# Genome-wide association study of Idiopathic Pulmonary Fibrosis susceptibility using clinically-curated European-ancestry datasets

Daniel Chin[1,2,3*], Tamara Hernandez-Beeftink[1,2,3*], Lauren Donoghue[4], Beatriz Guillen-Guio[1,2,3,5], Olivia C Leavy[1,2,3], Ayodeji Adegunsoye[6], Helen L Booth[7], CleanUP-IPF Investigators of the Pulmonary Trials Cooperative[8], William A Fahy[9], Tasha E Fingerlin[10], Bibek Gooptu[2,3], Ian P Hall[11], Simon P Hart[12], Mike R Hill[13], Nik Hirani[14], Simon R Johnson[15], Naftali Kaminski[16], Jose Miguel Lorenzo-Salazar[17], Shwu-Fan Ma[18], Robin J McAnulty[19], Mark I McCarthy[4], Amy D Stockwell[4], Toby M Maher[20,21], Ann B Millar[22], Philip L Molyneaux[23], Maria Molina-Molina[5,24,25], Vidya Navaratnam[26,27], Margaret Neighbors[4], Justin M Oldham[28], Helen Parfrey[29], Gauri Saini[30], Ian Sayers[11], X Rebecca Sheng[4], Iain D Stewart[31], Mary E Strek[6], Martin D Tobin[1], Moira KB Whyte[14], Maria C Zarcone[20], Yingze Zhang[32], Fernando Martinez[8], Brian L Yaspan[4], Carl J Reynolds[33], David A Schwartz[34], Carlos Flores[5,17,35,36], Imre Noth[18], R Gisli Jenkins[20], Richard J Allen[1,2,3], Louise V Wain[1,2,3#]

[1]Division of Public Health and Epidemiology, School of Medical Sciences, University of Leicester, Leicester, UK, [2]NIHR Leicester Biomedical Research Centre, Leicester, UK, [3]Centre for Fibrosis Research, University of Leicester, Leicester, UK, [4]Genentech, California, USA, [5]Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBERES), Instituto de Salud Carlos III, Madrid, Spain, [6]University of Chicago, Chicago, USA, [7]University College London Hospitals, London, UK, [8]Weill Cornell Medicine, New York, USA, [9]GlaxoSmithKline, London, UK, [10]National Jewish Health, Colorado, USA, [11]Centre for Respiratory Research, NIHR Nottingham Biomedical Research Centre, School of Medicine, Biodiscovery Institute, University of Nottingham, Nottingham, UK, [12]University of Hull, Hull, UK, [13]University of Oxford, Oxford, UK, [14]University of Edinburgh, Edinburgh, UK, [15]Centre for Respiratory Research, NIHR Biomedical Research Centre and Biodiscovery Institute, Translational Medical Sciences, School of Medicine, University of Nottingham, Nottingham, UK, [16]Yale School of Medicine, Connecticut, USA, [17]Genomics Division, Instituto Tecnologico y de Energias Renovables, Santa Cruz de Tenerife, Spain, [18]University of Virginia, Virginia, USA, [19]University College London, London, UK, [20]NIHR Imperial Biomedical Research Unit, National Heart and Lung Institute, Imperial College London, London, UK, [21]Keck Medicine of USC, University of Southern California, Los Angeles, California, 90033, USA, [22]University of Bristol, Bristol, UK, [23]National Heart and Lung Institute, Imperial College London, London, UK, [24]Servei de Pneumologia, Hospital Universitari de Bellvitge (HUB), Laboratori de Pneumologia Experimental, Institut de Investigació Biomèdica de Bellvitge (IDIBELL), Barcelona, Spain, [25]Facultat de Medicine, Universitat de Barcelona, Barcelona, Spain, [26]Department of Respiratory Medicine, Sir Charles Gardiner Hospital, Perth, Australia, [27]Centre for Respiratory Research, University of Western Australia, Perth, Australia, [28]University of Michigan, Michigan, USA, [29]Royal Papworth Hospital NHS Foundation Trust, Cambridge, UK, [30]University of Nottingham, Nottingham, UK, [31]Imperial College London, London, UK, [32]University of Pittsburgh, Pittsburgh, USA, [33]Imperial College, London, UK, [34]University of Colorado Medicine, Colorado, USA, [35]Research Unit, Hospital Universitario Nuestra Señora de Candelaria, Instituto de Investigación Sanitaria de Canarias, Santa Cruz de Tenerife, Spain, [36]Facultad de Ciencias de la Salud, Universidad Fernando Pessoa Canarias, Las Palmas de Gran Canaria, Spain

*Contributed equally

[#]**Corresponding author**: Louise V. Wain
Division of Public Health and Epidemiology,
School of Medical Sciences,
University of Leicester, UK Leicester LE1 7RH
E-mail address: lvw1@leicester.ac.uk

# ABSTRACT

**Background**

Idiopathic pulmonary fibrosis (IPF) is a rare, incurable lung disease with a median survival of 3-5 years after diagnosis. Treatment options are limited. Genetic association studies can identify new genes involved in disease that might represent potential new drug targets, and it has been shown that drug targets with support from genetic studies are more likely to be successful in clinical development. Previous genome-wide association studies (GWAS) of IPF susceptibility have identified more than 20 signals implicating genes involved in multiple mechanisms, including telomere dysfunction, cell-cell adhesion, host defence immunity, various signalling pathways and, more recently, mitotic spindle assembly complex.

**Aim**

To leverage new datasets and genotype imputation to discover further genes involved in development of IPF that could yield new pathobiological avenues for exploration and to guide future drug target discovery.

**Methods**

We conducted a GWAS of IPF susceptibility including seven IPF case-control studies comprising 5,159 IPF cases and 27,459 controls of European ancestry, where IPF diagnosis was made by a respiratory clinician according to international guidelines. Genotypes were obtained from Whole Genome Sequencing (WGS) or from array-based imputation to the TOPMed WGS reference panel. New signals were replicated in independent biobanks with IPF defined using Electronic Healthcare Records. Bayesian fine-mapping was performed to identify the most likely causal variant(s) and bioinformatic investigation undertaken to map associated variants to putative causal genes.

**Results**

We identified three novel genetic signals of association with IPF susceptibility. Genes prioritised by functional evidence at these signals included *MUC1*, which encodes a large transmembrane glycoprotein and known biomarker of lung fibrosis, and *NTN4* encoding Netrin-4 whose known roles include angiogenesis. The third signal may map to *SLC6A6,* a taurine and beta-alanine transporter gene, previously implicated in retinal, cardiac and kidney dysfunction.

**Conclusion**

Our study has identified new associations not previously identified by previous large biobank-based studies thereby highlighting the value of utilising clinically-curated IPF case-control studies, and new genotype imputation. We present new evidence for disease-driving roles of *MUC1* and of endothelial cell and vascular changes in IPF.


**Keywords:** Idiopathic Pulmonary Fibrosis, Genome-Wide Association Studies, genetics

## INTRODUCTION

Idiopathic pulmonary fibrosis (IPF) is a chronic, progressive lung disease thought to result from an aberrant response to lung injury, culminating in an exaggerated healing response with excessive deposition of extracellular matrix in the interstitium (1) (2). This incurable lung disease affects more than 7 in 100,000 people with poor survival (median survival of 3-5 years after diagnosis) (3) and has limited treatment options (4). IPF is incompletely understood and is significantly influenced by both genetic and environmental factors, with a high heritability and polygenic aetiology (5) (6).

Genome-wide association studies (GWAS) assess genetic variants from across the genome for their association with disease. Previous GWAS have highlighted more than 20 independent genetic signals linked to IPF susceptibility, implicating pathways such as telomere dysfunction, cell-cell adhesion, host defence, TGF-β signalling, and mitotic spindle assembly (5) (7) (8) (9) (10) (11). Whole genome sequencing (WGS) technologies provide comprehensive genomic coverage, but remain cost prohibitive for large studies. Instead, reference panels derived from large WGS datasets allow for improved imputation of unmeasured variants, including those with low allele frequencies, providing a cost-effective solution for increased genome-wide coverage (12).

Although each genetic signal has a small individual effect, except for the gain-of-function *MUC5B* promoter variant (6) (13), targeting the pathways conferred by genetic associations could offer significant therapeutic potential, as drug targets supported by genetic evidence studies have a higher likelihood of success in clinical development (14). This is driving efforts to increase the discovery of IPF-associated genes and to refine existing signals to identify the most likely causal variants for molecular and mechanistic evaluation.

For this study, we re-imputed previously published studies using a more recent imputation panel that enables measurement of three times as many variants as our previous studies and aggregated new datasets comprising clinically-curated IPF cases, and controls, to improve the quality and coverage of genotyping and to increase sample size for new gene discovery while offering more precise risk estimates. Our findings implicate new loci for IPF susceptibility.

## METHODS

Datasets

Seven independent case–control studies were analysed: Colorado (15), US (9), UK (10), IPF Job Exposures Study (IPF-JES) (16), Genentech (17), United States, United Kingdom, and Spain (UUS) (11) and Study of Clinical Efficacy of Antimicrobial Therapy Strategy Using Pragmatic Design in Idiopathic Pulmonary Fibrosis - University of California, Davis (CleanUP-UCD) (18) datasets (**Figure 1**). All studies included unrelated European-ancestry individuals, with cases of IPF diagnosed according to the American Thoracic Society and European Respiratory Society guidelines (19) (20) (21). Written informed consent and ethics approval were properly obtained for all studies, following the World Medical Association's Code of Ethics (Declaration of Helsinki) and approved by the appropriate institutional review

board or Research Ethics Committee. More information on each study can be found in the supplementary material.

Genotyping and quality controls

The Colorado, IPF-JES, UK, US, UUS, and CleanUP-UCD studies had been genotyped using SNP arrays (supplementary material). Quality control measures included filtering for low call rates, sex mismatches, heterozygosity, non-European genetic ancestry, relatedness, and ensuring no overlap between studies. These six studies were imputed using the TOPMed WGS reference panel (GRCh38) via the TOPMed Imputation Server (12). For the Genentech study, genotypes were obtained through WGS on the Illumina HiSeq X Ten platform with an average read depth of 30X. Related individuals and those with call rates below 10% were excluded from the analysis. Additional details are available in previously published studies (15) (9) (10) (16) (17) (11) (18) and the supplementary material.

Association analysis and meta-analysis

For six studies (Colorado, IPF-JES, UK, US, UUS, and CleanUP-UCD), genome-wide association analyses were conducted using logistic regression with PLINK v2 (22) adjusting each study for the first ten principal components to correct for population stratification. Variants with a poor imputation quality (r2<0.5) and minor allele count (MAC) ≤3 were removed. Association analysis for the Genentech study was performed using logistic regression with PLINK v1.9 (22) including sex, age, and five genetic-ancestry principal components as covariates.

The seven datasets were meta-analysed using inverse-variance weighted fixed effects meta-analysis with METAL (23), excluding variants that did not pass quality control in at least two studies. No minor allele frequency filter was applied. We estimated the genomic inflation factor using LDSC (24), and we applied a correction when this factor exceeded 1.10 in the meta-analysis. Data for chromosome X were available for four studies (UUS, Colorado, CleanUP-UCD, and UK) with GWAS and meta-analysis performed as described above (see supplementary material). For chromosome X, we applied a correction when classic lambda factor exceeded 1.15 in the meta-analysis.

Signal selection

Independent association signals were selected after performing conditional analysis with COJO-GCTA v1.90.2 (25), and based on a *p* value threshold of $<5.0\times10^{-8}$. Forest plots were prepared using the forestplot R v4.1.3 package and visually reviewed for outlying study signals. Signals were excluded if the significance (*p* value) of the association in any individual study was lower than the meta-analysis significance (i.e., if combining data across studies served to reduce the association rather than support it). We sought replication of new signals using an independent subset of the Global Biobank Meta-analysis Initiative (GBMI) IPF GWAS (26) comprising 6,257 cases defined using Electronic Healthcare Records (EHR), and 947,616 controls, all of European ancestry. We defined a replication threshold of $p<0.05$ with Bonferroni correction for the number of new signals tested. We annotated the sentinel variants with ANNOVAR v07.06.20 (27) and Ensembl Variant Effect Predictor (VEP) v.105, and we identified potentially causal variants using the Wakefield Bayes factor method to calculate the posterior inclusion probability (PIP) of each variant to define a 95% credible

set (28). We performed look-ups for all SNPs previously reported for association with IPF at genome-wide significance ($p<5x10^{-8}$).

*In silico* functional assessment

We performed a comprehensive functional impact assessment using empirical data from several integrated software tools and datasets for the sentinel SNPs and for each variant in the credible set with a PIP>0.1. We interrogated gene expression Quantitative Trait Locus (eQTLs) from the Genotype-Tissue Expression Project (GTEx) v8 (29) to identify whether IPF variants were associated with expression of any genes in lung, cultured fibroblasts or whole blood (see supplementary material for further details). We used coloc (30) R v4.1.3 package to investigate whether the same causal variant was driving both the genetic association with IPF risk and an association with gene expression changes and report shared signals with a colocalisation probability hypothesis 4 (H4) >0.7. We functionally annotated the highlighted regions based on active chromatin marks (i.e., DHS, eRNAs, ATACseq, CHIPseq, H3K27ac) (31) and searched for mouse orthologs of the human genes where knockout resulted in a measurable respiratory phenotype (https://www.mousephenotype.org/). We used Open Targets Genetics to identify additional supporting data for variant to gene mapping and for phenome-wide association study (PheWAS) analysis.

## RESULTS

We analysed 5,159 IPF cases and 27,459 controls of European ancestry, and a total of 25,290,839 autosomal variants (**Figure 2 & Table S1**). For X chromosome, association analyses were performed for 23,890 individuals (3,388 IPF cases and 20,502 controls), and 261,736 variants. There was no evidence of inflated test statistics (**Figure 2 & Figure S1**). After conditional analyses, we identified 37 autosomal independent significant signals ($p<5\times10^{-8}$), three of which were supported by multiple studies and had not been previously reported (**Table 1 & Figure 3**). All three signals were replicated in the independent GBMI IPF GWAS ($p<0.0125$ (**Table 1**). Only one signal (3p25.1) included a single SNP that accounted for more than 50% of the posterior probability of being causal (rs112271207 PIP=0.691) (**Table S2**). All signals overlapped open chromatin marks/regions (ATACseq or DHS) with evidence of enhancer activity (CHIPseq: H3K27ac) (**Table 2 & Table S3**). No signals reached the genome-wide significance threshold for the X chromosome analysis.

There were 32 previously reported IPF susceptibility signals and 31 were represented by variants in our dataset (**Table S4)** (an intronic signal, rs539683219, in the *PSKH1* gene could not be assessed and had previously been shown to be unique to East Asian ancestry populations (7)). All but three signals (in or near *DNAJB4*, *FSTL5* and *TPI1P2*) met a Bonferroni-corrected threshold of $p<1.56x10^{-3}$ in our study and all but seven signals reached genome-wide significance. These seven included five previously reported association signals (in or near *DNAJB4, FSTL5, TPI1P2, GPR157*, and *FKBP5*) whose initial discovery had been enabled by a multi-ancestry study design) (7) and two signals (at 10q25.1 and *STMN3*) previously reported in a meta-analysis of a subset of the studies included here (5).

The most statistically significant new signal (rs112271207, 3p25.1) was a common (MAF 6%) intergenic variant located 91.6 kb downstream of *LSM3* and 77.6 kb upstream of *LINC01267*. This SNP had previously been associated at near-genome-wide significance

with lung function ($p$=6.1x10$^{-8}$) with the allele associated with increased risk of IPF (T) associated with decreased lung function. The variant was not associated with gene expression in GTEx. The Open Targets Genetics V2G annotation prioritised a member of a family of sodium and chloride ion dependent transporters (*SLC6A6)* (132 kb downstream) as the most likely causal gene based on chromatin interaction data (**Table S3**).

The second most significant signal (rs9426886, 1q22) was located within an intron of the *TRIM46* gene, which is involved in the formation of parallel microtubule bundles. Although the sentinel variant was significantly associated with expression of several genes in cultured fibroblast, lung and whole blood (**Table S5**), the highest colocalisation probability was 0.616 for *MUC1* in whole blood. *MUC1* encodes a cell-surface glycoprotein. The IPF risk allele (T) associated with increased expression of *MUC1* in whole blood. This signal was also associated with urate levels ($p$=1.5x10$^{-101}$) and haematocrit measurement ($p$=4.7x10$^{-41}$) with the allele associated with increased IPF risk associated with increased haematocrit measurement, and decreased urate levels.

Finally, the signal at 12q23.1 (rs7957346) had the highest probability of colocalisation for expression of Netrin 4 (*NTN4*) in lung tissue (posterior probability = 0.649) and had previously been associated with lung function with the IPF risk increasing allele (A) associated with decreased lung function ($p$=7.8x10$^{-45}$). *NTN4* encodes Netrin 4; netrins are a family of proteins involved in cell and axon migration during development and link to angiogenesis and vascularization.

## DISCUSSION

Using our large clinically-curated IPF case-control resources, together with new imputation to the TOPMed Reference panel, we report novel associations that extend our previous genetic findings for IPF susceptibility and implicate new genes involved in its pathobiology.

The signal at 1q22 suggestively implicated *MUC1* which encodes a cell surface glycoprotein well known to be important in several lung diseases and infections (32). The MUC1 glycoprotein comprises three domains; an extracellular domain, a transmembrane domain, and a cytoplasmic tail. The extracellular domain becomes cleaved during alveolar epithelial damage and is commonly known as Krebs Von Den Lungen-6 (KL-6). KL-6 has been widely recognised as a potential prognostic biomarker of lung fibrosis as it is elevated in the serum and bronchioalveolar lavage fluid of patients with IPF. Studies have shown that KL-6 may also drive fibrosis through promotion of fibroblast activation and their differentiation to myofibroblasts. The MUC1 cytoplasmic tail (CT) has also been implicated in intracellular signalling in fibrosis (33). The biology of MUC1 is complex, however, early trials in cancer, where MUC1 is a well-studied tumour-associated antigen, suggest that monoclonal antibodies or inhibitors of MUC1 might represent a new therapeutic avenue for IPF (34). Although the highest colocalisation probability implicated *MUC1*, the evidence was from blood only, and we cannot not rule out the possibility that another nearby gene might be the true causal gene for this locus. Of particular note, thrombospondin-3 (*THBS3*) has a known role in cell-to-cell and cell-to-matrix interactions and has been implicated in cardiac fibrosis (35) and skin healing (36).

The signal on chromosome 12, overlapping the 3'-UTR of *SNRPF*, provided the strongest colocalisation with gene expression evidence in this study for *NTN4* which encodes the secreted protein Netrin-4, originally identified as a guide of axon migration but more recently established to have an important role in kidney and vascular development, and has also been implicated in lung morphogenesis (37). *NTN4* is relatively highly expressed in aberrant basaloid cells (IPFCellAtlas.com; (38)), a recently described patient-specific cell population (39)(40)(41). *NTN4* has also been implicated by GWAS of diffusing capacity for carbon monoxide (DLCO) in a population enriched for COPD patients (42). The role of Netrin-4 in angiogenesis (43)(44)(45) and the renewed interest in the relevance of endothelial cell and vascular abnormalities in IPF (46) make this an intriguing signal for further investigation. Previous genetic association studies of other respiratory traits have also implicated Small Nuclear Ribonucleoprotein Polypeptide F (*SNRPF*) (47)(48) and centrosome gene, *CCDC38* (49) at this locus.

The most significant new signal ($p$=2.23x$10^{-11}$) was located near to *LSM3*, encoding 'LSM3 Homolog, U6 Small Nuclear RNA And MRNA Degradation Associated', an sm-like protein with roles in RNA metabolism. The variant also mapped to *SLC6A6,* a taurine and beta-alanine transporter gene, implicated in Hypotaurinemic Retinal Degeneration and Cardiomyopathy (HTRDC) (50) and kidney fibrosis in diabetic knockout mice (51). Taurine may have a role in lung homeostasis and protection against oxidative stress (52)(53), a known driver of IPF pathogenesis.

All of the studies included in the genome-wide discovery stage of this analysis comprised cases that have been defined according to clinical criteria. Previous studies have noted an attenuation of effect sizes for IPF risk associated variants in studies where IPF case status has been defined using routine EHR (7)(54). Despite a smaller discovery sample size than previous studies that included IPF cases defined by EHR, we identified new signals that were replicated in an independent biobank EHR-based dataset (7). The use of the TOPMed Imputation reference panel enabled us to analyse over 25 million variants with a minor allele count greater than 3 and imputation quality greater than 0.5. This represented a more than 3-fold increase in variants analysed over our previous clinically-curated IPF GWAS (5). We included a larger number of variants than have been typically tested in GWAS; if we were to apply a more stringent genome-wide significance threshold of $p$<5x$10^{-9}$ (for example, as applied in other recent large GWAS (55), two of our three new signals would have met this p-value threshold in our discovery GWAS and the third would also have exceeded this threshold had we applied a two-stage design with meta-analysis of discovery GWAS and replication data. Nevertheless, the consistency of the signals across the participating studies (including replication) provides reassurance on the robust nature of the signals.

A key challenge in translating genetic association signals to mechanistic insight is robust variant-to-gene mapping to understand the functional, and ultimately clinical, consequence of the genetic perturbation (56). We applied several *in silico* approaches to map the associated signals to genes. However, none of our signals included either a gene coding region or splice elements or had a high probability of sharing a causal variant with an eQTL for nearby genes. Additional functional *in silico* approaches, as well as *in vitro* and *in vivo* studies are needed to map these new signals to genes, predict their function, and investigate their biological function and relevance to IPF.

Our study only included individuals of European ancestry which means that we cannot assess the generalisability of these association signals to other populations. We were unable to validate several previously reported signals that had been reported in multi-ancestry studies. There is a dearth of studies of clinically-curated IPF case-control datasets; these prior multi-ancestry studies were predominantly derived from large population biobanks. Future efforts should be focused on increasing availability of carefully phenotyped IPF datasets from non-European ancestry populations.

In conclusion, we have identified new signals of association with IPF risk that may provide further mechanistic insight into the pathobiology of IPF and support identification of potential new drug targets.

## FUNDING

This study was funded by a Medical Research Council Programme Grant (MR/V00235X/1) to RGJ and LVW by a Wellcome Trust PhD studentship for DC as part of the Wellcome Trust Genetic Epidemiology and Public Health Genomics Doctoral Training Programme (218505/Z/19/Z). LVW held a GlaxoSmithKline / Asthma + Lung UK Chair in Respiratory Research (C17-1). BGG is supported by Wellcome Trust grant 221680/Z/20/Z. JMO reports National Institutes of Health (NIH) National Heart, Lung, and Blood Institute grants R56HL158935 and K23HL138190. CF is supported by the Instituto de Salud Carlos III (PI20/00876, PI23/00980, CB06/06/1088, and PMP22/00083), co-financed by the European Regional Development Funds “A way of making Europe” from the EU, and by an agreement with Instituto Tecnológico y de Energías Renovables (ITER) to strengthen scientific and technological education, training, research, development and innovation in genomics, epidemiological surveillance based on massive sequencing, personalized medicine, and biotechnology (OA23/043). CJR was supported by Wellcome grant 201291/Z/16/Z. This work was partially supported by the National Institute for Health Research (NIHR) Leicester Biomedical Research Centre; the views expressed are those of the author(s) and not necessarily those of the National Health Service, the NIHR, or the Department of Health. This research used the ALICE and SPECTRE High Performance Computing Facility at the University of Leicester. For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

## COMPETING INTERESTS

LD is full-time employee of Genentech. AS hold stock options in Genentech. MN is full time employee of Genentech, Inc., a wholly owned subsidiary of Roche, and from which entity the data was sourced without fee/requirement. XRS, MM and BLY are full-time employees of Genentech and hold stock options in Roche. WAF is full-time employee of GSK. JMO reports personal fees from Boehringer Ingelheim, Genentech, United Therapeutics, AmMax Bio and Lupin Pharmaceuticals outside of the submitted work. DAS is a consultant for Vertex and the founder and chief scientific officer of Eleven P15, a company focused on the early detection and treatment of pulmonary fibrosis. BG has received in-kind research support from Galecto Biotech, and consultancy honoraria from GSK. AA reports personal fees from Boehringer Ingelheim, Genentech, Medscape, Abbvie, PatientMpower, Brainomix, PureTech, Gossamer Bio and Inogen outside of the submitted work. MES reports research

funding from Boehringer Ingelheim, consultancy for Boehringer Ingelheim, and participation on Adjudication Committee or Data Safety Board for Bristol Myers Squibb, Fibrinogen and Pliant. SH reports payment from Trevi therapeutics, Boehringer Ingelheim, and Chiesi. SJ reports payment from Boehringer Ingelheim. TMM reports personal fees from Boehringer Ingelheim, Roche/Genentech, Abbvie, Amgen, Astra Zeneca, Bayer, Bridge bio, Bristol-Myers Squibb, CSL Behring, Galapagos, Galecto, GSK, IQVIA, Merck, Pfizer, Pliant, PureTech, Sanofi, Trevi, Vicore; and participation in Fibrogen, United Therapeutics and Nerre. IN reports personal fees from Boehringer Ingelheim and Sanofi. HP reports personal fees from Boehringher Ingelheim Ltd and Trevi Therapeutics. PLM reports advisory board fees from Hoffman-La Roche, Boehringer Ingelheim, AstraZeneca, Trevi, Qureight, Endevour; and personal fees from United Therapeutics. MDT reports respiratory Drug Delivery 2023 speaker fee, unrelated to the manuscript. MM-M reports payment from Boehringer Ing, Ferrer, and Chiesi. FM reports consulting but no fees from AstraZeneca, Boehringer Ingelheim, Excalibur, GSK, Hoffmann-LaRoche, Lung Therapeutics, RS Biotherapeutics, Two XR; and travel expenses paid by Boehringer Ingelheim. CF declares funding Ministerio de Ciencia e Innovación, Instituto de Salud Carlos III and Instituto Tecnológico y de Energías Renovables; honoraria in educational events from Fundación Instituto Roche. RGJ reports honoraria from Chiesi, Roche, PatientMPower, AstraZeneca, GSK, Boehringer Ingelheim, and consulting fees from AdAlta, Abbvie, Arda Therapeutics, Bristol Myers Squibb, Veracyte, RedX, Pliant, Chiesi. LVW reports research funding from GlaxoSmithKline, Roche and Orion Pharma, and consultancy for GlaxoSmithKline and Galapagos, outside of the submitted work. The other authors declare no competing interests.

## ETHICS

This research was conducted with appropriate ethics approval. The PROFILE study (which provided samples for the UK and UUS studies) had institutional ethics approval at the University of Nottingham (NCT01134822 - ethics reference 10/H0402/2) and Royal Brompton and Harefield NHS Foundation Trust (NCT01110694 - ethics reference 10/H0720/12). UK samples were recruited across multiple sites with individual ethics approval (University of Edinburgh Research Ethics Committee [The Edinburgh Lung Fibrosis Molecular Endotyping (ELFMEN) Study NCT04016181] 17/ES/0075, NRES Committee South West - Southmead, Yorkshire and Humber Research Ethics Committee 08/H1304/54 and Nottingham Research Ethics Committee 09/H0403/59). Spanish samples were recruited under ethics approval by ethics committee from the Hospital Universitario N.S. de Candelaria (reference of the approval: PI-19/12). The UUS study also included individuals from clinical trials with ethics approval (ACE [NCT00957242] and PANTHER [NCT00650091]). For the UCSF cohort, sample and data collection were approved by the University of California San Francisco Committee on Human Research and all patients provided written informed consent. For the Vanderbilt cohort, the Institutional Review Boards from Vanderbilt University approved the study and all participants provided written informed consent before enrolment. For individuals recruited at the University of Chicago, consenting patients with IPF who were prospectively enrolled in the institutional review board-approved ILD registry (IRB#14163A) were included. Individuals recruited at the University of Pittsburgh Medical Centre had ethics approval from the University of Pittsburgh Human Research Protection Office (referenceSTUDY20030223: Genetic Polymorphisms in IPF). Individuals from the COMET (NCT01071707) and Lung Tissue Research Consortium (NCT02988388)

studies were also included in the Chicago study. All subjects in the Colorado study gave written informed consent as part of IRB-approved protocols for their recruitment at each site and the GWAS study was approved by the National Jewish Health IRB and Colorado Combined Institutional Review Boards (COMIRB). Subjects in the Genentech study provided written informed consent for whole-genome sequencing of their DNA. Ethical approval was provided as per the original clinical trials (INSPIRE [NCT00075998], RIFF [NCT01872689], CAPACITY [NCT00287729 and NCT00287716] and ASCEND [NCT01366209]). Individuals in the CleanUP-UCD study included individuals from clinical trials with ethics approval (NCT02759120). These samples were genotyped under University of Virginia ethics approval (IRB 20845). IPFJES involved human participants and was approved by East Midlands - Nottingham 1 Research Ethics Committee REC reference: 17/EM/0021IRAS project ID: 203355. Participants gave informed consent to participate in the study before taking part.

## DATA AVAILABILITY

Summary statistics (i.e., effect size estimates, standard errors, *p* values and basic variant information) for all variants included in the genome-wide meta-analysis can be accessed via https://github.com/genomicsITER/PFgenetics.

## AUTHORS CONTRIBUTION

LVW and RJA designed and supervised the study. DPC, TH-B, LD, BG-G, and OCL performed the analyses. AA, HLB, CleanUP-IPF Investigators of the Pulmonary Trials Cooperative, WAF, TEF, BG, IPH, SPH, MRH, NH, SRJ, NK, JML-S, S-FM, RJM, MIM, ADS, TMM, ABM, PLM, MM-M, VN, MN, JMO, HP, GS, IS, XRS, IDS, MES, MDT, MKBW, MCZ, YZ, FM, BLY, CJR, DAS, CF, IN, RGJ, and LVW participated in data collection. DPC, TH-B, and LVW wrote the first draft of the manuscript. All authors revised and approved the final version of the manuscript.

## REFERENCES


1. Undurraga Á. Fibrosis pulmonar idiopática. Revista médica clínica las Condes. 2015;26(3):292–301.

2. Lederer DJ, Martinez FJ. Idiopathic pulmonary fibrosis. N Engl J Med. 2018;378(19):1811–23.

3. Gupta RS, Koteci A, Morgan A, George PM, Quint JK. Incidence and prevalence of interstitial lung diseases worldwide: a systematic literature review. BMJ Open Respiratory Research. 2023;10(1):e001291.

4. Richeldi L, Azuma A, Cottin V, Kreuter M, Maher TM, Martinez FJ, et al. Nerandomilast in patients with idiopathic pulmonary fibrosis. N Engl J Med. 2025;392(22):2193–202.

5. Allen RJ, Stockwell A, Oldham JM, Guillen-Guio B, Schwartz DA, Maher TM, et al. Genome-wide association study across five cohorts identifies five novel loci associated with idiopathic pulmonary fibrosis. Thorax. 2022;77(8):829–33.

6. Peljto AL, Blumhagen RZ, Walts AD, Cardwell J, Powers J, Corte TJ, et al. Idiopathic pulmonary fibrosis is associated with common genetic variants and limited rare variants. American journal of respiratory and critical care medicine. 2023;207(9):1194–202.

7. Partanen JJ, Häppölä P, Zhou W, Lehisto AA, Ainola M, Sutinen E, et al. Leveraging global multi-ancestry meta-analysis in the study of idiopathic pulmonary fibrosis genetics. Cell genomics. 2022;2(10).

8. Mushiroda T, Wattanapokayakit S, Takahashi A, Nukiwa T, Kudoh S, Ogura T, et al. A genome-wide association study identifies an association of a common variant in TERT with susceptibility to idiopathic pulmonary fibrosis. J Med Genet. 2008;45(10):654–6.

9. Noth I, Zhang Y, Ma S, Flores C, Barber M, Huang Y, et al. Genetic variants associated with idiopathic pulmonary fibrosis susceptibility and mortality: a genome-wide association study. The Lancet respiratory medicine. 2013;1(4):309–17.

10. Allen RJ, Porte J, Braybrooke R, Flores C, Fingerlin TE, Oldham JM, et al. Genetic variants associated with susceptibility to idiopathic pulmonary fibrosis in people of European ancestry: a genome-wide association study. The Lancet respiratory medicine. 2017;5(11):869–80.

11. Allen RJ, Guillen-Guio B, Oldham JM, Ma S, Dressen A, Paynton ML, et al. Genome-wide association study of susceptibility to idiopathic pulmonary fibrosis. American journal of respiratory and critical care medicine. 2020;201(5):564–74.

12. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature. 2021;590(7845):290–9.

13. Seibold MA, Wise AL, Speer MC, Steele MP, Brown KK, Loyd JE, et al. A common MUC5B promoter polymorphism and pulmonary fibrosis. N Engl J Med. 2011;364(16):1503–12.

14. Minikel EV, Painter JL, Dong CC, Nelson MR. Refining the impact of genetic evidence on clinical success. Nature. 2024:1–6.

15. Fingerlin TE, Murphy E, Zhang W, Peljto AL, Brown KK, Steele MP, et al. Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. Nat Genet. 2013;45(6):613–20.

16. Reynolds CJ, Sisodia R, Barber C, Moffatt M, Minelli C, De Matteis S, et al. What role for asbestos in idiopathic pulmonary fibrosis? Findings from the IPF job exposures case–control study. Occup Environ Med. 2023;80(2):97–103.

17. Donoghue LJ, Stockwell AD, Neighbors M, Sheng RX, Prabhakaran R, Wolters PJ, et al. Identification of a Genetic Susceptibility Locus for Idiopathic Pulmonary Fibrosis in the 16p Subtelomere Using Whole-Genome Sequencing. American Journal of Respiratory and Critical Care Medicine. 2023;207(7):941–4.

18. Martinez FJ, Yow E, Flaherty KR, Snyder LD, Durheim MT, Wisniewski SR, et al. Effect of antimicrobial therapy on respiratory hospitalization or death in adults with idiopathic pulmonary fibrosis: the CleanUP-IPF randomized clinical trial. JAMA. 2021;325(18):1841–51.

19. Raghu G, Remy-Jardin M, Myers JL, Richeldi L, Ryerson CJ, Lederer DJ, et al. Diagnosis of idiopathic pulmonary fibrosis. An official ATS/ERS/JRS/ALAT clinical practice guideline. American journal of respiratory and critical care medicine. 2018;198(5):e44–68.

20. Society ER, American Thoracic Society. American Thoracic Society. American journal of respiratory and critical care medicine. 2002;165(2):277–304.

21. Raghu G, Collard HR, Egan JJ, Martinez FJ, Behr J, Brown KK, et al. An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. American journal of respiratory and critical care medicine. 2011;183(6):788–824.

22. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4(1):s13742–8.

23. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26(17):2190–1.

24. Bulik-Sullivan BK, Loh P, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics Consortium, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015;47(3):291–5.

25. Yang J, Ferreira T, Morris AP, Medland SE, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet. 2012;44(4):369–75.

26. Kurilshikov A, Medina-Gomez C, Bacigalupe R, Radjabzadeh D, Wang J, Demirkan A, et al. Large-scale association analyses identify host factors influencing human gut microbiome composition. Nat Genet. 2021;53(2):156–65.

27. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.

28. Wakefield J. Bayes factors for genome-wide association studies: comparison with P-values. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society. 2009;33(1):79–86.

29. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The genotype-tissue expression (GTEx) project. Nat Genet. 2013;45(6):580–5.

30. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS genetics. 2014;10(5):e1004383.

31. Biddie SC, Weykopf G, Hird EF, Friman ET, Bickmore WA. DNA-binding factor footprints and enhancer RNAs identify functional non-coding genetic variants. Genome Biol. 2024;25(1):208.

32. Ballester B, Milara J, Cortijo J. The role of mucin 1 in respiratory diseases. European Respiratory Review. 2021;30(159).

33. Milara J, Ballester B, Montero P, Escriva J, Artigues E, Alós M, et al. MUC1 intracellular bioactivation mediates lung fibrosis. Thorax. 2020;75(2):132–42.

34. Ballester B, Milara J, Cortijo J. Mucins as a new frontier in pulmonary fibrosis. Journal of Clinical Medicine. 2019;8(9):1447.

35. Ma Y, Han X, Zan S, Liu H, Zhou X, Zhao D, et al. Microrna363-5p targets thrombospondin3 to regulate pathological cardiac remodeling. Mol Cell Biochem. 2024:1–14.

36. Bergmeier V, Etich J, Pitzler L, Frie C, Koch M, Fischer M, et al. Identification of a myofibroblast-specific expression signature in skin wounds. Matrix Biology. 2018;65:59–74.

37. Liu Y, Stein E, Oliver T, Li Y, Brunken WJ, Koch M, et al. Novel role for Netrins in regulating epithelial behavior during lung branching morphogenesis. Current biology. 2004;14(10):897–905.

38. Neumark N, Cosme Jr C, Rose K, Kaminski N. The idiopathic pulmonary fibrosis cell atlas. American Journal of Physiology-Lung Cellular and Molecular Physiology. 2020;319(6):L887–92.

39. Adams TS, Schupp JC, Poli S, Ayaub EA, Neumark N, Ahangari F, et al. Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. Science advances. 2020;6(28):eaba1983.

40. Reyfman PA, Walter JM, Joshi N, Anekalla KR, McQuattie-Pimentel AC, Chiu S, et al. Single-cell transcriptomic analysis of human lung provides insights into the pathobiology of pulmonary fibrosis. American journal of respiratory and critical care medicine. 2019;199(12):1517–36.

41. Morse C, Tabib T, Sembrat J, Buschur KL, Bittar HT, Valenzi E, et al. Proliferating SPP1/MERTK-expressing macrophages in idiopathic pulmonary fibrosis. European Respiratory Journal. 2019;54(2).

42. Sakornsakolpat P, McCormack M, Bakke P, Gulsvik A, Make BJ, Crapo JD, et al. Genome-wide association analysis of single-breath DLCO. American journal of respiratory cell and molecular biology. 2019;60(5):523–31.

43. Lejmi E, Leconte L, Pédron-Mazoyer S, Ropert S, Raoul W, Lavalette S, et al. Netrin-4 inhibits angiogenesis via binding to neogenin and recruitment of Unc5B. Proceedings of the National Academy of Sciences. 2008;105(34):12491–6.

44. Wilson BD, Ii M, Park KW, Suli A, Sorensen LK, Larrieu-Lahargue F, et al. Netrins promote developmental and therapeutic angiogenesis. Science. 2006;313(5787):640–4.

45. Nacht M, Martin TBS, Byrne A, Klinger KW, Teicher BA, Madden SL, et al. Netrin-4 regulates angiogenic responses and tumor cell growth. Exp Cell Res. 2009;315(5):784–94.

46. May J, Mitchell JA, Jenkins RG. Beyond epithelial damage: vascular and endothelial contributions to idiopathic pulmonary fibrosis. J Clin Invest. 2023;133(18):e172058.

47. Manichaikul A, Hoffman EA, Smolonska J, Gao W, Cho MH, Baumhauer H, et al. Genome-wide study of percent emphysema on computed tomography in the general population. The Multi-Ethnic Study of Atherosclerosis Lung/SNP Health Association Resource Study. American journal of respiratory and critical care medicine. 2014;189(4):408–18.

48. Peat G, Jones W, Nuhn M, Marugán JC, Newell W, Dunham I, et al. The open targets post-GWAS analysis pipeline. Bioinformatics. 2020;36(9):2936–7.

49. Wain LV, Sayers I, Soler Artigas M, Portelli MA, Zeggini E, Obeidat M, et al. Whole exome re-sequencing implicates CCDC38 and cilia structure and function in resistance to smoking related airflow obstruction. PLoS genetics. 2014;10(5):e1004314.

50. Ansar M, Ranza E, Shetty M, Paracha SA, Azam M, Kern I, et al. Taurine treatment of retinal degeneration and cardiomyopathy in a consanguineous family with SLC6A6 taurine transporter deficiency. Hum Mol Genet. 2020;29(4):618–23.

51. Han X, Patters AB, Ito T, Azuma J, Schaffer SW, Chesney RW. Knockout of the TauT gene predisposes C57BL/6 mice to streptozotocin-induced diabetic nephropathy. PLoS One. 2015;10(1):e0117718.

52. Baliou S, Adamaki M, Ioannou P, Pappa A, Panayiotidis MI, Spandidos DA, et al. Protective role of taurine against oxidative stress. Molecular medicine reports. 2021;24(2):605.

53. Gurujeyalakshmi G, Wang Y, Giri SN. Taurine and niacin block lung injury and fibrosis by down-regulating bleomycin-induced activation of transcription nuclear factor-κB in mice. J Pharmacol Exp Ther. 2000;293(1):82–90.

54. Leavy OC, Allen RJ, Kraven LM, Morgan AD, Tobin MD, Quint JK, et al. No title. P32 Using genetic information to define idiopathic pulmonary fibrosis in UK biobank. 2022.

55. Shrine N, Izquierdo AG, Chen J, Packer R, Hall RJ, Guyatt AL, et al. Multi-ancestry genome-wide association analyses improve resolution of genes and pathways influencing lung function and chronic obstructive pulmonary disease risk. Nat Genet. 2023;55(3):410–22.

56. Lichou F, Trynka G. Functional studies of GWAS variants are gaining momentum. Nature communications. 2020;11(1):6283.

57. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature. 2012;489(7414):75–82.

58. Javierre BM, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, et al. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. Cell. 2016;167(5):1369,1384. e19.

## FIGURE LEGENDS

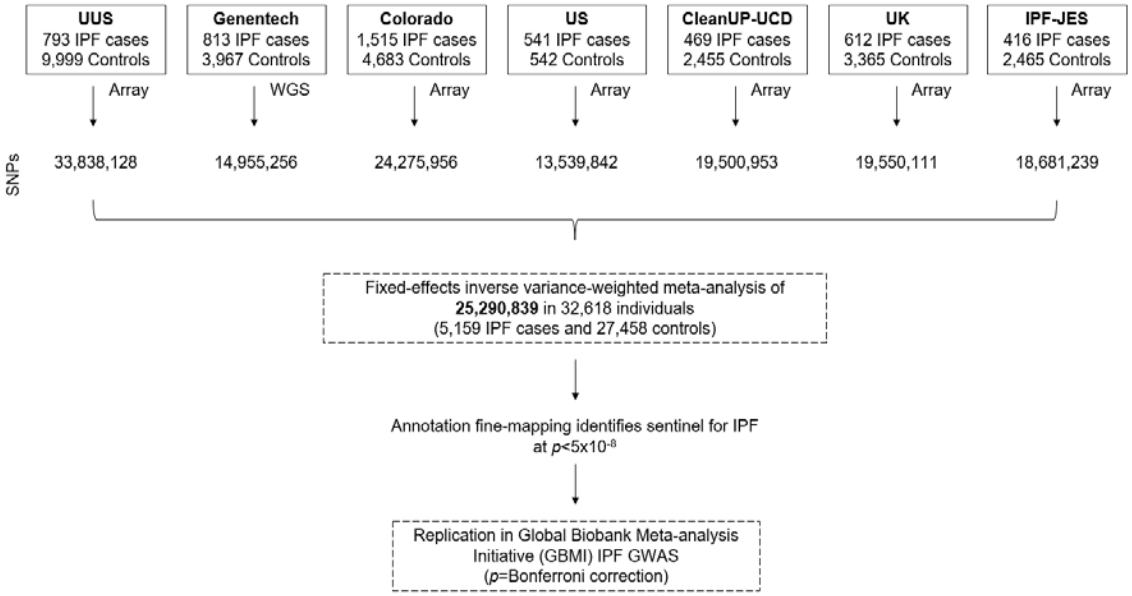


**Figure 1. Study overview.** SNPs: Single Nucleotide Polymorphisms. WGS: Whole-genome sequencing data.

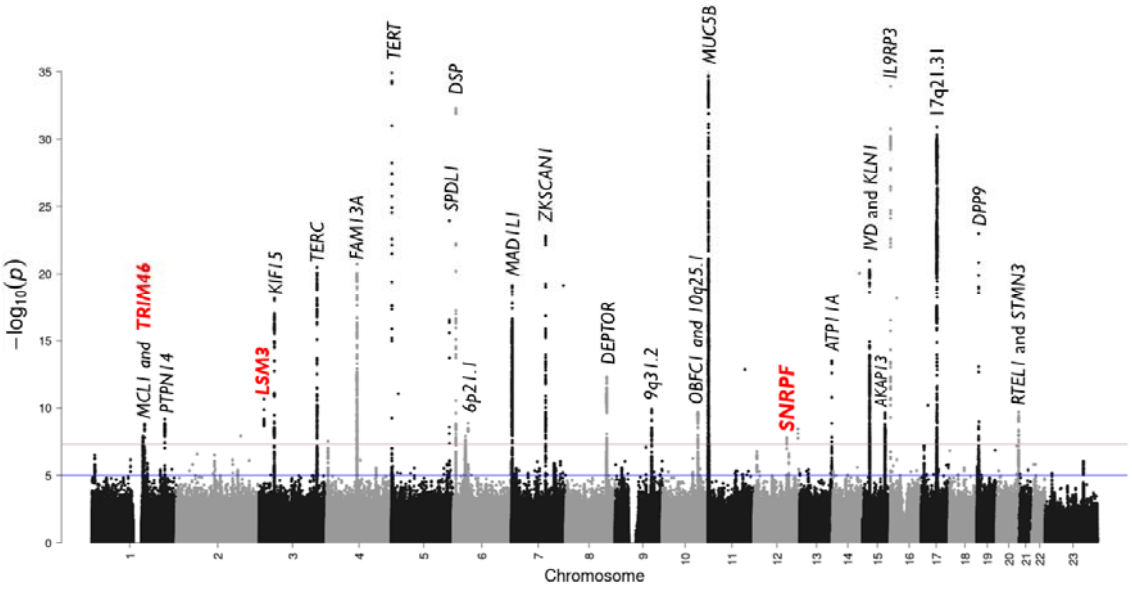


**Figure 2. Manhattan plot of meta-analysis results.** The y-axis shows the transformed p-values (–log10[p-value]) while the x-axis represents chromosome positions. The horizontal red line corresponds to the genome-wide threshold ($p$=5.0×10$^{-8}$) and the blue line shows the suggestive significance threshold ($p$=5.0×10$^{-5}$). The genomic inflation factor of the meta-analysis results did not show major deviations from the null hypothesis of no association. Newly discovered nearest annotated genes are highlighted in red and bold. The plot has been truncated at $p$=1.0x10$^{-35}$. The most significant previously reported *MUC5B* signal (rs35705950) has a p<5.0x10$^{-121}$.

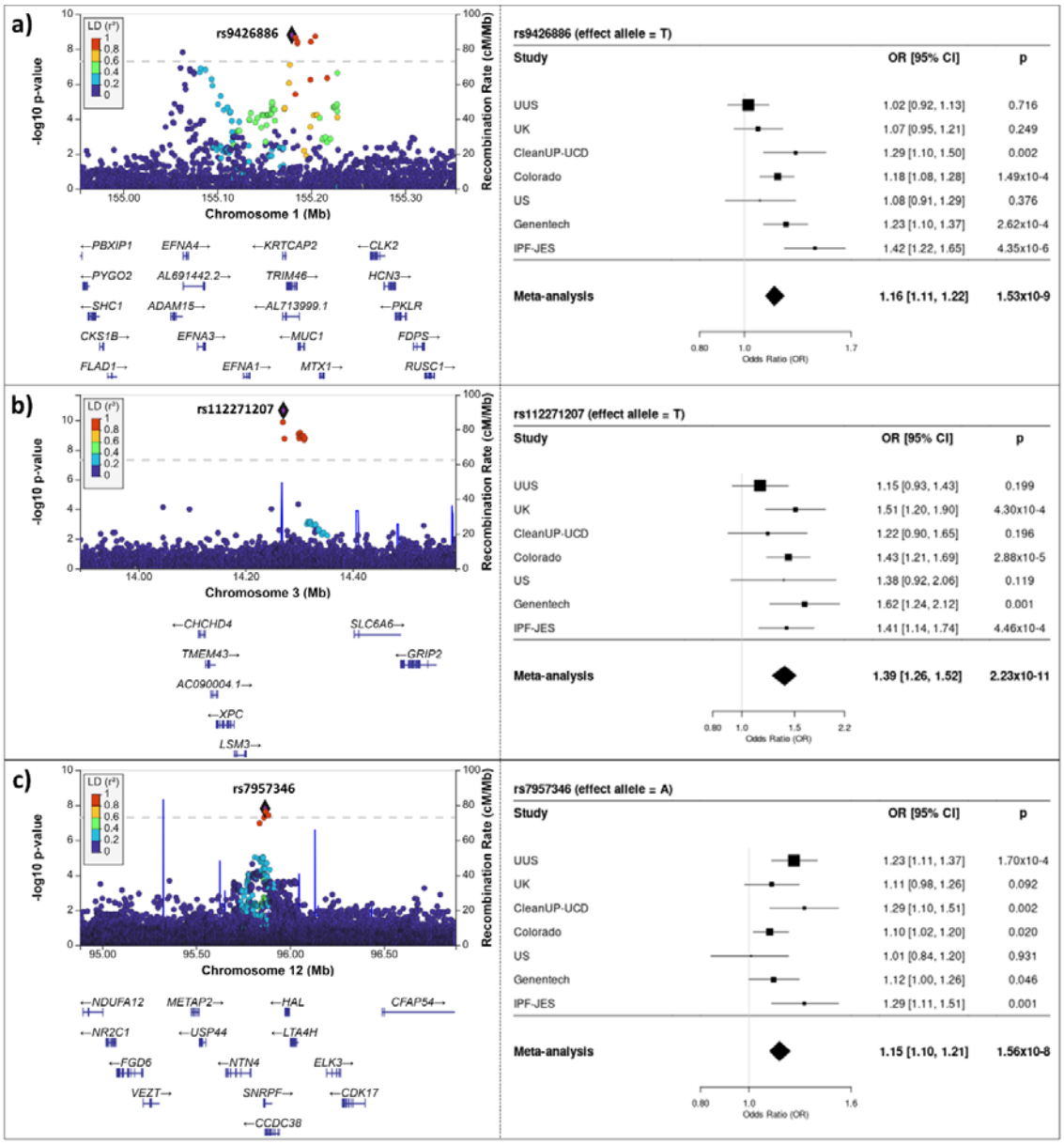


**Figure 3. Region plots and forest plots of the three signals of interest:** a) 1q22 region; b) 3p25.1 region; and c) 12q23.1 region. In region plots, the y-axis shows the transformed p-values (-log10[p-value]), while the x-axis represents chromosomal positions (GRCh38). The genome-wide significance threshold ($p$ =5.0x10$^{-8}$) is indicated by the horizontal dashed line. Linkage disequilibrium values (r2) are presented according to the LD colour scheme of the upper left legend. Plots were generated using LocusZoom (http://locuszoom.org/).

# TABLES

**Table 1.** New genome-wide significant association signals

| | | | | This study | | | | Replication | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Region (CHR:BP)** | **SNP** | **Location (nearest gene)** | **A1/A2** | **A1 Freq.** | **OR (CI95%)** | **P** | **N$^{o}$ SNPs in credible set (largest PIP)** | **A1 Freq.** | **OR (CI95%)** | **P** |
| 1q22 (1:155179278) | rs9426886 | Intronic (*TRIM46)* | T/A | 0.568 | 1.16 (1.11-1.22) | $1.53x10^{-9}$ | 7 (0.208) | 0.555 | 1.05 (1.01-1.09) | $8.02x10^{-3}$ |
| 3p25.1 (3:14270502) | rs112271207 | Intergenic (*LSM3\LINC01267)* | T/C | 0.060 | 1.39 (1.26-1.52) | $2.23x10^{-11}$ | 9 (0.691) | 0.069 | 1.11 (1.03-1.20) | $5.64x10^{-3}$ |
| 12q23.1 (12:95866696) | rs7957346 | 3' UTR (*SNRPF)* | A/C | 0.582 | 1.15 (1.10-1.21) | $1.56x10^{-8}$ | 5 (0.376) | 0.588 | 1.08 (1.04-1.12) | $4.69x10^{-5}$ |

A1: effect allele; A2: non-effect allele; Freq: Frequency; OR: Odds Ratio; PIP: posterior inclusion probability; Coordinates in GRCh38; CI: Confidence Interval.

**Table 2**. Summary of variant to gene mapping and functional information

| **Signal*** | **Location (nearest gene)** | **Gene implicated by eQTL (GTEx: tissue, highest posterior probability of shared causal variant)** | **Other gene evidence[#]** | **PheWAS (aligned to IPF risk allele)[#]** | **Open Chromatin**** |
|---|---|---|---|---|---|
| 1q22 (rs9426886) | Intronic (*TRIM46)* | *MUC1* (whole blood, 0.616) | *THBS3*, *ADAM15* (sQTL[1]; DHS-promoter correlation[2]) | increased haematocrit, decreased urate | DHS; H3K27ac |
| 3p25.1 (rs112271207) | Intergenic (*LSM3\LINC01267)* | NA | *SLC6A6* (PCHiC[3]) | decreased lung function | ATACseq; H3K27ac |
| 12q23.1 (rs7957346) | 3' UTR (*SNRPF)* | *NTN4* (lung, 0.649) | *NTN4* (sQTL[2]) | decreased lung function | ATACseq; H3K27ac |

[#]Open Targets Genetics: 1. Splice QTL (source: GTEx); 2. DHS-promoter capture (source: Thurman et al 2012 (57)); 3. Promotor Capture HiC in blood-derived endothelial precursors, macrophages, megakaryocytes, naive B cells and neutrophils (from Open Targets Genetics, source: Javierre et al 2016 (58).
**Biddie et al 2023 (31). *Minimum study-level imputation quality Rsq for each variant: rs9426886:0.936, rs112271207:0.929, rs7957346:0.991.
See Table S3 for full results.