Genetically increased circulating FUT3 level leads to reduced risk of idiopathic pulmonary fibrosis: a Mendelian randomisation study

Tomoko Nakanishi [1,2,3,4], 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] and J. Brent Richards [1,2,5,18,19]

[1]Dept of Human Genetics, McGill University, Montréal, QC, Canada. [2]Centre for Clinical Epidemiology, Dept of Medicine, Lady Davis Institute, Jewish General Hospital, McGill University, Montréal, QC, Canada. [3]Kyoto–McGill International Collaborative School in Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan. [4]Japan Society for the Promotion of Science, Tokyo, Japan. [5]Dept of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, QC, Canada. [6]Dept of Health Sciences, University of Leicester, Leicester, UK. [7]Dept of Cancer Biology, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [8]Dept of Medicine, Faculty of Medicine, McGill University, Montréal, QC, Canada. [9]Translational Research in Respiratory Diseases, Research Institute of the McGill University Health Centre (RI-MUHC), Montréal, QC, Canada. [10]National Heart and Lung Institute, Imperial College London, London, UK. [11]Dept of Interstitial Lung Disease, Royal Brompton and Harefield Hospitals, Guy's and St Thomas' NHS Foundation Trust, London, UK. [12]National Institute for Health Research, Leicester Respiratory Biomedical Research Centre, Glenfield Hospital, Leicester, UK. [13]Center for Genes, Environment and Health and Dept of Medicine, National Jewish Health, Denver, CO, USA. [14]Dept of Medicine, School of Medicine, University of Colorado Denver, Aurora, CO, USA. [15]McGill Genome Centre and Dept of Human Genetics, McGill University, Montréal, QC, Canada. [16]Dept of Medicine, School of Medicine, University of California, San Francisco, CA, USA. [17]Dept of Immunology, School of Medicine, University of Colorado Denver, Aurora, CO, USA. [18]Division of Endocrinology, Dept of Medicine, Jewish General Hospital, McGill University, Montréal, QC, Canada. [19]Dept of Twin Research, King's College London, London, UK.

Corresponding author: J. Brent Richards (brent.richards@mcgill.ca)

Shareable abstract (@ERSpublications)
After undertaking an efficient scan of 834 circulating proteins for their role in IPF risk using Mendelian randomisation, we found that individuals with genetically increased circulating FUT3 levels had lower risk of developing IPF https://bit.ly/3zCX8zf

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* 2022; 59: 2003979 [DOI: 10.1183/13993003.03979-2020].

Received: 27 Oct 2020
Accepted: 14 June 2021

Abstract

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

Methods To identify candidate IPF-influencing circulating proteins, we undertook an efficient screen of circulating proteins by applying a two-sample Mendelian randomisation (MR) approach with existing publicly available data. For instruments, we used genetic determinants of circulating proteins which reside *cis* to the encoded gene (*cis*-single nucleotide polymorphisms (SNPs)), identified by two genome-wide association studies (GWASs) in European individuals (3301 and 3200 subjects). We then applied MR methods to test if the levels of these circulating proteins influenced IPF susceptibility in the largest IPF GWAS (2668 cases and 8591 controls). We validated the MR results using colocalisation analyses to ensure that both the circulating proteins and IPF shared a common genetic signal.

Results MR analyses of 834 proteins found that a 1 SD increase in circulating galactoside 3(4)-L-fucosyltransferase (FUT3) and α-(1,3)-fucosyltransferase 5 (FUT5) was associated with a reduced risk of IPF (OR 0.81, 95% CI 0.74–0.88; $p=6.3\times10^{-7}$ and OR 0.76, 95% CI 0.68–0.86; $p=1.1\times10^{-5}$, respectively). 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; $p=9.8\times10^{-6}$ and 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; $p=1.4\times10^{-4}$ and MR-Egger OR 0.59, 95% CI 0.38–0.90; p=0.01). FUT3 and FUT5 signals colocalised with IPF signals, with posterior

probabilities of a shared genetic signal of 99.9% and 97.7%, respectively. Further transcriptomic investigations supported the protective effects of *FUT3* for IPF.

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

Introduction

Idiopathic pulmonary fibrosis (IPF) is a progressive, fatal fibrotic interstitial lung disease that affects adults, leading to decreased lung compliance, disrupted gas exchange and resultant respiratory failure [1]. The median survival time from diagnosis is 3–5 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 present, the only way to detect early disease is through high-resolution computed tomography scanning, which reveals interstitial lung abnormalities in up to 10% of the population aged >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 [6–9], 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 estimates of the effects 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 noninvasive biomarkers of disease risk. One way to estimate the causality of circulating biomarkers is using Mendelian randomisation (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 randomisation largely breaks the association with most confounding factors. Furthermore, 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 that efficiently screened hundreds of proteins. Bayesian colocalisation analyses were undertaken to ensure that candidate circulating proteins and IPF shared a common aetiological genetic signal and that the MR results were not biased by linkage disequilibrium (LD). Candidate IPF-influencing proteins identified through MR and colocalisation 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 aetiology 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 were 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 by 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 randomisation

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].

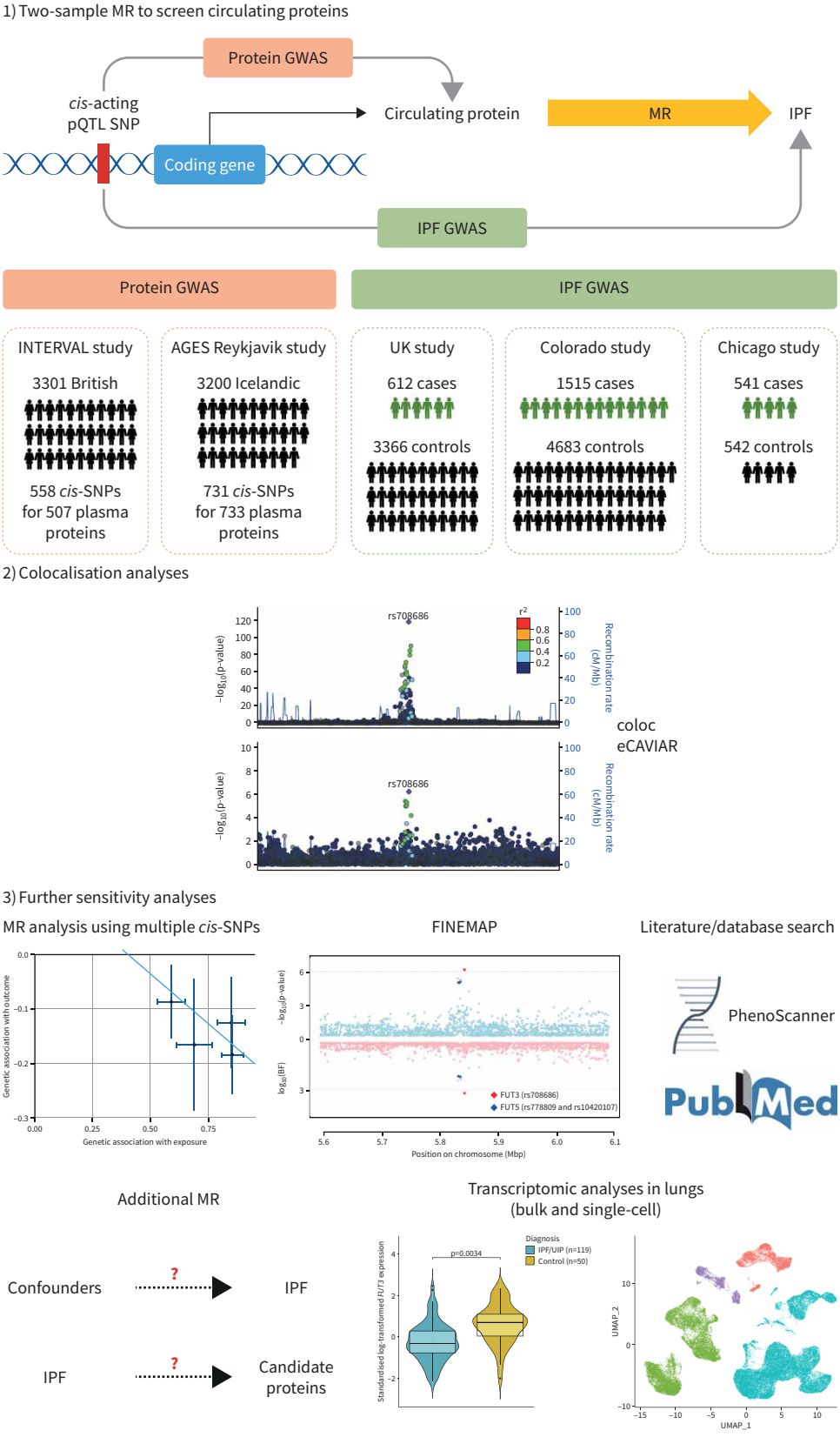

FIGURE 1 Overall study design. See the main text and supplementary material for full details. MR: Mendelian randomisation; GWAS: genome-wide association study; pQTL: protein quantitative trait loci; SNP: single nucleotide polymorphism; IPF: idiopathic pulmonary fibrosis; UIP: usual interstitial pneumonia; UMAP: uniform manifold approximation and projection.

Large-scale GWASs 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 single nucleotide polymorphisms (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 ($r^2 \leqslant 0.001$) *cis*-pQTL SNPs that are significantly associated with circulating proteins ($p<5\times10^{-8}$) from two pQTL GWASs [12, 13]. More details are provided in the supplementary material.

Statistical analysis

We performed MR using the 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 0.05 with 507 multiple testing for Sun *et al*. [12] and 733 multiple testing for Emilsson *et al*. [13]).

Colocalisation analysis

Candidate IPF-influencing proteins supported by MR were evaluated *via* colocalisation 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. Colocalisation 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, protein level and IPF) or they are distinct causal variants that are just in LD with each other. Detailed methods are described in the supplementary material. LocusZoom plots were created to visualise these colocalisations [22].

Sensitivity analysis

Sensitivity analyses were performed for proteins with support from MR and colocalisation 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 MendelianRandomisation R package [23, 24], because consistency of estimates could strengthen the hypothesised 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. Detailed methods are described in the supplementary material. 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 standardised log-transformed expressions, adjusted for age, sex and smoking status (ever *versus* 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 counts of *FUT3* were compared between IPF and control subjects, stratified by each cell type annotation according to the original publications. Detailed methods are described in the supplementary material.

Results

Cohort characteristics

The GWAS of circulating protein levels from the INTERVAL study by Sun *et al*. [12] consisted of 3301 participants of European descent in England (mean age 43.7 years) (table 1). The circulating protein GWAS from the AGES Reykjavik study by Emilsson *et al*. [13] recruited 3200 Icelanders with a mean age of 76.6 years (table 1).

The IPF GWAS was a meta-analysis of three distinct cohorts (UK-, Colorado- and Chicago-based studies), which in total consisted of 2668 cases and 8591 controls [11]. The mean age was 67.3 years for cases and 64.7 years for controls. It is highly unlikely that there was any overlap of participants between the proteome and IPF GWASs, since they largely included different geographical locations. Demographic characteristics from each study can be found in table 1 and the supplementary material.

TABLE 1 Demographic characteristics of the study cohorts

	Sample size (n)	Ethnicity	Age (mean) (years)	Males (%)	Smokers (%)	Assay	Sample
Proteome GWAS							
Sun *et al.* [12] (INTERVAL study)	3301	British	43.7	51.1	8.6[+]	SOMAscan	Plasma
Emilsson *et al.* [13] (AGES Reykjavik study)	3200	Icelandic	76.6[#]	42.7[#]	12[#]	SOMAscan	Serum
Allen *et al.* [11] (IPF GWAS)							
Cases	2668	European	67.3	69.3	72.5[§]		
Controls	8591	European	64.7[¶]	57.1	66.1[§]		

GWAS: genome-wide association study; IPF: idiopathic pulmonary fibrosis. [#]: demographic characteristics were calculated with total participants in the AGES Reykjavik study (n=5457) (for smoking status, there was insufficient data to differentiate between current or ever-smokers); [¶]: mean age was calculated with samples from the Chicago- and UK-based studies (n=3908) since this information was not available for the Colorado-based study (supplementary material); [+]: percentage of current smokers; [§]: percentage of ever-smokers was calculated with samples from the Chicago- and UK-based studies (n=1153 for cases and n=3908 for controls) since this information was not available for the Colorado-based study (supplementary material).

Mendelian randomisation

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: galactoside 3(4)-L-fucosyltransferase (FUT3), α-(1,3)-fucosyltransferase 5 (FUT5) and tumour necrosis factor receptor superfamily member 6B (TNFRSF6B) (table 2). FUT3 and FUT5 were replicated by the GWASs of both Sun *et al.* [12] and Emilsson *et al.* [13]. A 1 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; $p=6.3\times10^{-7}$ and OR 0.76, 95% CI 0.68–0.86; $p=1.1\times10^{-5}$), respectively (table 2). Some previously described biomarkers for IPF, namely MMP1, MMP7 [6, 7] and CCL18 [9], and other members of the fucosyltransferase family (FUT8, FUT10 and POFUT1) were also assessed in this MR study. None showed causal effects on IPF risk (table 3, and supplementary tables S1 and S2). Supplementary tables S1 and S2 also show the results of all proteins analysed.

Colocalisation analysis

We performed colocalisation analyses between the GWASs for candidate proteins (FUT3, FUT5 and TNFRSF6B) in Sun *et al.* [12] and the IPF GWAS to assess potential confounding due to LD. Both FUT3 and FUT5 were well colocalised with IPF by coloc with posterior probabilities of 99.9% and 97.7%, respectively, for a shared signal. TNFRSF6B had a lower posterior probability of 15.8% (figure 2). eCAVIAR estimated a high colocalisation joint posterior probability (CLPP) in FUT3 and FUT5 SNPs (0.28 and 0.016, respectively), but TNFRSF6B had a low CLPP of 4.3×10^{-6} (figure 2). Given the lack of clear colocalisation 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

TABLE 2 Mendelian randomisation (MR) analyses of the proteome for idiopathic pulmonary fibrosis

	Chr.	Position (hg19)	SNP	Effect allele	Protein GWAS					IPF GWAS			MR estimate per increase in protein levels	
					Protein	AF	Effect[#]	p-value	PVE (%)	AF	Effect	p-value	OR (95% CI)	p-value
Sun *et al.* [12]	19	5 840 619	rs708686	C	FUT3	0.73	0.85	3.1×10^{-273}	27.3	0.72	−0.18	6.3×10^{-7}	0.81 (0.74–0.88)	6.3×10^{-7}
(INTERVAL study)	19	5 830 302	rs778809	G	FUT5	0.70	0.58	1.3×10^{-118}	14.0	0.68	−0.16	1.1×10^{-5}	0.76 (0.68–0.86)	1.1×10^{-5}
Emilsson *et al.* [13]	19	5 840 619	rs708686	C	FUT3	0.77	0.66	2.8×10^{-126}	21.0	0.72	−0.18	6.3×10^{-7}	0.76 (0.68–0.84)	6.3×10^{-7}
(AGES Reykjavik study)	19	5 833 279	rs10420107	G	FUT5	0.77	0.56	1.8×10^{-91}	11.7	0.68	−0.16	9.2×10^{-6}	0.75 (0.66–0.85)	9.2×10^{-6}
	20	62 370 349	rs1056441	T	TNFRSF6B	0.39	0.14	2.0×10^{-8}	1.0	0.31	−0.14	1.4×10^{-4}	0.38 (0.23–0.62)	1.4×10^{-4}

Chr.: chromosome; SNP: single nucleotide polymorphism; GWAS: genome-wide association study; AF: allele frequency; PVE: phenotypic variance explained by the *cis*-protein quantitative trait loci SNP. [#]: in Sun *et al.* [12], each protein was first natural log-transformed and adjusted for age, sex, and duration between blood draw and processing, followed by rank-inverse normalisation; in Emilsson *et al.* [13], effect sizes were estimated for Yeo–Johnson-transformed protein level and thus we could not interpret the magnitude of the effect sizes.

TABLE 3 Mendelian randomisation (MR) analyses of known idiopathic pulmonary fibrosis circulating biomarkers

	Chr.	Position (hg19)	SNP	Effect allele	Protein GWAS					IPF GWAS			MR estimate per increase in protein levels	
					Protein	AF	Effect[#]	p-value	PVE (%)	AF	Effect	p-value	OR (95% CI)	p-value
Emilsson *et al.* [13]	11	102 697 731	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
(AGES Reykjavik study)	11	102 401 633	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	34 392 880	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; SNP: single nucleotide polymorphism; GWAS: genome-wide association study; AF: allele frequency; PVE: phenotypic variance explained by the *cis*-protein quantitative trait loci SNP. [#]: in Emilsson *et al.* [13], effect sizes were estimated for Yeo–Johnson-transformed protein level and thus we could not interpret the magnitude of the effect sizes.

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 harmonise 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) that are in partial LD ($r^2\leqslant0.54$) with the sentinel SNP (rs708686). For FUT5, we included additional two *cis*-SNPs (rs3760775 and rs4807054) that are in partial LD ($r^2\leqslant0.12$) with the leading SNP (rs778809). The SNPs used were all identified in Sun *et al.* [12] and are listed in supplementary 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; $p=9.8\times10^{-6}$ and 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; $p=1.4\times10^{-4}$ and MR-Egger OR 0.59, 95% CI 0.38–0.90; p=0.01) (table 4 and supplementary 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 (supplementary tables S4 and S5).

Although the FUT3/5 SNPs are on the same chromosome 19 as the genome-wide significant SNP in the IPF GWAS (rs12610495, near *DPP9*), they were not in LD (supplementary figure S2). However, given the LD between the FUT3 and FUT5 *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 the IPF GWAS [11]. The FUT3 SNP, rs708686, had the highest $\log_{10}$(Bayes factor (BF)) at 3.4 and the FUT5 SNPs, rs778809 and rs10420107, had a $\log_{10}$(BF) at 1.8, suggesting the FUT3 SNP had a higher probability of being causal for IPF (supplementary figure S3). Detailed methods are described in the supplementary material.

Other shared genetic associations

PhenoScanner searches identified that the FUT3 *cis*-SNP, rs708686, was also associated with an increased level of FUT5 [12] and decreased levels of vitamin B12 [31], lactoperoxidase [12], lithostathine-1-α [32] and FAM3B [12]. The FUT5 *cis*-SNPs, rs778809 and rs10420107, were associated with increased levels of FUT3 and decreased levels of FAM3B [12] (supplementary table S6). rs778809 was also associated with the plasma levels of CA19-9 and carcinoembryonic antigen (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 the FUT3/5 *cis*-SNPs ($p<5\times10^{-8}$). We identified moderate associations between the FUT3 pQTL SNP and asthma (rs708686 allele T which decreases FUT3 level also decreases the risk of asthma; $p=1.1\times10^{-3}$) and between the FUT5 pQTL SNP and asthma (rs778809 allele A which decreases FUT5 level also decreases the risk of asthma; $p=3.4\times10^{-3}$) in the UK Biobank ($n_{cases}=38\,791$).

Next, to reduce the possibility of biasing the MR estimates by horizontal pleiotropy of the FUT3/5 *cis*-SNPs, we performed MR to test if the aforementioned potential confounders, *i.e.* vitamin B12, lactoperoxidase, lithostathine-1-α, 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 (supplementary table S7). Figure 3 illustrates the overall findings. Detailed methods are described in the supplementary material.

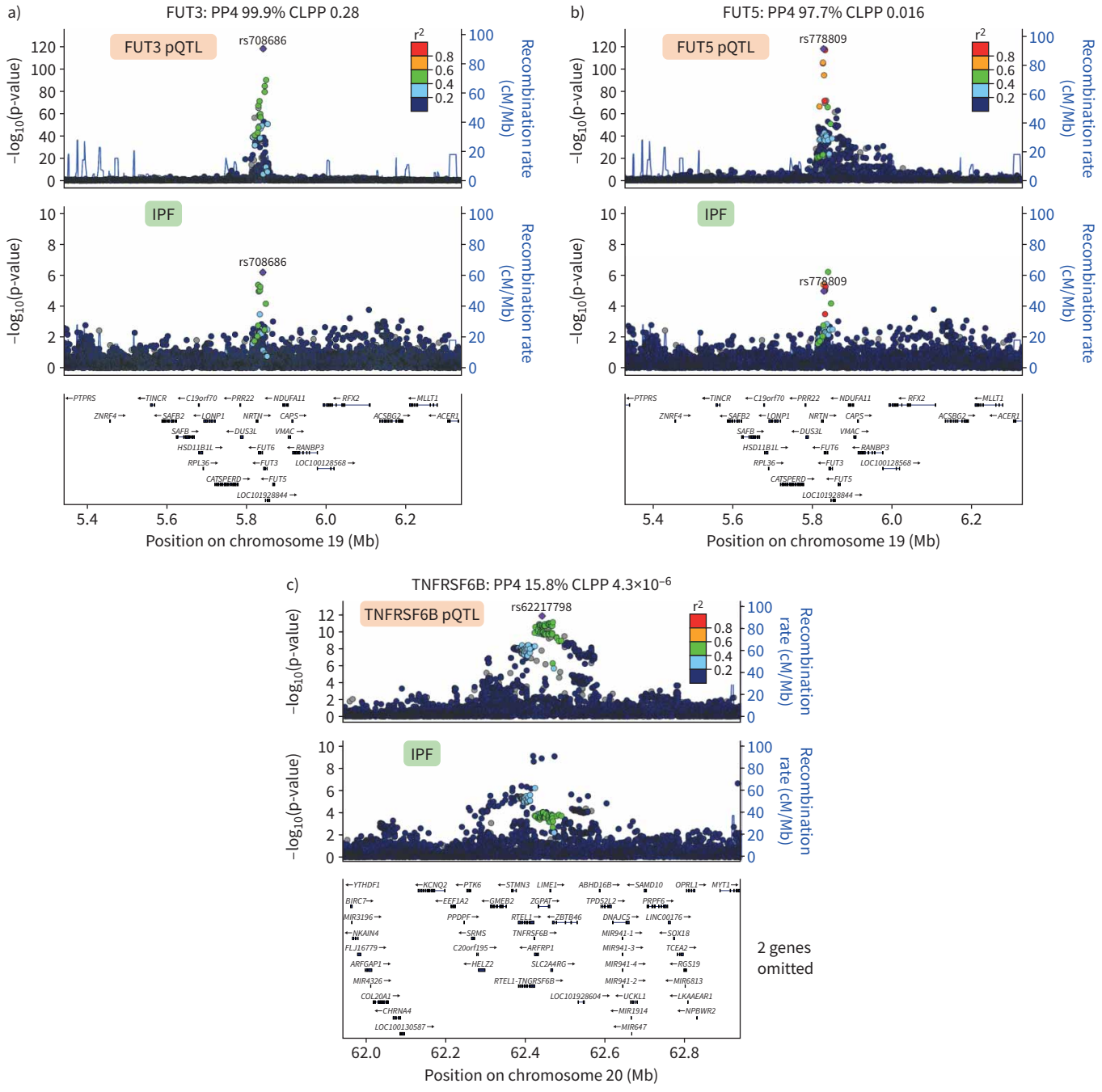

FIGURE 2 Regional LocusZoom plots and colocalisation analyses results. Regional LocusZoom plots of three candidate idiopathic pulmonary fibrosis-influencing proteins: a) FUT3, b) FUT5 and c) TNFRSF6B. Each point represents a variant with chromosomal position on the *x*-axis (within 500-kb regions of each sentinel variant for candidate proteins) and the $-\log_{10}$(p-value) on the *y*-axis. Variants are coloured by linkage disequilibrium with the sentinel variant. Blue lines show the recombination rate; gene locations are shown at the bottom of the plot. PP4: posterior probability that the two traits share causal variants calculated by the coloc R package; CLPP: colocalisation joint posterior probability that the variants are causal for two traits calculated by eCAVIAR; pQTL: protein quantitative trait loci.

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 ($\log_2$(fold change (FC)) −0.18; p=0.019) but no difference in FUT5 level ($\log_2$(FC) −0.024; p=0.76).

TABLE 4 Mendelian randomisation (MR) analyses considering linkage disequilibrium patterns using multiple *cis*-single nucleotide polymorphisms (SNPs) for FUT3 and FUT5

Protein	Method	MR estimate per 1 SD increase in protein level		Heterogeneity test		Intercept	
		OR (95% CI)	p-value	Test statistic	p-value	Intercept (95% CI)	p-value
FUT3	IVW	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	IVW	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 MendelianRandomisation version 0.4.3. Correlation matrices of SNPs were calculated using plink --r square with 503 individuals in the European subset of the 1000 Genomes Projects. We used a fixed effects inverse variance weighted (IVW) method and a random effects MR-Egger method.

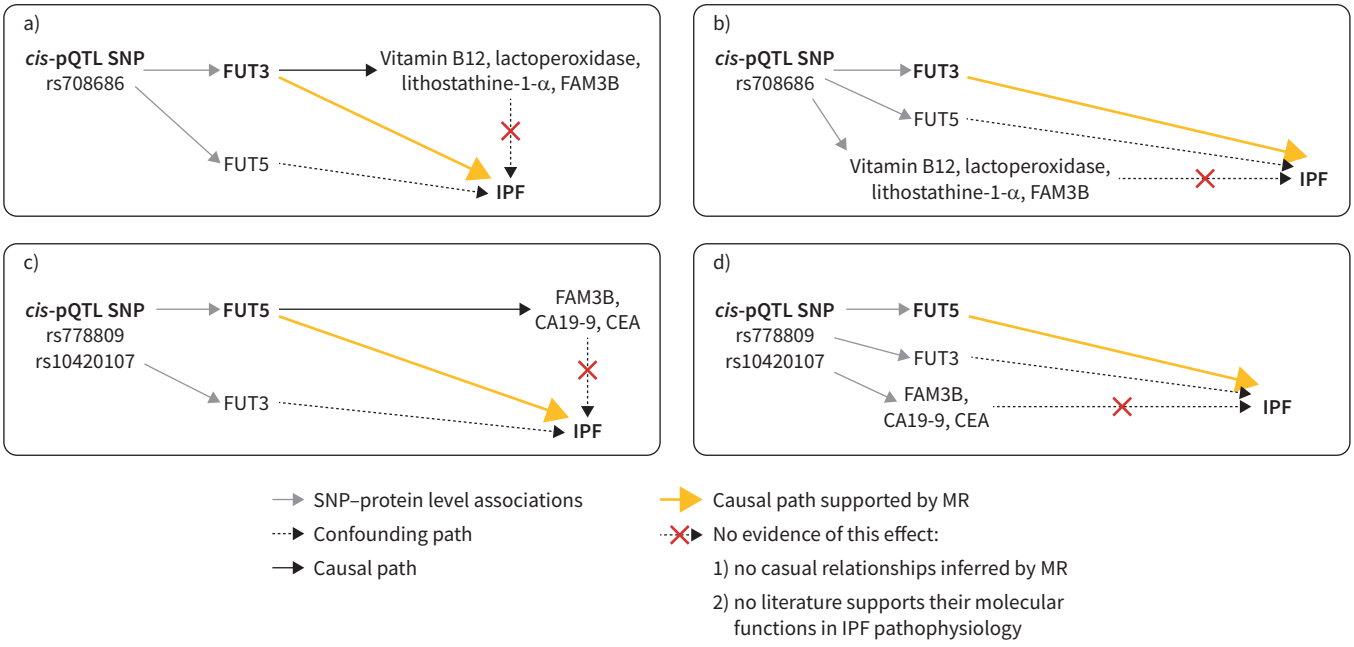

FIGURE 3 Directed acyclic graphs illustrating the Mendelian randomisation (MR) conclusions in four different scenarios. In all four scenarios, there was no evidence that the MR estimate of FUT3 and FUT5 on the idiopathic pulmonary fibrosis (IPF) risk was biased by violations of MR assumptions. Since we focused on *cis*-acting protein quantitative trait loci (pQTL) single nucleotide polymorphisms (SNPs) for FUT3 and FUT5, these pleiotropic effects on the levels of other molecules 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, in 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 linkage disequilibrium 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-α 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 (supplementary 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-α 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 (supplementary 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 carcinoembryonic antigen (CEA). This indirect effect represents vertical pleiotropy and would not bias FUT5's MR estimate. However, this indirect effect was not supported by either MR evidence (supplementary 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 (supplementary table S7) or literature/database searches.

To assess for potential horizontal pleiotropy, we next searched for articles using the search terms "idiopathic pulmonary fibrosis" and each potential confounding factor, *i.e.* vitamin B12, lactoperoxidase, lithostathine-1-α, 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 a high *FUT3* expression level was associated with reduced risk of IPF (OR 0.50 per 1 SD increase, 95% CI 0.31–0.80; $p=3.4\times10^{-3}$), but *FUT5* was not clearly associated with IPF (OR 0.72 per 1 SD increase, 95% CI 0.46–1.1; p=0.14; $n_{case}/n_{control}=119/50$) (figure 4 and supplementary table S8).

scRNA-seq analyses from two public datasets (GSE135893 and GSE136831) revealed that *FUT3* was mainly expressed in epithelial cells in lungs (supplementary figure S5). There were distinct patterns of epithelial cell types between IPF and normal lung tissue. Alveolar type 2 cells were decreased and ciliated cells were increased in IPF lungs, which was in line with previous studies (supplementary figure S6) [36, 37]. *FUT3* expression in alveolar type 2 cells tended to be lower in IPF lungs than normal lungs ($p=1.9\times10^{-48}$ in GSE135893 and p=0.16 in GSE136831) (supplementary figure S7). Detailed results are described in the supplementary material.

Discussion

We undertook MR analyses of 834 circulating proteins to assess their effect on susceptibility to IPF in the largest GWASs 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. Colocalisation of FUT3/5 and IPF genetic signals and the absence of evidence of MR violations after thorough sensitivity analyses provided robust support for an aetiological effect of FUT3/5 on IPF susceptibility.

MR evidence for FUT3/5 was independently replicated using the GWASs of SUN *et al*. [12] and EMILSSON *et al*. [13], which provide two distinct age distributions. SUN *et al*. [12] tested associations between protein levels and age, sex, BMI and estimated glomerular filtration rate (eGFR). They reported all proteins associated with either age, sex, BMI or eGFR with a significance threshold of $p<1\times10^{-5}$, whereby the positive association between age and FUT5 level ($p=1.6\times10^{-10}$) was described [12]. FUT3 level was not reported to be associated with any of the four demographic variables. In addition, neither *FUT3* nor *FUT5* was associated with age or sex among 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 among three original IPF cohorts in the IPF GWAS study (supplementary table S9).

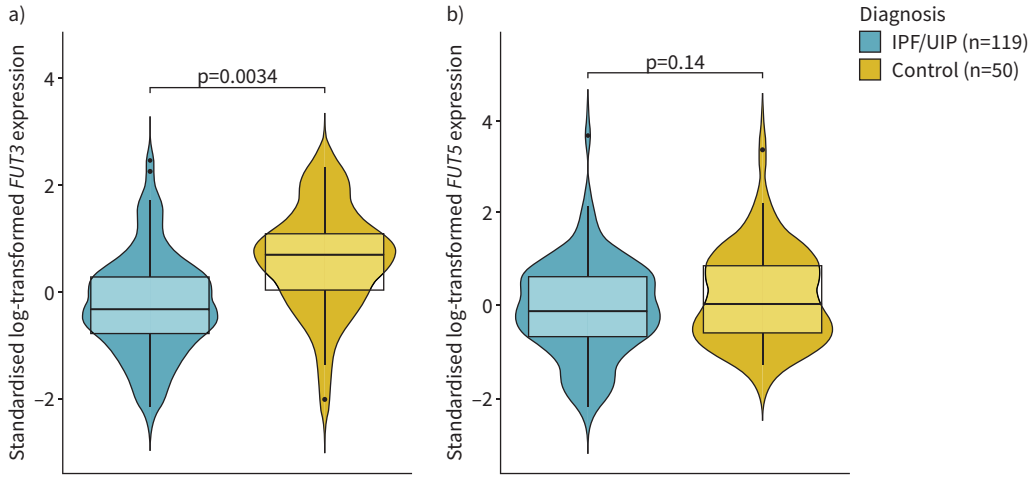

FIGURE 4 a) *FUT3* and b) *FUT5* expression in whole lung compared between idiopathic pulmonary fibrosis (IPF)/usual interstitial pneumonia (UIP) and controls. This figure is based on data from microarray-based lung transcriptomic dataset GSE32537. Standardised log-transformed expression levels were compared between IPF/UIP (n=119) and controls (n=50). p-values were calculated by logistic regressions adjusted for age, sex and smoking status.

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 prioritise 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 randomised controlled trials of pharmacological interventions to change biomarker levels [38–43]. Similarly, previous published biomarker studies have used the MR methodology 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 desmoplakin (*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 with 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* catalyse 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* gene cluster, whose corresponding enzymes share 85% 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 microorganisms and mucociliary clearance, consequently affecting the innate immune response and susceptibility to infections [52–54]. The gain-of-function mucin 5B (*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 that enhances lung fibrosis in a mouse model [56]. These lines of evidence suggest a plausible link between *MUC5B* and fucosylation where host defences 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 reduce susceptibility to IPF, which appears to contradict 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 levels of CA19-9. However, given that the pathology of IPF is characterised by microscopic honeycombing that is 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, which are backed by a biologically plausible rationale on protein levels and are unlikely to be mediated by other molecules. Furthermore, 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 colocalisation analyses, which additionally strengthened support for a shared genetic cause of FUT3/5 with IPF. Given the limited ethnicity of the current study population, further studies are needed to confirm the generalisability 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 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 the 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 homologues of these proteins in mice and 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-characterised, 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.

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

Author contributions: Conception and design: T. Nakanishi and J.B. Richards. Data analyses: T. Nakanishi and O.C. Leavy. Manuscript writing: T. Nakanishi and J.B. Richards. Data acquisition: R.J. Allen, R.G. Jenkins, L.V. Wain, P.J. Wolters and D.A. Schwartz. Interpretation of data: all authors. Intellectual contribution to the manuscript: all authors. All authors were involved in the 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. T. Nakanishi and J.B. Richards are the guarantors.

Conflict of interest: T. Nakanishi has nothing to disclose. A. Cerani has nothing to disclose. V. Forgetta has nothing to disclose. S. Zhou has nothing to disclose. R.J. Allen has nothing to disclose. O.C. Leavy has nothing to disclose. M. Koido has nothing to disclose. D. Assayag reports grants and personal fees for consultancy from Boehringer Ingelheim Canada, personal fees for consultancy from Hoffman-LaRoche Canada, personal fees for lectures from AstraZeneca Canada, outside the submitted work. R.G. Jenkins reports grants from AstraZeneca, Biogen, Galecto and GlaxoSmithKline, personal fees from Boehringer Ingelheim, Daewoong, Galapagos, Heptares, Promedior and Roche, nonfinancial support from NuMedii and Redx, grants and personal fees from Pliant, other (trustee) from Action for Pulmonary Fibrosis, outside the submitted work. L.V. Wain reports grants from GlaxoSmithKline, outside the submitted work. I.V. Yang reports grants from the NIH and personal fees from Eleven P15 related to research in pulmonary fibrosis, outside the submitted work; and has a patent "Circulating biomarkers of preclinical pulmonary fibrosis" pending. G.M. Lathrop has nothing to disclose. P.J. Wolters reports grants and personal fees from Boehringer Ingelheim and Roche/Genentech, personal fees from Gossamer Bio, Blade Therapeutics and Pliant, outside the submitted work. D.A. Schwartz reports grants from the NIH-NHLBI (R38-HL143511, T32-HL007085, UG3/UH3-HL151865, R01-HL149836, P01-HL0928701, UH2/3-HL123442, X01-HL134585 and R25-ES025476) and DOD Focused Program Grant W81XWH-17-1-0597, during the conduct of the study; personal fees from Eleven P15, outside the submitted work; and has a patent "Compositions and methods of treating or preventing fibrotic diseases" pending, a patent "Biomarkers for the diagnosis and treatment of fibrotic lung disease" pending and a patent "Methods and compositions for risk prediction, diagnosis, prognosis, and treatment of pulmonary disorders" issued. J.B. Richards has served as an advisor to GlaxoSmithKline and Deerfield Capital.

Support statement: T. Nakanishi is supported by Research Fellowships of the Japan Society for the Promotion of Science (JSPS) for Young Scientists and JSPS Overseas Challenge Program for Young Researchers. A. Cerani is supported by the Fonds de Recherche Québec Santé (FRQS) and Canadian Institutes of Health Research (CIHR) doctoral awards, and is a Queen Elizabeth Scholar. S. Zhou is supported by a CIHR postdoctoral fellowship. The Richards research group is supported by CIHR (365825; 409511), Lady Davis Institute of the Jewish General Hospital, Canadian Foundation for Innovation, NIH Foundation, Cancer Research UK and FRQS. J.B. Richards is supported by a FRQS Clinical Research Scholarship. Support from Calcul Québec and Compute Canada is acknowledged. TwinsUK is funded by the Wellcome Trust, Medical Research Council, European Union, and 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. R.J. Allen is supported by an Action for Pulmonary Fibrosis Mike Bray fellowship. L.V. Wain holds a GlaxoSmithKline/British Lung Foundation Chair in Respiratory Research. The research was partially supported by the NIHR Leicester Biomedical Research Centre; the views expressed are those of the author(s) and not necessarily those of the NHS, NIHR or Dept of Health. P.J. Wolters received funding from the Nina Ireland Program for Lung Health. D.A. Schwartz is supported by the NIH-NHLBI (R38-HL143511, T32-HL007085, UG3/UH3-HL151865, R01-HL149836, P01-HL0928701, UH2/3-HL123442, X01-HL134585 and R25-ES025476) and DOD Focused Program Grant (W81XWH-17-1-0597). These funding agencies had no role in the design, implementation or interpretation of this study. Funding information for this article has been deposited with the Crossref Funder Registry.

References

1 Richeldi L, Collard HR, Jones MG. Idiopathic pulmonary fibrosis. *Lancet* 2017; 389: 1941–1952.
2 Vancheri C, Failla M, Crimi N, *et al.* 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, *et al.* 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, *et al.* 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* 2019; 74: 506–511.
6 Rosas IO, Richards TJ, Konishi K, *et al.* MMP1 and MMP7 as potential peripheral blood biomarkers in idiopathic pulmonary fibrosis. *PLoS Med* 2008; 5: e93.
7 White ES, Xia M, Murray S, *et al.* 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, *et al.* New serum indicator of interstitial pneumonitis activity. Sialylated carbohydrate antigen KL-6. *Chest* 1989; 96: 68–73.
9 Neighbors M, Cabanski CR, Ramalingam TR, *et al.* 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* 2018; 6: 615–626.
10 Labrecque JA, Swanson SA. Interpretation and potential biases of Mendelian randomisation estimates with time-varying exposures. *Am J Epidemiol* 2019; 188: 231–238.
11 Allen RJ, Guillen-Guio B, Oldham JM, *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, *et al.* Genomic atlas of the human plasma proteome. *Nature* 2018; 558: 73–79.
13 Emilsson V, Ilkov M, Lamb JR, *et al.* Co-regulatory networks of human serum proteins link genetics to disease. *Science* 2018; 361: 769–773.
14 Davey Smith G, Davies NM, Dimou N, *et al.* STROBE-MR: guidelines for strengthening the reporting of Mendelian randomisation studies. *Peer J Preprints* 2019; 7: 27857v.
15 Tam V, Patel N, Turcotte M, *et al.* Benefits and limitations of genome-wide association studies. *Nat Rev Genet* 2019; 20: 467–484.
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 MV, Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ* 2018; 362: k601.
18 Hemani G, Zheng J, Elsworth B, *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *Elife* 2018; 7: e34408.
19 Burgess S, Butterworth A, Thompson SG. Mendelian randomisation analysis with multiple genetic variants using summarized data. *Genet Epidemiol* 2013; 37: 658–665.
20 Giambartolomei C, Vukcevic D, Schadt EE, *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* 2014; 10: e100438.
21 Hormozdiari F, van de Bunt M, Segrè AV, *et al.* Colocalization of GWAS and eQTL signals detects target genes. *Am J Hum Genet* 2016; 99: 1245–1260.
22 Pruim RJ, Welch RP, Sanna S, *et al.* LocusZoom: regional visualisation of genome-wide association scan results. *Bioinformatics* 2010; 26: 2336–2337.
23 Yavorska OO, Burgess S. MendelianRandomisation: an R package for performing Mendelian randomisation analyses using summarized data. *Int J Epidemiol* 2017; 46: 1734–1739.
24 Burgess S, Dudbridge F, Thompson SG. Combining information on multiple instrumental variables in Mendelian randomisation: comparison of allele score and summarized data methods. *Stat Med* 2016; 35: 1880–1906.
25 Staley JR, Blackshaw J, Kamat MA, *et al.* PhenoScanner: a database of human genotype–phenotype associations. *Bioinformatics* 2016; 32: 3207–3209.
26 Kamat MA, Blackshaw JA, Young R, *et al.* PhenoScanner V2: an expanded tool for searching human genotype–phenotype associations. *Bioinformatics* 2019; 35: 4851–4853.
27 Yang IV, Coldren CD, Leach SM, *et al.* Expression of cilium-associated genes defines novel molecular subtypes of idiopathic pulmonary fibrosis. *Thorax* 2013; 68: 1114–1121.
28 Habermann AC, Gutierrez AJ, Bui LT, *et al.* Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis. *Sci Adv* 2020; 6: eaba1972.
29 Adams TS, Schupp JC, Poli S, *et al.* Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. *Sci Adv* 2020; 6: eaba1983.

30 Benner C, Spencer CCA, Havulinna AS, *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* 2016; 32: 1493–1501.
31 Nongmaithem SS, Joglekar CV, Krishnaveni GV, *et al.* 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, *et al.* Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nat Commun* 2018; 9: 3268.
33 He M, Wu C, Xu J, *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* 2014; 63: 143–151.
34 Hemani G, Tilling K, Davey Smith G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS Genet* 2017; 13: e1007081.
35 Todd JL, Neely ML, Overton R, *et al.* Peripheral blood proteomic profiling of idiopathic pulmonary fibrosis biomarkers in the multicentre IPF-PRO Registry. *Respir Res* 2019; 20: 227.
36 Parimon T, Yao C, Stripp BR, *et al.* Alveolar epithelial type II cells as drivers of lung fibrosis in idiopathic pulmonary fibrosis. *Int J Mol Sci* 2020; 21: 2269.
37 Plantier L, Crestani B, Wert SE, *et al.* Ectopic respiratory epithelial cell differentiation in bronchiolised distal airspaces in idiopathic pulmonary fibrosis. *Thorax* 2011; 66: 651–657.
38 Manousaki D, Mokry LE, Ross S, *et al.* Mendelian randomisation 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, *et al.* Vitamin D supplements and prevention of cancer and cardiovascular disease. *N Engl J Med* 2019; 380: 33–44.
40 Holmes MV, Smith GD. Dyslipidaemia: revealing the effect of CETP inhibition in cardiovascular disease. *Nat Rev Cardiol* 2017; 14: 635–636.
41 Holmes MV, Richardson TG, Ference BA, *et al.* Integrating genomics with biomarkers and therapeutic targets to invigorate cardiovascular drug development. *Nat Rev Cardiol* 2021; 18: 435–453.
42 Landray M. Tocilizumab in patients admitted to hospital with COVID-19 (RECOVERY): preliminary results of a randomised, controlled, open-label, platform trial. *Lancet* 2021; 397: 1637–1645.
43 Larsson SC, Burgess S, Gill D. Genetically proxied interleukin-6 receptor inhibition: opposing associations with COVID-19 and pneumonia. *Eur Respir J* 2021: 57: 2003545.
44 Fanidi A, Carreras-Torres R, Larose TL, *et al.* Is high vitamin B12 status a cause of lung cancer? *Int J Cancer* 2019; 145: 1499–1503.
45 Mokry LE, Ahmad O, Forgetta V, *et al.* 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, *et al.* Desmoplakin variants are associated with idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2016; 193: 1151–1160.
47 Moore C, Blumhagen RZ, Yang IV, *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* 2019; 200: 199–208.
48 Nance T, Smith KS, Anaya V, *et al.* Transcriptome analysis reveals differential splicing events in IPF lung tissue. *PLoS One* 2014; 9: e92111.
49 Dupuy F, Germot A, Marenda M, *et al.* 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* 2015; 1850: 236–252.
52 Janssen WJ, Stefanski AL, Bochner BS, *et al.* Control of lung defence by mucins and macrophages: ancient defence mechanisms with modern functions. *Eur Respir J* 2016; 48: 1201–1214.
53 de Mattos LC. Structural diversity and biological importance of ABO, H, Lewis and secretor histo-blood group carbohydrates *Rev Bras Hematol Hemoter* 2016; 38: 331–340.
54 Kerr SC, Fischer GJ, Sinha M, *et al.* FleA expression in *Aspergillus fumigatus* is recognized by fucosylated structures on mucins and macrophages to prevent lung infection. *PLoS Pathogens* 2016; 12: e1005555.
55 Seibold MA, Wise AL, Speer MC, *et al.* A common MUC5B promoter polymorphism and pulmonary fibrosis. *N Engl J Med* 2011; 364: 1503–1512.
56 Hancock LA, Hennessy CE, Solomon GM, *et al.* Muc5b overexpression causes mucociliary dysfunction and enhances lung fibrosis in mice. *Nat Commun* 2018; 9: 5363.
57 Maher TM, Oballa E, Simpson JK, *et al.* An epithelial biomarker signature for idiopathic pulmonary fibrosis: an analysis from the multicentre PROFILE cohort study. *Lancet Respir Med* 2017; 5: 946–955.
58 Kawai S, Suzuki K, Nishio K, *et al.* 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, *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, *et al.* The significance of elevated tumor markers among patients with idiopathic pulmonary fibrosis before and after lung transplantation. *Chest* 2012; 141: 1047–1054.
61 Raffield LM, Dang H, Pratte KA, *et al.* Comparison of proteomic assessment methods in multiple cohort studies. *Proteomics* 2020; 20: 1900278.