ORIGINAL ARTICLE
INTERSTITIAL LUNG DISEASE

Novel idiopathic pulmonary fibrosis susceptibility variants revealed by deep sequencing

Jose M. Lorenzo-Salazar[1,10], Shwu-Fan Ma[2,10], Jonathan Jou[3,10], Pei-Chi Hou[2], Beatriz Guillen-Guio[4], Richard J. Allen[5], R. Gisli Jenkins[6], Louise V. Wain[5,7], Justin M. Oldham[8], Imre Noth[2] and Carlos Flores[1,4,9]

Affiliations: [1]Genomics Division, Instituto Tecnológico y de Energías Renovables (ITER), Santa Cruz de Tenerife, Spain. [2]Division of Pulmonary and Critical Care Medicine, University of Virginia, Charlottesville, VA, USA. [3]College of Medicine, University of Illinois, Chicago, IL, USA. [4]Research Unit, Hospital Universitario N.S. de Candelaria, Universidad de La Laguna, Santa Cruz de Tenerife, Spain. [5]Dept of Health Sciences, University of Leicester, Leicester, UK. [6]NIHR Biomedical Research Centre, Respiratory Research Unit, University of Nottingham, Nottingham, UK. [7]National Institute for Health Research, Leicester Respiratory Biomedical Research Centre, Glenfield Hospital, Leicester, UK. [8]Pulmonary and Critical Care Medicine, University of California at Davis, Sacramento, CA, USA. [9]CIBER de Enfermedades Respiratorias, Instituto de Salud Carlos III, Madrid, Spain. [10]These authors contributed equally to this work.

Correspondence: Carlos Flores, Unidad de Investigación, Hospital Universitario N.S. de Candelaria, Carretera del Rosario s/n, 38010 Santa Cruz de Tenerife, Spain. E-mail: cflores@ull.edu.es

ABSTRACT

Background: Specific common and rare single nucleotide variants (SNVs) increase the likelihood of developing sporadic idiopathic pulmonary fibrosis (IPF). We performed target-enriched sequencing on three loci previously identified by a genome-wide association study to gain a deeper understanding of the full spectrum of IPF genetic risk and performed a two-stage case–control association study.

Methods: A total of 1.7 Mb of DNA from 181 IPF patients was deep sequenced (>100×) across 11p15.5, 14q21.3 and 17q21.31 loci. Comparisons were performed against 501 unrelated controls and replication studies were assessed in 3968 subjects.

Results: 36 SNVs were associated with IPF susceptibility in the discovery stage ($p<5.0\times10^{-8}$). After meta-analysis, the strongest association corresponded to rs35705950 ($p=9.27\times10^{-57}$) located upstream from the mucin 5B gene (*MUC5B*). Additionally, a novel association was found for two co-inherited low-frequency SNVs (<5%) in *MUC5AC*, predicting a missense amino acid change in mucin 5AC (lowest $p=2.27\times10^{-22}$). Conditional and haplotype analyses in 11p15.5 supported the existence of an additional contribution of *MUC5AC* variants to IPF risk.

Conclusions: This study reinforces the significant IPF associations of these loci and implicates *MUC5AC* as another key player in IPF susceptibility.

@ERSpublications
Deep sequencing of genome-wide association study hits identified novel low-frequency variants associated with IPF susceptibility. http://bit.ly/2IF4AT8

Cite this article as: Lorenzo-Salazar JM, Ma S-F, Jou J, *et al*. Novel idiopathic pulmonary fibrosis susceptibility variants revealed by deep sequencing. *ERJ Open Res* 2019; 5: 00071-2019 [https://doi.org/10.1183/23120541.00071-2019].

This article has supplementary material available from openres.ersjournals.com

Received: March 18 2019 | Accepted: April 07 2019

Introduction

Idiopathic pulmonary fibrosis (IPF), a devastating interstitial lung disease with unknown aetiology, encompasses a highly heterogeneous and unpredictable clinical course [1]. IPF remains an incurable condition, although lung transplant can improve long-term survival [2] and two antifibrotic therapies are effective at slowing disease progression [3, 4]. Identifying genetic risk factors will allow a better understanding of the causative molecular pathways involved in disease pathogenesis and guide novel therapeutic approaches to support the development of precision medicine approaches in IPF.

The existence of a familial form of pulmonary fibrosis (FPF) and the recognition that pulmonary fibrosis occurs in several rare genetic disorders strongly suggest that genetic factors influence susceptibility and prognosis [5]. Rare variants in surfactant-encoding genes (*SFTPC* and *SFTPA2*) and telomere integrity genes (*TERT*, *TERC*, *RTEL1* and *PARN*) are associated with both FPF and IPF [6–11]. Additionally, common single nucleotide variants (SNVs) predicting risk of sporadic IPF have been identified at 17 independent loci by means of several genome-wide scale studies [12–16]. Reviewed in detail elsewhere [17], variants at these loci in aggregate currently explain roughly 25–30% of the disease risk, and support a major role of telomere maintenance, cell adhesion/wound healing, fibrogenic and immunity/host defence pathways in IPF development. In conducting one of the largest genome-wide association studies (GWASs) in IPF, our group [12] confirmed that the common SNV rs35705950 of *MUC5B* is the strongest known risk factor for the disease [12–16], and identified additional novel common susceptibility SNVs with milder effects in the genes encoding Toll-interacting protein (*TOLLIP*, 11p15.5) and signal peptide peptidase like 2C (*SPPL2C*, 17q21.31). One of the most striking results of this study was that it revealed allelic heterogeneity in 11p15.5 given the existence of replicable common independent risk SNVs in *MUC5B* and *TOLLIP*, and possibly other nearby genes. Additionally, a fourth locus involving the gene encoding MAM domain containing glycosylphosphatidylinositol anchor 2 (*MDGA2*, 14q21.3) reached genome-wide significance in the second stage of our study, but could not be replicated in a third case–control sample of our study [12] nor in other GWASs conducted to date. Because incomplete overlap of results across distinct GWASs is common, and since *MDGA2* is a paralogue for a potential biomarker of IPF disease activity [18], this locus remains of potential importance.

As progress has been made in identifying susceptibility loci across many diseases, it is increasingly being shown that multiple nearby, but independent, signals often underlie strong susceptibility loci [19]. This observation, along with increasingly available and affordable high-throughput sequencing technologies, provides a valuable opportunity to better characterise previously identified risk loci. Here, we use a fine mapping approach based on target-enriched DNA sequencing to assess the full spectrum of variants in three IPF-associated genomic loci previously identified in our GWAS.

Materials and methods

Institutional review boards and ethics committees at participating centres approved the study. All participants provided written informed consent (see supplementary methods).

Discovery study

Study subjects

A total of 181 IPF subjects were obtained from the University of Chicago Natural History study (n=138), the Correlating Outcomes with biochemical Markers to Estimate Time-progression study (COMET; n=22) and the AntiCoagulant Effectiveness in IPF study (ACE; n=21). The majority of these patients (60.8%) overlapped with those used for the discovery stage in our previous GWAS [12]. However, for this study, we prioritised the cases based on the existence of sufficient DNA quantity for the targeted next-generation sequencing (NGS) experiments and high DNA integrity. Subjects were European-Americans, had an average age of 67 years at diagnosis, and respiratory symptoms including dyspnoea on exertion and/or cough for at least 3 months. A high-resolution computed tomography scan with a probable or definite usual interstitial pneumonia (UIP) pattern was required according to published diagnostic guidelines [2]. A surgical lung biopsy was obtained in 37.3% of patients, all confirming UIP. None of them had a clinically significant exposure to known fibrogenic agents or suffered from other known causes of interstitial lung disease. Patient details are listed in table 1.

Sequencing, variant calling, validation and association testing

Sequencing (>100× depth) and variant calling was performed in regions of interest (ROIs) spanning 1.7 Mb (supplementary table S1 and supplementary methods). The dataset obtained from the 181 cases was used for a case–control association study, where unrelated European individuals from the 1000 Genomes Project (1KGP; www.internationalgenome.org) were used as controls (n=501; release May 2, 2013). Single-variant association tests are typically underpowered for rare variants [20]. However, given the previous reported large effect for some of the variants in IPF [14] and the design of the study, we were

TABLE 1 Clinical and demographic characteristics of idiopathic pulmonary fibrosis cases included in the discovery study

	University of Chicago	COMET	ACE	p-value
Subjects	138	22	21	
Age years	69±9	63±8	69±6	0.02
Male	108 (78.3)	14 (63.6)	16 (76.2)	0.46
Ever-smoker	95 (75.4)	17 (77.3)	15 (71.4)	0.91
FVC % pred	67.2±16.7	72.1±12.9	53.9±15.4	8.0×10^{-4}
***D*LCO % pred**	47.8±16.8	46.4±14.2	33.1±19.5	1.7×10^{-3}
Transplant	12 (8.7)	0 (0)		0.17
Death	68 (49.3)	1 (4.5)	3 (14.3)	2.7×10^{-5}
Follow-up months	38.5±24.8	36.3±2.8	35.1±4.0	0.41
Time to death months	27.8±17.9	9.57#	4.0±2.9	0.06

Data are presented as n, mean±SD or n (%), unless otherwise stated. FVC: forced vital capacity; *D*LCO: diffusing capacity of the lung for carbon monoxide. #: mean.

interested in identifying variants with large, similar effect sizes within ROIs and not in delineating the most likely causal gene(s). Therefore, association testing was performed individually for each SNV. Effect sizes (odds ratios) and 95% confidence intervals were assessed with PLINK version 1.07 (http://zzz.bwh.harvard.edu/plink) under logistic regression models for biallelic loci with call rates >95%. Principal components were derived with Eigensoft version 6.0.1 [21] using a subset of 2342 variants with reduced linkage disequilibrium ($r^2<0.15$). The first two principal components were used to project the genetic ancestry of cases in the 1KGP dataset for visual inspection of the clustering. In addition, the first five principal components were included in the regression models to account for the population stratification and no evidence for inflation of the association results was observed ($\lambda=1.00$). Variants were annotated according to the minor allele frequency (MAF) in 1KGP, classifying them in two tiers (common/low frequency) based on a 5% threshold in controls. Significantly associated low-frequency variants were subjected to validation by Sanger sequencing (supplementary table S2 and supplementary methods).

Conditional and haplotype analyses in 11p15.5

Including the newly identified risk variants from this study, the top hits for this locus have been described in three mucin-encoding genes (*MUC2*, *MUC5AC* and *MUC5B*) and the *TOLLIP* gene. However, as the top hit of *MUC2* (rs7934606) [16, 22] falls outside of the ROI targeted by our NGS experiments, seven risk variants from three genes were included in final analyses: 1) rs34474233 and rs34815853, the two tightly linked variants of *MUC5AC* identified in the current study; 2) rs12802931 (from this study) and rs35705950 [12–14, 16, 22], the two 5′-flanking variants of *MUC5B*; and 3) rs111521887, rs5743894 and rs5743890, the three GWAS hits mapping near or within *TOLLIP* [12]. A formal conditional analysis taking the linkage disequilibrium structure of 11p15.5 into account was applied using the GCTA-COJO method [23], conditioning the risk variants to rs35705950 of *MUC5B*. In addition, haplotype associations were conducted in PLINK comparing the frequency of combinations of the risk variants between cases and controls with logistic regressions adjusted for five principal components. Combinations with frequencies >1% were reconstructed from all seven variants together and from variants from each of the gene pairs. Statistical significance was set at $p<2.0\times10^{-3}$ after a Bonferroni correction considering all haplotypes tested.

Replication study and meta-analysis of results

Replication was assessed in data from a study consisting of 602 IPF cases and 3366 UK Biobank controls as described by ALLEN *et al.* [16] (see supplementary methods for additional information). Random effects meta-analysis was performed with METASOFT version 2.0.1 [24] to estimate the overall effect size of associated SNVs across the discovery and replication studies. Replication was declared for risk variants satisfying the same direction of effects as in the discovery study, with $p<0.0014$ in the replication stage (corresponding to a Bonferroni-like correction threshold of 0.05/36) and with a genome-wide significant association ($p<5\times10^{-8}$) in the meta-analysis of both stages.

Results

Quality control of called variants in the discovery study

A total of 18 234 variants (13 932 SNVs and 4302 indels) were identified among IPF samples. The Ti/Tv ratio (*i.e.* the ratio of numbers of transitions *versus* transversions) was 2.192, within the range of expected ratios for whole genomes (*i.e.* 2.1–2.3) [25, 26]. This is not unexpected as a large fraction of the ROIs are

nonexonic sequences. Based on this, we inferred a false discovery rate (FDR) of 3.3%. We also evaluated the MAF and concordance of genotypes of called variants from NGS with those from the array data of our GWAS [12]. For the 231 variants that had genotype data in both datasets, MAFs showed a near-perfect linear correlation (Pearson correlation R^2=0.998) and genotype concordance was 96.1% (95% CI 95.9–96.4%). The genotype discrepancies between the array and NGS were attributed to missing genotypes on the array and the FDR rate estimates. Association testing in the discovery study was conducted in the subset of 10 245 biallelic variants and had genotypes in >95% individuals, which implies a 86.9% overlap of imputed variants with our previous GWAS assessment in these loci [12]. Genetic ancestry projections of cases and 1KGP samples from all continents based on a subset of the biallelic variants demonstrated clustering of the patients with Europeans (supplementary figure S1), supporting their recorded ethnicity.

Association in the discovery study

36 variants reached genome-wide significance (nine in 11p15.5, 17 in 14q21.3 and 10 in 17q21.31). Only 14% of these variants were assessed in our previous GWAS [12]. Most of them were located in introns or flanking regions (table 2 and figure 1). The strongest signals corresponded to rs35705950 within *MUC5B* in 11p15.5 (MAF 10.8% in controls; p=2.69×10^{-22}), rs12586854 within *MDGA2* in 14q21.3 (MAF 42.8%

TABLE 2 Association results reaching genome-wide significance in the discovery study

SNV	Chr.	Position (hg19)	Effect allele	MAF[#]	OR (95% CI)	p-value	Nearby gene	Function/location
rs371630624	11p15.5	1 213 302	C	*0.001*	1942 (245.6–15 360)	7.18×10^{-13}	*MUC5AC*	Synonymous
rs34474233[¶]	11p15.5	1 219 152	A	*0.044*	4.08 (2.56–6.49)	2.99×10^{-9}	*MUC5AC*	Missense (Ala5353Lys)
rs34815853[¶]	11p15.5	1 219 153	A	*0.044*	4.01 (2.52–6.38)	4.15×10^{-9}	*MUC5AC*	Missense (Ala5353Lys)
rs12802931	11p15.5	1 236 164	G	0.183	3.76 (2.73–5.16)	3.72×10^{-16}	*MUC5B*	8.1 kb 5′ of *MUC5B*
rs35705950	11p15.5	1 241 221	T	0.108	6.18 (4.28–8.93)	2.69×10^{-22}	*MUC5B*	3.1 kb 5′ of *MUC5B*
rs200243273	11p15.5	1 266 716	C	0.227	0.27 (0.17–0.43)	3.55×10^{-8}	*MUC5B*/ RP11-532E4.2	Missense/intronic
rs4963073	11p15.5	1 362 949	G	0.300	3.23 (2.12–4.921)	4.91×10^{-8}	CTD-2245O6.1	31 kb 3′ of CTD-2245O6.1
rs4963072	11p15.5	1 362 953	G	0.300	3.34 (2.19–5.11)	2.63×10^{-8}	CTD-2245O6.1	31 kb 3′ of CTD-2245O6.1
rs71469892	11p15.5	1 416 119	G	0.491	0.22 (0.15–0.31)	2.15×10^{-16}	*BRSK2*	Intronic
rs145898170	14q21.3	47 574 913	G	0.458	0.47 (0.36–0.62)	4.71×10^{-8}	*MDGA2*	Intronic
rs199838022	14q21.3	47 574 922	C	0.458	0.45 (0.34–0.57)	7.14×10^{-9}	*MDGA2*	Intronic
rs12586854	14q21.3	47 576 151	T	0.428	0.18 (0.12–0.26)	6.81×10^{-19}	*MDGA2*	Intronic
rs11157543	14q21.3	47 576 203	C	0.300	0.13 (0.07–0.22)	5.97×10^{-14}	*MDGA2*	Intronic
rs11157544	14q21.3	47 576 205	C	0.427	0.30 (0.21–0.41)	1.37×10^{-13}	*MDGA2*	Intronic
rs12586856	14q21.3	47 576 217	G	0.304	0.18 (0.11–0.29)	3.77×10^{-13}	*MDGA2*	Intronic
rs11157545	14q21.3	47 576 231	T	0.463	0.44 (0.34–0.59)	7.81×10^{-9}	*MDGA2*	Intronic
rs183643415	14q21.3	47 576 246	A	0.182	0.04 (0.01–0.12)	2.77×10^{-8}	*MDGA2*	Intronic
rs150322840	14q21.3	47 576 252	T	0.216	0.13 (0.07–0.24)	3.90×10^{-10}	*MDGA2*	Intronic
rs8005465	14q21.3	47 716 040	A	0.461	0.37 (0.27–0.51)	4.41×10^{-10}	*MDGA2*	Intronic
rs543453148	14q21.3	47 751 911	A	*0.004*	25.22 (8.29–76.73)	1.30×10^{-8}	*MDGA2*	Intronic
rs12890180	14q21.3	47 788 012	G	0.393	0.39 (0.28–0.53)	2.91×10^{-9}	*MDGA2*	Intronic
rs73251857	14q21.3	47 800 734	G	0.154	0.06 (0.02–0.14)	9.59×10^{-10}	*MDGA2*	Intronic
rs7141653	14q21.3	47 828 946	C	0.363	0.25 (0.17–0.37)	1.65×10^{-11}	*MDGA2*	Intronic
rs7145329	14q21.3	47 931 577	T	0.376	0.34 (0.24–0.49)	2.59×10^{-9}	*MDGA2*	Intronic
rs4900770	14q21.3	47 938 755	A	0.498	0.38 (0.28–0.53)	9.10×10^{-9}	*MDGA2*	Intronic
rs58731325	14q21.3	48 009 745	G	0.469	0.34 (0.25–0.47)	7.35×10^{-11}	*MDGA2*	Noncoding transcript/intronic
rs115811519	17q21.31	43 677 790	C	0.070	4.93 (2.83–8.58)	1.68×10^{-8}	RP11-707O23.1	7 kb 3′ of RP11-707O23.1
rs56383763	17q21.31	43 682 323	C	0.242	0.07 (0.03–0.16)	1.75×10^{-9}	CTC-501O10.1	17 kb 5′ of *CRHR1*
rs373417	17q21.31	43 691 173	T	0.239	0.10 (0.05–0.20)	1.24×10^{-10}	*CRHR1*	6.5 kb 5′ of *CRHR1*
rs7221124	17q21.31	43 764 301	A	0.265	0.04 (0.02–0.09)	6.70×10^{-14}	*CRHR1*	Intronic
rs55938136	17q21.31	43 798 360	A	*0.018*	151.90 (62.14–371.50)	3.37×10^{-28}	*CRHR1*	Intronic
rs11870844	17q21.31	44 141 279	A	0.257	3.98 (2.82–5.62)	3.85×10^{-15}	*KANSL1*	Intronic
rs371996525	17q21.31	44 183 317	A	0.244	0.04 (0.02–0.11)	2.17×10^{-10}	*KANSL1*	Intronic
rs142920272	17q21.31	44 301 840	C	0.248	0.10 (0.05–0.20)	7.45×10^{-11}	*KANSL1*	Intronic
rs2668637	17q21.31	44 322 960	G	0.095	5.32 (3.09–9.13)	1.43×10^{-9}	*KANSL1/LRRC37A*	Intergenic
rs2696618	17q21.31	44 325 635	C	0.249	6.74 (4.02–11.31)	5.09×10^{-13}	*KANSL1/LRRC37A*	23 kb 5′ of *KANSL1*

SNV: single nucleotide variant; Chr.: chromosome; MAF: minor allele frequency. [#]: MAF in Europeans from the 1000 Genomes Project (low-frequency variants in italic); [¶]: because of their complete linkage disequilibrium, these variants can be merged into rs71464134. The functional information provided corresponds to the predicted change for the merged reference sequence.

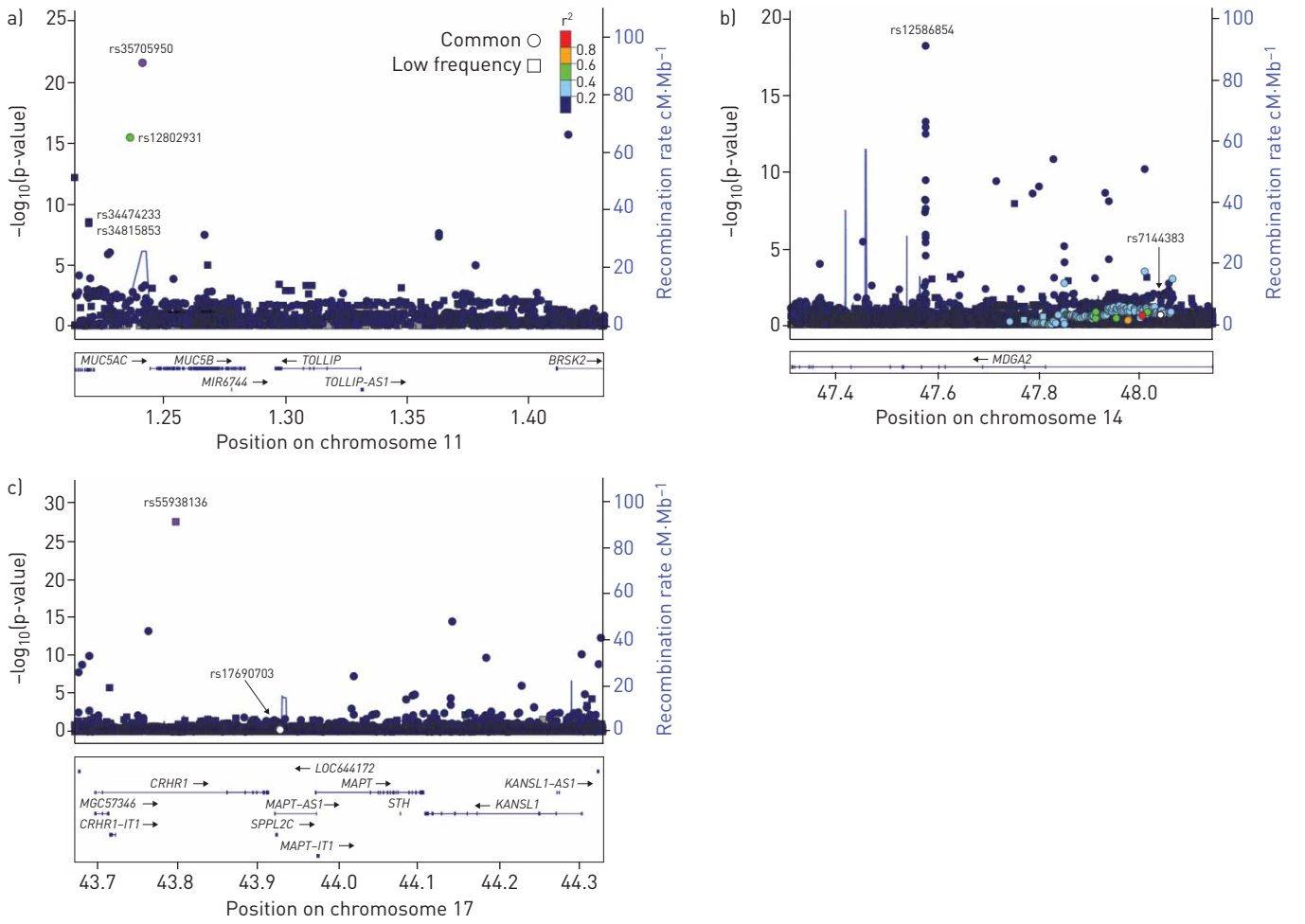

FIGURE 1 Regional association plots of a) 11p15.5, b) 14q21.3 and c) 17q21.31 with annotations of previously detected signals (rs35705950 in chromosome 11, rs7144383 in chromosome 14 and rs17690703 in chromosome 17). Chromosomal position is shown in Mb. Significance is represented on a $-\log_{10}$(p-value) scale. A threshold minor allele frequency in controls of 0.05 was used to stratify the results derived by common *versus* low-frequency variants. Colours reflect linkage disequilibrium (r^2) values against the top hit on each region according to the European population data from the 1000 Genomes Project.

in controls; $p=6.81\times10^{-19}$) and rs55938136 within *CRHR1* in 17q21.31 (MAF 1.8% in controls; $p=3.37\times10^{-28}$). Besides, another variant of *MUC5B* located ~8.1 kb away from the 5′ region of the gene was also strongly associated with IPF (rs12802931: MAF 18.3%; $p=3.72\times10^{-16}$), although it was not independent from rs35705950 (p=0.731 conditioning on rs35705950). Strikingly, three coding low-frequency variants of *MUC5AC* were among the significant findings ($p\leqslant4.15\times10^{-9}$): one with a synonymous prediction (rs371630624) and two others (rs34815853 and rs34474233) affecting the same codon leading to a missense amino acid change (p.Ala5353Lys: MAF 4.4% in controls) that was supported by the sequencing results (figure 2). Individually, they are predicted by PolyPhen (http://genetics.bwh.harvard.edu/pph2) to be benign (rs34815853) and possibly damaging (rs34474233), but the simultaneous effects of the two are unknown. Orthogonal validation by Sanger sequencing strongly supported that the two missense variants were true positives (figure 2); however, it did not support the existence of the variant with synonymous prediction (*i.e.* false positive). Besides these three, only two other low-frequency variants from 14q21.3 (rs543453148: MAF 0.4% in controls) and 17q21.31 (rs55938136: MAF 1.8% in controls) were significantly associated with IPF. Sanger sequencing of rs543453148 suggested the existence of variation but with alleles that were unaligned to those recorded by NGS. Sanger results were fully congruent with the NGS for rs55938136. In the context of our previous results [12], while several other SNVs reached genome-wide significance in 11p15.5, none of the three *TOLLIP* risk variants (rs111521887, rs5743894 and rs5743890) previously evidenced were significant in this study (figure 1). As for 14q21.3 and 17q21.31, none of the two top hits reported before were nominally significant in this study (rs7144383: p=0.181; rs17690703: p=0.639). The SNV at rs4898572, an intronic variant in strong linkage disequilibrium with rs7144383 in *MDGA2*, was also not significant in this study (p=0.191).

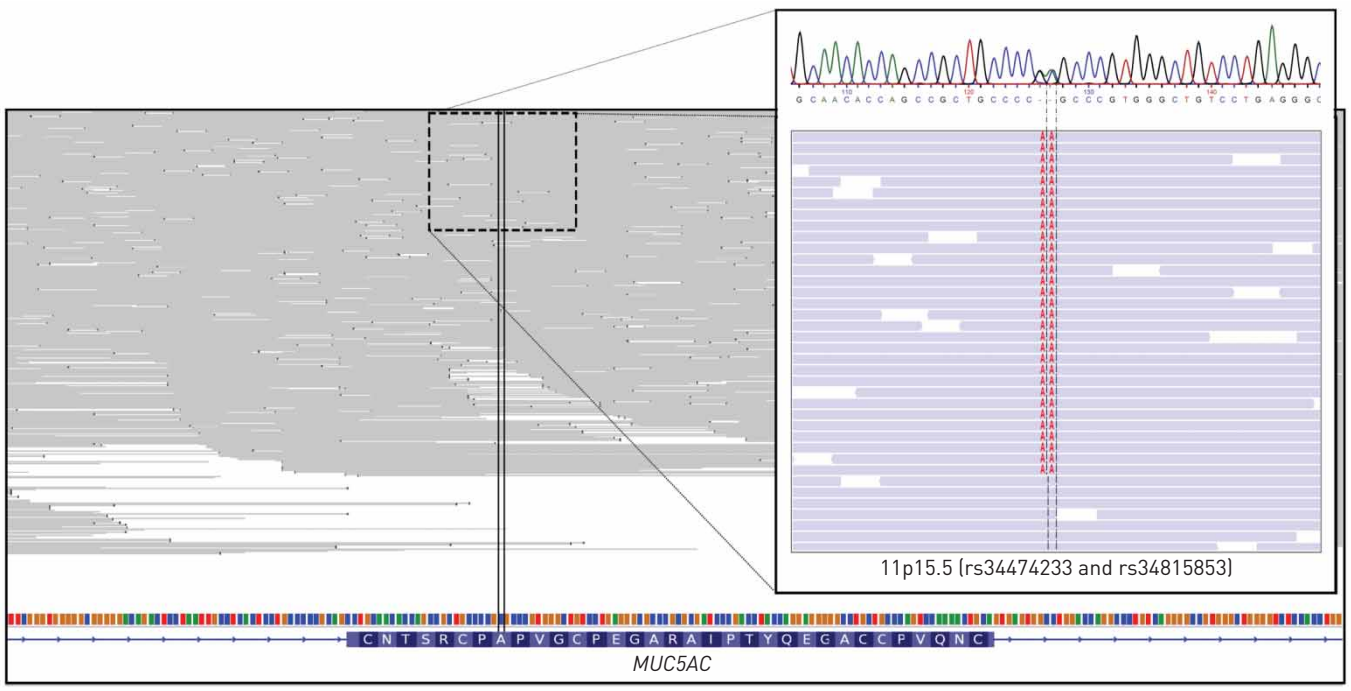

FIGURE 2 Detailed pile-up view of sequence reads mapping and Sanger sequencing results of the two *MUC5AC* variants affecting the missense change.

Elucidating distinctive gene contributions in the 11p15.5 region

Previous evidence highlights the importance of the 11p15.5 region harbouring mucin genes and *TOLLIP* in IPF susceptibility and mortality [12–16]. Given the novel finding of hits in *MUC5AC*, we performed further association analyses focusing on the topmost significant risk variant combinations from this region. In total, seven risk variants showing variable linkage disequilibrium relationships in the discovery study (figure 3) resided in the 11p15.5 captured by NGS experiments. These variants from three genes (*MU5AC*, *MUC5B* and *TOLLIP*) were used to reconstruct the 25 most common haplotypes as a result of distinct gene combinations (supplementary table S3). 12 of these were associated with IPF irrespective of the model adjustments, eight of them with statistically significant risk effects. Among all risk combinations, those defined by *MUC5AC* together with *MUC5B* variants showed the largest effect (OR 6.44; $p \leqslant 1.3\times10^{-11}$), while intermediate ORs in the range of 3.39–4.03 ($p \leqslant 1.8\times10^{-4}$) were generally found for combinations containing variants from each of these two genes separately. A formal association test of any of the genome-wide significant 11p15.5 variants in the discovery study conditioned to rs35705950 of *MUC5B* resulted in attenuation of all the signals (table 3). However, they remained nominally significant for the two *MUC5AC* variants (rs34474233 and rs34815853: $p \leqslant 6.27\times10^{-3}$), suggesting an additional contribution to IPF risk. The haplotypes of any of the *TOLLIP* risk variants with those from *MUC5AC* and/or *MUC5B* had no evident effects in terms of the odds ratios or significance.

Replication study and meta-analysis of results

Of the 36 variants that reached genome-wide significance in the discovery study, 10 variants had nominal significance in the replication study, had the same direction of effects as in the discovery study and resulted in a meta-analysis $p<5\times10^{-8}$: five were located on 11p15.5 and the remaining five on 17q21.31 (table 4). However, only four of them reached the adjusted significance threshold ($p<1.4\times10^{-3}$) in the replication study, all corresponding to *MUC5AC* and *MUC5B* genes. Replication was not supported for the 14q21.3 variants. In meta-analysis the most significant findings were those corresponding to *MUC5B*: rs35705950 (OR 4.90, 95% CI 3.30–7.28; $p=9.27\times10^{-57}$) and the linkage disequilibrium proxy rs12802931 (OR 2.96, 95% CI 1.93–4.53; $p=4.60\times10^{-35}$). Most importantly, these results strongly supported the association of the two *MUC5AC* variants rs34474233 (OR 3.39, 95% CI 2.65–4.32; $p=2.27\times10^{-22}$) and rs34815853 (OR 3.37, 95% CI 2.64–4.30; $p=3.02\times10^{-22}$) predicting a missense change in the protein.

Discussion

In recent years, there has been growing evidence that genetic factors play an important role in IPF. However, a large fraction of genetic risk remains unexplained [22]. Here, we screened 1.7 Mb from three

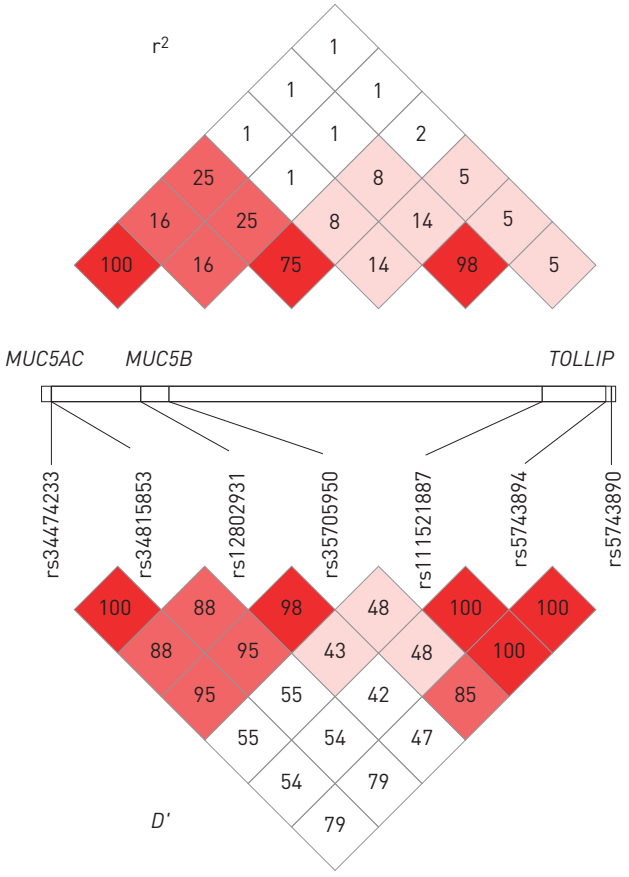

FIGURE 3 Linkage disequilibrium plot of r^2 and *D'* estimates in the discovery study for the risk variants in *MUC5AC*, *MUC5B* and *TOLLIP*. Each diamond of the linkage disequilibrium plot represents a pairwise comparison, with its values schematically symbolised by a colour gradient, ranging from red (stronger linkage disequilibrium) to white (reduced linkage disequilibrium).

loci of interest and tested association for IPF susceptibility in two stages comprising 4650 unrelated subjects. The initial stage identified 36 variants (average MAF 26.6% in controls) that reached genome-wide significance. Three of these constituted validated low-frequency SNVs (<5%), suggesting a minor impact of low-frequency variants in IPF susceptibility in these loci. By locus, the top signals at 11p15.5 reinforced that the strongest risk corresponds to the previously described *MUC5B* promoter variant rs35705950. Besides this, two tightly linked low-frequency SNVs at 11p15.5 (rs34474233 and rs34815853) that predicted the p.Ala5353Lys amino acid change in *MUC5AC* were associated with IPF for the first time. Studies conditioned on rs35705950 of *MUC5B* and replication of results in independent case–control samples further supported that the *MUC5AC* p.Ala5353Lys change has an additional contribution to IPF risk. Regarding the results of 14q21.3 and 17q21.31, we observed no evidence of

TABLE 3 Association results of 11p15.5 with or without conditioning on rs35705950

Nearby gene	Function/location	SNV	Unconditioned p-value	Conditioned p-value
MUC5AC	Missense (Ala5353Lys)	rs34474233#	2.99×10^{-9}	4.12×10^{-3}
MUC5AC	Missense (Ala5353Lys)	rs34815853#	4.15×10^{-9}	6.27×10^{-3}
MUC5B	8.1 kb 5′ of *MUC5B*	rs12802931	3.72×10^{-16}	0.731
***MUC5B*/RP11-532E4.2**	Missense/intronic	rs200243273	3.55×10^{-8}	1.44×10^{-4}
CTD-2245O6.1	31 kb 3′ of CTD-2245O6.1	rs4963073	4.91×10^{-8}	1.48×10^{-6}
CTD-2245O6.1	31 kb 3′ of CTD-2245O6.1	rs4963072	2.63×10^{-8}	2.66×10^{-6}
BRSK2	Intronic	rs71469892	2.15×10^{-16}	1.29×10^{-9}

The rs371630624 variant at *MUC5AC* was excluded from this analysis as it was not supported by Sanger sequencing. #: these variants can be merged into rs71464134.

TABLE 4 Variants showing nominal significance in the replication study, with the same direction of effects as in the discovery study and that met the genome-wide significance level in the meta-analysis

SNV	Chr.	Position (hg19)	Gene	Effect/ noneffect allele	MAF	Discovery		Replication		Meta-analysis	
						OR (95% CI)	p-value	OR (95% CI)	p-value	OR (95% CI)	p-value
rs34474233	11	1 219 152	*MUC5AC*	A/G	0.044	4.08 (2.56–6.49)	2.99×10^{-9}	3.15 (2.37–4.20)	4.10×10^{-14}	3.39 (2.65–4.32)	2.27×10^{-22}
rs34815853	11	1 219 153	*MUC5AC*	A/C	0.044	4.01 (2.53–6.37)	4.15×10^{-9}	3.16 (2.37–4.20)	4.13×10^{-14}	3.37 (2.64–4.30)	3.02×10^{-22}
rs12802931	11	1 236 164	*MUC5B*	G/A	0.183	3.76 (2.73–5.16)	3.72×10^{-16}	2.42 (2.02–2.90)	6.07×10^{-22}	2.96 (1.93–4.53)	4.60×10^{-35}
rs35705950	11	1 241 221	*MUC5B*	T/G	0.108	6.18 (4.28–8.94)	2.69×10^{-22}	4.11 (3.31–5.11)	1.86×10^{-37}	4.90 (3.30–7.28)	9.27×10^{-57}
rs4963072	11	1 362 953	CTD-2245O6.1	G/C	0.300	3.34 (2.18–5.11)	2.63×10^{-8}	1.29 (1.08–1.54)	5.30×10^{-3}	1.59 (0.38–6.65)	4.91×10^{-8}
rs56383763	17	43 682 323	CTC-501O10.1	C/T	0.242	0.07 (0.03–0.16)	1.75×10^{-9}	0.82 (0.68–0.97)	2.42×10^{-2}	0.24 (0.02–2.82)	2.13×10^{-8}
rs373417	17	43 691 173	*CRHR1*	T/C	0.239	0.10 (0.05–0.20)	1.24×10^{-10}	0.82 (0.69–0.98)	2.72×10^{-2}	0.29 (0.04–2.36)	1.59×10^{-9}
rs371996525	17	44 183 317	*KANSL1*	A/C	0.244	0.04 (0.02–0.11)	2.17×10^{-10}	0.80 (0.67–0.95)	1.26×10^{-2}	0.19 (0.01–3.44)	1.98×10^{-9}
rs142920272	17	44 301 840	*KANSL1*	C/T	0.248	0.10 (0.05–0.20)	7.45×10^{-11}	0.83 (0.69–0.98)	3.07×10^{-2}	0.29 (0.04–2.35)	1.11×10^{-9}
rs2696618	17	44 325 635	*KANSL1/LRRC37A*	C/G	0.249	6.74 (4.02–11.31)	5.09×10^{-13}	1.25 (1.05–1.49)	1.05×10^{-2}	2.28 (0.28–18.51)	4.40×10^{-12}

SNV: single nucleotide variant; Chr.: chromosome; MAF: minor allele frequency.

replication based on an independent study with larger sample size. Besides this, none of the two top hits that have been linked to *TOLLIP* in the literature [12] were nominally significant in the discovery study despite the large overlap of samples (60.8%). This would have been determined by the statistical power of the discovery, as it was <35% for detecting the reported effects of these variants (not shown).

Thousands of genetic variants have been reported for association with complex traits [27], the majority being frequent in the population (>5%) [28]. The contribution of low-frequency variants (<5%) in diseases such as those affecting blood lipid levels and cardiovascular disease [29], among others, has just started to be unearthed, facilitated by exome and whole-genome sequencing. There are only a few GWASs of IPF completed so far, all showing risk loci linked to common variants (MAF 11–54% for European ancestry populations) [6, 12, 16, 22]. Besides rs35705950 of *MUC5B*, which has a strong effect in the disease [12–16], other common SNVs associated thus far have milder effects with regard to IPF risk. One of the possibilities underlying these GWAS signals is the existence of underlying low-frequency variants with strong disease effects, as has been recently demonstrated for other well-known IPF genes by exome sequencing experiments (*TERT*, *RTEL1* and *PARN*) [30]. In that scenario, such variation would be better ascertained for disease significance through NGS as conducted in this study. Thus, we focused on three genomic loci to uncover low-frequency variants with strong effects in IPF. Notably, we recognise the limitation to provide precise evaluations of low-frequency variants in IPF due to the small discovery sample size. However, our results support some contribution from low-frequency SNVs to IPF susceptibility in these loci, given that among the 36 genome-wide significant hits, only three of them with validation support (two in 11p15.5 and one in 17q21.31) were low-frequency variants in the controls. Despite that, two of these variants result in the same missense amino acid change for MUC5AC, encoded by another mucin gene located in 11p15.5. As *MUC5AC* p.Ala5353Lys was observed in 4.4% of controls and in as much as 13.8% of the cases, it associates with a relatively strong effect on IPF susceptibility.

Hypersecretion of mucins, most abundantly the glycoproteins MUC5AC and MUC5B, is common during respiratory tract inflammation *via* cytokine stimulation (interleukin-13 and epidermal growth factor) [31]. Chronic hypersecretion and changes in the mucus viscosity can promote its accumulation in the airways, compromising the immune response and perpetuating tissue damage, leading to disease exacerbations [17]. Despite that the common variant rs35705950 of *MUC5B* results in increased mucin gene expression in lung tissues [14, 32] and is the strongest known risk factor for IPF, the exact mechanistic links between the enhanced production of this mucin and the development of IPF are incompletely understood. In IPF, overexpression of MUC5B and reduced expression of MUC5AC have been described in goblet cells located in the lung lesions in comparison with controls [32, 33]. Regulation of MUC5AC derives from the activation of cellular stress, damage and repair pathways, suggesting a key role during disrupted homeostasis [31]. Its activity has been involved in epithelial wound healing after mucosal injury [34]. Therefore, aberrant upregulation of MUC5B and downregulation or activity alterations of MUC5AC may synergise to alter mucus cell differentiation [35] and disrupt epithelial organisation. We speculate that the *MUC5AC* p.Ala5353Lys variant may, therefore, be promoting mucus production, either by directly increasing MUC5AC or indirectly by triggering further increases of MUC5B in the bronchiole [36]. Alternatively, altered glycosylation of this mucin could also contribute to impaired tissue remodelling and promote the disease [37]. Collectively, this evidence along with the results from our study mark *MUC5AC* as another biologically plausible IPF susceptibility gene. Further experiments will be needed to evaluate the potentially relevant cellular mechanisms.

One of the strengths of this study is that we have provided fine-grained variant information from entire and well-recognised IPF loci, enlarging the spectrum of frequencies for SNVs in entire genes and flanking regions involved in the previously evidenced GWAS hits [12]. This is an important contribution as the bulk (>90%) of genetic risk factors involved in complex traits are located in noncoding sequences, supporting the weight of variation regulating transcription in the susceptibility of complex diseases [38]. Moreover, despite the challenge of sequencing the inaccessible repetitive mucin-encoding regions [39], our analytic procedures maintained false variant calls at low levels. These robust results were possible by the high mean depth of coverage reached in the sequencing experiments (>100×). This challenging task, however, imposed some major limitations. First, the discovery study was greatly facilitated by the use of a public database of controls for association testing, which is suitable and advantageous in NGS-based disease mapping approaches [40]. However, because the sequencing depth in cases and controls was different, the quality of sequencing results most likely differed between them, which can lead to considerable risk of sequencing artefacts and other technical issues that can introduce systematic errors [40]. To minimise such a possibility, we used an orthogonal sequencing method to validate the key findings and replicated the results in independent cases and controls. Further NGS studies with larger sample sizes will help to assess the impact of known and unknown genetic variation in these regions, as well as the role of other types of variants besides SNVs.

Acknowledgements: We would like to give special thanks to the clinicians, research nurses and assistants who identified and enrolled patients at each participating medical centre.

Author contributions: S.F. Ma and J. Jou collected DNA samples, designed the experiments, and generated the sequencing data. S.F. Ma and P-C. Hou performed the validation experiments and interpreted the trace results. J.M. Lorenzo-Salazar, J.M. Oldham and C. Flores analysed and interpreted data. B. Guillen-Guio, R.J. Allen, R.G. Jenkins and L.V. Wain performed the replication study. I. Noth and C. Flores were the principal investigators of the discovery study and conceived the project. R.G. Jenkins was the principal investigator on the replication study. All authors contributed to the drafting, revision and coordination of the manuscript. All authors read and approved of the final manuscript. C. Flores takes full responsibility for the content of the manuscript, including the data and analysis.

Conflict of interest: J.M. Lorenzo-Salazar has nothing to disclose. S-F. Ma has nothing to disclose. J. Jou has nothing to disclose. P-C. Hou has nothing to disclose. B. Guillen-Guio has nothing to disclose. R.J. Allen has nothing to disclose. R.G. Jenkins reports grants from GlaxoSmithKline, during the conduct of the study; grants from Biogen and Galecto, personal fees from Boehringer Ingelheim, Galapagos, Heptares, Pliant and Roche, grants and personal fees from GlaxoSmithKline and MedImmune, and has been on advisory boards for NuMedii and Redex, outside the submitted work. He is a trustee of the British Thoracic Society and Action for Pulmonary Fibrosis, and is an NIHR Research Professor (RP-2017-08-014). L.V. Wain reports grants from GlaxoSmithKline, outside the submitted work, and holds a GlaxoSmithKline/British Lung Foundation Chair in Respiratory Research. J.M. Oldham has nothing to disclose. I. Noth reports personal fees and consultancy fees from Boehringer Ingelheim, and personal fees from Genentech/Hoffman La Roche, Global Blood therapeutics, Sanofi Aventis and Zambon, outside the submitted work. In addition, I. Noth has a patent TOLLIP and NAC in IPF pending. C. Flores reports grants from Instituto de Salud Carlos III and Instituto Tecnológico y de Energías Renovables (ITER), during the conduct of the study.

Support statement: This research was funded by the Instituto de Salud Carlos III (grant PI17/00610) and the Spanish Ministry of Science, Innovation and Universities (grant RTC-2017-6471-1; MINECO/AEI/FEDER, UE), which were co-financed by the European Regional Development Funds "A Way of Making Europe" from the European Union, and by the agreement OA17/008 with Instituto Tecnológico y de Energías Renovables (ITER) (C. Flores). Funding was also received from the Pulmonary Fibrosis Foundation, Coalition for Pulmonary Fibrosis grants, and 1RO1HL130796-01A1 (I. Noth); Core Subsidy Mini Awards of the Institute of Translational Medicine and Clinical and Translational Science Award (UL1 RR024999) (S-F. Ma); and a fellowship from Agencia Canaria de Investigación, Innovación y Sociedad de la Información (TESIS2015010057) co-funded by European Social Fund (B. Gillen-Guio). The research was partially supported by the NIHR Nottingham Biomedical Research Centre; the views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Dept of Health. The replication study has been conducted using the UK Biobank Resource under application 8389. Funding information for this article has been deposited with the Crossref Funder Registry.

References

1 Ley B, Collard HR, King TE Jr. Clinical course and prediction of survival in idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2011; 183: 431–440.
2 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.
3 King TE Jr, 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.
4 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.
5 Spagnolo P, Luppi F, Cerri S, *et al.* Genetic testing in diffuse parenchymal lung disease. *Orphanet J Rare Dis* 2012; 7: 79.
6 Mushiroda T, Wattanapokayakit S, Takahashi A, *et al.* A genome-wide association study identifies an association of a common variant in *TERT* with susceptibility to idiopathic pulmonary fibrosis. *J Med Genet* 2008; 45: 654–656.
7 Alder JK, Chen JJ-L, Lancaster L, *et al.* Short telomeres are a risk factor for idiopathic pulmonary fibrosis. *Proc Natl Acad Sci USA* 2008; 105: 13051–13056.
8 Armanios MY, Chen JJ-L, Cogan JD, *et al.* Telomerase mutations in families with idiopathic pulmonary fibrosis. *N Engl J Med* 2007; 356: 1317–1326.
9 Tsakiri KD, Cronkhite JT, Kuan PJ, *et al.* Adult-onset pulmonary fibrosis caused by mutations in telomerase. *Proc Natl Acad Sci USA* 2007; 104: 7552–7557.
10 Maitra M, Wang Y, Gerard RD, *et al.* Surfactant protein A2 mutations associated with pulmonary fibrosis lead to protein instability and endoplasmic reticulum stress. *J Biol Chem* 2010; 285: 22103–22113.
11 Wang W-J, Mulugeta S, Russo SJ, *et al.* Deletion of exon 4 from human surfactant protein C results in aggresome formation and generation of a dominant negative. *J Cell Sci* 2003; 116: 683–692.
12 Noth I, Zhang Y, Ma S-F, *et al.* Genetic variants associated with idiopathic pulmonary fibrosis susceptibility and mortality: a genome-wide association study. *Lancet Respir Med* 2013; 1: 309–317.
13 Zhang Y, Noth I, Garcia JGN, *et al.* A variant in the promoter of *MUC5B* and idiopathic pulmonary fibrosis. *N Engl J Med* 2011; 364: 1576–1577.
14 Seibold MA, Wise AL, Speer MC, *et al.* A common *MUC5B* promoter polymorphism and pulmonary fibrosis. *N Engl J Med* 2011; 364: 1503–1512.
15 Peljto AL, Zhang Y, Fingerlin TE, *et al.* Association between the *MUC5B* promoter polymorphism and survival in patients with idiopathic pulmonary fibrosis. *JAMA* 2013; 309: 2232–2239.
16 Allen RJ, Porte J, Braybrooke R, *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.
17 Evans CM, Fingerlin TE, Schwarz MI, *et al.* Idiopathic pulmonary fibrosis: a genetic disease that involves mucociliary dysfunction of the peripheral airways. *Physiol Rev* 2016; 96: 1567–1591.

18 Richards TJ, Kaminski N, Baribaud F, *et al.* Peripheral blood proteins predict mortality in idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2012; 185: 67–76.
19 Cannon ME, Mohlke KL. Deciphering the emerging complexities of molecular mechanisms at GWAS loci. *Am J Hum Genet* 2018; 103: 637–653.
20 Asimit J, Zeggini E. Rare variant association analysis methods for complex traits. *Annu Rev Genet* 2010; 44: 293–308.
21 Price AL, Patterson NJ, Plenge RM, *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006; 38: 904–909.
22 Fingerlin TE, Murphy E, Zhang W, *et al.* Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. *Nat Genet* 2013; 45: 613–620.
23 Yang J, Ferreira T, Morris AP, *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* 2012; 44: 369–375.
24 Han B, Eskin E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am J Hum Genet* 2011; 88: 586–598.
25 DePristo MA, Banks E, Poplin R, *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011; 43: 491–498.
26 Li B, Liu DJ, Leal SM. Identifying rare variants associated with complex traits *via* sequencing. *Curr Protoc Hum Genet* 2013; 1: 1.26.
27 Welter D, MacArthur J, Morales J, *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 2014; 42: D1001–D1006.
28 Hindorff LA, Sethupathy P, Junkins HA, *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 2009; 106: 9362–9367.
29 Do R, Stitziel NO, Won H-H, *et al.* Exome sequencing identifies rare *LDLR* and *APOA5* alleles conferring risk for myocardial infarction. *Nature* 2015; 518: 102–106.
30 Petrovski S, Todd JL, Durheim MT, *et al.* An exome sequencing study to assess the role of rare genetic variation in pulmonary fibrosis. *Am J Respir Crit Care Med* 2017; 196: 82–93.
31 Young HWJ, Williams OW, Chandra D, *et al.* Central role of Muc5ac expression in mucous metaplasia and its regulation by conserved 5′ elements. *Am J Respir Cell Mol Biol* 2007; 37: 273–290.
32 Seibold MA, Smith RW, Urbanek C, *et al.* The idiopathic pulmonary fibrosis honeycomb cyst contains a mucocilary pseudostratified epithelium. *PLoS One* 2013; 8: e58658.
33 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.
34 Buisine M-P, Desreumaux P, Leteurtre E, *et al.* Mucin gene expression in intestinal epithelial cells in Crohn's disease. *Gut* 2001; 49: 544–551.
35 Thai P, Loukoianov A, Wachi S, *et al.* Regulation of airway mucin gene expression. *Annu Rev Physiol* 2008; 70: 405–429.
36 Nakano Y, Yang IV, Walts AD, *et al. MUC5B* promoter variant rs35705950 affects MUC5B expression in the distal airways in idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2016; 193: 464–466.
37 Nicolaou N, Kevelam SH, Lilien MR, *et al.* Gain-of-glycosylation impairs heterodimerization and maturation of alpha3beta1 and causes interstitial lung disease and congenital nephrotic syndrome. *J Clin Invest* 2012; 122: 4375–4387.
38 Maurano MT, Humbert R, Rynes E, *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* 2012; 337: 1190–1195.
39 Spagnolo P, Cottin V. Genetics of idiopathic pulmonary fibrosis: from mechanistic pathways to personalised medicine. *J Med Genet* 2017; 54: 93–99.
40 Guo MH, Plummer L, Chan Y-M, *et al.* Burden testing of rare variants identified through exome sequencing *via* publicly available control data. *Am J Hum Genet* 2018; 103: 522–534.