EUROPEAN RESPIRATORY *journal*

FLAGSHIP SCIENTIFIC JOURNAL OF ERS

Early View

Original research article

Genetically increased circulating FUT3 level leads to reduced risk of Idiopathic Pulmonary Fibrosis: a Mendelian Randomisation Study

Tomoko Nakanishi, Agustin Cerani, Vincenzo Forgetta, Sirui Zhou, Richard J. Allen, Olivia C. Leavy, Masaru Koido, Deborah Assayag, R. Gisli Jenkins, Louise V. Wain, Ivana V. Yang, G. Mark Lathrop, Paul J. Wolters, David A. Schwartz, J. Brent Richards

Please cite this article as: Nakanishi T, Cerani A, Forgetta V, *et al*. Genetically increased circulating FUT3 level leads to reduced risk of Idiopathic Pulmonary Fibrosis: a Mendelian Randomisation Study. *Eur Respir J* 2021; in press (https://doi.org/10.1183/13993003.03979-2020).

This manuscript has recently been accepted for publication in the *European Respiratory Journal*. It is published here in its accepted form prior to copyediting and typesetting by our production team. After these production processes are complete and the authors have approved the resulting proofs, the article will move to the latest issue of the ERJ online.

Genetically increased circulating FUT3 level leads to reduced risk of Idiopathic Pulmonary Fibrosis: a Mendelian Randomization Study

Tomoko Nakanishi[1–4], Agustin Cerani[2,5], Vincenzo Forgetta[2], Sirui Zhou[2,5], Richard J. Allen[6], Olivia C. Leavy[6], Masaru Koido[7], Deborah Assayag[8,9], R. Gisli Jenkins[10,11], Louise V. Wain[6,12], Ivana V. Yang[13,14], G. Mark Lathrop[15], Paul J. Wolters[16], David A. Schwartz[14,17], J. Brent Richards[1,2,5,18,19]

1. Department of Human Genetics, McGill University, Montréal, Québec, Canada.
2. Centre for Clinical Epidemiology, Department of Medicine, Lady Davis Institute, Jewish General Hospital, McGill University, Montréal, Québec, Canada.
3. Kyoto-McGill International Collaborative School in Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan.
4. Research Fellow, Japan Society for the Promotion of Science, Tokyo, Japan.
5. Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada.
6. Department of Health Sciences, University of Leicester, Leicester, United Kingdom.
7. Department of Cancer Biology, Institute of Medical Science, The University of Tokyo, Tokyo, Japan.
8. Department of Medicine, Faculty of Medicine, McGill University, Montréal, Québec, Canada.
9. Translational Research in Respiratory Diseases, Research Institute McGill University Health Centre, Montréal, Québec, Canada.
10. National Institute for Health Research, Nottingham Biomedical Research Centre Nottingham University Hospitals NHS Trust, Nottingham, United Kingdom.
11. Division of Respiratory Medicine, University of Nottingham, Nottingham, United Kingdom.
12. National Institute for Health Research, Leicester Respiratory Biomedical Research Centre, Glenfield Hospital, Leicester, United Kingdom.
13. Center for Genes, Environment and Health and Department of Medicine, National Jewish Health, Denver, Colorado, USA.
14. Department of Medicine, University of Colorado Denver, School of Medicine, Aurora, Colorado, USA.
15. McGill Genome Centre and Department of Human Genetics, McGill University, Montréal, Québec, Canada.
16. Department of Medicine, School of Medicine, University of California, San Francisco, California, USA.
17. Department of Immunology, University of Colorado Denver, School of Medicine, Aurora, Colorado, USA.

18. Division of Endocrinology, Departments of Medicine, Jewish General Hospital, McGill University, Montréal, Québec, Canada.
19. Department of Twin Research, King's College London, London, United Kingdom.

Corresponding Author:
J. Brent Richards, MD, MSc
Professor of Medicine, McGill University
Senior Lecturer, King's College London (Honorary)
Contact:
Pavillon H-413, Jewish General Hospital
3755 Cote Ste Catherine
Montréal, Québec, Canada, H3T 1E2
T: +1 514 340 8222 x24362 F: +1 514 340 7529
E: brent.richards@mcgill.ca www.mcgill.ca/genepi

Take home message:
Undertaking an efficient scan of 834 circulating proteins for their role in IPF risk using Mendelian randomization (MR), we found that those with genetically increased circulating FUT3 levels had lower risk of developing IPF.

Author contributions:
Conception and design: TN and JBR. Data analyses: TN and OCL. Manuscript writing: TN and JBR. Data acquisition: RJA, RGJ, LVW, PJW and DAS. Interpretation of data: TN, AC, VF, SZ, DA, RJA, OCL, KM, RGJ, LVW, IVY, GML, PJW, DAS and JBR. Intellectual contribution to the manuscript: TN, AC, VF, SZ, DA, RJA, OCL, KM, RGJ, LVW, IVY, GML, PJW, DAS and JBR. All authors were involved in preparation of the further draft of the manuscript and revising it critically for content. All authors gave final approval of the version to be published. TN and JBR are the guarantors. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding:
TN is supported by Research Fellowships of Japan Society for the Promotion of Science (JSPS) for Young Scientists and JSPS Overseas Challenge Program for Young Researchers. AC is supported by the Fonds de Recherche Québec Santé (FRQS) and Canadian Institutes of Health Research (CIHR) Doctoral awards and Queen Elizabeth Scholar. SZ is supported by a CIHR postdoctoral fellowship. The Richards research group is supported by CIHR(365825; 409511), the Lady Davis Institute of the Jewish General Hospital, the Canadian Foundation for Innovation, the NIH Foundation, Cancer Research UK and FRQS. JBR is supported by a FRQS Clinical

Research Scholarship. Support from Calcul Québec and Compute Canada is acknowledged. TwinsUK is funded by the Welcome Trust, Medical Research Council, European Union, the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London. RJA is supported by an Action for Pulmonary Fibrosis Mike Bray fellowship. LVW holds a GSK/British Lung Foundation Chair in Respiratory Research. The research 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 (NHS), the NIHR or the Department of Health. PJW received funding from the Nina Ireland Program for Lung Health. DAS is supported by National Institutes of Health, National Heart, Lung, and Blood Institute (NIH-NHLBI: R38-HL143511, T32-HL007085, UG3/UH3-HL151865, R01-HL149836, P01-HL0928701, UH2/3-HL123442, X01-HL134585, R25-ES025476), and DOD Focused Program Grant (Number: W81XWH-17-1-0597). These funding agencies had no role in the design, implementation or interpretation of this study.

Abstract

Idiopathic pulmonary fibrosis (IPF) is a progressive, fatal fibrotic interstitial lung disease. Few circulating biomarkers have been identified to have causal effects on IPF.

To identify candidate IPF-influencing circulating proteins, we undertook an efficient screen of circulating proteins by applying a two-sample Mendelian randomization (MR) approach with existing publicly available data. For instruments we used genetic determinants of circulating proteins which reside *cis* to the encoded gene (*cis*-SNPs), identified by two genome-wide association studies (GWASs) in European individuals (3,301 and 3,200 subjects). We then applied MR methods to test if the levels of these circulating proteins influenced IPF susceptibility in the largest IPF GWAS (2,668 cases and 8,591 controls). We validated the MR results using colocalization analyses to ensure that both the circulating proteins and IPF shared a common genetic signal.

MR analyses of 834 proteins found that a one SD increase in circulating FUT3 and FUT5 was associated with a reduced risk of IPF (OR: 0.81, 95%CI: 0.74–0.88, $p=6.3x10^{-7}$, and OR: 0.76, 95%CI: 0.68–0.86, $p=1.1x10^{-5}$). Sensitivity analyses including multiple-*cis* SNPs provided similar estimates both for FUT3 (inverse variance weighted [IVW] OR: 0.84, 95%CI: 0.78–0.91, $p=9.8x10^{-6}$, MR-Egger OR: 0.69, 95%CI: 0.50 - 0.97, p=0.03) and FUT5 (IVW OR: 0.84, 95%CI: 0.77–0.92, $p=1.4x10^{-4}$, MR-Egger OR: 0.59, 95%CI: 0.38 - 0.90, p=0.01) FUT3 and FUT5 signals colocalized with IPF signals, with posterior probabilities of a shared genetic signal of 99.9% and 97.7%. Further transcriptomic investigations supported the protective effects of *FUT3* for IPF.

An efficient MR scan of 834 circulating proteins provided evidence that genetically increased circulating FUT3 level is associated with reduced risk of IPF.

Key words: “Idiopathic Pulmonary Fibrosis”, “3-galactosyl-N-acetylglucosaminide 4-alpha-L-fucosyltransferase”, “fucosyltransferase 5”, “Mendelian Randomization Analysis”, “Proteome”

Introduction

Idiopathic pulmonary fibrosis (IPF) is a progressive, fatal fibrotic interstitial lung disease (ILD) that affects adults, leading to decreased lung compliance, disrupted gas exchange and resultant respiratory failure[1]. The median survival time from diagnosis is 3 to 5 years, which is worse than the prognosis of most types of cancers[2]. Early detection or prevention of IPF is important as the currently available therapies are anti-fibrotic agents that have been shown to slow disease progression[3][4]. At the current time the only way to detect early disease is through high resolution CT scanning which reveals interstitial lung abnormalities in up to 10% of the population aged over 60 years in whom only a small minority progress to develop IPF[5]. Therefore, a serum biomarker that can predict or refine disease risk through a causal relationship is urgently required.

Although several serum biomarkers for IPF have been identified[6][7][8][9], these biomarkers still lack strong evidence of disease causality and are more useful at defining prognosis once IPF has occurred. Causal inference in IPF through traditional observational studies is challenging due to potential confounding and reverse causation that can bias estimate of the effect of biomarkers on IPF. For example, smoking, a known risk factor for IPF is confounded by its association with many other lifestyle choices. Similarly, IPF itself may influence the level of the biomarker—a phenomenon known as reverse causation. This last source of bias is

particularly difficult to rule out since the timing of IPF onset is most often unknown.

Despite these challenges, identifying IPF-influencing circulating proteins is helpful as such markers could serve as both drug targets to decrease susceptibility and non-invasive biomarkers of disease risk. One way to estimate the causality of circulating biomarkers is using Mendelian randomization (MR), which uses germline genetic variants as instrumental variables to assess the role of risk factors in disease susceptibility. Since genetic variants are randomly assigned at conception, this process of randomization largely breaks association with most confounding factors. Further, since germline genetic variants are always assigned prior to disease onset, reverse causation can be avoided. A further advantage of MR studies is that they can provide an assessment of a lifetime of risk factor exposure assuming the effect of the genetic variant on the risk factor is stable throughout an individual's life[10].

The goal of this study was therefore to identify circulating proteins which influence the risk for IPF by applying a MR design which efficiently screened hundreds of proteins. Bayesian colocalization analyses were undertaken to ensure that candidate circulating proteins and IPF shared a common etiological genetic signal and that the MR results were not biased by linkage disequilibrium (LD). Candidate IPF-influencing proteins identified through MR and colocalization analyses were further evaluated via literature and genetic-phenotype database searches, and transcriptomic investigations. The results from these experiments could provide a better understanding of the etiology of IPF, and could potentially identify targets for future therapies.

Materials and methods

Study design and data sources

We applied a two-sample MR design to identify circulating proteins associated with risk of IPF. For this, summary data was obtained from the largest IPF genome-wide association study (GWAS) to date in individuals of European ancestry[11] and from the two protein quantitative trait loci (pQTL) GWASs, Sun *et al*.[12] and Emilsson *et al*.[13]. Detailed methods of protein assays are described in each study[12][13]. See **Figure 1** for a schema of our study design.

Ethical approval

No separate ethical approval was required due to the use of publicly available data.

Mendelian randomization

MR relies upon three major assumptions[14]. First, the genetic variants must reliably associate with the exposure. With the advent of large-scale modern GWASs, genetic variants associating with exposure can be identified in large datasets[15]. Second, the genetic variants must not be associated with confounders of the exposure-outcome relationship. A potential violation of this assumption can occur due to confounding by LD and/or population ancestry[16]. Lastly, genetic variants must not affect the outcome, except through the exposure of interest (referred to as a lack of horizontal pleiotropy)[17].

Large-scale GWAS for circulating proteins[12][13] have often found that the genetic determinants of circulating proteins reside *cis* (in close proximity) to the encoding genes. The use of *cis*-acting SNPs for MR reduces potential horizontal pleiotropy and increases the validity of MR assumptions, because a *cis*-SNP strongly associated with the protein is likely to directly influence the gene's transcription and consequently the circulating protein level. We selected independent ($r^2 \leq 0.001$) *cis*-pQTL SNPs which are significantly associated with circulating proteins ($p<5x10^{-8}$) from two pQTL GWASs[12][13]. More details are described in the online supplement.

Statistical analysis

We performed MR using "TwoSampleMR" R package[18]. For proteins with a single *cis*-SNP, the Wald estimator ($\beta_{IPF}/\beta_{protein}$) was used to estimate the effect of the protein on IPF risk. Where multiple SNPs were available, our primary analyses used an inverse variance weighted (IVW) estimator[19]. Benjamini-Hochberg correction was applied to adjust for the multiple proteins tested, which is likely to be conservative because some protein levels are partially correlated with each other. (False discovery rate [FDR] of 0.05 with 507 multiple testing for Sun *et al*. and 733 multiple testing for Emilsson *et al*.)

Colocalization analysis

Candidate IPF-influencing proteins supported by MR were evaluated via colocalization analyses using the "coloc" R package[20] and eCAVIAR[21] for the proteins in Sun *et al.*[12], which

provided genome-wide summary statistics for each protein. Colocalization analysis is a way to estimate the posterior probability of whether the same genetic variants are responsible for the two GWAS signals (in this case the protein level and IPF) or they are distinct causal variants that are just in LD with each other. The detailed methods are in the online supplement. LocusZoom[22] plots were created to visualize these colocalizations.

Sensitivity analysis

Sensitivity analyses were performed for proteins with support from MR and colocalization analyses. Multiple *cis*-SNPs in weak LD (r^2<0.6) with the leading *cis*-SNPs for candidate proteins were included in IVW and MR-Egger analyses that considered correlated variants using the "MendelianRandomization" R package[23][24], because consistency of estimates could strengthen the hypothesized effects. MR-Egger allows for a y-intercept term using a random-effects model. An intercept different from zero indicates directional horizontal pleiotropy, suggestive of a violation of the third MR assumption. The detailed methods are in the online supplement. Bidirectional MR was also conducted to test whether IPF had an effect on candidate protein levels.

To further test for the presence of horizontal pleiotropy, potential pleiotropic effects of each protein-associated SNP were searched using Phenoscanner[25][26], a database with over 65 billion associations and over 150 million unique genetic variants.

Transcriptomic data in lung tissue

We further investigated *FUT3* and *FUT5* using microarray-based transcriptomic data in whole lungs; GSE32537[27]. Logistic regression was fitted to assess the associations between IPF and standardized log-transformed expressions, adjusted for age, sex and smoking status (ever vs never). We additionally explored the expression profiles using two single-cell RNA sequencing (scRNA-seq) datasets; GSE135893[28] and GSE136831[29]. The unique molecular identifier (UMI) counts of *FUT3* was compared between IPF and control subjects, stratified by each cell type annotation according to the original manuscripts. Detailed methods are described in the online supplement.

Results

Cohort characteristics

The GWAS of circulating protein levels from the INTERVAL study[12] (Sun *et al.*) consisted of 3,301 participants of European descent in England (mean age: 43.7 years, **Table 1**). The circulating protein GWAS from the AGES Reykjavik study[13] (Emilsson *et al.*) recruited 3,200 Icelanders with a mean age of 76.6 years (**Table 1**).

The IPF GWAS was a meta-analysis of three distinct cohorts, which in total consisted of 2,668 cases and 8,591 controls[11]. The mean age was 67.3 years for cases and 64.7 years for controls, respectively. It is highly unlikely that there was any overlap of participants between the proteome and IPF GWAS, since they largely included different geographical locations. Demographic characteristics from each study can be found in **Table 1** and the online supplement.

Mendelian randomization

After MR scanning across 507 and 733 proteins from the two separate pQTL GWASs (834 total proteins, 406 of which were overlapped) for their association with IPF, three candidate proteins survived Benjamini-Hochberg correction. These proteins were: galactoside 3(4)-L-fucosyltransferase (FUT3), alpha-(1,3)-fucosyltransferase 5 (FUT5), and tumor necrosis factor receptor superfamily member 6B (TNFRSF6B) (**Table 2**). FUT3 and FUT5 were replicated by both Sun *et al*. and Emilsson *et al*. GWASs. A one SD genetically determined higher plasma FUT3 and FUT5 had on average 19% and 24% lower risk of developing IPF (OR: 0.81, 95%CI: 0.74–0.88, $p=6.3x10^{-7}$, and OR: 0.76, 95%CI: 0.68–0.86, $p=1.1x10^{-5}$, respectively) (**Table 2**). Some previously described biomarkers for IPF, namely MMP-1, MMP-7[6][7], and CCL-18[9], and other members of fucosyltransferase family; FUT8, FUT10, and POFUT1 were also

assessed in this MR study. None showed causal effects on IPF risk (**Table 3, S1-2**). **Tables S1-2** show the results of all proteins analyzed.

Colocalization analysis

We performed colocalization analyses between the GWASs for candidate proteins (FUT3, FUT5 and TNFRSF6B) in Sun *et al.* and IPF GWAS to assess potential confounding due to LD. Both FUT3 and FUT5 were well-colocalized with IPF by coloc with posterior probabilities of 99.9% and 97.7% for a shared signal, respectively. TNFRSF6B had a lower posterior probability of 15.8% (**Figure 2**). eCAVIAR estimated high colocalization joint-posterior probabilities (CLPP) in FUT3 and FUT5 SNPs (0.28 and 0.016, respectively) but TNFRSF6B had a low CLPP with $4.3x10^{-6}$ (**Figure 2**). Given the lack of clear colocalization for TNFRSF6B, remaining analyses were focused on FUT3 and FUT5.

Sensitivity analyses

In Sun *et al.*[12], three *cis*-SNPs (rs104097772, rs12982233, and rs812936) were independently associated with FUT3 level when conditioned on the lead SNP; rs708686. One *trans*-SNP (rs679574) was also identified for FUT3 level. Two *cis*-SNPs (rs3760775 and rs4807054) were identified for FUT5, which were independently associated when conditioned on the lead SNP; rs778809. FUT3's *trans*-SNP (rs679574) was removed from analyses because it is palindromic and has a minor allele frequency of 0.49, making it impossible to harmonize with the IPF GWAS statistics. By using a method that can incorporate SNPs in LD[23], we included the other three

cis-SNPs (rs104097772, rs12982233, and rs812936) which are in partial LD ($r^2 \leq 0.54$) with the sentinel SNP; rs708686. For FUT5, we included additional two *cis*-SNPs (rs3760775 and rs4807054) that are in partial LD ($r^2 \leq 0.12$) with the leading SNP; rs778809. The SNPs used were all identified in Sun *et al.* and are listed in **Table S3**. MR analyses, accounting for LD, using multiple *cis*-SNPs showed similar estimates both for FUT3 (IVW OR: 0.84, 95%CI: 0.78-0.91, $p=9.8 \times 10^{-6}$, MR-Egger OR: 0.69, 95%CI: 0.50-0.97, p=0.03) and FUT5 (IVW OR: 0.84, 95%CI: 0.77-0.92, $p=1.4 \times 10^{-4}$, MR-Egger OR: 0.59, 95%CI: 0.38-0.90, p=0.01) (**Table 4**, **Figure S1**). The MR-Egger intercept estimates were close to the null, suggesting no detected evidence of directional pleiotropy (**Table 4**). Bidirectional MR provided no evidence that IPF influences FUT3 and FUT5 levels. (**Table S4-5**).

Although FUT3/5 SNPs are on the same chr19 as the genome-wide significant SNP in the IPF GWAS (rs12610495, near *DPP9*), they were not in LD (**Figure S2**). However, given the LD between FUT3's and FUT5's *cis*-SNPs (rs708686 and rs778809/rs10420107, $r^2=0.49$), we performed statistical fine-mapping on the locus using FINEMAP[30] to explore the most important causal SNPs in IPF GWAS[11]. FUT3's SNP; rs708686 had the highest $\log_{10}$(Bayes factor [BF]) at 3.4 and FUT5's SNPs; rs778809/rs10420107 had a $\log_{10}$BF at 1.8, suggesting FUT3's SNP had a higher probability of being causal for IPF (**Figure S3**). Detailed methods are in the online supplement.

Other shared genetic associations

Phenoscanner searches identified that FUT3's *cis*-SNP, rs708686, was also associated with an increased level of FUT5[12] and decreased levels of vitamin B12[31], lactoperoxidase[12], lithostathine-1-alpha[32] and FAM3B[12]. FUT5's *cis*-SNPs, rs778809 and rs10420107, were associated with increased levels of FUT3 and decreased levels of FAM3B[12] (**Table S6**). Rs778809 was also associated with the plasma levels of CA19-9 and CEA in individuals of Asian ancestry but the directions of the effects were not mentioned in the report[33]. Since we used *cis*-SNPs for FUT3 and FUT5, these pleiotropic effects on other molecules were more likely to represent vertical pleiotropy, where SNPs influencing levels of FUT3 and FUT5 in turn affect levels of the other molecules. Vertical pleiotropy does not violate the assumptions of MR. No other respiratory diseases or smoking habits were identified to be genome-wide significantly associated with FUT3/5 *cis*-SNPs ($p<5x10^{-8}$). We identified moderate associations between the FUT3 pQTL SNP and asthma (rs708686 allele T which decreases FUT3 level also decreases the risk of asthma, $P=1.1x10^{-3}$) and between the FUT5 pQTL SNP and asthma (rs778809 allele A which decreases FUT5 level also decreases the risk of asthma, $P=3.4x10^{-3}$) in UK Biobank (Ncases=38,791).

Next, to reduce the possibility of biasing the MR estimates by horizontal pleiotropy of FUT3/5 *cis*-SNPs, we performed MR to test if the potential confounders described above, namely vitamin B12, lactoperoxidase, lithostathine-1-alpha, FAM3B, CA19-9 and CEA, could have an

effect on IPF risk[34]. For these traits, only genetic determinants of each molecule identified in European ancestries were used. None of these potential confounders had evidence of their effects on IPF risk using MR (**Table S7**). **Figure 3** illustrates the overall findings. The detailed methods are in the online supplements.

Literature search

Further assessment for external validation of our findings involved a literature review by searching PubMed for reports published in English. The largest blood proteomic SOMAscan profiling study to date[35], involving 300 IPF patients and 100 matched controls for sex and smoking status, indicated that those with IPF had 0.89-fold lower level of FUT3 (log_2FC: -0.18, p=0.019) but no difference in FUT5 level (log_2FC: -0.024, p=0.76).

To assess the potential horizontal pleiotropy, we next searched for articles using the search terms “idiopathic pulmonary fibrosis” and each potential confounding factor, namely, vitamin B12, lactoperoxidase, lithostathine-1-alpha, FAM3B, CA19-9 and CEA. No previously published articles were found to describe the molecular mechanism of these factors in IPF pathophysiology.

Transcriptomic data of lung tissue

Using microarray-based transcriptomic data in whole lungs (GSE32537), we confirmed that high FUT3 expression level was associated with reduced risk of IPF (OR: 0.50 per 1 SD increase, 95%CI: 0.31-0.80, $p=3.4x10^{-3}$), but FUT5 was not clearly associated with IPF (OR: 0.72 per 1

SD increase, 95%CI: 0.46-1.1, p=0.14, Ncase/Ncontrol=119/50, **Figure 4, Table S8**).

scRNA-seq analyses from two public datasets (GSE135893 or GSE136831) revealed that FUT3 was mainly expressed in epithelial cells in lungs (**Figure S5**). There were distinct patterns of epithelial cell types between IPF and normal lung tissues. Alveolar type 2 cells were decreased and ciliated cells were increased in IPF lungs, which was in line with previous studies[36, 37] (**Figure S6**). *FUT3* expression in alveolar type 2 cells tended to be lower in IPF lungs than normal lungs ($p=1.9x10^{-48}$ in GSE135893 and p=0.16 in GSE136831, **Figure S7**). Detailed results are described in the online supplement.

Discussion

We undertook MR analyses of 834 circulating proteins to assess their effect on susceptibility to IPF in the largest GWAS studies of these traits available to date. Our analyses showed that subjects with genetically-determined higher circulating levels of FUT3 and FUT5 had lower susceptibility to IPF. Colocalization of FUT3/5 and IPF genetic signals and the absence of evidence of MR violations after thorough sensitivity analyses provided robust support of an etiologic effect of FUT3/5 on IPF susceptibility.

MR evidence for FUT3/5 was independently replicated using Sun *et al.* and Emilsson *et al.* GWASs, which provide two distinct age distributions. Sun *et al*. tested associations between protein levels and age, sex, BMI and eGFR. They reported all proteins associated with either age, sex, BMI or eGFR with a significance threshold of $p<1x10^{-5}$, whereby the positive association between age and FUT5 level ($p=1.6x10^{-10}$) was described[12]. FUT3 level was not reported to be associated with any of the four demographic variables. In addition, neither *FUT3* or *FUT5* was associated with age and sex amongst control samples (N=50) in publicly available bulk transcriptomic data in lungs (GSE32537). The genetic signals for IPF at the FUT3/5 locus were also consistent amongst three original IPF cohorts in the IPF GWAS study (**Table S9**).

Given that the cost of measuring hundreds of proteins in adequately powered IPF studies involving samples collected years before disease onset is currently prohibitive, our approach provides an opportunity to prioritize candidate causal protein biomarkers by repurposing available data from large GWASs. MR studies for circulating biomarkers have often replicated or predicted the results of large-scale randomized controlled trials of pharmacological interventions to change biomarker levels[38–43]. Similarly, previous published biomarker studies have used MR methods to strengthen conclusions reported in the observational literature due to its robustness to reverse causation and most sources of confounding[44][45]. Observational evidence sometimes provides opposite directions of effects to genetic findings, which is also the

case for IPF. For example, rs207695 has been repeatedly shown to be associated with increased risk of IPF and the same variant is also known to decrease the expression of *DSP* in lungs and epithelial cells[11, 46, 47]. Taken together, this suggests that genetically low *DSP* expression leads to increased risk of IPF. On the other hand, some studies had identified that *DSP* is overexpressed in IPF lung tissue compared to normal lungs[46, 48], providing an opposite direction of effect. However, these observational results may be influenced by reverse causation, where IPF may influence the transcription of *DSP*. Nevertheless, an independent observational study demonstrated lower levels of circulating FUT3 in IPF patients[35], and our transcriptomic analyses also supported that increased *FUT3* expression was associated with reduced risk of IPF.

It is still unclear how FUT3 may influence IPF risk. The fucosyltransferases encoded by *FUT3* catalyze the formation of α-1,4 fucosylated-glycoconjugates and are present only in two hominids, humans and chimpanzees. These genes are closely related, belonging to the Lewis *FUT5-FUT3-FUT6* genes cluster, whose corresponding enzymes share 85% of sequence similarity due to duplications of ancestral Lewis gene events[49]. Both FUT3 and FUT5 allow the synthesis of Lewis blood-group antigens in exocrine secretions from precursor oligosaccharides[49]. Fucosylation is a post-translational modification that attaches fucose residues to polysaccharides, which partly determines mucin size and charge heterogeneity[50][51]. PTS domain fucosylation in mucins could influence both the affinity to

bind microorganism and mucocliliary clearance, consequently affecting the innate immune response and susceptibility to infections[52][53][54]. The gain of function *MUC5B* promoter SNP, rs35705950, has been repeatedly demonstrated to be associated with IPF risk[11][55]. Overexpression of MUC5B in lungs was also shown to cause mucociliary dysfunction which enhances lung fibrosis in a mouse model[56]. These lines of evidence suggest a plausible link between MUC5B and fucosylation where host defenses influence the pathophysiology of pulmonary fibrosis.

Elevated levels of CA19-9 had been shown to be associated with severity of pulmonary fibrosis[57]. However, our results found no evidence of this biomarker being causal for IPF. We observed that increased levels of FUT3 reduces susceptibility to IPF, which appears to contradict to the previous studies since the FUT3 (Lewis) enzyme is known to be essential for biosynthesis of CA19-9[58] and low levels of FUT3 lead to decreased level of CA19-9. However, given that the pathology of IPF is characterized by microscopic honeycombing that are filled with mucus and inflammatory cells[59], this leads to overproduction of glycans, precursors of CA19-9. Concentrations of CA19-9 had been also noted to decline in IPF patients after lung transplantation[60]. Elevated levels of CA19-9 are therefore likely to be a consequence of IPF.

Like all methods, our approach has important limitations. MR results may be biased by potential violations of its assumptions, which are not always confirmable, except for the SNP-exposure associations. However, our study design reduced potential horizontal pleiotropy by using *cis*-SNPs backed by a biologically plausible rationale on protein levels and are unlikely to be mediated by other molecules. Further, we undertook multiple sensitivity analyses to evaluate potential pleiotropic effects and did not identify evidence of horizontal pleiotropy for FUT3/5 and IPF. We also undertook colocalization analyses, which additionally strengthened support of a shared genetic cause of FUT3/5 with IPF. Given the limited ethnicity of the current study population, further studies are needed to confirm generalizability of these findings to non-European ancestry. Last, it was not ruled out in Sun *et al*[12] that the association between *cis*-SNP rs708686 and the FUT3 level measured by SOMAscan was influenced by potential epitope-binding artefacts driven by protein-altering variants. The negative MR findings of the causal relationships between established IPF biomarkers and IPF susceptibility could be attributed to the known evidence of modest correlations between some proteins measured by aptamer-based technology and those measured by immunoassay[61]. Such lack of correlation can lead to false-negative findings.

As FUT3/5 pQTL SNPs were in LD and pleiotropic to each other, we could not confirm whether FUT3 and FUT5 had independent roles on IPF or whether they are influenced by each other. However, our sensitivity analyses and transcriptomic investigations suggested that FUT3 had a higher probability of being protective for IPF. There are no direct homologs of these proteins in mice, therefore *in-vivo* functional follow-ups were not possible. Alternatively, to test our results in a traditional observational study scenario, molar measurement of FUT3 in pre-diagnostic blood samples in larger, well-characterized, independent populations would be required. Unfortunately, at present, such samples are limited, given IPF's low incidence rate, but these should become more widely available with the development of large-scale population-based longitudinal biobanks.

In summary, undertaking an efficient MR scan of circulating proteins, our study demonstrated that genetically increased circulating FUT3 level is associated with reduced risk of IPF. These findings provide insights into the pathophysiology of this life-threatening disease, which may have potential translational relevance by identifying new targets for needed interventions.

Acknowledgment

We appreciate the benevolence of individuals who participated in all cohorts.

References

1. Richeldi L, Collard HR, Jones MG. Idiopathic pulmonary fibrosis. *Lancet* [Internet] Elsevier Ltd; 2017; 389: 1941–1952Available from: http://dx.doi.org/10.1016/S0140-6736(17)30866-8.
2. Vancheri C, Failla M, Crimi N, Raghu G. Idiopathic pulmonary fibrosis: A disease with similarities and links to cancer biology. *Eur. Respir. J.* 2010; 35: 496–504.
3. Richeldi L, Du Bois RM, Raghu G, Azuma A, Brown KK, Costabel U, Cottin V, Flaherty KR, Hansell DM, Inoue Y, Kim DS, Kolb M, Nicholson AG, Noble PW, Selman M, Taniguchi H, Brun M, Le Maulf F, Girard M, Stowasser S, Schlenker-Herceg R, Disse B, Collard HR. Efficacy and safety of nintedanib in idiopathic pulmonary fibrosis. *N. Engl. J. Med.* 2014; 370: 2071–2082.
4. King TE, Bradford WZ, Castro-Bernardini S, Fagan EA, Glaspole I, Glassberg MK, Gorina E, Hopkins PM, Kardatzke D, Lancaster L, Lederer DJ, Nathan SD, Pereira CA, Sahn SA, Sussman R, Swigris JJ, Noble PW. A phase 3 trial of pirfenidone in patients with idiopathic pulmonary fibrosis. *N. Engl. J. Med.* 2014; 370: 2083–2092.
5. Hunninghake GM. Interstitial lung abnormalities: Erecting fences in the path towards advanced pulmonary fibrosis. *Thorax* [Internet] BMJ Publishing Group; 2019 [cited 2021 Apr 4]; 74: 506–511Available from: http://thorax.bmj.com/.
6. Rosas IO, Richards TJ, Konishi K, Zhang Y, Gibson K, Lokshin AE, Lindell KO, Cisneros J, MacDonald SD, Pardo A, Sciurba F, Dauber J, Selman M, Gochuico BR, Kaminski N. MMP1 and MMP7 as potential peripheral blood biomarkers in idiopathic pulmonary fibrosis. *PLoS Med.* 2008; 5: 0623–0633.
7. White ES, Xia M, Murray S, Dyal R, Flaherty CM, Flaherty KR, Moore BB, Cheng L, Doyle TJ, Villalba J, Dellaripa PF, Rosas IO, Kurtis JD, Martinez FJ. Plasma surfactant protein-D, matrix metalloproteinase-7, and osteopontin index distinguishes idiopathic pulmonary fibrosis from other idiopathic interstitial pneumonias. *Am. J. Respir. Crit. Care Med.* 2016; 194: 1242–1251.
8. Kohno N, Kyoizumi S, Awaya Y, Fukuhara H, Yamakido M, Akiyama M. New serum indicator of interstitial pneumonitis activity. Sialylated carbohydrate antigen KL-6. *Chest* 1989; 96: 68–73.
9. Neighbors M, Cabanski CR, Ramalingam TR, Sheng XR, Tew GW, Gu C, Jia G, Peng K, Ray JM, Ley B, Wolters PJ, Collard HR, Arron JR. Prognostic and predictive biomarkers for patients with idiopathic pulmonary fibrosis treated with pirfenidone: post-hoc assessment of the CAPACITY and ASCEND trials. *Lancet Respir. Med.* [Internet] Elsevier Ltd; 2018; 6: 615–626Available from: http://dx.doi.org/10.1016/S2213-2600(18)30185-1.

10. Labrecque JA, Swanson SA. Interpretation and Potential Biases of Mendelian Randomization Estimates with Time-Varying Exposures. *Am. J. Epidemiol.* 2019; 188: 231–238.
11. Allen RJ, Guillen-Guio B, Oldham JM, Ma SF, Dressen A, Paynton ML, Kraven LM, Obeidat M, Li X, Ng M, Braybrooke R, Molina-Molina M, Hobbs BD, Putman RK, Sakornsakolpat P, Booth HL, Fahy WA, Hart SP, Hill MR, Hirani N, Hubbard RB, McAnulty RJ, Millar AB, Navaratnam V, Oballa E, Parfrey H, Saini G, Whyte MKB, Zhang Y, Kaminski N, et al. Genome-wide association study of susceptibility to idiopathic pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* 2020; 201: 564–574.
12. Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, Burgess S, Jiang T, Paige E, Surendran P, Oliver-Williams C, Kamat MA, Prins BP, Wilcox SK, Zimmerman ES, Chi A, Bansal N, Spain SL, Wood AM, Morrell NW, Bradley JR, Janjic N, Roberts DJ, Ouwehand WH, Todd JA, Soranzo N, Suhre K, Paul DS, Fox CS, Plenge RM, et al. Genomic atlas of the human plasma proteome. *Nature* [Internet] Springer US; 2018; 558: 73–79Available from: http://dx.doi.org/10.1038/s41586-018-0175-2.
13. Emilsson V, Ilkov M, Lamb JR, Finkel N, Gudmundsson EF, Pitts R, Hoover H, Gudmundsdottir V, Horman SR, Aspelund T, Shu L, Trifonov V, Sigurdsson S, Manolescu A, Zhu J, Olafsson Ö, Jakobsdottir J, Lesley SA, To J, Zhang J, Harris TB, Launer LJ, Zhang B, Eiriksdottir G, Yang X, Orth AP, Jennings LL, Gudnason V. Co-regulatory networks of human serum proteins link genetics to disease. *Science (80-.).* 2018; 361: 769–773.
14. Davey Smith G, Davies NM, Dimou N, Egger M, Gallo V, Golub R, Higgins JP, Langenberg C, Loder EW, Brent Richards J, Richmond RC, Skrivankova VW, Swanson SA, Timpson NJ, Tybjaerg-Hansen A, VanderWeele TJ, Woolf BA, Yarmolinsky J. STROBE-MR: Guidelines for strengthening the reporting of Mendelian randomization studies. PeerJ Inc.; 2019 [cited 2020 Aug 5]; Available from: https://doi.org/10.7287/peerj.preprints.27857v1.
15. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* [Internet] Springer US; 2019; 20: 467–484Available from: http://dx.doi.org/10.1038/s41576-019-0127-1.
16. Swanson SA, Hernan MA. The challenging interpretation of instrumental variable estimates under monotonicity. *Int. J. Epidemiol.* 2018; 47: 1289–1297.
17. Davies NM, Holmes M V., Davey Smith G. Reading Mendelian randomisation studies: A guide, glossary, and checklist for clinicians. *BMJ* 2018; 362.

18. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, Laurin C, Burgess S, Bowden J, Langdon R, Tan VY, Yarmolinsky J, Shihab HA, Timpson NJ, Evans DM, Relton C, Martin RM, Davey Smith G, Gaunt TR, Haycock PC. The MR-Base platform supports systematic causal inference across the human phenome. Loos R, editor. *Elife* [Internet] eLife Sciences Publications, Ltd; 2018; 7: e34408Available from: https://doi.org/10.7554/eLife.34408.
19. Burgess S, Butterworth A, Thompson SG. Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data. *Genet. Epidemiol.* [Internet] Wiley; 2013; 37: 658–665Available from: http://10.0.3.234/gepi.21758.
20. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, Plagnol V. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.* 2014; 10.
21. Hormozdiari F, van de Bunt M, Segrè A V., Li X, Joo JWJ, Bilow M, Sul JH, Sankararaman S, Pasaniuc B, Eskin E. Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am. J. Hum. Genet.* 2016; 99: 1245–1260.
22. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, Boehnke M, Abecasis GR, Willer CJ. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* [Internet] Oxford University Press (OUP); 2010; 26: 2336–2337Available from: http://10.0.4.69/bioinformatics/btq419.
23. Yavorska OO, Burgess S. MendelianRandomization: An R package for performing Mendelian randomization analyses using summarized data. *Int. J. Epidemiol.* 2017; 46: 1734–1739.
24. Library WO, Burgess S, Dudbridge F, Thompson SG. Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. 2015; .
25. Staley JR, Blackshaw J, Kamat MA, Ellis S, Surendran P, Sun BB, Paul DS, Freitag D, Burgess S, Danesh J, Young R, Butterworth AS. PhenoScanner: a database of human genotype–phenotype associations. *Bioinformatics* [Internet] Oxford University Press (OUP); 2016; 32: 3207–3209Available from: http://10.0.4.69/bioinformatics/btw373.
26. Kamat MA, Blackshaw JA, Young R, Surendran P, Burgess S, Danesh J, Butterworth AS, Staley JR. PhenoScanner V2: an expanded tool for searching human genotype–phenotype associations. *Bioinformatics* [Internet] Oxford University Press (OUP); 2019; 35: 4851–4853Available from: http://10.0.4.69/bioinformatics/btz469.

27. Yang I V., Coldren CD, Leach SM, Seibold MA, Murphy E, Lin J, Rosen R, Neidermyer AJ, McKean DF, Groshong SD, Cool C, Cosgrove GP, Lynch DA, Brown KK, Schwarz MI, Fingerlin TE, Schwartz DA. Expression of cilium-associated genes defines novel molecular subtypes of idiopathic pulmonary fibrosis. *Thorax* [Internet] Thorax; 2013 [cited 2021 Mar 10]; 68: 1114–1121Available from: https://pubmed.ncbi.nlm.nih.gov/23783374/.
28. Habermann AC, Gutierrez AJ, Bui LT, Yahn SL, Winters NI, Calvi CL, Peter L, Chung MI, Taylor CJ, Jetter C, Raju L, Roberson J, Ding G, Wood L, Sucre JMS, Richmond BW, Serezani AP, McDonnell WJ, Mallal SB, Bacchetta MJ, Loyd JE, Shaver CM, Ware LB, Bremner R, Walia R, Blackwell TS, Banovich NE, Kropski JA. Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis. *Sci. Adv.* [Internet] American Association for the Advancement of Science; 2020 [cited 2021 Mar 12]; 6: eaba1972Available from: http://advances.sciencemag.org/.
29. Adams TS, Schupp JC, Poli S, Ayaub EA, Neumark N, Ahangari F, Chu SG, Raby BA, DeIuliis G, Januszyk M, Duan Q, Arnett HA, Siddiqui A, Washko GR, Homer R, Yan X, Rosas IO, Kaminski N. Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. *Sci. Adv.* [Internet] American Association for the Advancement of Science; 2020 [cited 2021 Mar 16]; 6: eaba1983Available from: www.ipfcellatlas.com.
30. Benner C, Spencer CCA, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. FINEMAP: Efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* 2016; 32: 1493–1501.
31. Nongmaithem SS, Joglekar C V., Krishnaveni G V., Sahariah SA, Ahmad M, Ramachandran S, Gandhi M, Chopra H, Pandit A, Potdar RD, Fall CHD, Yajnik CS, Chandak GR. GWAS identifies population-specific new regulatory variants in FUT6 associated with plasma B12 concentrations in Indians. *Hum. Mol. Genet.* 2017; 26: 2551–2564.
32. Yao C, Chen G, Song C, Keefe J, Mendelson M, Huan T, Sun BB, Laser A, Maranville JC, Wu H, Ho JE, Courchesne P, Lyass A, Larson MG, Gieger C, Graumann J, Johnson AD, Danesh J, Runz H, Hwang SJ, Liu C, Butterworth AS, Suhre K, Levy D. Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nat. Commun.* 2018; 9.

33. He M, Wu C, Xu J, Guo H, Yang H, Zhang X, Sun J, Yu D, Zhou L, Peng T, He Y, Gao Y, Yuan J, Deng Q, Dai X, Tan A, Feng Y, Zhang H, Min X, Yang X, Zhu J, Zhai K, Chang J, Qin X, Tan W, Hu Y, Lang M, Tao S, Li Y, Li Y, et al. A genome wide association study of genetic loci that influence tumour biomarkers cancer antigen 19-9, carcinoembryonic antigen and alpha fetoprotein and their associations with cancer risk. *Gut* England; 2014; 63: 143–151.
34. Hemani G, Tilling K, Davey Smith G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. Li J, editor. *PLOS Genet.* [Internet] Public Library of Science; 2017 [cited 2020 Jul 31]; 13: e1007081Available from: https://dx.plos.org/10.1371/journal.pgen.1007081.
35. Todd JL, Neely ML, Overton R, Durham K, Gulati M, Huang H, Roman J, Newby LK, Flaherty KR, Vinisko R, Liu Y, Roy J, Schmid R, Strobel B, Hesslinger C, Leonard TB, Noth I, Belperio JA, Palmer SM, Asi W, Baker A, Beegle S, Belperio JA, Condos R, Cordova F, Culver DA, De Andrade JAM, Dilling D, Flaherty KR, Glassberg M, et al. Peripheral blood proteomic profiling of idiopathic pulmonary fibrosis biomarkers in the multicentre IPF-PRO Registry. *Respir. Res.* Respiratory Research; 2019; 20: 1–13.
36. Parimon T, Yao C, Stripp BR, Noble PW, Chen P. Alveolar Epithelial Type II Cells as Drivers of Lung Fibrosis in Idiopathic Pulmonary Fibrosis. *Int. J. Mol. Sci.* [Internet] MDPI AG; 2020 [cited 2021 Mar 17]; 21: 2269Available from: https://www.mdpi.com/1422-0067/21/7/2269.
37. Plantier L, Crestani B, Wert SE, Dehoux M, Zweytick B, Guenther A, Whitsett JA. Ectopic respiratory epithelial cell differentiation in bronchiolised distal airspaces in idiopathic pulmonary fibrosis. [cited 2021 Mar 17]; Available from: http://thorax.bmj.com/.
38. Manousaki D, Mokry LE, Ross S, Goltzman D, Brent Richards J. Mendelian Randomization Studies Do Not Support a Role for Vitamin D in Coronary Artery Disease. *Circ. Cardiovasc. Genet.* 2016; 9: 349–356.
39. Manson JAE, Cook NR, Lee IM, Christen W, Bassuk SS, Mora S, Gibson H, Gordon D, Copeland T, D'Agostino D, Friedenberg G, Ridge C, Bubes V, Giovannucci EL, Willett WC, Buring JE. Vitamin D supplements and prevention of cancer and cardiovascular disease. *N. Engl. J. Med.* 2019; 380: 33–44.
40. Holmes M V., Smith GD. Dyslipidaemia: Revealing the effect of CETP inhibition in cardiovascular disease. *Nat. Rev. Cardiol.* [Internet] Nature Publishing Group; 2017; 14: 635–636Available from: http://dx.doi.org/10.1038/nrcardio.2017.156.
41. Holmes M V., Richardson TG, Ference BA, Davies NM, Davey Smith G. Integrating genomics with biomarkers and therapeutic targets to invigorate cardiovascular drug development. *Nat. Rev. Cardiol.* [Internet] Nature Publishing Group; 2021 [cited 2021 Mar 12]; : 1–19Available from: http://www.nature.com/articles/s41569-020-00493-1.

42. Landray M. Tocilizumab in patients admitted to hospital with COVID-19 (RECOVERY): preliminary results of a randomised, controlled, open-label, platform trial. *medRxiv* [Internet] Cold Spring Harbor Laboratory Press; 2021 [cited 2021 Mar 9]; : 2021.02.11.21249258Available from: https://doi.org/10.1101/2021.02.11.21249258.
43. Larsson SC, Burgess S, Gill D. Genetically proxied interleukin-6 receptor inhibition: Opposing associations with COVID-19 and pneumonia [Internet]. Eur. Respir. J. European Respiratory Society; 2021 [cited 2021 Mar 9].Available from: https://doi.org/10.1183/13993003.03545-2020.
44. Fanidi A, Carreras-Torres R, Larose TL, Yuan J-M, Stevens VL, Weinstein SJ, Albanes D, Prentice R, Pettinger M, Cai Q, Blot WJ, Arslan AA, Zeleniuch-Jacquotte A, McCullough ML, Le Marchand L, Wilkens LR, Haiman CA, Zhang X, Stampfer MJ, Smith-Warner SA, Giovannucci E, Giles GG, Hodge AM, Severi G, Johansson M, Grankvist K, Langhammer A, Brumpton BM, Wang R, Gao Y-T, et al. Is high vitamin B12 status a cause of lung cancer? *Int. J. cancer* United States; 2019; 145: 1499–1503.
45. Mokry LE, Ahmad O, Forgetta V, Thanassoulis G, Richards JB. Mendelian randomisation applied to drug development in cardiovascular disease: A review. *J. Med. Genet.* 2015; 52: 71–79.
46. Mathai SK, Pedersen BS, Smith K, Russell P, Schwarz MI, Brown KK, Steele MP, Loyd JE, Crapo JD, Silverman EK, Nickerson D, Fingerlin TE, Yang I V., Schwartz DA. Desmoplakin Variants Are Associated with Idiopathic Pulmonary Fibrosis. *Am. J. Respir. Crit. Care Med.* [Internet] American Thoracic Society; 2016 [cited 2021 Mar 9]; 193: 1151–1160Available from: http://www.atsjournals.org/doi/10.1164/rccm.201509-1863OC.
47. Moore C, Blumhagen RZ, Yang I V., Walts A, Powers J, Walker T, Bishop M, Russell P, Vestal B, Cardwell J, Markin CR, Mathai SK, Schwarz MI, Steele MP, Lee J, Brown KK, Loyd JE, Crapo JD, Silverman EK, Cho MH, James JA, Guthridge JM, Cogan JD, Kropski JA, Swigris JJ, Bair C, Kim DS, Ji W, Kim H, Song JW, et al. Resequencing study confirms that host defense and cell senescence gene variants contribute to the risk of idiopathic pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* [Internet] American Thoracic Society; 2019 [cited 2021 Mar 9]; 200: 199–208Available from: https://www.atsjournals.org/doi/10.1164/rccm.201810-1891OC.
48. Nance T, Smith KS, Anaya V, Richardson R, Ho L, Pala M, Mostafavi S, Battle A, Feghali-Bostwick C, Rosen G, Montgomery SB. Transcriptome Analysis Reveals Differential Splicing Events in IPF Lung Tissue. Buratti E, editor. *PLoS One* [Internet] Public Library of Science; 2014 [cited 2021 Mar 10]; 9: e92111Available from: https://dx.plos.org/10.1371/journal.pone.0092111.

49. Dupuy F, Germot A, Marenda M, Oriol R, Blancher A, Julien R, Maftah A. α1,4-fucosyltransferase activity: A significant function in the primate lineage has appeared twice independently. *Mol. Biol. Evol.* 2002; 19: 815–824.
50. Johnson DC. Airway mucus function and dysfunction. *N. Engl. J. Med.* 2011; 364: 978.
51. Corfield AP. Mucins: A biologically relevant glycan barrier in mucosal protection. *Biochim. Biophys. Acta - Gen. Subj.* 2015; 1850: 236–252.
52. Janssen WJ, Stefanski AL, Bochner BS, Evans CM. Control of lung defence by mucins and macrophages: Ancient defence mechanisms with modern functions. *Eur. Respir. J.* [Internet] European Respiratory Society; 2016 [cited 2020 Aug 6]; 48: 1201–1214Available from: http://ow.ly/s4t7301FLWl.
53. de Mattos LC. Structural diversity and biological importance of ABO, H, Lewis and secretor histo-blood group carbohydrates. *Rev. Bras. Hematol. Hemoter.* [Internet] Associação Brasileira de Hematologia, Hemoterapia e Terapia Celular; 2016; 38: 331–340Available from: http://dx.doi.org/10.1016/j.bjhh.2016.07.005.
54. Kerr SC, Fischer GJ, Sinha M, Mccabe O, Palmer JM, Choera T, Lim Y, Wimmerova M, Carrington SD, Yuan S, Lowell CA, Oscarson S, Keller NP, Fahy J V. FleA Expression in Aspergillus fumigatus Is Recognized by Fucosylated Structures on Mucins and Macrophages to Prevent Lung Infection. 2016 [cited 2020 Aug 6]; Available from: http://www.nih.gov/.
55. Max A Seibold, Anastasia L Wise, Marcy C Speer, Mark P Steele, Kevin K Brown, James E Loyd, Tasha E Fingerlin, Weiming Zhang, Gunnar Gudmundsson, Steve D Groshong, Kenneth B Adler, Burton F Dickey, Roland M Bois, Ivana V Yang, Aretha Herron, Dolly Kervits D a S. A Common MUC5B Promoter Polymorphism and Pulmonary Fibrosis. *N Engl J Med* 2011; 364: 1503–1512.
56. Hancock LA, Hennessy CE, Solomon GM, Dobrinskikh E, Estrella A, Hara N, Hill DB, Kissner WJ, Markovetz MR, Grove Villalon DE, Voss ME, Tearney GJ, Carroll KS, Shi Y, Schwarz MI, Thelin WR, Rowe SM, Yang I V, Evans CM, Schwartz DA. Muc5b overexpression causes mucociliary dysfunction and enhances lung fibrosis in mice. *Nat. Commun.* [Internet] 2018; 9: 5363Available from: https://doi.org/10.1038/s41467-018-07768-9.
57. Maher TM, Oballa E, Simpson JK, Porte J, Habgood A, Fahy WA, Flynn A, Molyneaux PL, Braybrooke R, Divyateja H, Parfrey H, Rassl D, Russell AM, Saini G, Renzoni EA, Duggan AM, Hubbard R, Wells AU, Lukey PT, Marshall RP, Jenkins RG. An epithelial biomarker signature for idiopathic pulmonary fibrosis: an analysis from the multicentre PROFILE cohort study. *Lancet Respir. Med.* [Internet] Elsevier Ltd; 2017; 5: 946–955Available from: http://dx.doi.org/10.1016/S2213-2600(17)30430-7.

58. Kawai S, Suzuki K, Nishio K, Ishida Y, Okada R, Goto Y, Naito M, Wakai K, Ito Y, Hamajima N. Smoking and serum CA19-9 levels according to Lewis and secretor genotypes. *Int. J. Cancer* 2008; 123: 2880–2884.
59. Raghu G, Collard HR, Egan JJ, Martinez FJ, Behr J, Brown KK, Colby T V., Cordier JF, Flaherty KR, Lasky JA, Lynch DA, Ryu JH, Swigris JJ, Wells AU, Ancochea J, Bouros D, Carvalho C, Costabel U, Ebina M, Hansell DM, Johkoh T, Kim DS, King TE, Kondoh Y, Myers J, M??ller NL, Nicholson AG, Richeldi L, Selman M, Dudden RF, et al. An Official ATS/ERS/JRS/ALAT Statement: Idiopathic pulmonary fibrosis: Evidence-based guidelines for diagnosis and management. *Am. J. Respir. Crit. Care Med.* 2011; 183: 788–824.
60. Rusanov V, Kramer MR, Raviv Y, Medalion B, Guber A, Shitrit D. The significance of elevated tumor markers among patients with idiopathic pulmonary fibrosis before and after lung transplantation. *Chest* [Internet] The American College of Chest Physicians; 2012; 141: 1047–1054Available from: http://dx.doi.org/10.1378/chest.11-0284.
61. Raffield LM, Dang H, Pratte KA, Jacobson S, Gillenwater LA, Ampleford E, Barjaktarevic I, Basta P, Clish CB, Comellas AP, Cornell E, Curtis JL, Doerschuk C, Durda P, Emson C, Freeman CM, Guo X, Hastie AT, Hawkins GA, Herrera J, Johnson WC, Labaki WW, Liu Y, Masters B, Miller M, Ortega VE, Papanicolaou G, Peters S, Taylor KD, Rich SS, et al. Comparison of Proteomic Assessment Methods in Multiple Cohort Studies. *Proteomics* [Internet] Wiley-VCH Verlag; 2020 [cited 2021 Mar 18]; 20: 1900278Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/pmic.201900278.

Table 1: Demographic characteristics of the study cohorts.

	Sample size	Ethnicity	Age (mean)	Sex (% males)	Smokers (%)	Assay	Sample
Proteome GWAS							
Sun *et al.* (INTERVAL study)	3,301	British	43.7	51.1	8.6‡	SOMAscan	plasma
Emilsson *et al.* (AGES Reykjavik study)	3,200	Icelandic	76.6*	42.7*	12*	SOMAscan	serum
IPF GWAS Allen *et al.*							
cases	2,668	European	67.3	69.3	72.5§	-	-
controls	8,591	European	64.7†	57.1	66.1§	-	-

* demographic characteristics were calculated with total participants in AGE Reykjavik study (n = 5,457). For smoking status, there is insufficient data to differentiate between current or ever smokers.

† mean age was calculated with samples from the Chicago-based and UK-based studies (n = 3,908) since this information was not available for the Colorado-based study.

‡ % of current smokers.

§ % of ever smokers (calculated with samples from the Chicago-based and UK-based studies [n = 1,153 for cases and n = 3,908 for controls] since this information was not available for the Colorado-based study.

Table 2: Mendelian randomization analyses of proteome for IPF.

	CHR	POS	SNP	Effect Allele	Protein GWAS					IPF GWAS			MR estimate per increase in protein levels	
					Protein	Allele freq	Effect*	P-value	PVE† (%)	Allele freq	Effect	P-value	OR (95% CI)	P-value
Sun *et al.*	19	5840619	rs708686	C	FUT3	0.73	0.85	$3.1x10^{-273}$	27.3	0.72	-0.18	$6.3x10^{-7}$	0.81 (0.74 - 0.88)	$6.3x10^{-7}$
(INTERVAL study)	19	5830302	rs778809	G	FUT5	0.70	0.58	$1.3x10^{-118}$	14.0	0.68	-0.16	$1.1x10^{-5}$	0.76 (0.68 - 0.86)	$1.1x10^{-5}$
Emilsson *et al.*	19	5840619	rs708686	C	FUT3	0.77	0.66	$2.8x10^{-126}$	21.0	0.72	-0.18	$6.3x10^{-7}$	0.76 (0.68 - 0.84)	$6.3x10^{-7}$
(AGES Reykjavik study)	19	5833279	rs10420107	G	FUT5	0.77	0.56	$1.8x10^{-91}$	11.7	0.68	-0.16	$9.2x10^{-6}$	0.75 (0.66 - 0.85)	$9.2x10^{-6}$
	20	62370349	rs1056441	T	TNFRSF6B	0.39	0.14	$2.0x10^{-8}$	1.0	0.31	-0.14	$1.4x10^{-4}$	0.38 (0.23 - 0.62)	$1.4x10^{-4}$

CHR=chromosome; POS=position (hg19), MR=mendelian randomization

* Effect= in *Sun et al, each* protein was first natural log-transformed and adjusted for age, sex, and duration between blood draw and processing, followed by rank-inverse normalization.

In *Emilsson et al,* effect sizes were estimated for Yeo-Johnson transformed protein level, and thus we could not interpret the magnitude of the effect sizes.

† PVE=phenotypic variance explained by the *cis*-pQTL SNP.

Table 3: Mendelian randomization analyses of known IPF circulating biomarkers.

	CHR	POS	SNP	Effect Allele	Protein GWAS					IPF GWAS			MR estimate per increase in protein levels	
					Protein	Allele freq	Effect*	P-value	PVE† (%)	Allele freq	Effect	P-value	OR (95% CI)	P-value
Emilsson *et al.* (AGES Reykjavik study)	11	102697731	rs471994	G	MMP1	0.66	0.55	7.0×10^{-107}	19.1	0.65	-0.01	0.84	0.99 (0.87 - 1.12)	0.84
	11	102401633	rs11568819	G	MMP7	0.95	-0.50	5.0×10^{-21}	3.0	0.94	-0.04	0.59	1.08 (0.82 - 1.42)	0.59
	17	34392880	rs712042	T	CCL18	0.89	-0.89	7.0×10^{-124}	13.4	0.86	-0.04	0.42	1.05 (0.94 - 1.16)	0.42

CHR=chromosome; POS=position (hg19), MR=mendelian randomization

* Effect= In Emilsson et al, effect sizes were estimated for Yeo-Johnson transformed protein level, and thus we could not interpret the magnitude of the effect sizes.

† PVE=phenotypic variance explained by the *cis*-pQTL SNP.

Table 4: MR analyses considering LD patterns using multiple *cis*-SNPs for FUT3 and FUT5.

Protein	Method	MR estimate per 1 SD increase in protein levels		Heterogeneity test		Intercept	
		OR (95% CI)	P-value	Test Statistic	P-value	Intercept (95% CI)	P-value
FUT3	Inverse variance weighted	0.84 (0.78 - 0.91)	9.8×10^{-6}	6.06	0.11	-	-
	MR Egger	0.69 (0.50 - 0.97)	0.03	3.98	0.14	0.15 (-0.09 - 0.38)	0.23
FUT5	Inverse variance weighted	0.84 (0.77 - 0.92)	1.4×10^{-4}	7.19	0.03	-	-
	MR Egger	0.59 (0.38 - 0.90)	0.01	2.52	0.11	0.19 (-0.03 – 0.40)	0.09

MR was performed using “mr_inv” and “mr_egger” functions in “MendelianRandomization” v0.4.3.

Correlation matrices of SNPs were calculated using plink --r square with 503 individuals in the European subset of 1000 Genome projects.

We used a fixed-effects IVW method and a random-effects MR-Egger method.

Legends

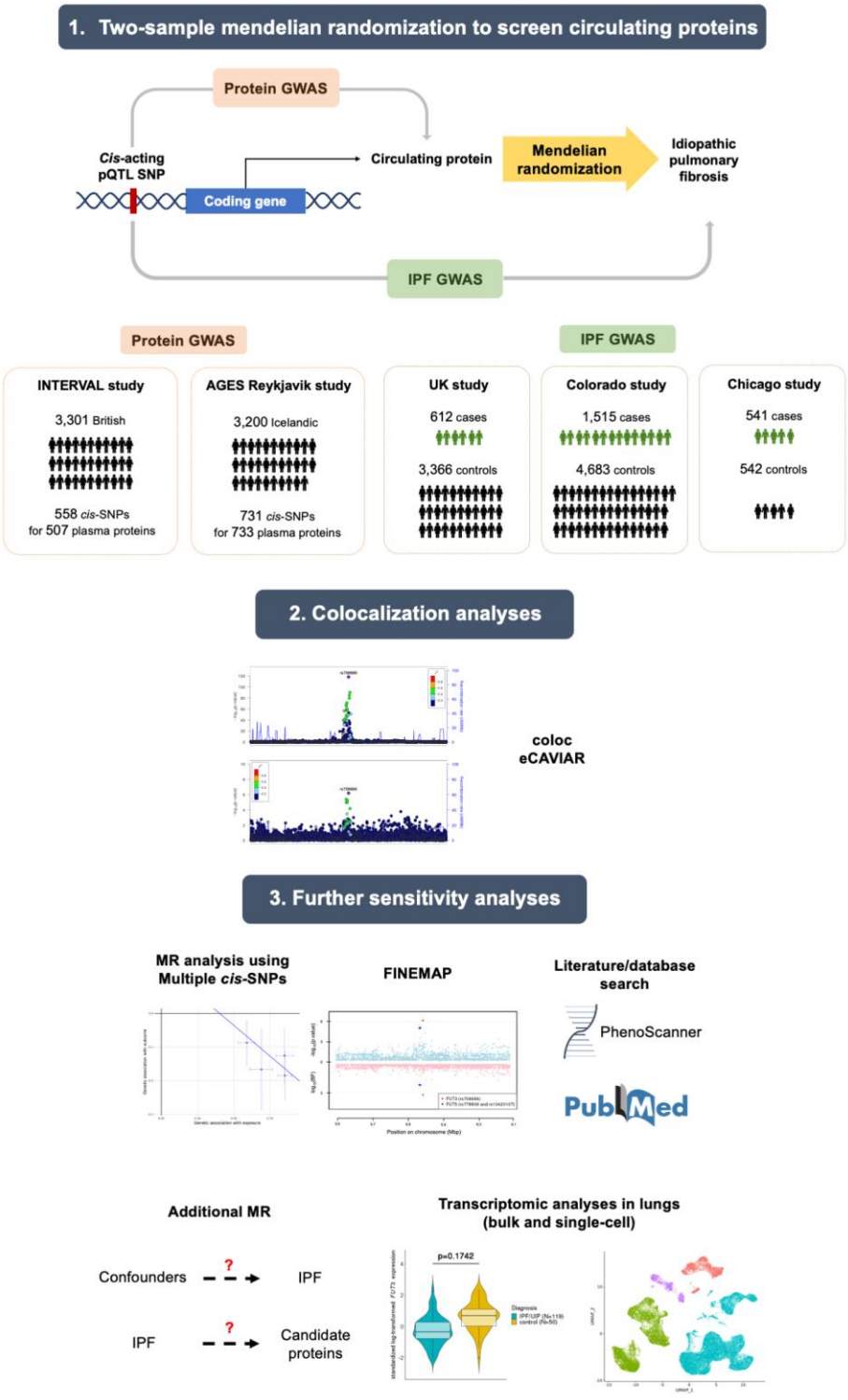
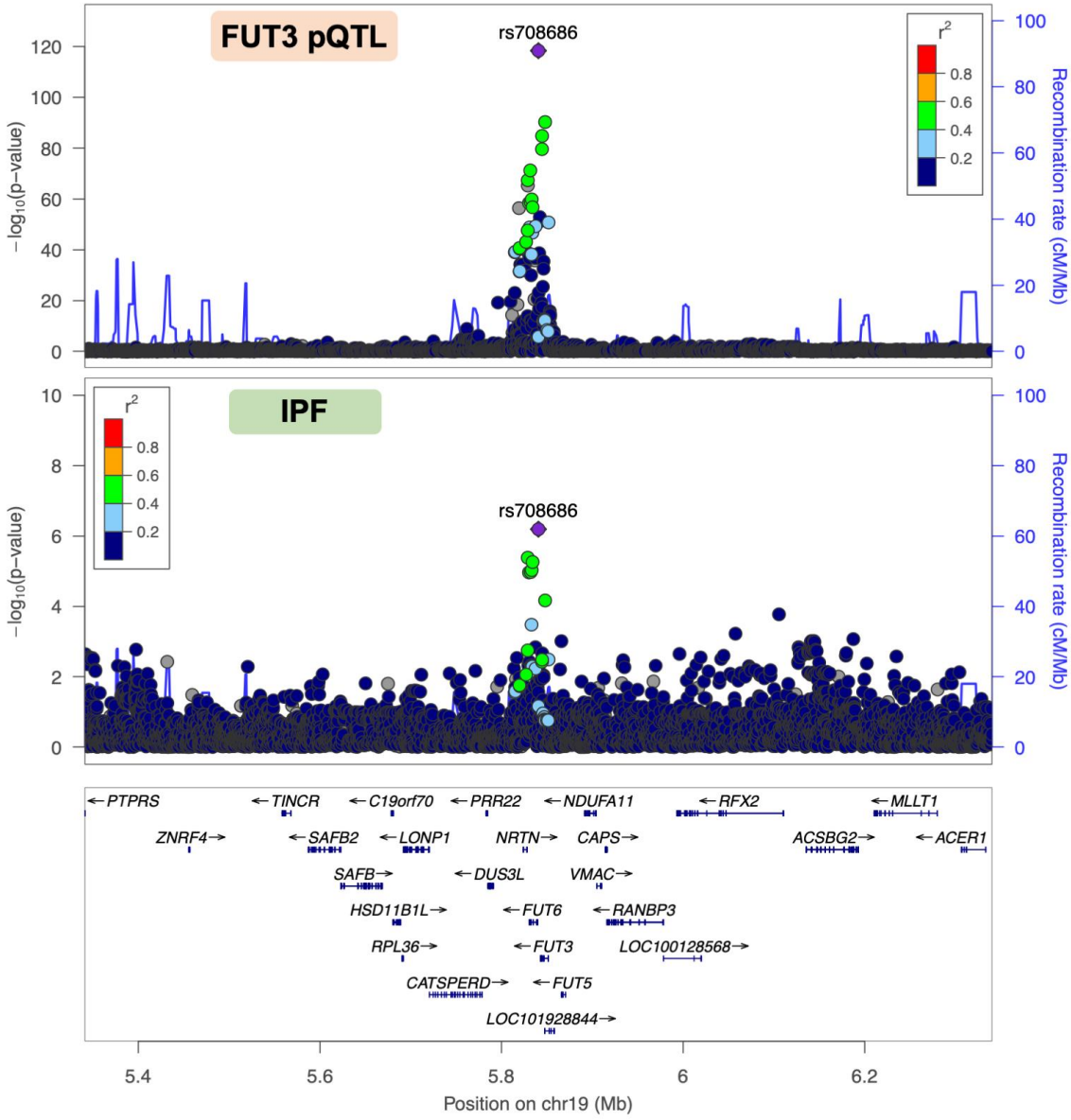
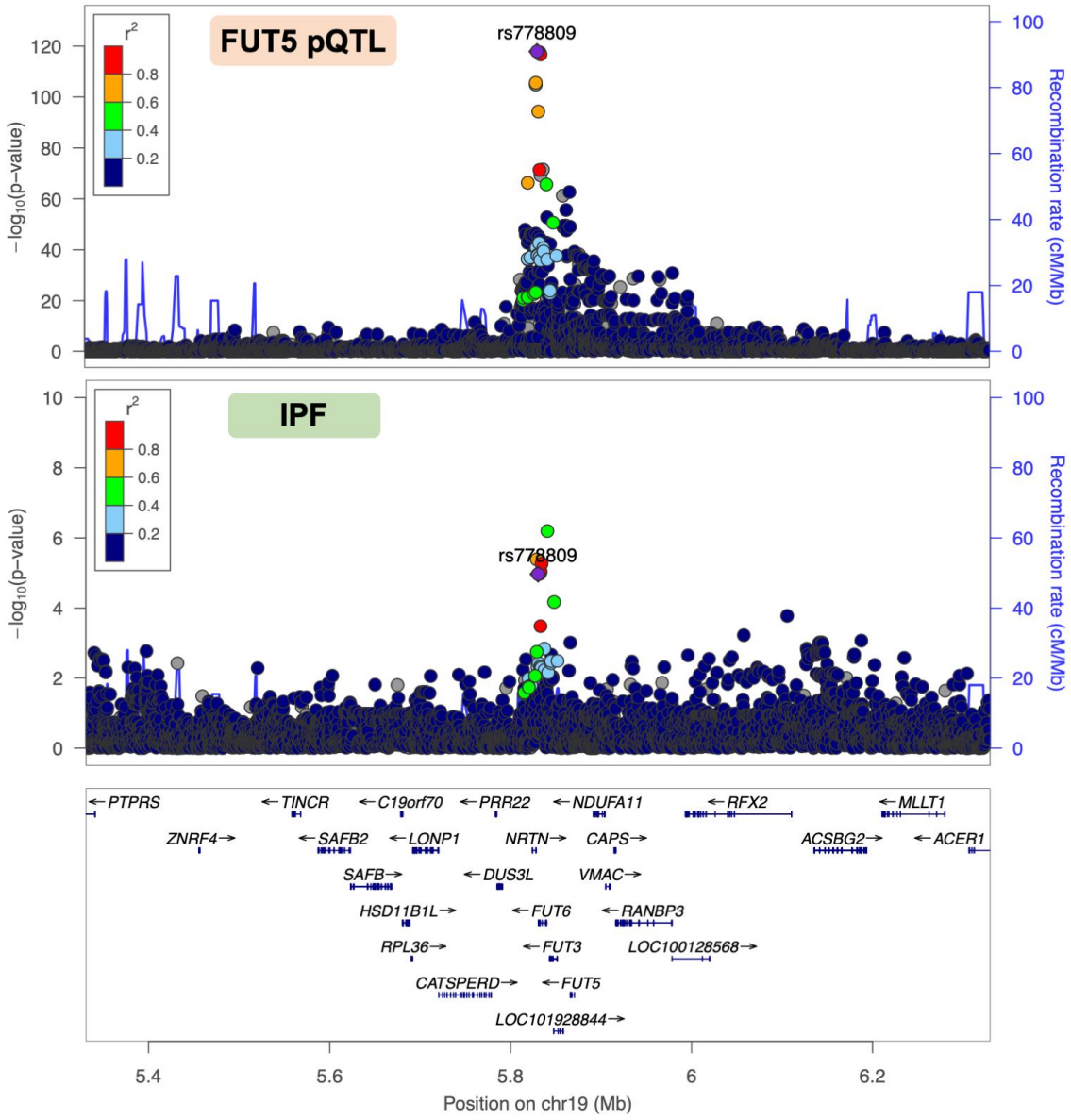


Figure 1. Overall study design.

(A) FUT3: PP4 99.9% CLPP 0.28



(B) FUT5: PP4 97.7% CLPP 0.016



(C) TNFRSF6B: PP4 15.8% CLPP 4.3×10^{-6}

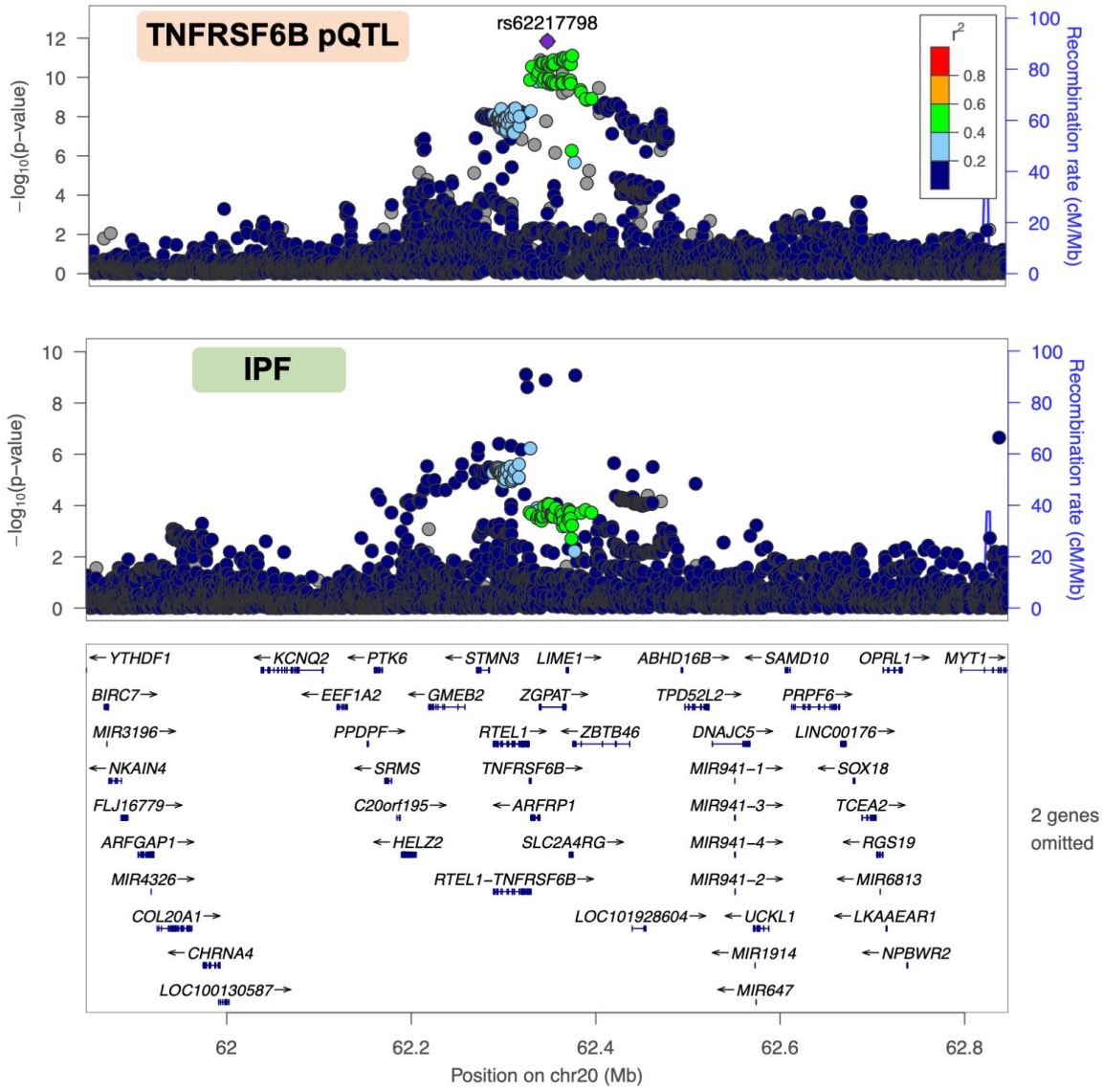


Figure 2. Regional LocusZoom plots and the colocalization analyses results.

Regional LocusZoom plots of three candidate IPF-influencing proteins by using LocusZoom. Each point represents a variant with chromosomal position on the x axis (within 500kb regions of each sentinel variants for candidate proteins) and the $-\log_{10}$(p-value) on the y axis. Variants are colored in by linkage disequilibrium with the sentinel variant. Blue lines show the recombination rate, and gene locations are shown at the bottom of the plot. PP4=posterior probability that the two traits share causal variants calculated by “coloc” R package. CLPP=the colocalization joint posterior probability (CLPP) that the variants are causal for two traits calculated by eCAVIAR. (A) FUT3 (B) FUT5 (C) TNFRSF6B

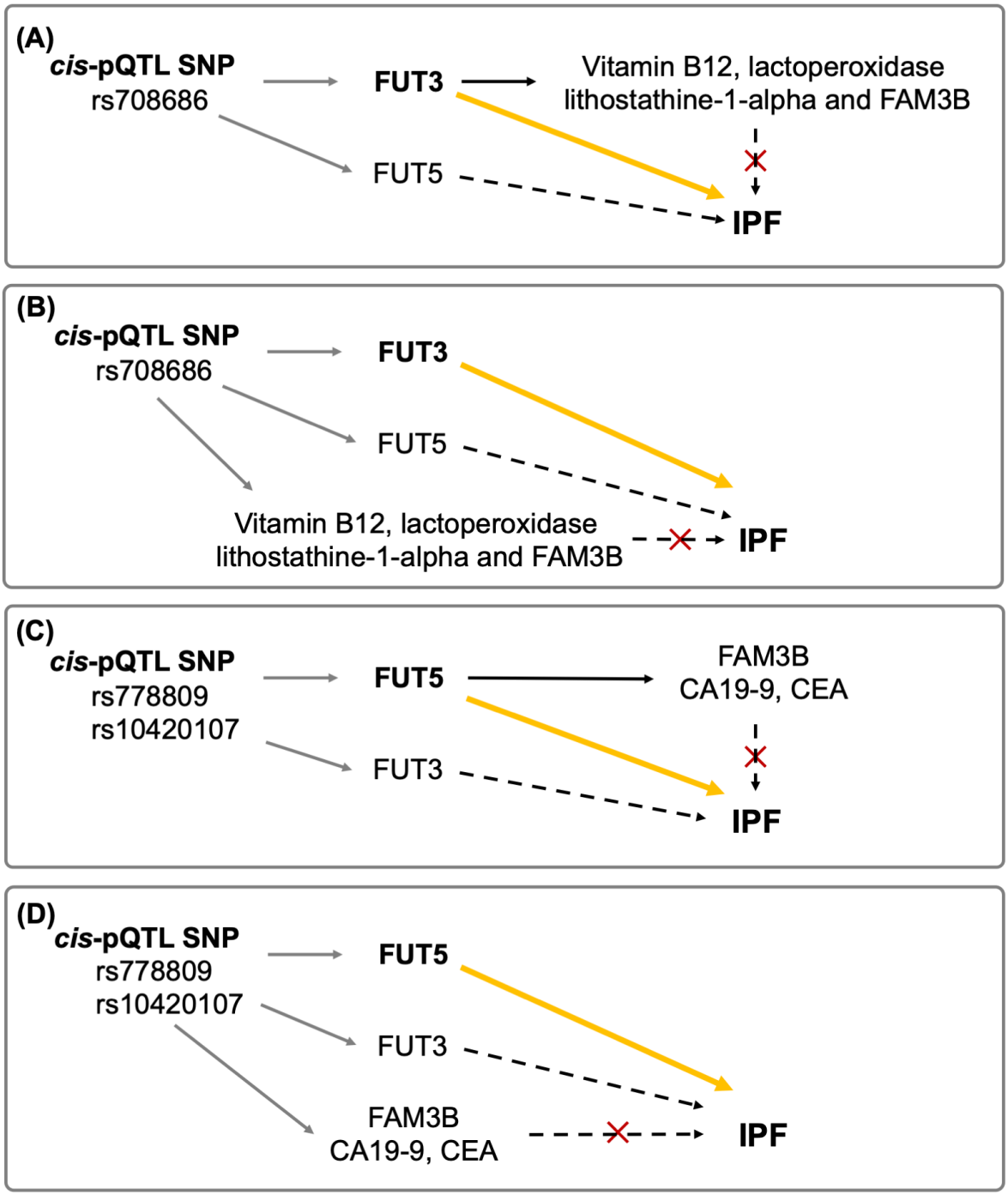

(A)
cis-pQTL SNP
rs708686
FUT3
Vitamin B12, lactoperoxidase
lithostathine-1-alpha and FAM3B
FUT5
IPF
(B)
cis-pQTL SNP
rs708686
FUT3
FUT5
Vitamin B12, lactoperoxidase
lithostathine-1-alpha and FAM3B
IPF
(C)
cis-pQTL SNP
rs778809
rs10420107
FUT5
FAM3B
CA19-9, CEA
FUT3
IPF
(D)
cis-pQTL SNP
rs778809
rs10420107
FUT5
FUT3
FAM3B
CA19-9, CEA
IPF
SNP-protein level associations
Causal path supported by MR
Confounding path
No evidence of this effect by
1. No causal relationships inferred by MR. (Table S7)
2. No literature supports their molecular functions
in IPF pathophysiology.
Causal path

Figure 3. Directed acyclic graphs illustrating the MR conclusions in four different scenarios. In all four scenarios, there was no evidence that the MR estimate of FUT3 and FUT5 on the IPF risk were biased by the violations of MR assumptions. Since we focused on *cis*-acting pQTL SNPs for FUT3 and FUT5, these pleiotropic effects on other molecules' levels are more likely to be vertical pleiotropy, rather than horizontal pleiotropy. Vertical pleiotropy occurs when *cis*-pQTL SNPs influence levels of FUT3 and FUT5 and these two proteins affect the levels of other molecules, which does not bias MR estimates. Moreover, MR analysis using possible confounders as the exposure and IPF as the outcome, no causal relationships were validated. As FUT3/5 pQTL SNPs were in LD and pleiotropic to each other, we could not confirm whether FUT3 and FUT5 had independent roles on IPF susceptibility.

(A) FUT3-associated *cis*-pQTL SNP rs708686 has an effect on IPF via FUT3 and FUT5. FUT3 has a direct effect on IPF and an indirect effect via Vitamin B12, lactoperoxidase, lithostathine-1-alpha and FAM3B, which is an example of vertical pleiotropy that would not bias FUT3's MR estimate. However, this indirect effect was not supported by either MR evidence (Table S7) or literature/database searches.

(B) FUT3-associated *cis*-pQTL SNP rs708686 has an effect on IPF via FUT3, FUT5 and potential confounding variables; Vitamin B12, lactoperoxidase, lithostathine-1-alpha and FAM3B. These confounders represent an example of horizontal pleiotropy that would bias FUT3's MR estimates. However, horizontal pleiotropic effects via these confounders were

not supported by either MR analysis (Table S7) or literature/database searches.

(C) FUT5-associated *cis*-pQTL SNPs rs778809 and rs10420107 have a direct effect on IPF via FUT5 and FUT3, and an indirect effect via FAM3B, CA19-9 and CEA. This indirect effect represents vertical pleiotropy and would not bias FUT5's MR estimate. This indirect effect, however, was not supported by either MR evidence (Table S7) or literature/database searches.

(D) FUT5-associated *cis*-pQTL SNPs rs778809 and rs10420107 have a direct effect on IPF via FUT5, FUT3, and potential confounding variables: FAM3B, CA19-9 and CEA. These confounders represent an example of horizontal pleiotropy that would bias FUT5's MR estimates. However, horizontal pleiotropic effects via these confounders were not supported by either MR analysis (Table S7), or literature/database searches.

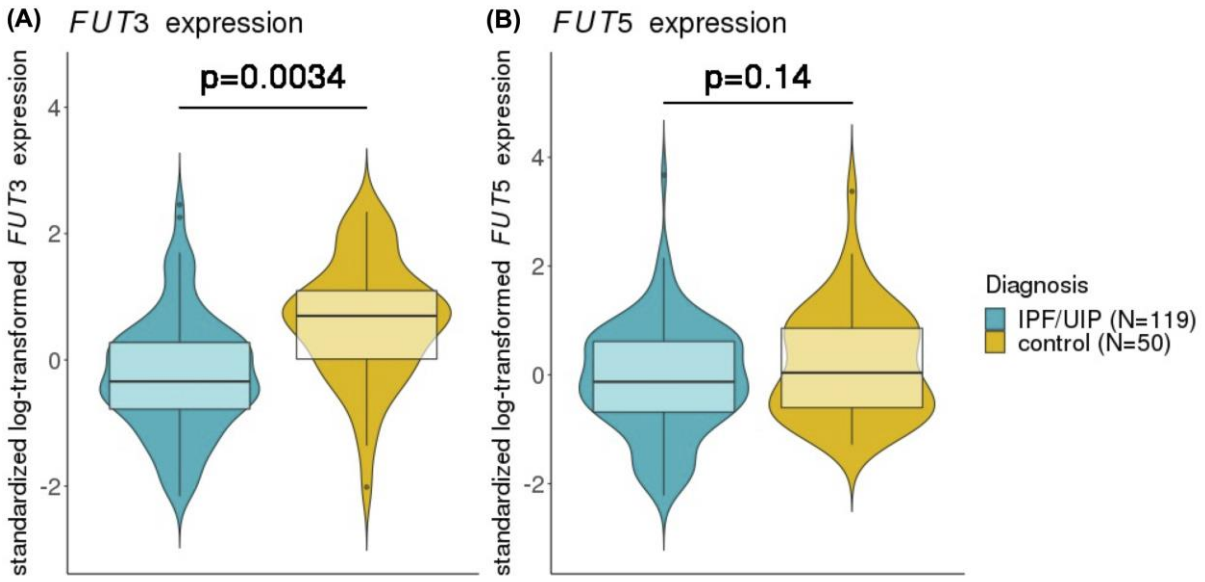


Figure 4. *FUT3* and *FUT5* expression in whole lung, compared between IPF/UIP and controls. This figure is based on data from micro-array based lung transcriptomic dataset (GSE32537). Standardized log-transformed expression levels were compared between IPF/UIP (N=119) and controls (N=50). P-values were calculated by logistic regressions adjusted for age, sex and smoking status.

Table of Contents

SUPPLEMENTARY MATERIAL AND METHODS

Selecting genetic determinants of circulating protein levels

We first identified genome-wide significant single nucleotide polymorphisms (SNPs) ($p<5x10^{-8}$) associated with circulating protein levels (referred to as protein quantitative trait loci (pQTL) SNPs) in the two studies[1][2] and then limited these SNPs to those that were *cis*-acting. The definitions of "*cis*-pQTL SNPs" were different between the two studies and within 1 Mb of the transcription start site of genes encoding the corresponding protein in Sun *et al.*[1] and within 300 kb window across the corresponding protein-coding sequence in Emilsson *et al*.[2] Since multiple independent pQTL SNPs were reported by conditional analyses in Sun *et al.*, we selected multiple independent, i.e. not in linkage disequilibrium (LD) with $r^2 \leq 0.001$, *cis*-pQTL SNPs within the 500 kb of the leading *cis*-pQTL SNPs[3]. The pairwise correlation of SNPs were calculated using the 503 individuals in the European subset of 1000 Genome projects[4].
Both pQTL GWASs used the SOMAscan assay which uses aptamers to measure protein levels. Each protein has its own detection reagent selected from chemically modified DNA libraries, referred to as Slow-Off rate Modified Aptamers (SOMAmers). Detailed methods of SOMAscan assay are described elsewhere[1][2]. The median variation in protein levels explained by pQTL SNPs was 5.8% (interquartile range: 2.6 – 12.4%) in Sun *et al.* Phenotypic variance explained by the *cis*-pQTL SNP was calculated by using the formula described elsewhere[5]. From Emilsson *et al.*, we selected the SNPs with the lowest p-values when the SNPs for the same protein with the different SOMAmers were available. Next, we assessed whether these same SNPs had been analyzed in the IPF GWAS[6], matching on rs number and position. When they were not, we identified LD proxies for these SNPs using an r^2 threshold of > 0.8 using 1000G European reference panel[4]. Both the pQTL GWASs and the IPF GWAS were built on Genome Reference Consortium Human Build 37 (GRCh37). We detected the allele flipping by inferring the forward strand alleles using allele frequency information with the "harmonise_data" function from "TwoSampleMR" R package from MR-base[7] and discarded the palindromic and ambiguous SNPs and the SNPs matched by LD proxies with minor allele frequency > 0.42, since we could not infer these strands correctly. This identified 558 SNPs associated with 507 plasma protein levels (540 different SOMAmers) and 731 SNPs associated with 733 plasma protein levels, respectively from the two pQTL GWAS studies. 406 proteins were overlapped and repeatedly tested by using Sun *et al*. and Emilsson *et al*. The list of proteins we analyzed is described in **Table S1** and **S2**.

Colocalization analysis

We conducted two sets of colocalization analysis using "coloc" R package[8] and eCAVIAR[9]. Coloc is a Bayesian approach that allows us to understand whether the same variants are responsible for the two GWAS signals (in this case the protein level and IPF) or they are distinct causal variants that are just in LD with each other. eCAVIAR is another probabilistic approach that accounts for the marginal statistics (i.e. Z score) obtained from the GWASs and LD

structure of each locus and has been demonstrated to have higher accuracy and precision than coloc[9]. In coloc analysis, we selected the regions within 1 Mb of the lead SNPs for FUT3 (rs708686), FUT5 (rs778809) and TNFRSF6B (rs1056441) both from Sun *et al.*[1] and the IPF GWAS[6]. As allele frequency information was not provided in Sun *et al.*[1], we used 503 individuals in the European subset of 1000 Genome projects to estimate the allele frequency. We selected the exactly same regions as coloc analyses and performed eCAVIAR[9] by setting the maximum number of causal SNPs as one with otherwise the default setting. Whereas the posterior probability was estimated for each locus with coloc, the colocalization posterior probabilities (CLPP) score in eCAVIAR was assigned to each variant within the locus. CLPP is a joint-probability that the variant is causal both in the protein GWAS and the IPF GWAS. The cut-offs for colocalization we applied were 80% in coloc and 0.01 in eCAVIAR, as described previously[8][9].

MR analysis using multiple *cis*-SNPs

As a sensitivity analysis, we included multiple *cis*-SNPs to perform MR using “mr_inv” and “mr_egger” functions in “MendelianRandomization” v0.4.3[10]. Correlation matrices of SNPs were calculated using plink --r square with 503 individuals in the European subset of 1000 Genome projects. We used a fixed-effects IVW method and a random-effects MR-Egger method.

Fine-mapping

FINEMAP is a stochastic search algorithm to explore a set of the most important causal SNPs. To assess the posterior probability of causality of the SNPs for IPF in 19p13.3 locus, we first applied GCTA-COJO[11][12] with parameters of --cojo-wind 20000, --cojo-slct, --cojo-collinear 0.9, and --cojo-p 1e-06 to define the conditionally independent SNP using the IPF GWAS summary statistics[6]. Rs708686 was defined as the conditionally independent SNP. We next calculated the posterior probability of causality of all the SNPs within 500 kb of rs708686 using FINEMAP v1.3.1[13] with parameters of --n-causal-snps 20, --corr-config 0.9, --corr-group 0.9 and –prior-std 0.21.

MR analysis to test the causal relationships between possible confounders and IPF

To reduce the possibility of biasing the MR estimates by horizontal pleiotropy of FUT3/5 *cis*-SNPs, we performed MR to test if potential confounders, namely vitamin B12, lactoperoxidase, lithostathine-1-alpha, FAM3B, CA19-9 and CEA could have an effect on IPF risk[14]. For these traits, only genetic determinants of each molecule identified in European ancestries were used. (**Table S7**). Since the underlying biology of these SNPs is not fully understood, we performed MR Steiger[14] using “mr_steiger” function in “TwoSampleMR” to orient the direction of causality.

Transcriptomic data in lung tissue

We identified several publicly available transcriptomic data performed both in IPF and control lung tissue published in peer-reviewed journals; two RNA sequencing data (SRP033095 [Ncase=8, Ncontrols=7], SRP010041[Ncase=3, Ncontrols=3]) and four microarray data (GSE21411 [Ncase=23, Ncontrol=6], GSE24206 [Ncase=11, Ncontrol=5], GSE32537(Ncase=119, Ncontrol=50), GSE35147 (Ncase=4, Ncontrol=4)). Since GSE32537 had the largest sample size and provided detailed information of phenotypes (age, sex, smoking status, pulmonary function tests), we decided to use GSE32537[15] for the analysis with "GEOquery v2.50.5" R package.

According to the original manuscript, "intensity data was log2-transformed, quantile normalized using robust multi-array average (RMA), and expression levels were summarized on a transcript level using the mean value of all probesets mapping to a transcript. Non-expressed and invariant transcripts were removed using a median variance filter, corrected by a Benjamini-Hochberg false discovery rate (FDR) of 0.10, resulting in a final dataset of 11,950 transcript measurements across 217 samples"[15].
As a quality control, we checked if *FUT3* and *FUT5* expression levels were associated with age, sex, or smoking status amongst controls (N=50). Logistic regression models were fitted to assess if *FUT3* and *FUT5* expression levels was associated with IPF, adjusted for age, sex, and smoking status (ever vs never).

Single-cell RNA sequencing data of lung tissue

We further investigated the expression profiles of *FUT3* and *FUT5* in lungs at single-cell resolution. We used two publicly available datasets, GSE136831 (Ncase=12, Ncontrol=10) [16] and GSE135893 (Ncase=32, Ncontrol=28)[17], both of which were sequenced with 10x Genomics Chromium platform. For GSE136831 data, we created Seurat object by applying "Read10x" function to GSE136831_AllCells.GeneIDs.txt.gz, GSE136831_AllCells.cellBarcodes.txt.gz and GSE136831_RawCounts_Sparse.mtx.gz, followed by "CreateSeuratObject" function in "Seurat v3.2.3" package. Meta data was obtained from GSE136831_AllCells.Samples.CellType.MetadataTable.txt.gz. This dataset had already been pre-processed and cells were kept if >12% of transcriptome was from intron-spanning reads, <20% were mitochondrial origin, and with at least 1,000 unique genes captured. For GSE135893 data, GSE135893_ILD_annotated_fullsize.rds was used for the downstream analysis. This dataset had already been pre-processed and cells containing less than 1,000 nFeature_RNA and more than 25% percentage of mitochondrial genes were filtered out.

FUT3 and *FUT5* were not well detected in either GSE135893 or GSE136831 (*FUT3*: 2.9% out of total cells had non-zero counts in GSE135893 and 0.66% in GSE136831, *FUT5*: 0% in GSE135893 and 0.13% in GSE136831). Although clearly such data should be interpreted with

caution, we decided to analyze *FUT3*, which were relatively more expressed.

We applied three sets of statistical analyses to compare *FUT3* expression levels between IPF and controls, stratified by cell types annotated in the original manuscripts. First, we compared *FUT3* expression level treating each individual cell as an independent sample by applying Wilcoxon rank sum test using "wilcox.test" R package, whose results were described in the main text. Second, we averaged the *FUT3* expression for each subject to create a single "sample" representative for each cell type. Last, we applied linear mixed model to account for the dependency of subjects using "lme4" R package with the following formula.

$$glmer(\ expression \sim diagnosis + (1 \mid subject), family = gaussian)$$

, where $expression$ denotes UMI counts transformed by using "SCTransform" function.

SUPPLEMENTARY RESULT

Cohort characteristics

The IPF GWAS was a meta-analysis of three distinct cohorts, which in total consisted of 2,668 cases and 8,591 controls[6]; a Chicago-based study with 541 IPF cases and 542 controls[18], a Colorado-based study with 1,515 fibrotic idiopathic interstitial pneumonia cases and 4,683 controls[19][20], and a UK-based study with 612 IPF cases and 3,366 controls[21]. The mean age was 67.3 years for cases and 64.7 years for controls, respectively, 69.3% of cases were males and 57.1% of controls were males. 72.5% of cases were ever smokers and 66.1% of controls were ever smokers (**Table 1**).

In Chicago study[18], IPF cases were selected from the University of Chicago and University of Pittsburgh via the Lung Tissue Research Consortium (LTRC), and the Correlating Outcomes with biomedical Markers to Estimate Time-progression in IPF (COMET) study and the controls were selected from the database of genotypes and phenotypes (dbGaP) and healthy individuals recruited from the University of Pittsburgh. All individuals were unrelated, of European-American ancestry.

In the Colorado Study[19][20], 1,515 fibrotic idiopathic interstitial pneumonia cases were recruited from the National Jewish Health IIP population, InterMune IPF trials, UCSF, Vanderbilt University IIP population and the National Heart, Lung and Blood Institute Lung Tissue Research Consortium. 4,683 controls were generated at Centre d'Etude du Polymorphisme Humain and approved for use as controls in other studies. Controls were selected such that they were genetically similar to the cases based on IBS (identical by state) estimates. All individuals were self-reported as non-Hispanic white.

In UK study[21], 612 IPF cases recruited from nine different centres in the UK. All diagnoses were made in accordance with accepted ATS/ERS criteria[22][23]. 3,366 controls selected from UK Biobank such that they had no history of any interstitial lung disease (defined by hospital episode statistics and cause of death) and followed a similar age, sex and smoking distribution to the cases.

Transcriptomic data in lung tissue

Using microarray-based transcriptomic data in whole lungs (GSE32537), both *FUT3* and *FUT5* expression levels were not associated with age, sex, and smoking status (**Figure S4, Table S10**) in control lung tissue (N=50). Next, we confirmed that low *FUT3* expression level was associated with increased risk of IPF (OR: 0.50 per 1 SD increase, 95%CI: 0.31-0.80, $p=3.4x10^{-3}$), but *FUT5* was not significantly associated with IPF (OR: 0.72 per 1 SD increase, 95%CI: 0.46-1.1, p=0.14, **Figure 4, Table S8**).

Single-cell RNA sequencing data of lung tissue

FUT3 were mainly expressed in epithelial cells in two datasets according to the annotation of the original manuscripts (**Figure S5**). Both in GSE136831 and GSE135893, there were distinct patterns of subgroups in epithelial cells between IPF and control lung tissues; alveolar type 2 cells (AT2) were decreased and ciliated cells and basal cells were increased in IPF lungs, which is in line with previous studies[24, 25] (**Figure S6**).

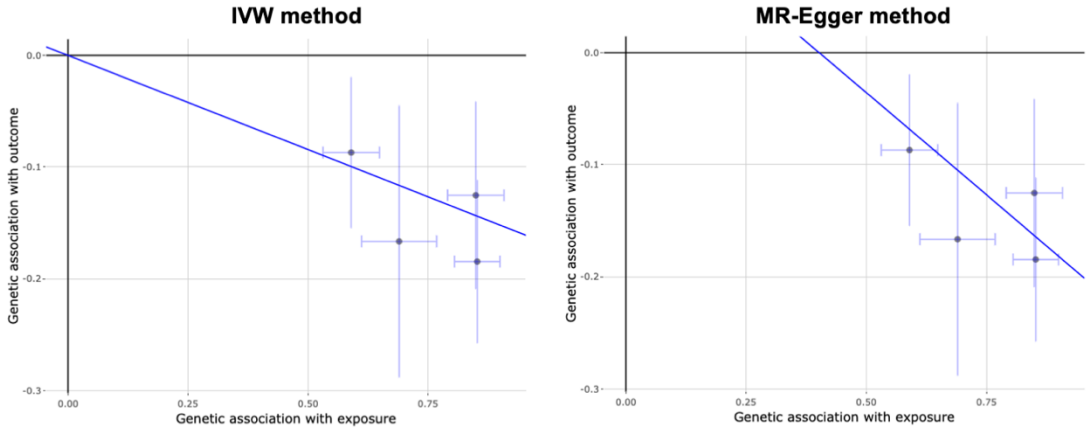
FUT3 expression in AT2 cells tended to be lower in IPF lungs than normal lungs ($p=1.9x10^{-48}$ in GSE135893 and p=0.16 in GSE136831), which is concordant with our MR evidence (**Figure S7, Table S11**).

On the other hand, *MUC5B* positive cells defined in GSE135893 and ciliated cells defined in GSE136831 had modestly higher *FUT3* expression in IPF than in controls. (**Figure S7,8**), although further validation is required.

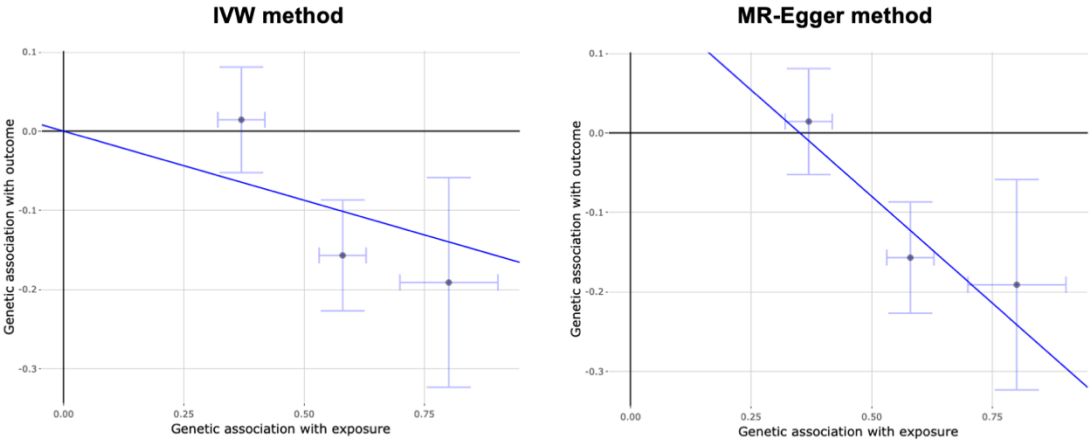
SUPPLEMENTARY FIGURES

Figure S1: Bivariate plots of effect sizes of the SNPs for the exposure and the outcome.

(A) FUT3 multiple *cis*-SNPs

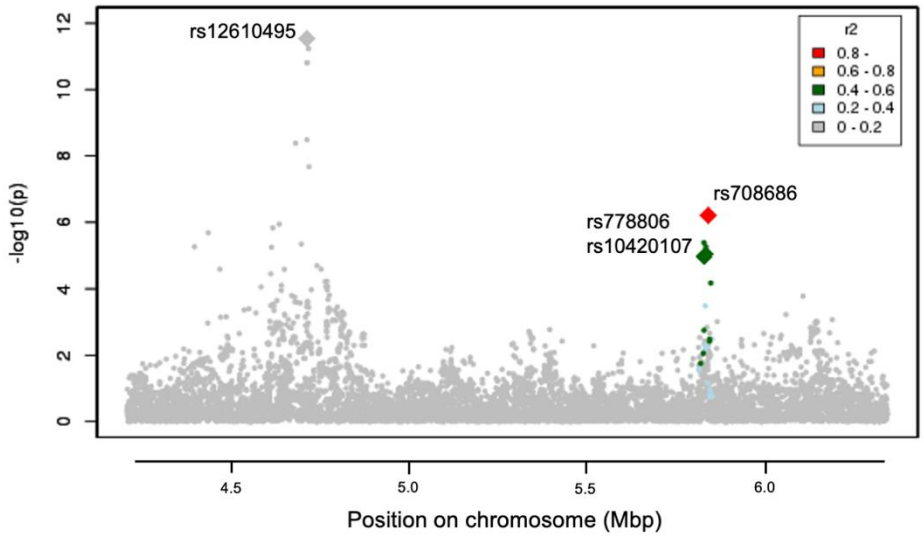


(B) FUT5 multiple *cis*-SNPs



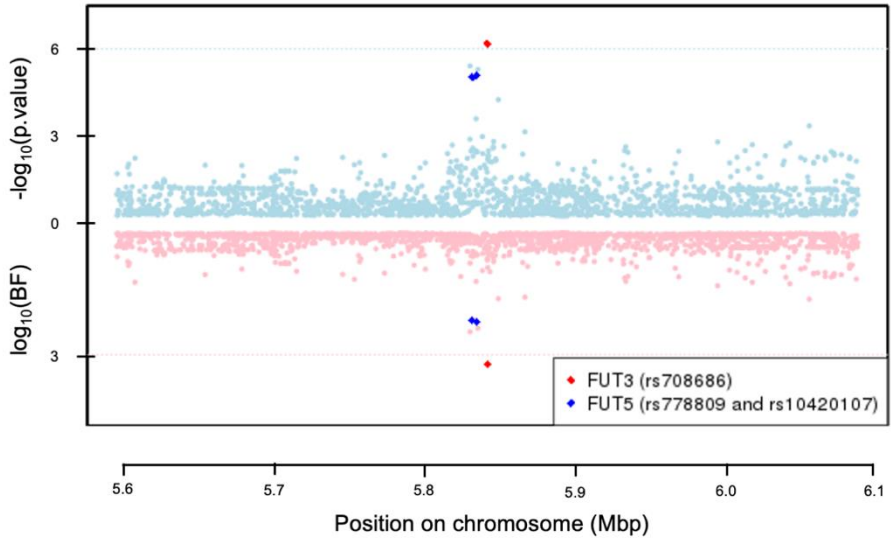
Each point represents the per allele effect size estimate of a SNP (lines from each point are 95% CI for the effect size) Both (A) FUT3 and (B) FUT5 demonstrated consistent estimates of the slope by IVW and MR-Egger methods accounting for correlated variants.

Figure S2. Regional Manhattan plot of the IPF GWAS at the 19p13.3 locus.



Each point represents a variant with chromosomal position on the x axis and the -log10(P value) on the y axis. Variants are colored in by linkage disequilibrium with rs708686. Blue=rs12610495 (top hit on chr19, which was near DPP9 gene.) Red = rs708686 (cis-pQTL SNP for FUT3) Green = rs778806 (cis-QTL SNPs for FUT5.)

Figure S3. Fine-mapping results of the IPF GWAS at the 19p13.3 locus.



For 500 kbp region around the lead SNP; rs708686 on 19p13.3 locus, we applied statistical fine mapping to calculate log10 Bayes factors (BF) for each SNP as a measure of their posterior probability for causality. Conditional independence testing was implemented using GCTA-COJO and log10BF were estimated using FINEMAP.

Figure S4: Scatter plots of standardized log-transformed *FUT3* and *FUT5* expression amongst control lung tissue (N=50).

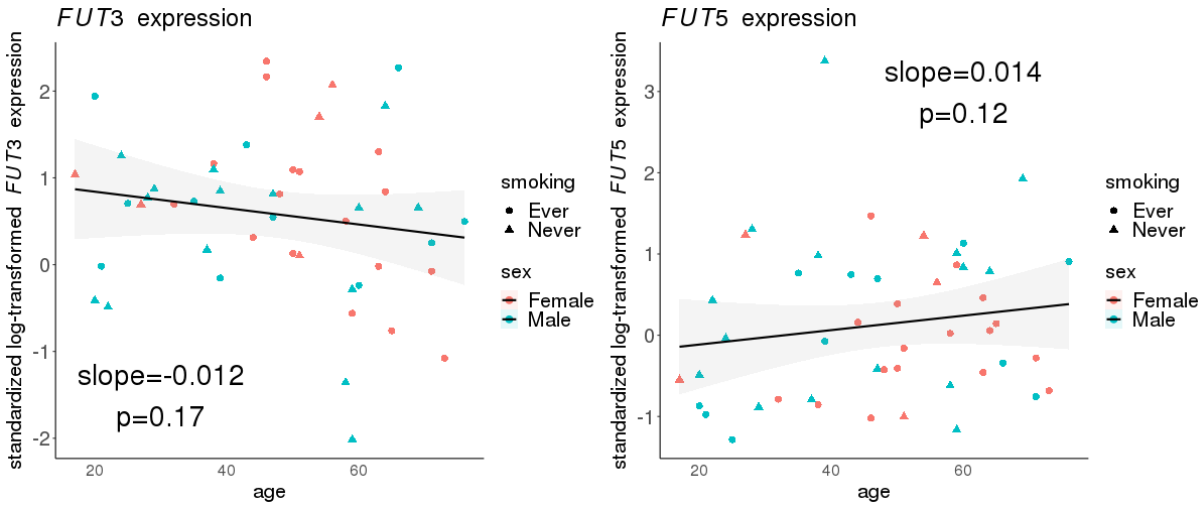
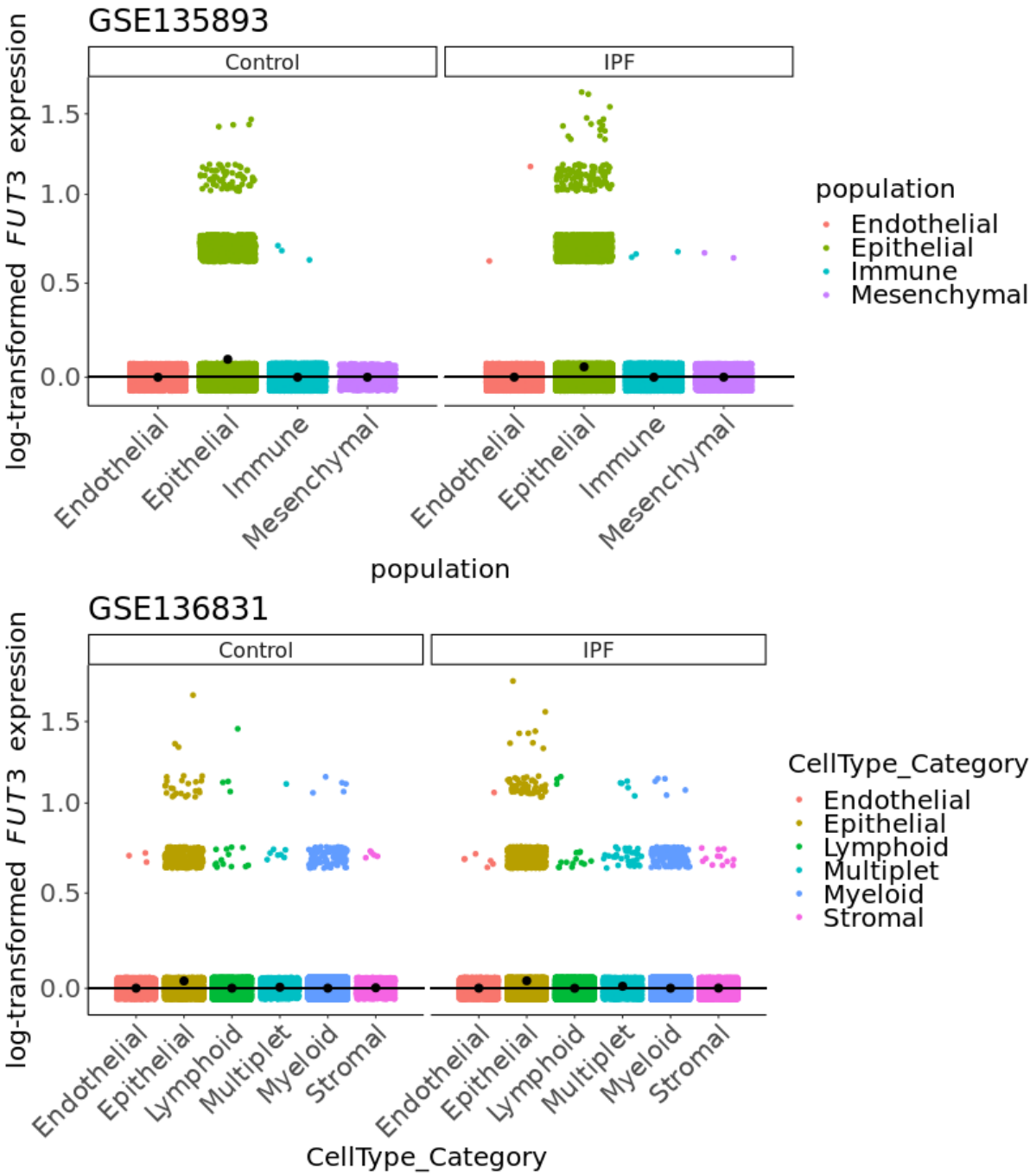
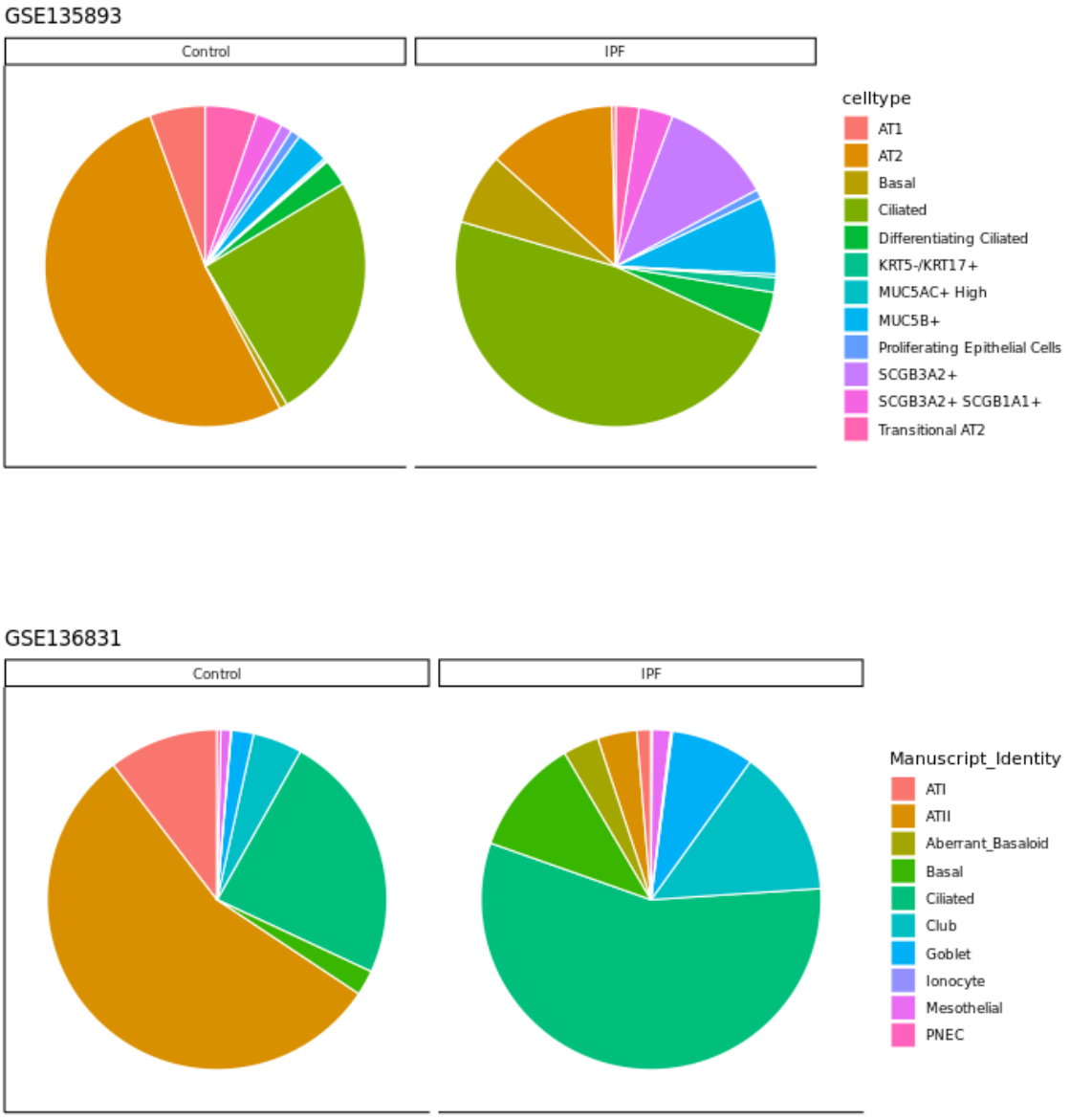


Figure S5: *FUT3* expression (UMI counts) per cell stratified by annotated cellular type.



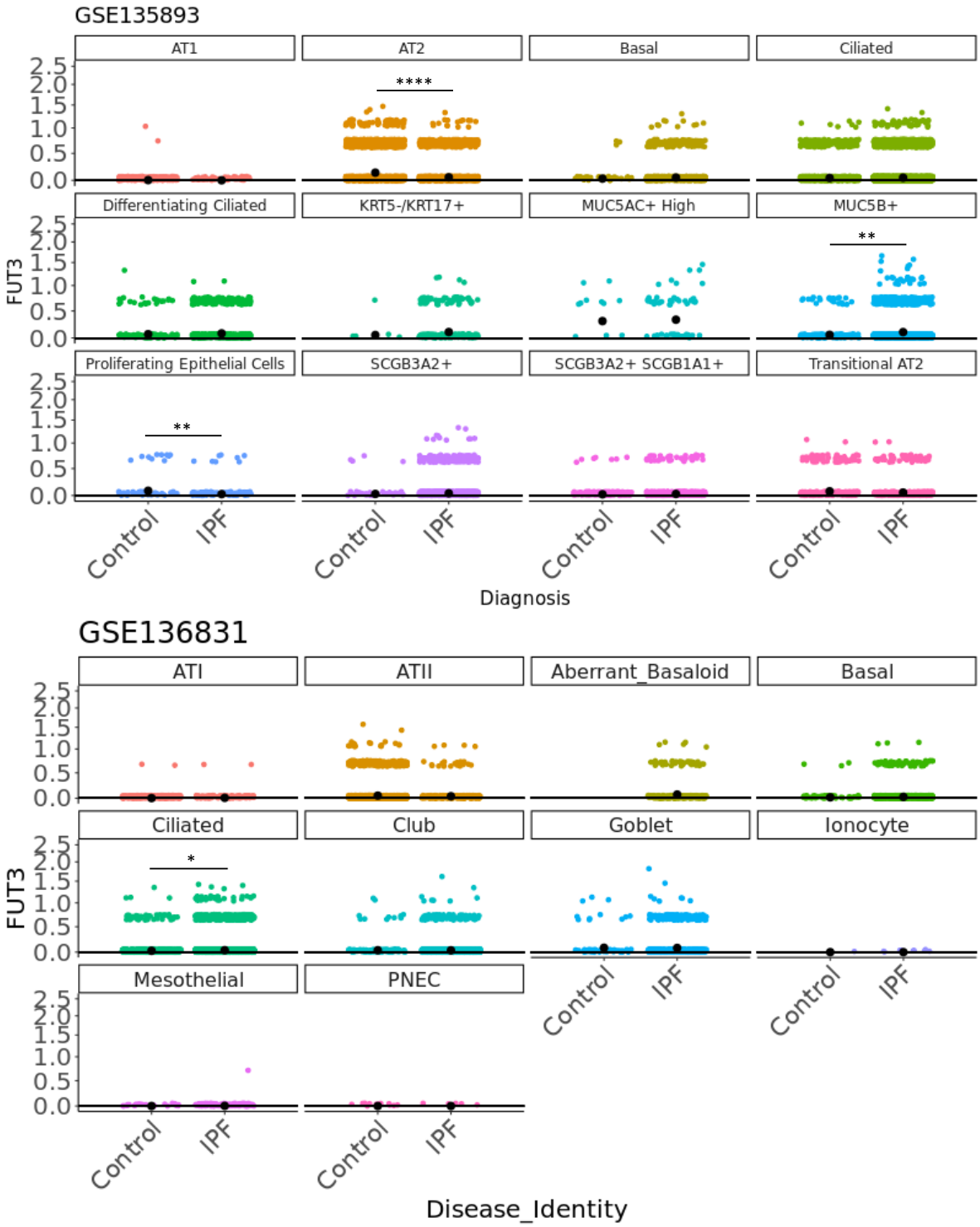
Y axis is UMI read counts per each cell. The read counts were normalized using “SCTransform” function in “Seurat” package. Black dots represent the mean value per each cell type.

Figure S6: Cell type proportions of epithelial cells in two scRNA-seq datasets in IPF and control lungs.



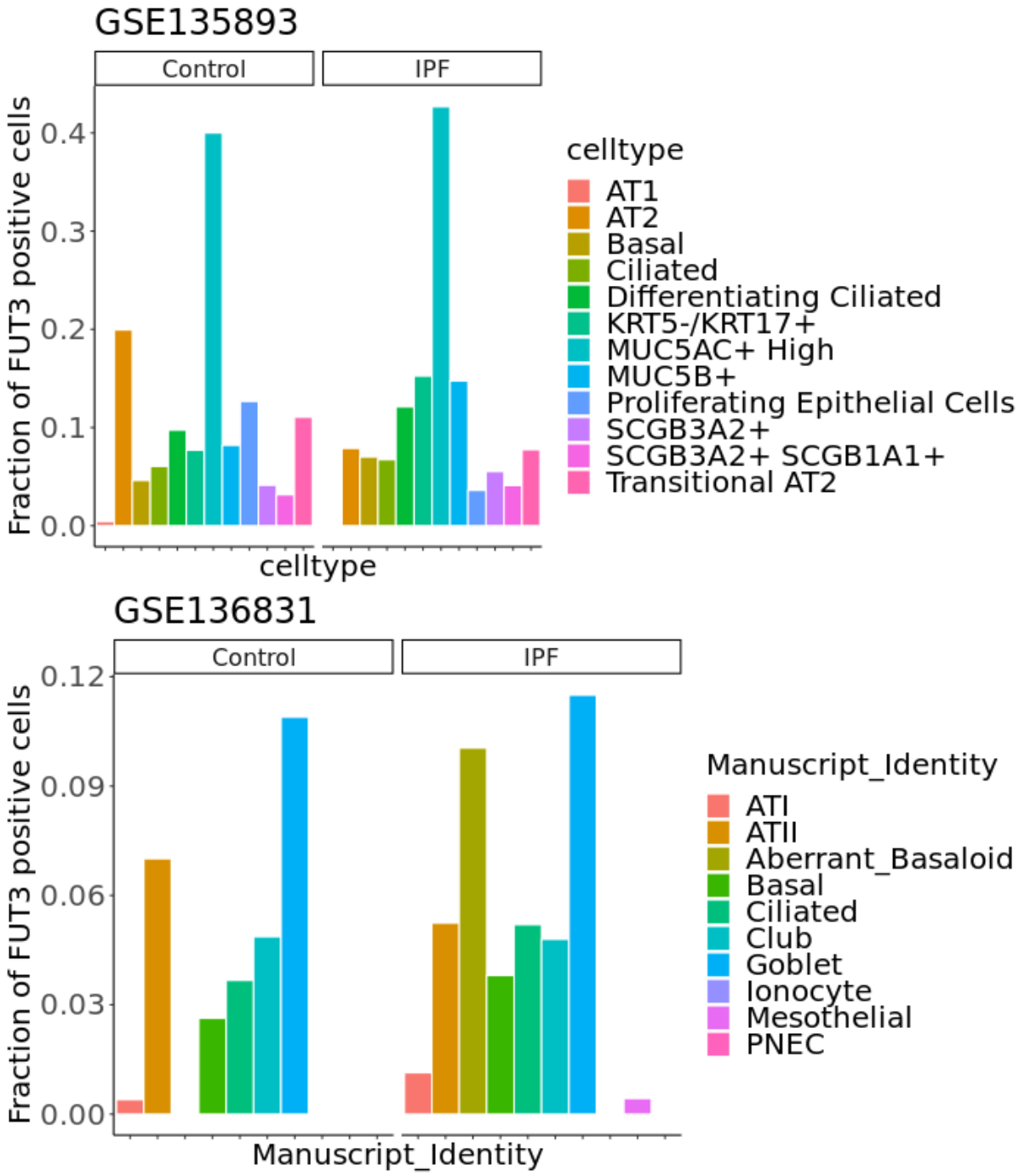
The cell type annotations were defined by clustering analyses in the original manuscripts.

Figure S7: *FUT3* expression comparison between IPF and control lung epithelial cells.



Y axis is read counts per each cell. The read counts were normalized using “SCTransform” function in “Seurat” package. Black dots represent the mean value per each cell type. *: p-value<0.05, **: p-value<0.005, ***: p-value<0.0005, ****: p-value<0.00005. P-values were calculated by Mann-Whitney’s U test using “wilcox.test” R function, treating each individual cell as an independent sample. For other sensitivity analyses, please refer to **Table S11**.

Figure S8: Comparison of the fraction of *FUT3* positive cells between IPF and control lungs.



REFERENCES

1. Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, Burgess S, Jiang T, Paige E, Surendran P, Oliver-Williams C, Kamat MA, Prins BP, Wilcox SK, Zimmerman ES, Chi A, Bansal N, Spain SL, Wood AM, Morrell NW, Bradley JR, Janjic N, Roberts DJ, Ouwehand WH, Todd JA, Soranzo N, Suhre K, Paul DS, Fox CS, Plenge RM, et al. Genomic atlas of the human plasma proteome. *Nature* [Internet] Springer US; 2018; 558: 73–79Available from: http://dx.doi.org/10.1038/s41586-018-0175-2.
2. Emilsson V, Ilkov M, Lamb JR, Finkel N, Gudmundsson EF, Pitts R, Hoover H, Gudmundsdottir V, Horman SR, Aspelund T, Shu L, Trifonov V, Sigurdsson S, Manolescu A, Zhu J, Olafsson Ö, Jakobsdottir J, Lesley SA, To J, Zhang J, Harris TB, Launer LJ, Zhang B, Eiriksdottir G, Yang X, Orth AP, Jennings LL, Gudnason V. Co-regulatory networks of human serum proteins link genetics to disease. *Science (80-.).* 2018; 361: 769–773.
3. Zheng J, Haberland V, Baird D, Walker V, Haycock P, Gutteridge A, Richardson TG, Staley J, Elsworth B, Burgess S, Sun BB, Danesh J, Runz H, Maranville JC, Martin HM, Yarmolinsky J, Laurin C, Holmes M V., Liu J, Estrada K, McCarthy L, Hurle M, Waterworth D, Nelson MR, Butterworth AS, Smith GD, Hemani G, Scott RA, Gaunt TR. Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *bioRxiv* [Internet] 2019; : 627398Available from: https://www.biorxiv.org/content/10.1101/627398v1.
4. Consortium T 1000 GP, Auton A, Abecasis GR, Altshuler (Co-Chair) DM, Durbin (Co-Chair) RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, Gabriel SB, Gibbs RA, Green ED, Hurles ME, Knoppers BM, Korbel JO, Lander ES, Lee C, Lehrach H, Mardis ER, Marth GT, McVean GA, Nickerson DA, Schmidt JP, Sherry ST, Wang J, Wilson RK, Gibbs (Principal Investigator) RA, et al. A global reference for human genetic variation. *Nature* [Internet] The Author(s); 2015; 526: 68–74Available from: https://doi.org/10.1038/nature15393.
5. Shim H, Chasman DI, Smith JD, Mora S, Ridker PM, Nickerson DA, Krauss RM, Stephens M. A multivariate genome-wide association analysis of 10 LDL subfractions, and their response to statin treatment, in 1868 Caucasians. *PLoS One* 2015; 10: 1–20.
6. Allen RJ, Guillen-Guio B, Oldham JM, Ma SF, Dressen A, Paynton ML, Kraven LM, Obeidat M, Li X, Ng M, Braybrooke R, Molina-Molina M, Hobbs BD, Putman RK, Sakornsakolpat P, Booth HL, Fahy WA, Hart SP, Hill MR, Hirani N, Hubbard RB, McAnulty RJ, Millar AB, Navaratnam V, Oballa E, Parfrey H, Saini G, Whyte MKB, Zhang Y, Kaminski N, et al. Genome-wide association study of susceptibility to idiopathic pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* 2020; 201: 564–574.
7. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, Laurin C, Burgess S, Bowden J, Langdon R, Tan VY, Yarmolinsky J, Shihab HA, Timpson NJ, Evans DM,

Relton C, Martin RM, Davey Smith G, Gaunt TR, Haycock PC. The MR-Base platform supports systematic causal inference across the human phenome. Loos R, editor. *Elife* [Internet] eLife Sciences Publications, Ltd; 2018; 7: e34408Available from: https://doi.org/10.7554/eLife.34408.

8. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, Plagnol V. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.* 2014; 10.
9. Hormozdiari F, van de Bunt M, Segrè A V., Li X, Joo JWJ, Bilow M, Sul JH, Sankararaman S, Pasaniuc B, Eskin E. Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am. J. Hum. Genet.* 2016; 99: 1245–1260.
10. Yavorska OO, Burgess S. MendelianRandomization: An R package for performing Mendelian randomization analyses using summarized data. *Int. J. Epidemiol.* 2017; 46: 1734–1739.
11. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 2011; 88: 76–82.
12. Yang J, Ferreira T, Morris AP, Medland SE, Madden PAF, Heath AC, Martin NG, Montgomery GW, Weedon MN, Loos RJ, Frayling TM, McCarthy MI, Hirschhorn JN, Goddard ME, Visscher PM. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* [Internet] Nature Publishing Group; 2012; 44: 369–375Available from: http://dx.doi.org/10.1038/ng.2213.
13. Benner C, Spencer CCA, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. FINEMAP: Efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* 2016; 32: 1493–1501.
14. Hemani G, Tilling K, Davey Smith G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. Li J, editor. *PLOS Genet.* [Internet] Public Library of Science; 2017 [cited 2020 Jul 31]; 13: e1007081Available from: https://dx.plos.org/10.1371/journal.pgen.1007081.
15. Yang I V., Coldren CD, Leach SM, Seibold MA, Murphy E, Lin J, Rosen R, Neidermyer AJ, McKean DF, Groshong SD, Cool C, Cosgrove GP, Lynch DA, Brown KK, Schwarz MI, Fingerlin TE, Schwartz DA. Expression of cilium-associated genes defines novel molecular subtypes of idiopathic pulmonary fibrosis. *Thorax* [Internet] Thorax; 2013 [cited 2021 Mar 10]; 68: 1114–1121Available from: https://pubmed.ncbi.nlm.nih.gov/23783374/.
16. Adams TS, Schupp JC, Poli S, Ayaub EA, Neumark N, Ahangari F, Chu SG, Raby BA, DeIuliis G, Januszyk M, Duan Q, Arnett HA, Siddiqui A, Washko GR, Homer R, Yan X, Rosas IO, Kaminski N. Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. *Sci. Adv.* [Internet] American Association for the Advancement of Science; 2020 [cited 2021 Mar 16]; 6:

eaba1983Available from: www.ipfcellatlas.com.

17. Habermann AC, Gutierrez AJ, Bui LT, Yahn SL, Winters NI, Calvi CL, Peter L, Chung MI, Taylor CJ, Jetter C, Raju L, Roberson J, Ding G, Wood L, Sucre JMS, Richmond BW, Serezani AP, McDonnell WJ, Mallal SB, Bacchetta MJ, Loyd JE, Shaver CM, Ware LB, Bremner R, Walia R, Blackwell TS, Banovich NE, Kropski JA. Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis. *Sci. Adv.* [Internet] American Association for the Advancement of Science; 2020 [cited 2021 Mar 12]; 6: eaba1972Available from: http://advances.sciencemag.org/.
18. Noth I, Zhang Y, Ma SF, Flores C, Barber M, Huang Y, Broderick SM, Wade MS, Hysi P, Scuirba J, Richards TJ, Juan-Guardela BM, Vij R, Han MLK, Martinez FJ, Kossen K, Seiwert SD, Christie JD, Nicolae D, Kaminski N, Garcia JGN. Genetic variants associated with idiopathic pulmonary fibrosis susceptibility and mortality: A genome-wide association study. *Lancet Respir. Med.* [Internet] Elsevier Ltd; 2013; 1: 309–317Available from: http://dx.doi.org/10.1016/S2213-2600(13)70045-6.
19. Fingerlin TE, Murphy E, Zhang W, Peljto AL, Brown KK, Steele MP, Loyd JE, Cosgrove GP, Lynch D, Groshong S, Collard HR, Wolters PJ, Bradford WZ, Kossen K, Seiwert SD, Du Bois RM, Garcia CK, Devine MS, Gudmundsson G, Isaksson HJ, Kaminski N, Zhang Y, Gibson KF, Lancaster LH, Cogan JD, Mason WR, Maher TM, Molyneaux PL, Wells AU, Moffatt MF, et al. Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. *Nat. Genet.* 2013; 45: 613–620.
20. Fingerlin TE, Zhang W, Yang I V., Ainsworth HC, Russell PH, Blumhagen RZ, Schwarz MI, Brown KK, Steele MP, Loyd JE, Cosgrove GP, Lynch DA, Groshong S, Collard HR, Wolters PJ, Bradford WZ, Kossen K, Seiwert SD, Bois RM, Garcia CK, Devine MS, Gudmundsson G, Isaksson HJ, Kaminski N, Zhang Y, Gibson KF, Lancaster LH, Maher TM, Molyneaux PL, Wells AU, et al. Genome-wide imputation study identifies novel HLA locus for pulmonary fibrosis and potential role for auto-immunity in fibrotic idiopathic interstitial pneumonia. *BMC Genet.* [Internet] BMC Genetics; 2016; 17: 1–12Available from: http://dx.doi.org/10.1186/s12863-016-0377-2.
21. Allen RJ, Porte J, Braybrooke R, Flores C, Fingerlin TE, Oldham JM, Guillen-Guio B, Ma SF, Okamoto T, John AE, Obeidat M, Yang I V., Henry A, Hubbard RB, Navaratnam V, Saini G, Thompson N, Booth HL, Hart SP, Hill MR, Hirani N, Maher TM, McAnulty RJ, Millar AB, Molyneaux PL, Parfrey H, Rassl DM, Whyte MKB, Fahy WA, Marshall RP, et al. Genetic variants associated with susceptibility to idiopathic pulmonary fibrosis in people of European ancestry: a genome-wide association study. *Lancet Respir. Med.* 2017; 5: 869–880.
22. Travis WD, Costabel U, Hansell DM, King TE, Lynch DA, Nicholson AG, Ryerson CJ, Ryu JH, Selman M, Wells AU, Behr J, Bouros D, Brown KK, Colby T V., Collard HR, Cordeiro CR, Cottin V, Crestani B, Drent M, Dudden RF, Egan J, Flaherty K,

Hogaboam C, Inoue Y, Johkoh T, Kim DS, Kitaichi M, Loyd J, Martinez FJ, Myers J, et al. An official American Thoracic Society/European Respiratory Society statement: Update of the international multidisciplinary classification of the idiopathic interstitial pneumonias. *Am. J. Respir. Crit. Care Med.* 2013; 188: 733–748.

23. Raghu G, Collard HR, Egan JJ, Martinez FJ, Behr J, Brown KK, Colby T V., Cordier JF, Flaherty KR, Lasky JA, Lynch DA, Ryu JH, Swigris JJ, Wells AU, Ancochea J, Bouros D, Carvalho C, Costabel U, Ebina M, Hansell DM, Johkoh T, Kim DS, King TE, Kondoh Y, Myers J, M??ller NL, Nicholson AG, Richeldi L, Selman M, Dudden RF, et al. An Official ATS/ERS/JRS/ALAT Statement: Idiopathic pulmonary fibrosis: Evidence-based guidelines for diagnosis and management. *Am. J. Respir. Crit. Care Med.* 2011; 183: 788–824.
24. Parimon T, Yao C, Stripp BR, Noble PW, Chen P. Alveolar Epithelial Type II Cells as Drivers of Lung Fibrosis in Idiopathic Pulmonary Fibrosis. *Int. J. Mol. Sci.* [Internet] MDPI AG; 2020 [cited 2021 Mar 17]; 21: 2269Available from: https://www.mdpi.com/1422-0067/21/7/2269.
25. Plantier L, Crestani B, Wert SE, Dehoux M, Zweytick B, Guenther A, Whitsett JA. Ectopic respiratory epithelial cell differentiation in bronchiolised distal airspaces in idiopathic pulmonary fibrosis. [cited 2021 Mar 17]; Available from: http://thorax.bmj.com/.