genes across cell types and disease status, z-scores were calculated from average expression values of genes within each annotated cell type regardless of disease status and plotted for IPF samples. Gene Ontology (GO) molecular function term enrichment was assessed using g:Profiler, the top ten significant terms were plotted using -log10 transformed p-values.

Results

Pathogenic rare variant burden by case

After pre-processing of 541 WGS PROFILE samples, 507 case samples were used in downstream analysis. A total of 451 PF case samples from GE100KGP were used in downstream analysis. The average age of people in PROFILE cases was 70.5 years (±8.4), which was older than the GE100KGP cases of 65.5 years (±14.5), whilst PROFILE cases were also represented by more males than GE100KGP cases, 76.3% vs 51.4%, respectively (Supplementary Table S3).

In the PROFILE cohort, 371 genes contained PAVs and were significantly associated with PF (Figure 1A). In the GE100KGP cases, 361 genes contained PAVs and were significantly associated with PF (Figure 1B). Initially, there were 87 genes that were overlapping across both case cohorts (Figure 1C), which were reduced to 77 genes meeting coverage criteria. Overlapping genes were represented by 997 rare PAVs in PROFILE and 920 in GE100KGP, including 206 variants that were concordant across both cohorts (Figure 1D).

A significant burden of protein-altering variants was observed in both PROFILE and GE100KGP cohorts for the known telomere-associated risk genes *TERT* (p=7.87E-15, p=1.36E-18, respectively) and *RTEL1* (p=4.08E-17, p=8.42E-15, respectively) (Supplementary Table S4). Following exclusion based on gnomAD coverage, there were 5 unique *RTEL1* variants in PROFILE (case n=6) and 3 in GE100KGP (case n=3). *TERT* was retained in PROFILE (2 unique variants, case n=2), but not GE100KGP.

Exon-level summary burden

There were 40 genes that were represented by 68 exons with mean summary exon-level pathogenicity ≥1.5 and at least one concordant PAV between case cohorts. In PROFILE, 168 variants were retained and in GE100KGP, 162 variants were retained. The greatest mean summary pathogenicity of 4.39 was observed for *DNAH7* exon 43 (Figure 2A). Twenty-two exons (32%) had a mean pathogenicity ≥3. Twelve gene exons included >1 concordant variant across cohorts, *FAT4* exon-2 included four concordant variants (mean pathogenicity 2.32). In the PROFILE cohort, the gene exons with the most frequent case burden of variants were *KIF26A* exon-12 (case n=15), followed by FAT3 exon-10 (case n=12), and *RANBP2* exon-20 (case n=10) (Figure 2B). In the GE100KGP cohort, *KIF26A* exon-12 variants and *FAT3* exon-10 variants were similarly represented by the most frequent number of cases (9 and 8, respectively), whilst *MYO15A* exon-2 variants were also observed in 8 cases (Figure 2B).

Association between exon-level burden and prognosis

Outcome modelling in the PROFILE cohort was restricted to carriers of qualifying variants within the 68 exons, representing 164 participants, of which 75 participants (45.7%) exhibited disease progression within 12 months and for whom median survival was 3.9 years (95%CI 2.8 to 4.8). Seventeen exons, representing 15 genes, clustered together in weighted association with greater risk of worse clinical outcome (Figure 3, Table 1, Supplementary Figure S5). Three exons, in three distinct genes, were associated with both disease progression and survival, including *FAT4* exon-10, *DYSF* exon-29, and *MYH6* exon-29. *DYSF* exon-19 was primarily associated with disease progression. An association was observed between *KIF26A* and both disease progression and survival endpoints, with exon-14 being associated with disease progression and exon-12 associated with survival. Strong association was suggested between *DNAH7* exon-43 and disease progression alone, whilst associations were also

observed with *PCDH15* exon-38, *DYSF* exon-19, *SBF1* exon-13, *PCM1* exon-4, *MYOM2* exon-37, *VPS13D* exon-58, *VPS13B* exon-13, and *COL6A3* exon-12. *DNAH12* exon-32 was associated with survival alone, associations were also observed with *MYO7A* exon-25 and *MYOF* exon-47.

Dysregulated expression and enriched pathways of burden genes

Differential gene expression across lung cell types supported dysregulation of the 15 exon-level burden genes in IPF lung tissue. *COL6A3* and *FAT4* were upregulated in IPF stromal lineages (Supplementary Figure S6), including fibroblasts and myofibroblasts (Figure 4A). *DNAH7*, *DNAH12*, *PCM1* were primarily upregulated in IPF epithelial lineages (Supplementary Figure S7), particularly multiciliated cells, whilst *MYOF* was upregulated in AT1 cells (Figure 4A). *KIF26A*, *DYSF*, *PCDH15*, as well as *FAT4*, were upregulated in IPF endothelial cells, particularly lymphatic differentiating endothelial cells (Figure 4A, Supplementary Figure S8). *SBF1* was upregulated in IPF lymphatic proliferating endothelial cells; *VPS13B*, *VPS13D* and *MYOM2* were upregulated in IPF 'SM activated stress response' cells (Figure 4A, Supplementary Figure S9). *MYO7A* and *MYH6* demonstrated the lowest expression across lung cells (Supplementary Figure S10).

Enrichment analysis performed on 40 genes with exon-level burden identified molecular function of cytoskeletal motor activity (p=4.63E-10) and polypeptide conformation or assembly activity (p=7.38E-10) within the top pathways, as well as related functions such as microtubule motor activity (p=9.70E-05) and dynein binding, which was consistent with 77 overlapping PAV burden genes (Supplementary Figure S11). The top enriched molecular functions for the 15 genes with exons associated with worse prognosis continued to be cytoskeletal motor activity (p=4.05E-06), polypeptide conformation and assembly activity (p=5.23E-06), and other functions including microtubule motor activity (Figure 4B).

Discussion

Approximately 75% of FPF cases have no known genetic cause, whilst 10% of sporadic PF cases are thought to be hereditary, suggesting a spectrum of genetic risk with substantial unexplained genetic architecture and pathogenesis[13, 29]. By focusing on reproducible rare PAV burden in susceptibility across two case cohorts and consolidating variant effects at the

exon level that were associated with poorer outcomes, we prioritised 15 genes. These genes were consistently dysregulated in IPF lung single-cell datasets and converged on cytoskeletal motor and microtubule-related molecular functions.

We identified that *COL6A3* and *FAT4* were genes with exon-level burden of pathogenic rare variants associated with poor outcomes along with dysregulated gene expression in stromal lineages of fibroblasts and myofibroblasts. *COL6A3* encodes collagen-alpha-1(VI), an extracellular matrix (ECM) component and an established biomarker in IPF[30, 31]. Greater levels of collagen in the IPF ECM leads to increased stiffness, whilst mechanical strain releases ECM-sequestered profibrotic cytokine TGFβ. *FAT4* encodes an atypical cadherin protein that is part of the canonical Hippo signalling pathway involved in proliferation, differentiation, and increasingly implicated in IPF pathogenesis[32, 33]. FAT4 is localised to primary cilia, a signalling structure distinct from motile cilia that hold critical functions in orientation of cell division and fibroblast migration during wound healing [34, 35], and mediate responses to ECM perturbations[36]. Compared with normal lung, greater numbers of primary cilia have been observed in IPF lung fibroblasts, which are also characterised by short primary cilia morphology that can be induced by TGFβ and restored by microtubule stabilisation[37, 38]. The rare variant observations support a genetic architecture in which the interactions between fibroblast organisation and ECM properties are dysregulated.

Microtubule organisation components showed increased gene expression in IPF multi-ciliated epithelial cells. Dynein axonemal heavy chain-7 and -12, encoded by *DNAH7*, *DNAH12* respectively, are microtubule-associated motor proteins that control cilium movement, whilst microtubule anchoring protein pericentriolar material-1, encoded by *PCM1*, is involved in recruiting proteins to centriolar satellites, influencing microtubule organisation, ciliogenesis and cilia motility[39]. Mutations in outer dynein arms have been associated with primary ciliary dyskinesia, a rare recessive disorder in which defective motile cilia result in poor mucociliary clearance, leading to chronic respiratory problems including persistent infection, inflammation, and increasing bronchiectasis with age. As PF is frequently diagnosed in late adulthood, the penetrance of rare variants may be incomplete without accumulation of other lifetime exposures[40]. Mucus hypersecretion and ciliary impairment in airways have been linked to the development of mucus plugs in alveolar regions of IPF patients[41], whilst lung fibrosis development in a bleomycin model was reduced by modulating cilia structure in Krt5+ basal cells[42]. The genetic architecture of microtubule disorganisation in multiciliated cells may predispose individuals to impaired mucociliary transport, chronic proinflammatory

signalling and recurrent microinjury, manifesting fibrotic remodelling over the life course of an individual. With established evidence that the *MUC5B* common variant is the strongest associated genetic risk factor for IPF susceptibility[6], this rare variant burden suggests airway mucociliary function may be more important in fibrogenesis than currently attributed.

Dysregulated expression of burden genes *KIF26A*, *DYSF*, *PCDH15*, *FAT4*, and *SBF1*, was observed in lymphatic endothelial cells. Studies have shown that endothelial cells can transition and contribute to the myofibroblast pool in pulmonary fibrosis models, corroborated through cell fate mapping in human IPF scRNAseq[43]. Lymphangiogenesis has been implicated in fibrotic lung remodelling, with the diameter of lymphatic vessels in alveolar spaces of IPF lung tissue correlating with disease severity[44, 45]. Protocadherin-15, kinesin family member 26A, and SET binding factor-1, encoded by *PCDH15*, *KIF26A* and *SBF1* respectively, as well as FAT4 described above, have roles in cell adhesion, cytoskeletal organisation and cellular morphology that could indicate a role in endothelial and lymphatic remodelling[46]. *DYSF,* encoding dysferlin, is an essential regulator of vascular homeostasis through its control of endothelial cell adhesion and membrane integrity [47], highlighting vascular remodelling in the genetic architecture of PF.

Myoferlin, encoded by burden gene *MYOF* and member of the Ferlin family with homology to dysferlin, was dysregulated in IPF AT1 cells. In IPF, AT2 to AT1 transition is impaired during alveolar epithelial repair, whilst epithelial metaplasia can lead to the recruitment of myofibroblasts and formation of fibrotic foci[48]. Myoferlin has been shown to induce the transition of epithelial cells into myofibroblasts through control of TGFβ signalling[49], and has previously featured in the top 20 genes during exploratory rare variant burden testing in an independent FPF cohort, although enrichment did not reach statistical significance[13].

Rare variant burden aggregated at exon-level has identified genes involved in cytoskeletal organisation and lung tissue remodelling, including regulation of fibroblastic properties, ECM properties, and mucociliary clearance. This study used variant-level, exon-level, gene-level and pathway-level approaches to iteratively prioritise biologically plausible mechanisms of perturbation within the genetic architecture of PF. However, the approach does not identify rare variant burden independently related to clinical outcomes and, whilst gene expression was evaluated, it is possible that identified PAVs do not alter transcript levels.

Two large PF cohorts were included to support reproducible risk susceptibility variants and exon burden. Inclusion of both familial and sporadic disease can provide insights into

converging pathogenic mechanisms. Identification of qualifying variants in *TERT* and *RTEL1* was consistent with previous studies although genomic coverage in gnomAD could not confirm *TERT* variants observed in GE100KGP. No restriction on ancestry was applied to summary level reference data from gnomAD, which may impact rare variant burden in susceptibility, however, associations with disease outcomes were performed within cases of European ancestry with adjustment for age thereby minimising the effect of population stratification on gene prioritisation. No exclusion criteria were applied to cases based on carriers of known causal variants, potentially reducing power to identify novel variants. Reducing technical artifacts and neutral background by excluding genes with significant synonymous burden may have limited identification of biologically relevant signals in very large genes. The findings require functional in-vitro and in-vivo studies to validate the role of genes, variants and their clinical relevance.

Conclusion

Through selection of reproducible PAV burden in coding regions across two large PF case cohorts, aggregating burden over exons and modelling associations between pathogenicity and disease outcomes, this study identified 15 genes implicated in the genetic architecture of PF susceptibility and prognosis. Using integrated IPF lung transcriptomics, cell-type specific dysregulation was observed in all 15 genes, particularly multiciliated epithelial cells, fibroblasts, and lymphatic endothelial cells, whilst molecular functions relating to cytoskeletal dynamics were enriched. Preclinical functional genomic studies should establish the role of the prioritised genes in the pathogenesis of PF.

Acknowledgements

This work was funded by an MRC Programme Grant (MR/V00235X/1) awarded to RGJ and a NIHR Imperial Biomedical Research Centre Springboard Award (PSR547) obtained by IS. DW is supported by the Taishan Scholars Program of Shandong Province (tsqn202312110) and the Fundamental Research Funds for the Central Universities (25CX06001A). This research is partly funded by the National Institute for Health and Care Research (NIHR) and Leicester (BRC). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. This research used the ALICE High

Performance Computing Facility at the University of Leicester. This research was made possible through access to data in the National Genomic Research Library, which is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The National Genomic Research Library holds data provided by patients and collected by the NHS as part of their care and data collected as part of their participation in research. The National Genomic Research Library is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure.

Contributions

DW, RGJ and IS conceived and designed the study. KB, DW, HF, AAB, YN, IS performed the data analysis. RGJ and IS secured funding. SRJ, TMM, PLM, GS contributed to the patient recruitment and data collection. KB, DW and IS wrote the initial draft. BL, AJ, HF, YN, AAB, LVW, DMR, RGJ, TMM, PLM, EAR, GS and provided input and editing.

Competing interests

LVW receives research funding from Orion Pharma, GSK, and Genentech, consulting fees from Galapagos, Boehringer Ingelheim, and GSK, and travel support from Genentech, has research collaboration with AstraZeneca, Nordic Bioscience, and Sysmex (OGT), and serves on advisory board for Galapagos. SRJ receives research funding from Medical Research Council, LifeArc, LAM Action, and La Morato de TV3, and travel support from Ferrer, and is the Trustee of LAM Action. TMM receives consulting fees from Boehringer Ingelheim, Roche/Genentech, Abbvie, Amgen, Astra Zeneca, Bayer, Bridge bio, Bristol-Myers Squibb, CSL Behring, Galapagos, Galecto, GlaxoSmithKline, IQVIA, Merck, Pfizer, Pliant, PureTech, Sanofi, Trevi, and Vicore, serves on advisory board for Fibrogen, United Therapeutics, and Nerre, and holds stock in Qureight. PLM receives research funding from AstraZeneca, GSK, Asthma & lung UK, and Action for pulmonary Fibrosis, consulting fees from Hoffman-La Roche, Boehringer Ingelheim, AstraZeneca, Trevi, Qureight, Endevour, and Redx, and payments or honoraria from Boehringer Ingelheim, and Hoffman-La Roche, serves on advisory board for United Therapeutics, and holds stock in Qureight. RGJ received grant funding to institution from AstraZeneca, Biogen, Galecto, GlaxoSmithKline, Nordic Bioscience, RedX,

and Pliant Therapeutics, consulting fees from AstraZeneca, Brainomix, Bristol Myers Squibb, Chiesi, Cohbar, Daewoong, GlaxoSmithKline, Veracyte, Resolution Therapeutics, and Pliant Therapeutics, payments or honoraria from Boehringer Ingelheim, Chiesi, Roche, PatientMPower, and AstraZeneca, payment for expert testimony from Pinsent Masons LLP, is the president of Action for Pulmonary Fibrosis and in the leadership of NuMedii. EAR has received lecture and consulting fees from Boehringer Ingelheim, Abbvie and Bristol-Myers Squibb paid into her institution.

References

1. Ng, N., et al., *Genetics of interstitial lung diseases: a state-of-the-art review.* Eur Respir J, 2025. **66**(3).
2. Finnerty, J.P., et al., *Efficacy of antifibrotic drugs, nintedanib and pirfenidone, in treatment of progressive pulmonary fibrosis in both idiopathic pulmonary fibrosis (IPF) and non-IPF: a systematic review and meta-analysis.* BMC Pulm Med, 2021. **21**(1): p. 411.
3. Lee, H.L., et al., *Familial idiopathic pulmonary fibrosis: clinical features and outcome.* Chest, 2005. **127**(6): p. 2034-41.
4. Garcia, C.K., *Idiopathic pulmonary fibrosis: update on genetic discoveries.* Proc Am Thorac Soc, 2011. **8**(2): p. 158-62.
5. Oh, C.K., L.A. Murray, and N.A. Molfino, *Smoking and idiopathic pulmonary fibrosis.* Pulm Med, 2012. **2012**: p. 808260.
6. Chin, D., et al., *Genome-wide association study of Idiopathic Pulmonary Fibrosis susceptibility using clinically-curated European-ancestry datasets.* medRxiv, 2025.
7. Oldham, J.M., et al., *PCSK6 and Survival in Idiopathic Pulmonary Fibrosis.* Am J Respir Crit Care Med, 2023. **207**(11): p. 1515-1524.
8. Allen, R.J., et al., *Longitudinal lung function and gas transfer in individuals with idiopathic pulmonary fibrosis: a genome-wide association study.* Lancet Respir Med, 2023. **11**(1): p. 65-73.
9. Dhindsa, R.S., et al., *Identification of a missense variant in SPDL1 associated with idiopathic pulmonary fibrosis.* Commun Biol, 2021. **4**(1): p. 392.
10. Zhang, D., et al., *Rare and Common Variants in KIF15 Contribute to Genetic Risk of Idiopathic Pulmonary Fibrosis.* Am J Respir Crit Care Med, 2022. **206**(1): p. 56-69.
11. Sutton, R.M., et al., *Rare surfactant-related variants in familial and sporadic pulmonary fibrosis.* Hum Mutat, 2022. **43**(12): p. 2091-2101.
12. Alonso-Gonzalez, A., et al., *Rare variants and survival of patients with idiopathic pulmonary fibrosis: analysis of a multicentre, observational cohort study with independent validation.* Lancet Respir Med, 2025. **13**(6): p. 495-504.
13. Liu, Q., et al., *The Genetic Landscape of Familial Pulmonary Fibrosis.* Am J Respir Crit Care Med, 2023. **207**(10): p. 1345-1357.
14. Maher, T.M., et al., *Prospective study of fibrosis in the lung endpoints (PROFILE): characteristics of an incident cohort of patients with idiopathic pulmonary fibrosis.* BMJ Open Respir Res, 2026. **13**(1).
15. Karczewski, K.J., et al., *The mutational constraint spectrum quantified from variation in 141,456 humans.* Nature, 2020. **581**(7809): p. 434-443.

16. Kircher, M., et al., *A general framework for estimating the relative pathogenicity of human genetic variants.* Nat Genet, 2014. **46**(3): p. 310-5.
17. Guo, M.H., et al., *Burden Testing of Rare Variants Identified through Exome Sequencing via Publicly Available Control Data.* Am J Hum Genet, 2018. **103**(4): p. 522-534.
18. Adzhubei, I.A., et al., *A method and server for predicting damaging missense mutations.* Nat Methods, 2010. **7**(4): p. 248-9.
19. Sim, N.L., et al., *SIFT web server: predicting effects of amino acid substitutions on proteins.* Nucleic Acids Res, 2012. **40**(Web Server issue): p. W452-7.
20. Rentzsch, P., et al., *CADD: predicting the deleteriousness of variants throughout the human genome.* Nucleic Acids Res, 2019. **47**(D1): p. D886-D894.
21. Cheng, J., et al., *Accurate proteome-wide missense variant effect prediction with AlphaMissense.* Science, 2023. **381**(6664): p. eadg7492.
22. Alirezaie, N., et al., *ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants.* Am J Hum Genet, 2018. **103**(4): p. 474-483.
23. Ioannidis, N.M., et al., *REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants.* Am J Hum Genet, 2016. **99**(4): p. 877-885.
24. Prokopenko, D., et al., *Region-based analysis of rare genomic variants in whole-genome sequencing datasets reveal two novel Alzheimer's disease-associated genes: DTNB and DLG2.* Mol Psychiatry, 2022. **27**(4): p. 1963-1969.
25. Adams, T.S., et al., *Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis.* Sci Adv, 2020. **6**(28): p. eaba1983.
26. Habermann, A.C., et al., *Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis.* Sci Adv, 2020. **6**(28): p. eaba1972.
27. Morse, C., et al., *Proliferating SPP1/MERTK-expressing macrophages in idiopathic pulmonary fibrosis.* Eur Respir J, 2019. **54**(2).
28. Korsunsky, I., et al., *Fast, sensitive and accurate integration of single-cell data with Harmony.* Nat Methods, 2019. **16**(12): p. 1289-1296.
29. Kropski, J.A., T.S. Blackwell, and J.E. Loyd, *The genetic basis of idiopathic pulmonary fibrosis.* Eur Respir J, 2015. **45**(6): p. 1717-27.
30. Specks, U., et al., *Increased expression of type VI collagen in lung fibrosis.* Am J Respir Crit Care Med, 1995. **151**(6): p. 1956-64.
31. Jessen, H., et al., *Longitudinal serological assessment of type VI collagen turnover is related to progression in a real-world cohort of idiopathic pulmonary fibrosis.* BMC Pulm Med, 2021. **21**(1): p. 382.
32. Gokey, J.J., et al., *Active epithelial Hippo signaling in idiopathic pulmonary fibrosis.* JCI Insight, 2018. **3**(6).
33. Gokey, J.J., S.D. Patel, and J.A. Kropski, *The Role of Hippo/YAP Signaling in Alveolar Repair and Pulmonary Fibrosis.* Front Med (Lausanne), 2021. **8**: p. 752316.
34. Saburi, S., et al., *Loss of Fat4 disrupts PCP signaling and oriented cell division and leads to cystic kidney disease.* Nat Genet, 2008. **40**(8): p. 1010-5.
35. Christensen, S.T., et al., *Analysis of primary cilia in directional cell migration in fibroblasts.* Methods Enzymol, 2013. **525**: p. 45-58.
36. Trempus, C.S., et al., *A novel role for primary cilia in airway remodeling.* Am J Physiol Lung Cell Mol Physiol, 2017. **313**(2): p. L328-L338.
37. Lee, J., et al., *Increased Primary Cilia in Idiopathic Pulmonary Fibrosis.* Mol Cells, 2018. **41**(3): p. 224-233.

38. Verma, P., et al., *Morphological reprogramming of primary cilia length mitigates the fibrotic phenotype in fibroblasts across diverse fibrotic conditions.* J Cell Sci, 2025. **138**(20).
39. Begar, E., E. Seyrek, and E.N. Firat-Karalar, *Navigating centriolar satellites: the role of PCM1 in cellular and organismal processes.* FEBS J, 2025. **292**(4): p. 688-708.
40. Krauss, E., et al., *Clinical characteristics of patients with familial idiopathic pulmonary fibrosis (f-IPF).* BMC Pulm Med, 2019. **19**(1): p. 130.
41. Peng, Y., et al., *Mucus Hypersecretion and Ciliary Impairment in Conducting Airway Contribute to Alveolar Mucus Plugging in Idiopathic Pulmonary Fibrosis.* Front Cell Dev Biol, 2021. **9**: p. 810842.
42. Kim, E., et al., *Aberrant Multiciliogenesis in Idiopathic Pulmonary Fibrosis.* Am J Respir Cell Mol Biol, 2022. **67**(2): p. 188-200.
43. Wu, X., et al., *Regulating the cell shift of endothelial cell-like myofibroblasts in pulmonary fibrosis.* Eur Respir J, 2023. **61**(6).
44. Baluk, P., et al., *Lymphatic Proliferation Ameliorates Pulmonary Fibrosis after Lung Injury.* Am J Pathol, 2020. **190**(12): p. 2355-2375.
45. El-Chemaly, S., et al., *Abnormal lymphangiogenesis in idiopathic pulmonary fibrosis with insights into cellular and molecular mechanisms.* Proc Natl Acad Sci U S A, 2009. **106**(10): p. 3958-63.
46. Schoofs, H., et al., *Dynamic cytoskeletal regulation of cell shape supports resilience of lymphatic endothelium.* Nature, 2025. **641**(8062): p. 465-475.
47. Sharma, A., et al., *A new role for the muscle repair protein dysferlin in endothelial cell adhesion and angiogenesis.* Arterioscler Thromb Vasc Biol, 2010. **30**(11): p. 2196-204.
48. Pan, X., et al., *TRbeta activation confers AT2-to-AT1 cell differentiation and anti-fibrosis during lung repair via KLF2 and CEBPA.* Nat Commun, 2024. **15**(1): p. 8672.
49. Barnhouse, V.R., et al., *Myoferlin regulates epithelial cancer cell plasticity and migration through autocrine TGF-beta1 signaling.* Oncotarget, 2018. **9**(27): p. 19209-19222.

Tables

Table 1: Prioritised burden genes involved in PF pathogenesis.

Gene	Exon	Mean exon burden	All cases	Disease progression	Below median survival	Concordant PAV	rsID	Summary variant pathogenicity score
DNAH7	43	4.39	5	6.06	0.11	2-195858535-A-C	rs201115799	4.31
DYSF	29	3.78	3	3.68	4.18	2-71570698-G-A	rs779159879	3.92
MYH6	29	3.76	7	6.37	1.72	14-23388898-G-A	rs145611185	3.19
MYOF	47	3.11	5	0.20	1.87	10-93323283-G-A	rs199815755	3.21
VPS13D	58	3.04	3	1.65	0.42	1-12383150-C-T	rs200100914	3.10
KIF26A	12	3.00	24	0.08	1.54	14-104175154-C-T	rs765014920	3.67
DYSF	19	2.98	4	2.93	0.76	2-71551712-C-T	rs149386446	2.44
KIF26A	14	2.93	6	2.96	0.79	14-104179372-G-A	rs374991880	2.75
MYO7A	25	2.7	5	-0.10	2.44	11-77182598-G-A	rs199810429	1.88
MYOM2	37	2.67	4	1.65	0.46	8-2144911-C-T	rs201286207	2.40
DNAH12	32	2.55	4	0.05	2.51	3-57433396-T-C	rs200594823	2.64
PCM1	4	2.55	4	1.69	0.09	8-17937310-G-C	rs200940134	2.35
FAT4	10	2.31	14	5.41	6.03	4-125452703-C-T	rs138275098	1.97
COL6A3	12	2.14	5	1.61	-0.29	2-237365838-C-T	rs749438477	2.26
VPS13B	13	2.06	5	1.54	-0.15	8-99143022-G-A	rs141046414	1.95
SBF1	13	1.97	6	2.89	-0.11	22-50464824-T-C	rs9617016	2.03
PCDH15	38	1.83	4	3.35	0.04	10-53807091-C-T	rs776833175	2.28

Association of mean exon burden with disease progression and survival below median survival presented as F-statistics. Disease progression defined as ≥10% relative decline in percent predicted forced vital capacity over 12 months, median survival defined as 3.9 years. Associations tested in 164 individuals with qualifying protein altering variants (PAV) in 68 burden exons. Where more than one concordant PAV was identified, the PAV with the greatest summary variant pathogenicity score was presented. All cases identified within PROFILE and GE100KGP with qualifying PAVs within gene exon presented.

Figure legends

Figure 1. Overlapping rare variant burden of pulmonary fibrosis susceptibility.

Coding regions using whole genome sequencing in TRAPD pipeline and gnomAD control in A) PROFILE IPF cohort B) Genomics England 100K Genome Project (GE100KGP) PF cohort. C) Venn diagram of overlapping genes with significant rare protein altering variant burden between A and B. D) Venn diagram of overlapping rare protein altering variants with sufficient gnomAD coverage; represents 77/87 shared genes in C.

Figure 2. Exon-level pathogenicity and case numbers across case cohorts.

A) Exon mean summary variant pathogenicity score presented for 68 gene exons. B) Exon mean summary variant pathogenicity score and case number represented in PROFILE and GE100KGP for 68 gene exons.

Figure 3. Exon-level pathogenicity weighted association with prognosis in IPF.

Heatmap of standardised effect size of exon mean summary pathogenicity (red greater risk of outcome, blue lower risk of outcome) on outcomes of disease progression (prog: 10% relative decline in ppFVC or mortality within 12 months) or survival below median (surv: median survival 3.9 years) in PROFILE cases (n=164) derived from SKAT-O based kernel regression. 17/68 gene exons, representing 15 genes, met F-statistic ≥1.5 for risk of at least one outcome without a protective effect on other.

Figure 4. IPF lung cell expression and enriched functions of burden genes.

A) Heatmap of differential gene expression IPF z-scores in integrated single cell RNA sequencing for 15 genes with F-statistics ≥1.5, annotated at Human Lung Cell Atlas level 3. Red represents strongly upregulated relative to respective gene average, blue represents strongly downregulated. B) Enrichment analysis for 15 genes, Gene Ontology presents significant molecular function terms.

Figures

Figure 1. Overlapping rare variant burden of pulmonary fibrosis susceptibility.

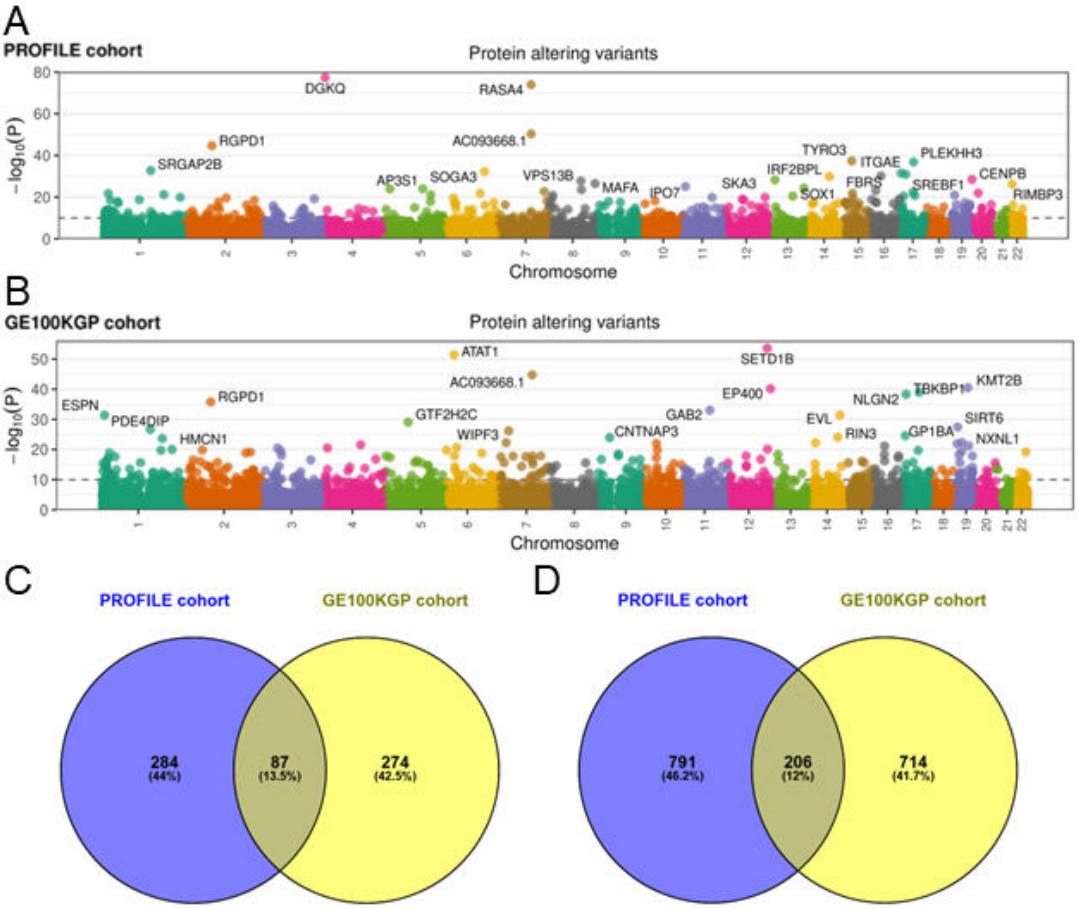

Figure 2. Exon-level pathogenicity and case numbers across case cohorts.

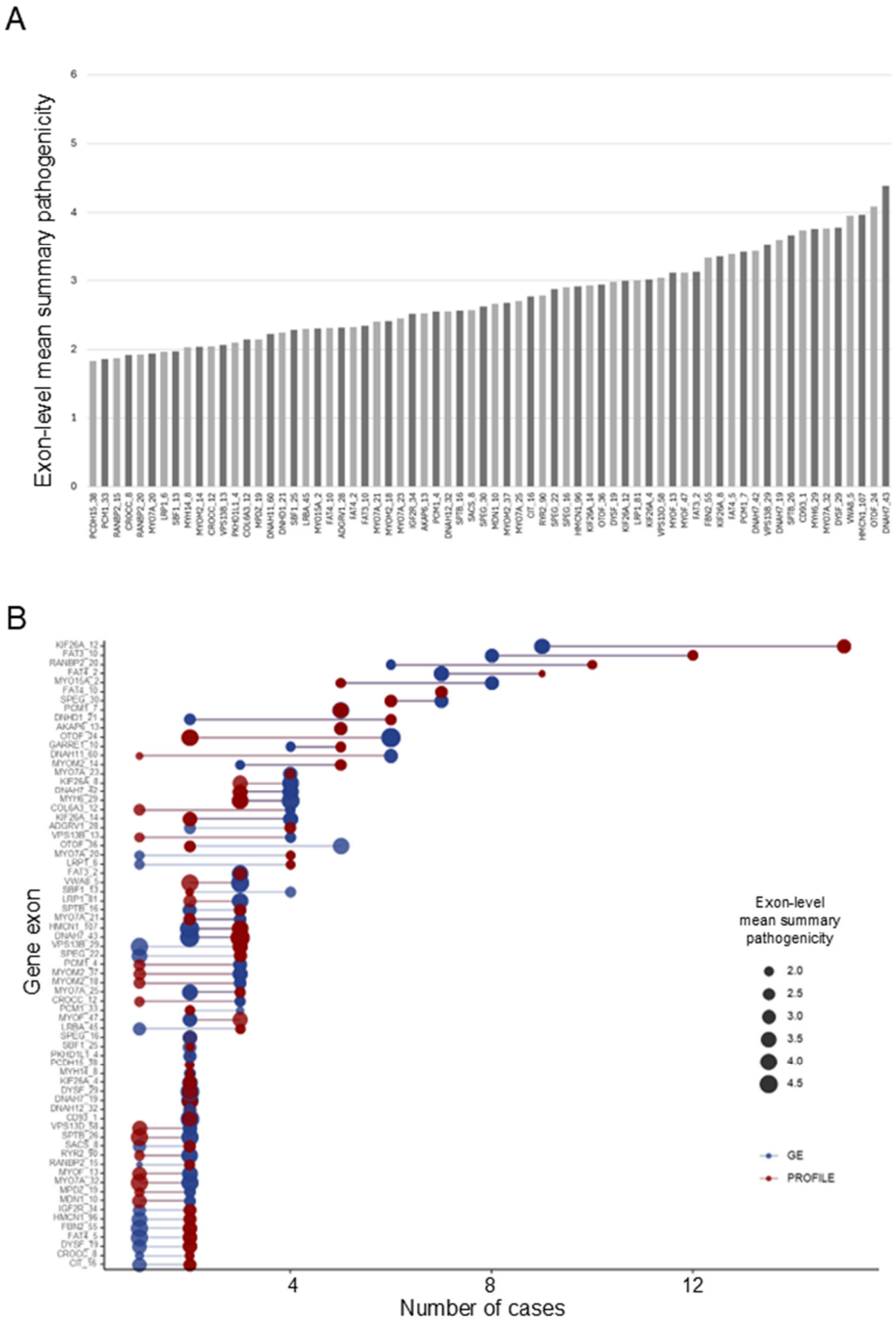

Figure 3. Exon-level pathogenicity weighted association with prognosis in IPF.

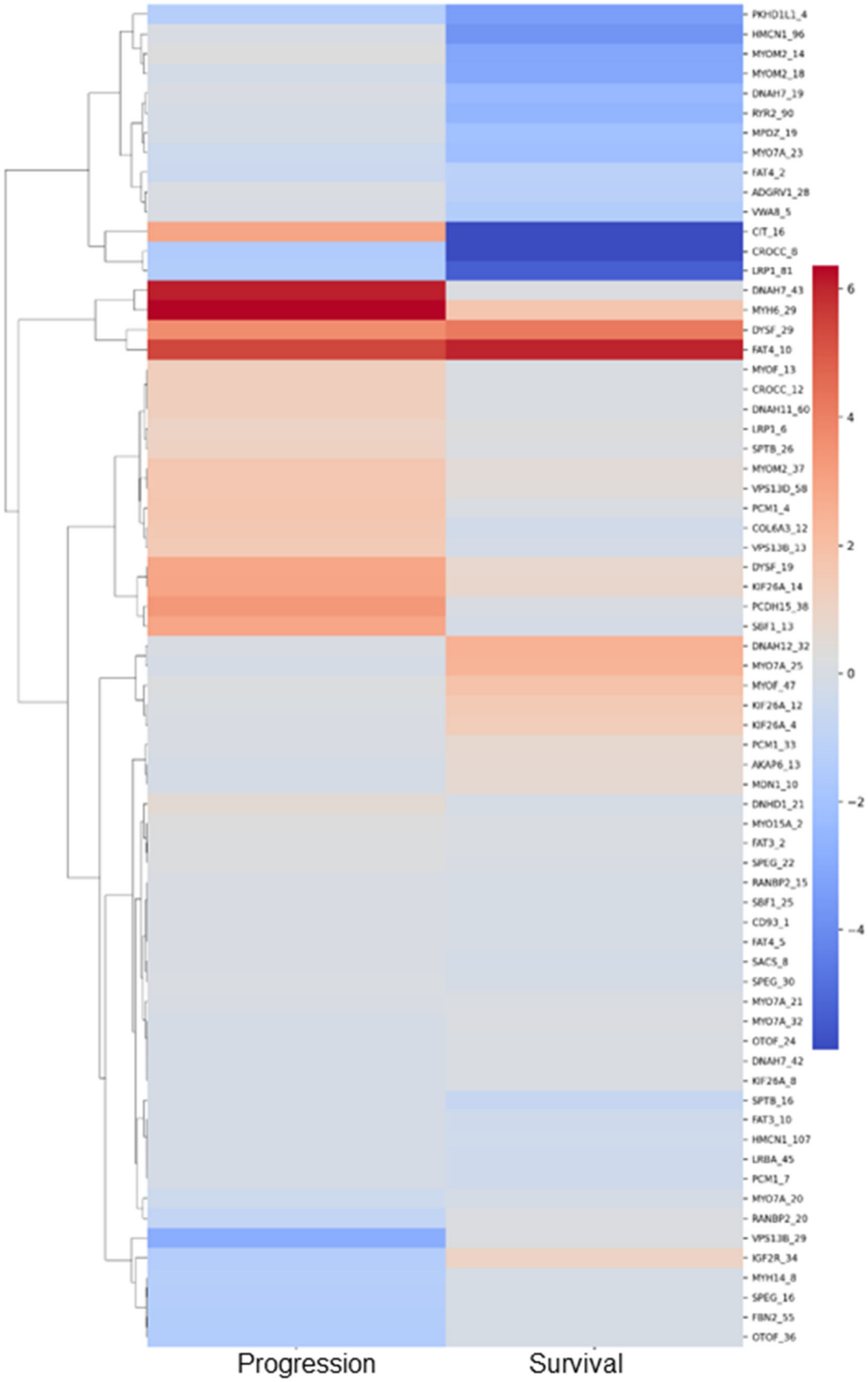
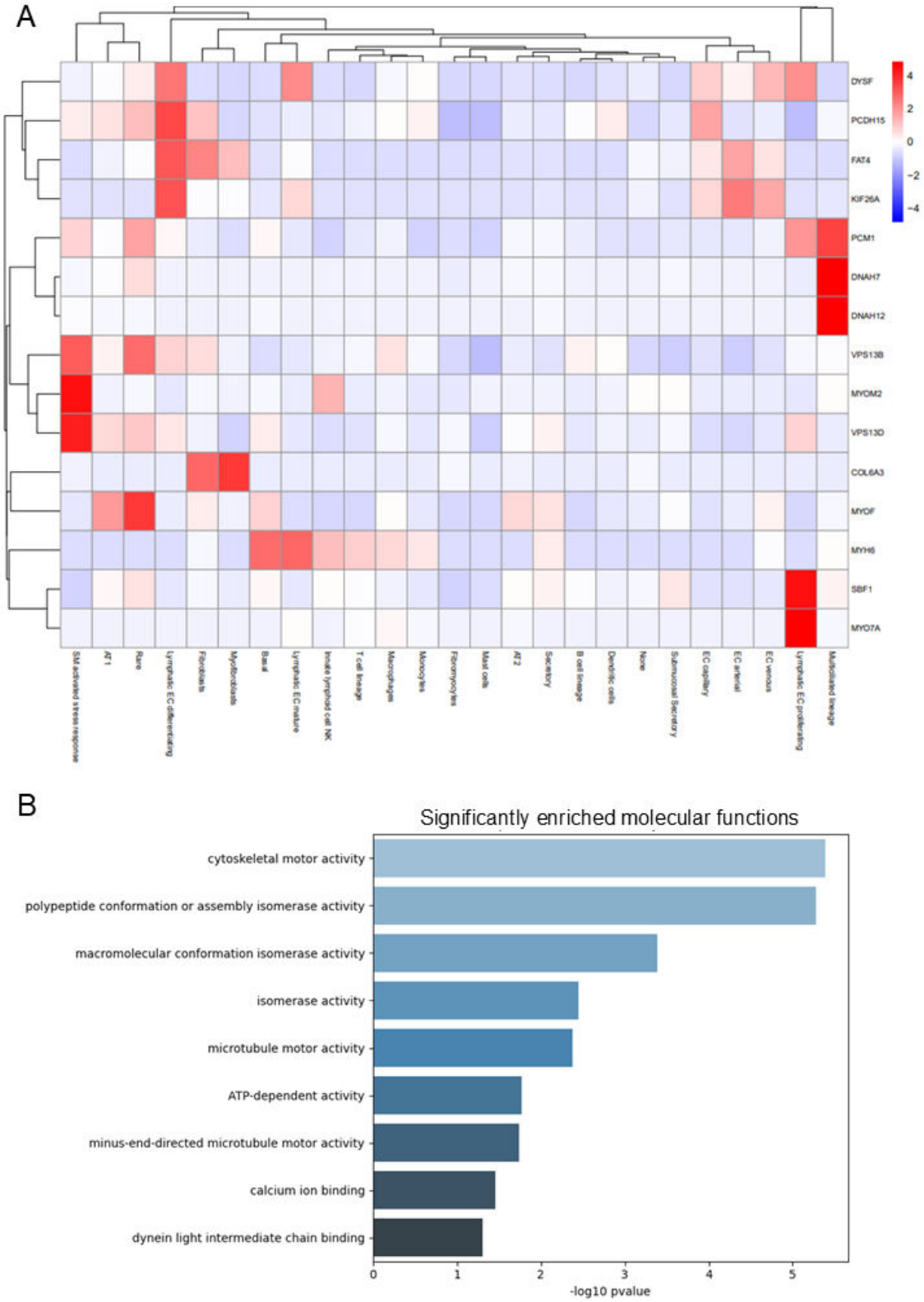
Figure 4. IPF lung cell expression and enriched functions of burden genes.