Genetic burden of dysregulated cytoskeletal organisation in the pathogenesis of pulmonary fibrosis

Dapeng Wang[1,*], Yujie Ni[1], Bin Liu[1], Huan Ding[1], Alison John[1,2], Louise V Wain[3,4], Simon R. Johnson[5], Toby M Maher[1,6], Philip L Molyneaux[1,2], Elizabeth Renzoni[1,7], Gauri Saini[5], Athol U Wells[1,7], Deborah Morris-Rosendahl[1,8], R Gisli Jenkins[1,2], and Iain Stewart[1,2]

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

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

[3]Department of Population Health Sciences, University of Leicester, Leicester, UK

[4]NIHR Leicester Biomedical Research Centre, Leicester, UK

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

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

[7]The Interstitial Lung Disease Unit, Royal Brompton Hospital, London, UK

[8]Clinical Genetics and Genomics Laboratory, Royal Brompton and Harefield Hospitals, Guy's and St. Thomas' NHS Foundation Trust, London, UK

[*]Corresponding author:

Dapeng Wang, National Heart and Lung Institute, Imperial College London, London, SW3 6LY, UK

Email: dapeng.wang@imperial.ac.uk

ORCID: 0000-0002-9925-4574

Abstract

Background

Pulmonary fibrosis (PF) is a shared characteristic of chronic interstitial lung diseases of mixed aetiology. Previous studies on PF highlight a pathogenic role for common and rare genetic variants. This study aimed to identify rare pathogenic variants that are enriched in distinct biological pathways and dysregulated gene expression.

Methods

Rare variants were identified using whole genome sequencing (WGS) from two independent PF cohorts, the PROFILE study and the Genomics England 100K (GE100KGP) cohort, with the gnomAD database as a reference. Four pathogenic variant categories were defined: loss of function variants, missense variants, protein altering variants, and protein truncating variants. Gene burden testing was performed for rare variants defined as having a minor allele frequency <0.1%. Overrepresentation analysis of gene ontology terms and gene concept network analysis were used to interpret functional pathways. Integration of publicly available transcriptomic datasets was performed using weighted gene co-expression network analysis of idiopathic pulmonary fibrosis (IPF) lung tissue compared with healthy controls.

Results

Burden testing was performed on 507 patients from the PROFILE study and 451 PF patients from GE100KGP cohort, compared with 76,156 control participants from the gnomAD database. Ninety genes containing significantly more pathogenic rare variants in cases than in controls were observed in both cohorts. Fifty-six genes included missense variants and 87 genes included protein altering variants. For missense variants, *HMCN1*, encoding hemicentin-1, and *RGPD1*, encoding a protein with a RanBD1 domain, were highly associated with PF in both PROFILE (p=5.70E-22 and p=4.48E-51, respectively) and GE100KGP cohorts (p=2.27E-24 and p=1.59E-36, respectively). 56 of 90 genes with significant burden were observed within modules correlated with disease in transcriptomic analysis, including *HMCN1* and *RGPD1*. Enriched functional categories from genetic and transcriptomic analyses included pathways involving extracellular matrix constituents, cell adhesion properties and microtubule organisation.

Conclusions

Rare pathogenic variant burden testing and weighted gene co-expression network analysis of transcriptomic data provided complementary evidence for pathways regulating cytoskeletal dynamics in PF pathogenesis. Functional validation of candidates could provide novel targets for intervention strategies.

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]. Treatments including Pirfenidone and Nintedanib are available to slow down the progression of fibrosis, but patients with PF have a poor prognosis and relatively short life expectancy after diagnosis[2]. There are considerable overlapping clinical characteristics, genetic variation and environmental risk factors between IPF and FPF[1,3,4], suggesting a potential common molecular mechanism leading to the disease pathogenesis. The prevalence of FPF has been underestimated because of the undiagnosed FPF in sporadic cases[5]. Furthermore, many novel variants may be restricted to specific families in FPF[1,6].

Genetic studies have the potential to elucidate disease mechanisms of PF and to support targeted and personalised medicine. Genome-wide association studies (GWAS) have identified common variants in coding and non-coding regulatory regions that have resulted in specific genes being associated with IPF. For example polymorphisms in, *MUC5B*, *TERT*, *TERC*, *AKAP13* and *DSP* are associated with increased IPF risk[7,8], *PCSK6* has been associated with IPF survival[9], and *PKN2* with longitudinal progression of IPF[10]. These studies have extended insights into the genetic contribution to PF, yet the analysis of variants with array-based techniques is restricted to common variants and some rare variants in predefined regions of the genome. The majority of the variants identified through GWAS analysis are located within non-coding regulatory regions of the genome and thus likely to lead to dysregulated gene transcription.

Whole genome sequencing (WGS) analysis takes advantage of next-generation sequencing technologies to facilitate the assessment of rare variants in an unbiased manner. WGS supports identification of rare pathogenic variants in coding regions, such as those which may lead to truncation and loss of function, reducing gene expression, or missense and other variants that could alter protein function. Whole exome sequencing (WES) has highlighted a rare missense variant in *SPDL1* associated with IPF, implicating mitotic checkpoint signalling as a dysregulated pathway in fibrotic development[11]. Application of WES to FPF identified rare variants in predominantly telomere related genes that could explain up to 23% of risk[12], whilst telomerase activity and telomere length is associated with prognosis[13,14]. Similarly, WGS supported the association of telomerase activity in IPF pathogenesis through identification of rare variants in *TERT*, *RTEL1*[15] and *ADA*[16]. Despite advances in understanding the role of

telomeropathy in PF, there is limited evidence regarding other rare variants, novel genes or pathways in the development of PF.

Transcriptomic analysis can offer independent evidence for the validation of the gene signals from genetic analysis. RNA sequencing (RNA-Seq) has been widely applied to IPF, providing a wealth of information regarding the genome-wide gene expression signature and alternative splicing events in the pathogenesis of disease, as well as their associated molecular pathways and regulatory networks[17-21]. Due to the availability of a number of open-source bulk RNA-Seq datasets, integration of these data facilitates a comprehensive and consistent picture of the gene expression changes, increasing the sample size and enhancing the statistical power.

This study aims to conduct gene burden testing of rare pathogenic variants in coding regions using WGS from two cohorts of patients with PF, compared with a publicly available control cohort. Integrative analysis of publicly accessible transcriptomic data from IPF lung samples was compared with healthy control samples and used to assess enriched functional pathways of candidate genes through network analysis, and interpret up-regulation or down-regulation of expression. A pipeline of iterative and integrative methodologies was applied to assess evidence of novel genes and biological pathways underlying PF development.

Methods

Populations included in rare variant gene burden testing

WGS data were obtained from the PROFILE cohort, a UK based prospective study of anti-fibrotic naïve PF patients described previously[11], 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 the cancer and rare diseases. Analysis of GE100KGP data was performed within Research Environment of Genomics England.

Quality control procedures were applied to PROFILE PF samples, including sex concordance, sample contamination, European ancestry and relatedness of individuals[11]. 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 (Table S1). The samples with concordance between phenotypic sex and

inferred ploidy, unrelated individuals, and European ancestry (probability > 0.8) available in GRCh38 reference genome were retained.

The summary level VCF files were downloaded from gnomAD (v3.1.2)[22] for 76,156 individuals generated through WGS and mapping against GRCh38 reference genome, used as control samples without enrichment of any specific diseases.

Analysis of pathogenic rare variants

Rare variants and ultra-rare variants were defined as variants with an allele frequency < 0.1% and allele frequency < 0.01% in the control cohort respectively. Additional filtering for internal allele frequency cut-off was also applied (Table S2). To explore different combinations of potentially pathogenic variants, four pathogenic variant categories were defined based on the consequence predicted by the Ensembl Variant Effect Predictor (VEP)[23] and Combined Annotation–Dependent Depletion (CADD)[24] score provided by the gnomAD VCF files for variants that are shared between case and controls and calculated for case specific variants: loss of function (LoF), missense, protein altering variants (PAV), and protein truncating variants (PTV). Synonymous variants were used as controls, genes that were significant when synonymous variants were used were excluded from subsequent analysis (Table S2).

BCFtools[25] was used for the processing of VCF files into a single VCF file per cohort, keeping only variants that passed quality controls. Multi-allelic variants were split and indels were left-aligned and normalised. The missing genotypes were replaced with reference alleles. For both case and control samples, variants from 22 autosomes were selected and unique variant identifier was constructed to link the variants across cohorts. To support discovery analysis using burden testing, variants that were shared between case and controls and case specific variants were included.

Testing Rare vAriants using Public Data (TRAPD) pipeline[26] was used for the burden testing between the individual level case data and summary level control data. Potentially pathogenic rare or ultra-rare variants were identified and the relations between the variants and genes were constructed. 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. The top 20 significant genes that were replicated across both cohorts within each category were selected for *in-silico* functional genomics.

Transcriptomic data processing

Bulk RNA-Seq raw count data from lung tissue that comprised both healthy and IPF groups were obtained from Gene Expression Omnibus (GEO)[27] (https://www.ncbi.nlm.nih.gov/geo/): GSE213001, GSE199949, GSE150910, GSE99621, GSE52463.

Transcriptomic raw count data from each study were filtered to keep protein coding genes. In studies that took multiple samples from the same subject, the average value was used to represent the overall gene expression level for that subject. After datasets were merged into one single matrix, missing values were filled using K-nearest-neighbour imputation, which was performed using impute package (v1.76.0) based on k=10, maximum gene missing in 50% of participants and maximum participant missing in 80% of genes. Genes with raw count <10 in more than 90% of samples were removed. Library size normalisation and variance-stabilising transformation were performed using DESeq2 package[28] (v1.42.0). Batch effects were removed using limma package[29] (v3.58.1) and confirmed with principal component analysis.

Identification of significant modules using weighted gene co-expression network analysis

Weighted gene co-expression network analysis (WGCNA) was implemented using WGCNA package[30] (v1.72-1) to integrate transcriptomic datasets and identify genes that were significantly associated with IPF. 'Signed' network was constructed to only consider positive correlations. Soft threshold power, based on scale-free topology fit index (0.90), was selected to construct a weighted adjacency matrix. The connectivity measure per gene was calculated according to the connection strengths with other genes. The expression profiles of each module were summarised by the module eigengene, and each module eigengene was regressed on sample trait: IPF or control. Modules of interest were selected based on the p-value <0.05. Overlapping genes from the rare variant burden analysis were identified as those classified within significant WGCNA modules.

To assess differential expression of individual overlapping genes, t-tests were performed between IPF and control lung samples on the integrated normalised data and plotted with fold change. A Bonferroni corrected p-value and fold change effect size ≥0.025 were used as criteria to highlight important differences in gene expression.

Pathway analysis

Gene Ontology (GO) terms were used for pathway analysis of results from transcriptomic analysis and rare variant burden tests. To identify the functional categories and pathways relating to the significant genes and facilitate the data visualisation, over-representation analysis of gene ontology terms for genes at a $p<0.05$ and qvalue<0.2, and gene-concept network analysis were conducted using clusterProfiler package[31].

In-silico functional genomics

UCSC Genome Browser was used to identify the consequence of variants on the amino acid sequence in the top hits[32], and UniProtKB was used to determine protein domains based on amino acid sequence[33]. The ColabFold platform was used for prediction of protein structures and complexes using MMseqs2 with AlphaFold2 on Google Colaboratory[34,35], and PyMOL (v3.0) was used to visualise and interpret functional consequence of predicted changes in protein sequences[36].

Results

Pathogenic rare variant burden testing

After quality control of 541 WGS PROFILE samples, 507 samples were used in downstream analysis, including 451 participants with IPF and 56 with idiopathic non-specific interstitial pneumonia (iNSIP). A total of 451 PF case samples from GE100KGP were used in downstream analysis, including 111 people with FPF (Figure S1). 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 was also represented by more males than GE100KGP, 76.3% vs 51.4%, respectively (Table S3).

In the PROFILE cases, a total of 50 significantly associated genes had rare LoF variants, 53 had rare PTVs, 263 had rare missense variants, and 371 had rare PAVs (Figure 1A). In the GE100KGP cases, a total of 74 significantly associated genes had LoF variants, 74 had PTVs, 225 had missense variants, and 361 had PAVs (Figure 2A). Between both cohorts, there were 90 shared significantly associated genes: 56 shared genes had missense variants and 87 shared genes had PAVs (Figure 3A). Results were similar when restricted to ultra-rare variants (Figure S2, S3, S4).

RGPD1 (RANBP2 like and GRIP domain containing 1), *ATAT1* (alpha tubulin acetyltransferase-1) and *HMCN1* (hemicentin-1) were independently observed in the top 20 significant genes in both PROFILE and GE100KGP cohorts. *RGPD1* was the top replicated gene with significant burden of missense variants (PROFILE p=4.48E-51, cases n=23, unique variants n=8; GE100KGP p=1.59E-36, cases n=16, unique variants n=5) and PAVs (Figure 1B, 2B, Table S4). *ATAT1* was the top replicated gene with significant burden of LoF variants and PTVs (PROFILE p=1.06E-35, cases n=16, unique variants n=12; GE100KGP p=1.54E-63, cases n=28, unique variants n=12). *HMCN1* was in the top 20 for missense variants in both cohorts (PROFILE, p=5.70E-22; cases n=39; unique variants n=36 and GE100KGP p=2.27E-24; cases n=35; unique variants n=32), reaching the top 20 for PAVs in GE100KGP alone.

In the PROFILE cohort specifically, LoF variants, missense variants, PAVs, and PTVs were observed in *DGKQ* (diacylglycerol kinase theta) and *RASA4* (RAS p21 protein activator 4) (Figure 1B). LoF variants, PAVs, and PTVs were significantly increased in *ITGAE* (integrin subunit alpha E). In the GE100KGP cohort five genes demonstrated greatest significance for LoF variants, PAVs and PTVs: *SETD1B* (SET domain containing 1B, histone lysine methyltransferase); *ATAT1* (alpha tubulin acetyltransferase 1); *NLGN2* (neuroligin 2); *KMT2B* (lysine methyltransferase 2B); and *TBKBP1* (TBK1 binding protein 1) (Figure 2B). Two telomere associated genes *TERT* (for both missense variants and PAVs) and *RTEL1* (for PAVs) were identified in both PROFILE and GE100KGP cohort (Table S4). *FBRS* (Fibrosin) was identified as a gene with significant burden of LoF variants (PROFILE p=3.97E-29, cases n=13, unique variants n=6; GE100KGP p=7.84E-21, cases n=9, unique variants n=3), as well as PAVs and PTVs, although it was not in the top 20 in both cohorts.

Gene ontologies and pathways

The gene ontology terms in the overrepresentation analysis for 90 shared significant genes between the two cohorts of disease cases included “actin filament-based process” and “actin cytoskeleton organization” for biological process category; “ATP-dependent activity”, “cytoskeletal motor activity”, “actin filament binding”, “microfilament motor activity” and “extracellular matrix structural constituent” for molecular function category; “myosin complex”, “supramolecular fiber” and “myosin filament” for cellular component category (Figure 3, S7). Similar GO terms were significantly enriched when PROFILE and GE100KGP were assessed separately, including “microtubule-based process” within the category of

biological processes; “ATP-dependent activity”, “cytoskeletal protein binding”, “cytoskeletal motor activity” and “actin binding” within the category of molecular function; “supramolecular fiber”, “basement membrane” and “myosin filament” within the cellular component category (Figure S5).

Gene expression analysis in IPF lung tissue

A total of five cohorts with bulk RNA-Seq data were included from GEO, comprising lung tissue from 147 IPF participants and 135 controls. Following quality control procedures, WGCNA identified 17 co-expressed modules of genes. For IPF disease status compared with control, 14 modules had significant module-trait-relationship ($p<0.05$) of which 6 modules had positive correlations and 8 modules had negative (Figure 4A).

56 of the 90 genes identified in the genetic analysis were classified within 11 of the 14 significantly correlated modules (Figure S8A). The gene ontology overrepresentation analysis of 56 overlapping genes identified “microtubule-based process” and “microtubule organizing center organization for biological processes”, and “ATP-dependent activity”, “cytoskeletal motor activity” and “extracellular matrix structural constituent” for molecular functions, and “supramolecular fiber”, “microtubule” and “brush border” for cellular components (Figure 4B). Top gene hits with missense and PAVs, *HMCN1* and *RGPD1*, were classified within significantly upregulated modules (Table S4).

Within the 56 overlapping genes present in WGCNA modules, 17 met individual criteria for upregulated expression in IPF lung tissue; *COL6A3* (collagen type VI alpha 3 chain), *CADPS* (calcium dependent secretion activator), *HMCN1* (hemicentin-1), *ESPN* (espin), *SPEG* (striated muscle enriched protein kinase), *ADAMTSL1* (a disintegrin and metalloproteinase with thrombospondin motifs like-1), *MYOF* (myoferlin), *NLGN2* (neuroligin 2), *SORD* (sorbitol dehydrogenase), *WDR90* (tryptophan-aspartic acid repeat domain 90), CROCC (ciliary rootlet coiled-coil), as well as *DYNC2H1* (dynein 2 complex) and *DNAH5*, *DNAH6*, *DNAH7*, *DNAH11* and *DNAH12* (dynein axonemal heavy chain 5, 6, 7, 11, 12). A total of three overlapping genes had significantly down regulated expression in IPF lung tissue; *ADGRV1* (adhesion G protein-coupled receptor V1), *MYH14* (myosin heavy chain 14), and *PCDH15* (protocadherin related 15) (Figure 4C).

Functional annotation of top hits

A combined 74 participants across both cohorts had a rare missense variant in *HMCN1* within 454 kbp of each other from exon-12 to exon-107 of 107 exons. A total of 61 independent variants were observed, with the most frequent missense variant observed in four participants, followed by three missense mutations observed at a frequency of three participants, six at a frequency of two participants, and the remaining 51 observed in single participants (Figure 5A). Variant rs371095992 was predicted to change the amino acid from cysteine to tryptophan in the TSP-1 domain, leading to a predicted interaction with the mature form of transforming growth factor beta (TGF-β) (Figure 5B, 5C).

Across PROFILE and GE100KGP, a total of 39 participants had missense variants in *RGPD1* for 11 unique variants. Nine of the 11 variants were within exon 20, predicted to encode the Ran binding domain type 1 domain 2 (RanBD1-D2) (Figure 6A). *RANBP2* (Ran binding protein 2) was also observed as an overlapping gene in significant gene co-expression modules (Table S4).

In the combined cohort of PROFILE and GE100KGP, a total of 44 participants had a LoF variant in *ATAT1*. There were 24 unique variants, with 23 leading to a predicted stop codon within the same disordered protein domain region (Figure 6B). All variants were indels within 12 base pairs of each other on exon-10 of 13 exons at 5 independent positions, and all predicted to lead to an early stop codon.

Discussion

The iterative pipeline observed a total of 90 genes observed with significant rare variant burden of missense variants, PAVs, LoF variants and PTVs that were overlapping between two PF case cohorts. *HMCN1*, *RGPD1* and *ATAT1* were observed within the top 20 genes in pathogenic categories across both PF cohorts. *RGPD1* included the greatest burden regarding missense and PAVs, whilst *ATAT1* included the greatest burden in LoF variants and PTVs. A significant rare variant burden in *HMCN1* was observed for categories of missense variants and PAVs, as well as significantly upregulated expression of *HMCN1* in IPF lung tissue compared to control lung tissue. Gene ontology analysis of WGS across both independent cohorts supports the importance of cell adhesion, actin cytoskeleton dynamics, and extracellular matrix components in the development of PF. Integrative analysis of five published lung transcriptomic datasets identified co-regulated gene modules associated with IPF compared with healthy controls, frequently enriched in genes within similar biological pathways.

Complementary findings between genetic and transcriptomic analyses highlight important pathogenic mechanisms in PF.

HMCN1 has been implicated in diseases including age-related macular degeneration[37] and Fraser Syndrome[38], however little is known about its role in PF pathogenesis. *HMCN1* has previously been reported to be highly expressed in lung tissues in IPF patients in comparison with healthy controls using integrative analysis of array gene expression profiling data[39,40]. TGF-β is considered the master regulator of fibrogenic mechanisms, whilst hemicentin-1 is an ECM component involved in TGF-β regulated organisation of the actin cytoskeleton and cellular adhesion[41], belonging to the fibulin family that is associated with basement membrane and elastic fibre formation[42]. Consistently, fibulin-1 is associated with IPF disease progression[43], fibulin-4 is related to collagen fibre assembly in lung fibrosis[44], and functional compensation effects have been demonstrated between *HMCN1* and other fibulin family members[45]. *HMCN1* has a regulatory role in enhancing the TGF-β signalling pathway mediated pro-fibrotic response of neonatal mouse cardiac fibroblasts and in cardiac fibroblasts migration during myocardial remodelling[41,46], as well as in activating RhoA signalling mediated invasiveness of ovarian cancer-associated fibroblasts[47]. Hemicentin-1 is the only fibulin member with a TSP-1 domain, allowing it to regulate TGF-β bioavailability by an interaction between TSP-1 and the latency association peptide at the C-terminus of pro-TGF-β[42]. We observed a variant that was predicted to lead to a conformational change in the structure of the TSP-1 domain of hemicentin-1 and a direct interaction with mature TGF-β. *HMCN1* knock-down studies demonstrate reduced TGF-β signalling, suggesting greater expression in IPF may be linked to the high levels of TGF-β in disease[41,46]. Further studies should validate the aberrant interaction in-vitro, determine its impact on TGF-β activity, and assess the functional consequence of variants encoding changes to amino acids in other hemicentin-1 domains.

This study observed a burden of pathogenic variants in genes controlling actin filament-based processes, cytoplasmic microtubule organization and actin cytoskeleton organization. Previous studies have identified a common variant in *KNL1*[48] and a rare variant in *SPDL1*[11], with both genes important in microtubule interaction at the mitotic spindle. Substantial actin remodelling of the cytoskeleton is necessary for differentiation of alveolar epithelium AT1 cells from AT2 cells[49], as well as myofibroblast differentiation in human lung fibroblasts[50]. LoF variants were observed in *ATAT1*, a gene encoding enzyme α-tubilin acetyltransferase that is specifically involved in the acetylation of lysine residue of α-tubulin to destabilise microtubules and intricately linked to substrate rigidity and mechanosensing[51,52]. PAVs and differential

expression were observed in *WDR90,* a core component of the microtubule central wall[53], dynein genes *DNAH5*, *DNAH6*, *DNAH7*, *DNAH11*, *DNAH12* and *DYNC2H1*, encoding microtubule-associated motor proteins, and *CROCC*, which encodes rootletin and is involved in centrosome adhesion and cilium assembly. The majority of variants in *RGPD1* were observed in a RanBD1 domain, which may have roles in the organisation of microtubules at the kinetochore and in binding to α-tubulin acetyltransferase[54,55]. *RGPD1* gene expression was classified within a significantly upregulated transcriptomic module. *RANBP2* has a related function and was observed to have an overlapping significant burden, although classified in a downregulated module within transcriptomic analyses. These findings further implicate microtubule stability in the aetiology of PF, which could be targeted by small molecules[56].

Aberrant collagen deposition in the extracellular matrix contributes to stiffness and is a major feature of PF pathogenesis[57,58]. Corroborating evidence for rare variants and gene expression were observed for additional components of the extracellular matrix including *COL6A3*, which encodes collagen-alpha-1(VI) chain and is a major component of microfibrils, as well as *ADAMTSL1*, which encodes a member of the *ADAMTS* (A Disintegrin And Metalloproteinase with ThromboSpondin motifs) family that are involved in collagen assembly and modification of proteoglycans in the extracellular matrix[59]. Myoferlin is a membrane-anchored protein involved in cell-adhesion, for which rare variants were similarly identified in a recent gene burden test within FPF specifically[12]. A common IPF risk variant has been previously identified in the intronic region of desmoplakin, which is involved in intercellular junctions and cell adhesion, leading to reduced basement membrane integrity and increased expression of extracellular matrix genes[60]. Extracellular matrix constituents are produced by fibroblasts, with proliferation of fibroblasts a major characteristic in the development of pulmonary fibrosis[57]. A significant burden in LoF variants, PAVs and PTVs were observed in *FBRS*, which encodes fibrosin and is involved in fibroblast proliferation and myofibroblast differentiation[61]. Together, these findings highlight extracellular matrix components and cell adhesion processes that could be important targets in precision medicine strategies for families and individuals.

The iterative focus on overlapping genes and pathways across independent cohorts can mitigate the possibility of false positive findings. This bioinformatics pipeline is hypothesis generating and requires functional *in-vitro* and *in-vivo* studies to validate the role of specific genetic variants and their clinical relevance. Genome coverage and read misalignment may lead to artefacts in estimating the burden of indels in cases compared to controls, which may be particularly relevant for *ATAT1* as many indels were observed in the same region. The analysis

can be impacted by FPF patients with a known causal variant, or cases classified with PF who might carry qualifying variants relating to their original reason for recruitment to the study in GE100KGP cohort. This study considered univariable relationships of disease compared with control, whilst traits that may not have been controlled for in the sampling or the analytical pipeline can confound the observed findings, such as smoking history, comorbidities or treatment.

In conclusion, this study using burden testing of rare pathogenic variants in two independent disease cohorts implicated a number of genes with novel involvement in the disease pathogenesis of PF, including *HMCN1* and *RGPD1*, which was supported by findings from transcriptomic analysis across five open-source lung tissue datasets. Identification of rare variants in genes controlling actin cytoskeleton and microtubule assembly, extracellular matrix and cell adhesion, with supporting evidence of differential gene expression, provides genetic evidence of their involvement in fibrogenesis. Further preclinical studies should assess the role of extracellular matrix component hemicentin-1 as well as microtubule organisation in the pathogenesis of PF.

Acknowledgements

This work was funded by an MRC Programme Grant (MR/V00235X/1) awarded to RGJ. This research is partly funded by the National Institute for Health and Care Research (NIHR) Leicester Biomedical Research Centre (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. DW, YN, BL, HD, AJ, LVW and DMR performed the data analysis. SRJ, TMM, PLM, ER, GS and AUW contributed to the patient recruitment and data collection. DW and IS wrote the manuscript with input from all authors.

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. TMM, via his institution, has received industry-academic funding from Astra Zeneca and GlaxoSmithKline R&D; and consultancy or speaker fees from Abbvie, Amgen, Astra Zeneca, Bayer, Boehringer Ingelheim, BMS, CSL Behring, Endeavor, Fibrogen, Galapagos, Galecto, GlaxoSmithKline, IQVIA, Merck, Pliant, Pfizer, Qureight, Roche, Sanofi-Aventis, Structure Therapeutics, Trevi and Vicore. He is supported by an NIHR Clinician Scientist Fellowship (NIHR Ref: CS-2013-13-017) British Lung Foundation Chair in Respiratory Research (C17-3). 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.

References

1. Garcia, C.K. Idiopathic pulmonary fibrosis: update on genetic discoveries. *Proc Am Thorac Soc* **8**, 158-62 (2011).
2. Finnerty, J.P., Ponnuswamy, A., Dutta, P., Abdelaziz, A. & Kamil, H. 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* **21**, 411 (2021).
3. Lee, H.L. *et al.* Familial idiopathic pulmonary fibrosis: clinical features and outcome. *Chest* **127**, 2034-41 (2005).
4. Oh, C.K., Murray, L.A. & Molfino, N.A. Smoking and idiopathic pulmonary fibrosis. *Pulm Med* **2012**, 808260 (2012).
5. Fernandez, B.A. *et al.* A Newfoundland cohort of familial and sporadic idiopathic pulmonary fibrosis patients: clinical and genetic features. *Respir Res* **13**, 64 (2012).
6. Kropski, J.A. Familial Interstitial Lung Disease. *Semin Respir Crit Care Med* **41**, 229-237 (2020).
7. Fingerlin, T.E. *et al.* Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. *Nat Genet* **45**, 613-20 (2013).
8. Allen, R.J. *et al.* Genetic variants associated with susceptibility to idiopathic pulmonary fibrosis in people of European ancestry: a genome-wide association study. *Lancet Respir Med* **5**, 869-880 (2017).
9. Oldham, J.M. *et al.* PCSK6 and Survival in Idiopathic Pulmonary Fibrosis. *Am J Respir Crit Care Med* **207**, 1515-1524 (2023).
10. 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* **11**, 65-73 (2023).
11. Dhindsa, R.S. *et al.* Identification of a missense variant in SPDL1 associated with idiopathic pulmonary fibrosis. *Commun Biol* **4**, 392 (2021).
12. Liu, Q. *et al.* The Genetic Landscape of Familial Pulmonary Fibrosis. *Am J Respir Crit Care Med* **207**, 1345-1357 (2023).
13. Stuart, B.D. *et al.* Effect of telomere length on survival in patients with idiopathic pulmonary fibrosis: an observational cohort study with independent validation. *Lancet Respir Med* **2**, 557-65 (2014).
14. Duckworth, A. *et al.* Telomere length and risk of idiopathic pulmonary fibrosis and chronic obstructive pulmonary disease: a mendelian randomisation study. *Lancet Respir Med* **9**, 285-294 (2021).
15. Peljto, A.L. *et al.* Idiopathic Pulmonary Fibrosis Is Associated with Common Genetic Variants and Limited Rare Variants. *Am J Respir Crit Care Med* **207**, 1194-1202 (2023).
16. Donoghue, L.J. *et al.* Identification of a Genetic Susceptibility Locus for Idiopathic Pulmonary Fibrosis in the 16p Subtelomere Using Whole-Genome Sequencing. *Am J Respir Crit Care Med* **207**, 941-944 (2023).
17. Huang, Y. *et al.* Central lung gene expression associates with myofibroblast features in idiopathic pulmonary fibrosis. *BMJ Open Respir Res* **10**(2023).
18. Furusawa, H. *et al.* Chronic Hypersensitivity Pneumonitis, an Interstitial Lung Disease with Distinct Molecular Signatures. *Am J Respir Crit Care Med* **202**, 1430-1444 (2020).
19. Jaffar, J. *et al.* Matrix metalloproteinase-7 is increased in lung bases but not apices in idiopathic pulmonary fibrosis. *ERJ Open Res* **8**(2022).
20. Nance, T. *et al.* Transcriptome analysis reveals differential splicing events in IPF lung tissue. *PLoS One* **9**, e92111 (2014).

21. Luzina, I.G. *et al.* Transcriptomic evidence of immune activation in macroscopically normal-appearing and scarred lung tissues in idiopathic pulmonary fibrosis. *Cell Immunol* **325**, 1-13 (2018).
22. Karczewski, K.J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443 (2020).
23. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).
24. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).
25. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**(2021).
26. Guo, M.H., Plummer, L., Chan, Y.M., Hirschhorn, J.N. & Lippincott, M.F. Burden Testing of Rare Variants Identified through Exome Sequencing via Publicly Available Control Data. *Am J Hum Genet* **103**, 522-534 (2018).
27. Edgar, R., Domrachev, M. & Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**, 207-10 (2002).
28. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
29. Ritchie, M.E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).
30. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
31. Yu, G., Wang, L.G., Han, Y. & He, Q.Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284-7 (2012).
32. Nassar, L.R. *et al.* The UCSC Genome Browser database: 2023 update. *Nucleic Acids Res* **51**, D1188-D1195 (2023).
33. UniProt, C. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* **51**, D523-D531 (2023).
34. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* **19**, 679-682 (2022).
35. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).
36. Schrodinger, LLC. The PyMOL Molecular Graphics System, Version 3.0. (2024).
37. Thompson, C.L. *et al.* Complement factor H and hemicentin-1 in age-related macular degeneration and renal phenotypes. *Hum Mol Genet* **16**, 2135-48 (2007).
38. Carney, T.J. *et al.* Genetic analysis of fin development in zebrafish identifies furin and hemicentin1 as potential novel fraser syndrome disease genes. *PLoS Genet* **6**, e1000907 (2010).
39. Wan, H. *et al.* Identification of Hub Genes and Pathways Associated With Idiopathic Pulmonary Fibrosis via Bioinformatics Analysis. *Front Mol Biosci* **8**, 711239 (2021).
40. Fu, C. *et al.* Identification of immune biomarkers associated with basement membranes in idiopathic pulmonary fibrosis and their pan-cancer analysis. *Front Genet* **14**, 1114601 (2023).
41. Chowdhury, A. *et al.* Fibulin-6 regulates pro-fibrotic TGF-beta responses in neonatal mouse ventricular cardiac fibroblasts. *Sci Rep* **7**, 42725 (2017).
42. Argraves, W.S., Greene, L.M., Cooley, M.A. & Gallagher, W.M. Fibulins: physiological and disease perspectives. *EMBO Rep* **4**, 1127-31 (2003).
43. Jaffar, J. *et al.* Fibulin-1 predicts disease progression in patients with idiopathic pulmonary fibrosis. *Chest* **146**, 1055-1063 (2014).
44. Nakamura, T. Roles of short fibulins, a family of matricellular proteins, in lung matrix assembly and disease. *Matrix Biol* **73**, 21-33 (2018).

45. Lin, M.H. *et al.* Mammalian hemicentin 1 is assembled into tracks in the extracellular matrix of multiple tissues. *Dev Dyn* **249**, 775-788 (2020).
46. Chowdhury, A. *et al.* Expression of fibulin-6 in failing hearts and its role for cardiac fibroblast migration. *Cardiovasc Res* **103**, 509-20 (2014).
47. Liu, C.L., Pan, H.W., Torng, P.L., Fan, M.H. & Mao, T.L. SRPX and HMCN1 regulate cancer-associated fibroblasts to promote the invasiveness of ovarian carcinoma. *Oncol Rep* **42**, 2706-2715 (2019).
48. Allen, R.J. *et al.* Genome-wide association study across five cohorts identifies five novel loci associated with idiopathic pulmonary fibrosis. *Thorax* **77**, 829-833 (2022).
49. Wu, H. *et al.* Progressive Pulmonary Fibrosis Is Caused by Elevated Mechanical Tension on Alveolar Stem Cells. *Cell* **180**, 107-121 e17 (2020).
50. Sandbo, N. *et al.* Control of myofibroblast differentiation by microtubule dynamics through a regulated localization of mDia2. *J Biol Chem* **288**, 15466-73 (2013).
51. Montagnac, G. *et al.* alphaTAT1 catalyses microtubule acetylation at clathrin-coated pits. *Nature* **502**, 567-70 (2013).
52. Seetharaman, S. *et al.* Microtubules tune mechanosensitive cell responses. *Nat Mater* **21**, 366-377 (2022).
53. Steib, E. *et al.* WDR90 is a centriolar microtubule wall protein important for centriole architecture integrity. *Elife* **9**(2020).
54. Wu, K. *et al.* GCN5L1 interacts with alphaTAT1 and RanBP2 to regulate hepatic alpha-tubulin acetylation and lysosome trafficking. *J Cell Sci* **131**(2018).
55. Joseph, J. & Dasso, M. The nucleoporin Nup358 associates with and regulates interphase microtubules. *FEBS Lett* **582**, 190-6 (2008).
56. Hideshima, T. *et al.* Discovery of selective small-molecule HDAC6 inhibitor for overcoming proteasome inhibitor resistance in multiple myeloma. *Proc Natl Acad Sci U S A* **113**, 13162-13167 (2016).
57. Upagupta, C., Shimbori, C., Alsilmi, R. & Kolb, M. Matrix abnormalities in pulmonary fibrosis. *Eur Respir Rev* **27**(2018).
58. Clarke, D.L., Carruthers, A.M., Mustelin, T. & Murray, L.A. Matrix regulation of idiopathic pulmonary fibrosis: the role of enzymes. *Fibrogenesis Tissue Repair* **6**, 20 (2013).
59. Lamande, S.R., Morgelin, M., Adams, N.E., Selan, C. & Allen, J.M. The C5 domain of the collagen VI alpha3(VI) chain is critical for extracellular microfibril formation and is present in the extracellular matrix of cultured cells. *J Biol Chem* **281**, 16607-14 (2006).
60. Hao, Y. *et al.* Genome-Wide Association Study: Functional Variant rs2076295 Regulates Desmoplakin Expression in Airway Epithelial Cells. *Am J Respir Crit Care Med* **202**, 1225-1236 (2020).
61. Prakash, S., Paul, W.E. & Robbins, P.W. Fibrosin, a novel fibrogenic cytokine, modulates expression of myofibroblasts. *Exp Mol Pathol* **82**, 42-8 (2007).

Figures

Figure 1. The results of burden testing of rare variants for PROFILE cohort.

(A) Number of significant genes in each variant category after the removal of significant genes in synonymous variant category. (B) Manhattan plot for each variant category after the removal of significant genes in synonymous variant category. Only top 20 significant genes were labelled in the gene names.

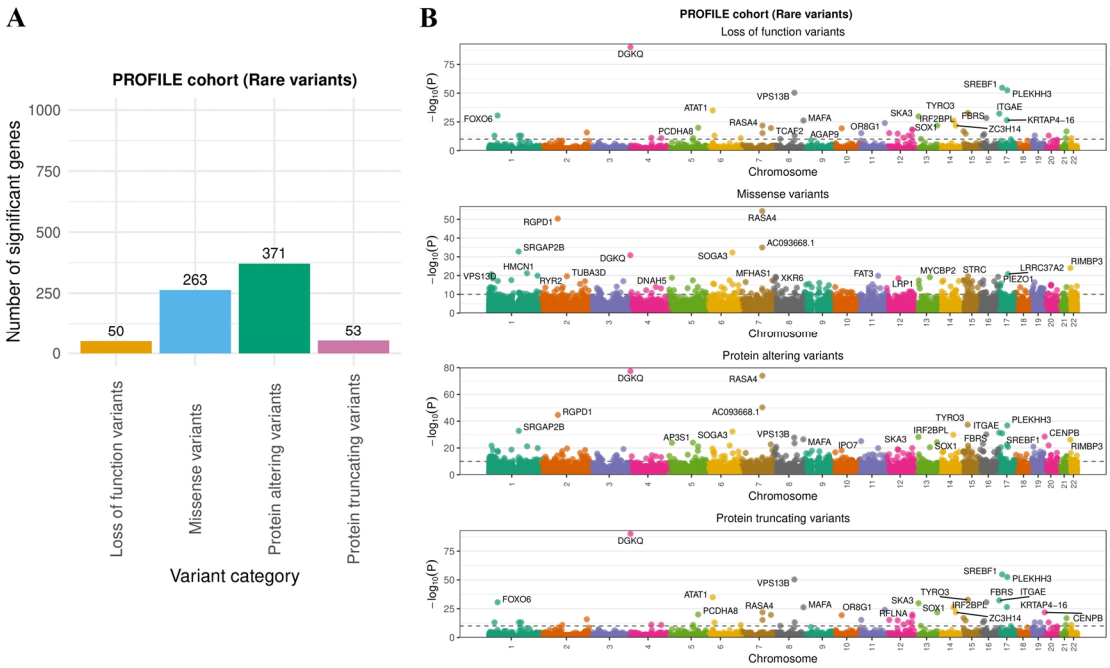


Figure 2. The results of burden testing of rare variants for GE100KGP cohort.

(A) Number of significant genes in each variant category after the removal of significant genes in synonymous variant category. (B) Manhattan plot for each variant category after the removal of significant genes in synonymous variant category. Only top 20 significant genes were labelled in the gene names.

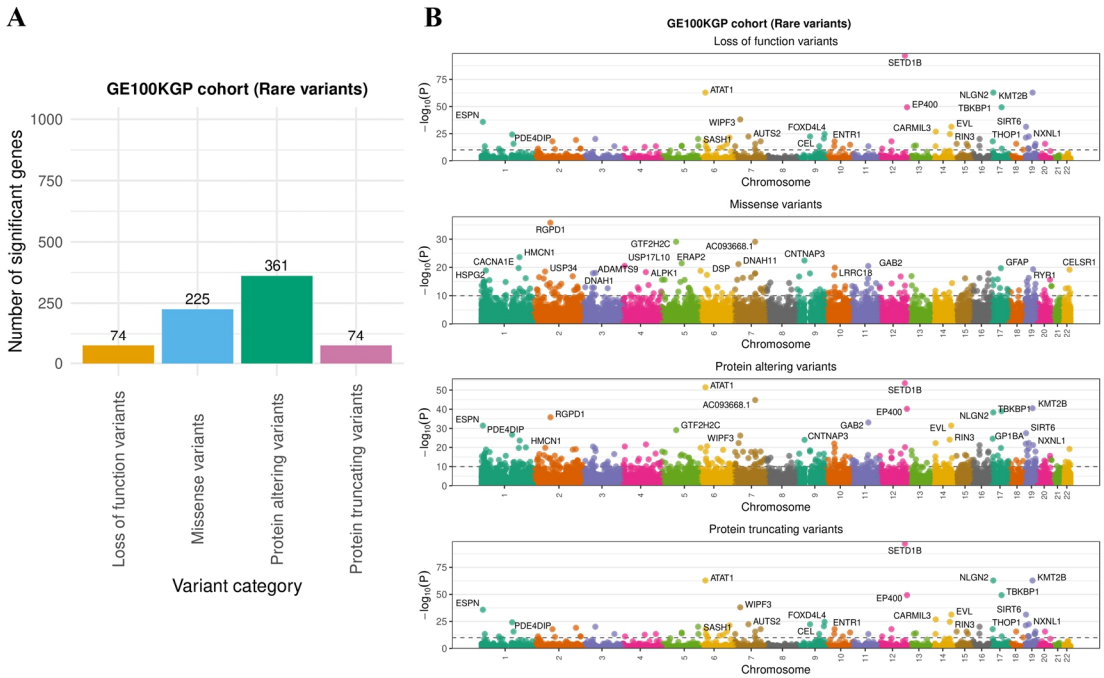


Figure 3. The shared significant genes between PROFILE cohort and GE100KGP cohort for the analysis of rare variants.

(A) Venn diagram of significant genes from the two cohorts. (B) Overrepresentation analysis of gene ontology terms for shared significant genes between the two cohorts. Top 15 terms were shown for each variant category.

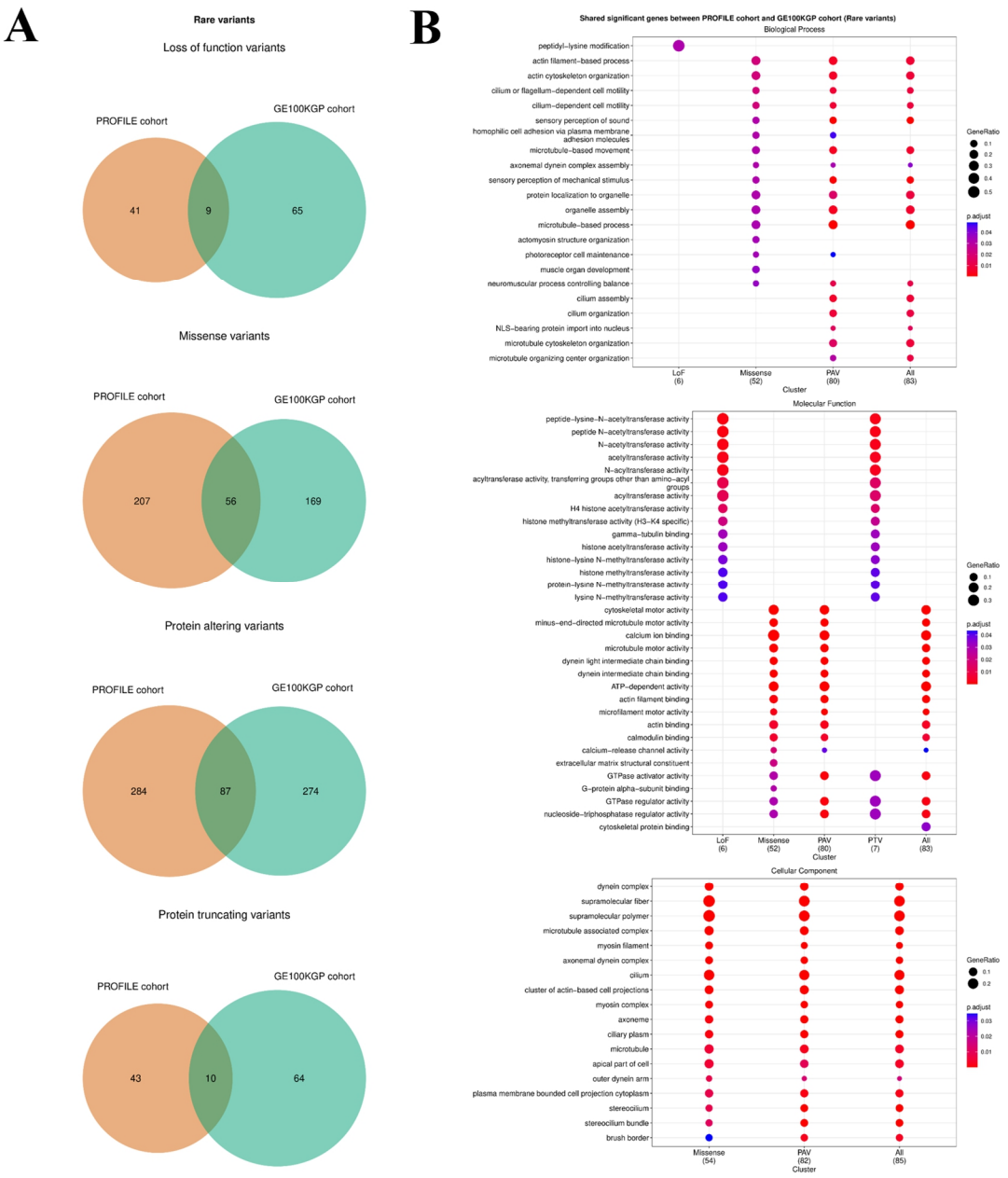


Figure 4. The comparison between genetic and transcriptomic analysis.

(A) The module-trait relationships evaluated by correlating module eigengenes with clinical traits for WGCNA analysis of transcriptomic datasets. Correlations and p values were presented. 14 modules were significant at a $p<0.05$. (B) Overrepresentation analysis of gene ontology terms for shared genes identified in the two analyses. Top 15 terms were shown. (C) Volcano plot for integrative transcriptomic analysis. *HMCN1* was among those upregulated genes in IPF lung tissue.

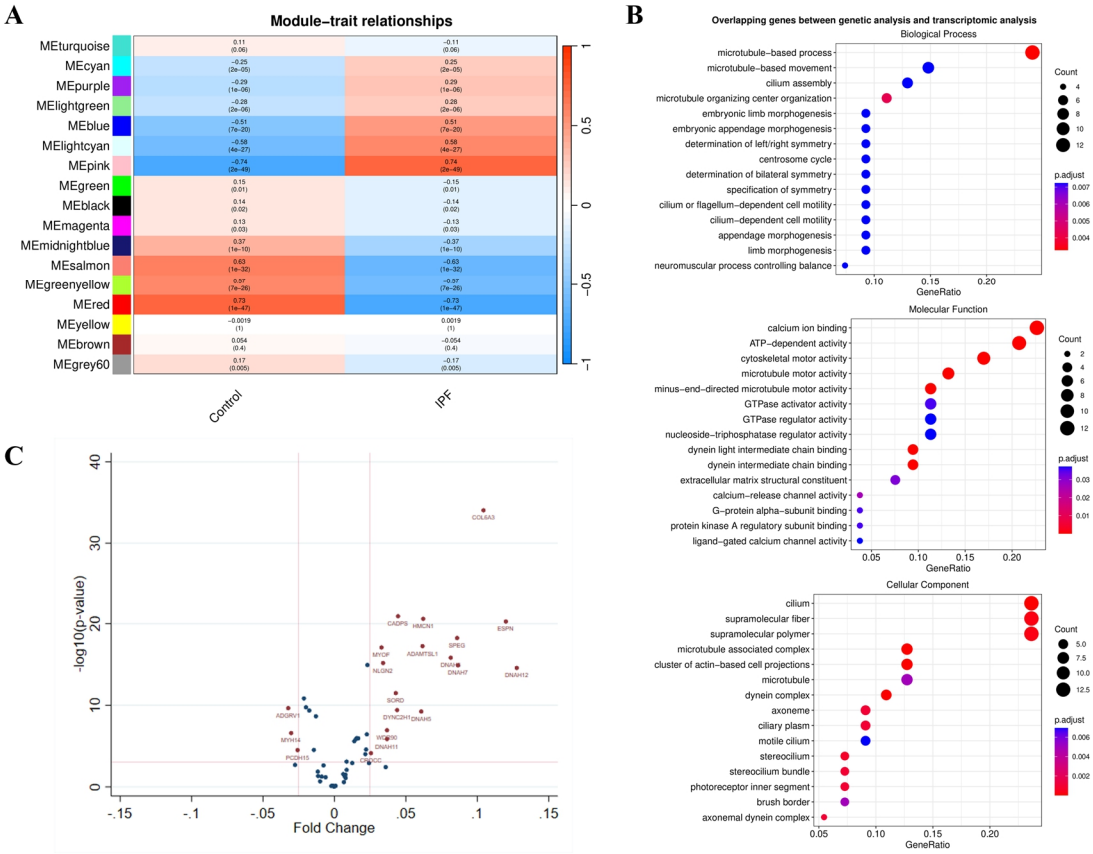


Figure 5. Functional and structural annotation of variants for *HMCN1*.

(A) Lollipop plot of the distribution and mutation frequency of pathogenic variants for *HMCN1* in PROFILE (solid-circle) and GE100KGP (hollow-circle) or both (solid-triangle) cases across the Von Willebrand factor type A (VWA_2), Immunoglobulin-like I-set (I-set), Thrombospondin type 1 repeat (TSP-1), Nidogen-like G2 (G2F), EGF-like calcium-binding (EGF_GA) and Complement Clr-like EGF-like (cEGF) domains. Black arrow highlights variant rs371095992 predicted to change the amino acid from cysteine to tryptophan. (B, C) Structure of *HMCN1* (Q96RW7) TSP-1 domain binding to pro-TGF-β (P01137) predicted by AlphaFold2 and visualised with PyMOL represents the mature TGF-β molecule (white), the latency associated peptide (LAP, blue) and the wildtype TSP-1 domain (B, green) or the mutant TSP-1 domain based on variant rs371095992 (C, pink). The wildtype TSP-1 domain interacts with LAP (B), the mutant TSP-1 domain suggests interaction with TGF-β and LAP (C).

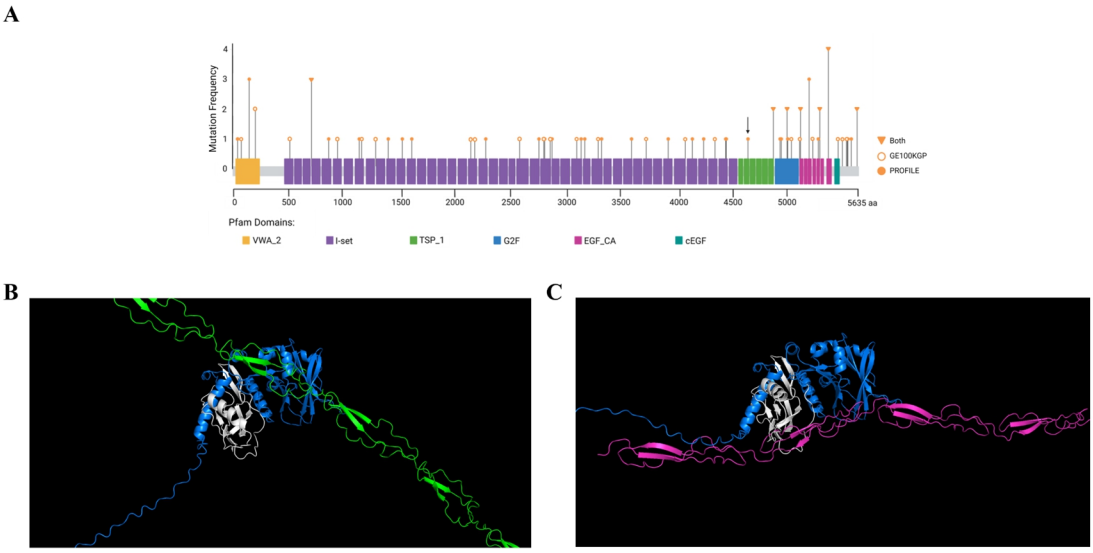
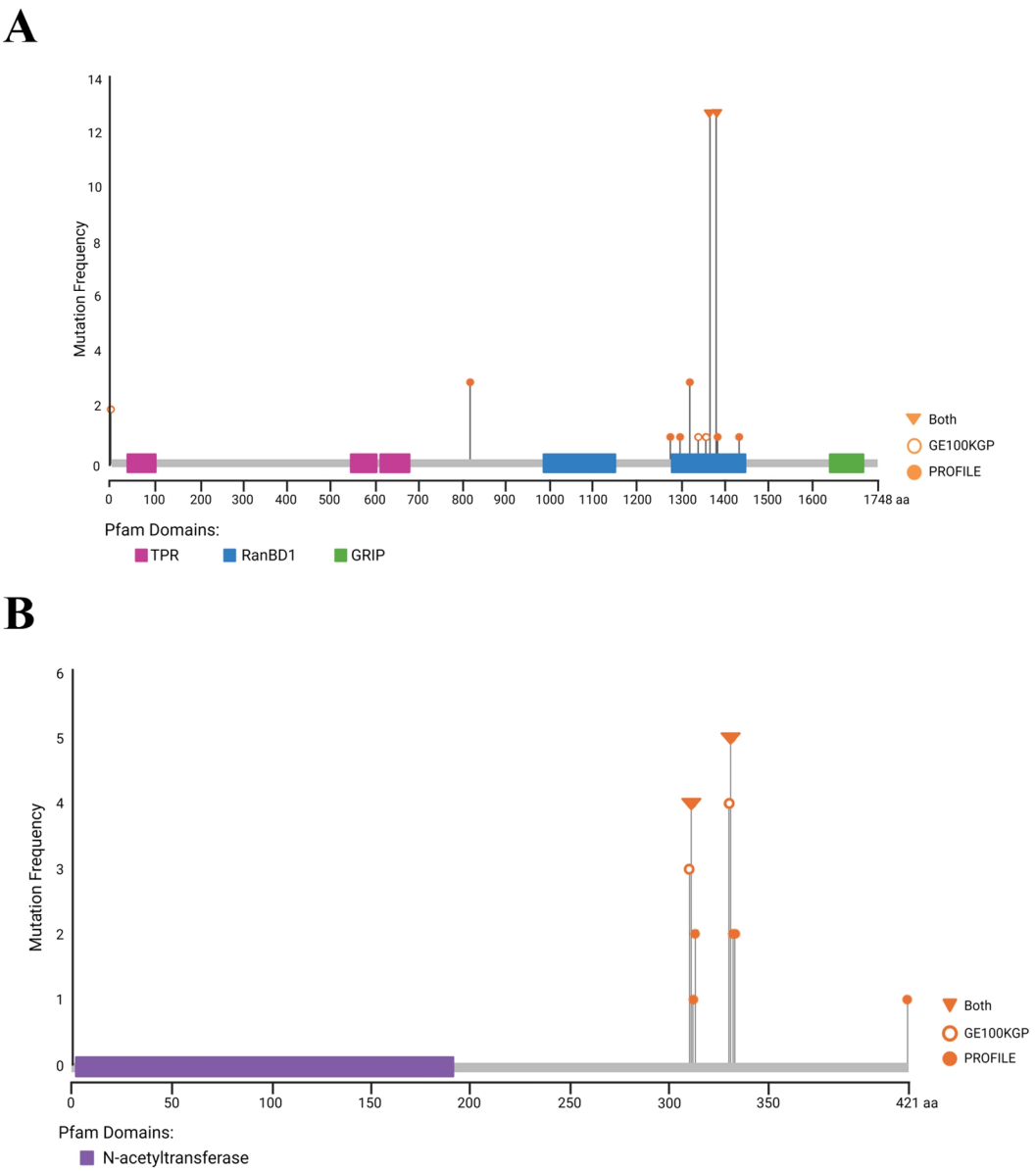


Figure 6. Lollipop plots of the distribution of potentially pathogenic variants for *RGPD1* and *ATAT1*.

(A) Lollipop plot of the distribution and mutation frequency of pathogenic variants for *RGPD1* in PROFILE (solid-circle) and GE100KGP (hollow-circle) or both (solid-triangle) cases across RanBD1 domain. (B) Lollipop plot of the distribution and mutation frequency of pathogenic variants predicted to lead to stop codons in *ATAT1* in PROFILE (solid-circle) and GE100KGP (hollow-circle) or both (solid-triangle).