Genetic overlap between idiopathic pulmonary fibrosis and COVID-19

Richard J Allen1*, Beatriz Guillen-Guio1*, Emma Croot2, Luke M Kraven1, Samuel Moss3, Iain Stewart3, R Gisli Jenkins3, Louise V Wain1,4

1 Department of Health Sciences, University of Leicester, Leicester, UK
2 Department of Genetics and Genome Biology, University of Leicester, Leicester, UK
3 National Heart and Lung Institute, Imperial College London, London, UK
4 National Institute for Health Research, Leicester Respiratory Biomedical Research Centre, Glenfield Hospital, Leicester, UK
* These authors contributed equally

Abstract
Genome-wide association studies (GWAS) of coronavirus disease 2019 (COVID-19) and idiopathic pulmonary fibrosis (IPF) have identified genetic loci associated with both traits, suggesting possible shared biological mechanisms. Using updated GWAS of COVID-19 and IPF, we evaluated the genetic overlap between these two diseases and identified four genetic loci (including one novel) with likely shared causal variants between severe COVID-19 and IPF. Although there was a positive genetic correlation between COVID-19 and IPF, two of these four shared genetic loci had an opposite direction of effect. IPF-associated genetic variants related to telomere dysfunction and spindle assembly showed no association with COVID-19 phenotypes. Together, these results suggest there are both shared and distinct biological processes driving IPF and severe COVID-19 phenotypes.

Introduction
Coronavirus disease 2019 (COVID-19) is an infectious disease potentially leading to long lasting respiratory symptoms and has resulted in over 4 million deaths worldwide. Idiopathic pulmonary fibrosis (IPF) is a chronic interstitial lung disease (ILD) characterised by an aberrant response to alveolar injury leading to progressive scarring of the lungs. Individuals with ILD are at a higher risk of death from COVID-191.

Large genome-wide association studies (GWAS) have identified multiple genetic signals associated with severe COVID-192, including a signal within the DPP9 gene that is also associated with increased IPF risk3. GWAS have identified 20 genome-wide significant signals of association with IPF risk4,5 with the largest genetic risk factor being a common variant located in the promoter region of MUC5B (rs35705950, odds ratio>4). Previous analyses suggest IPF is a causal risk factor for severe COVID-19 but noted that the effect of rs35705950 was in the opposite direction (i.e. the allele associated with increased risk of IPF was protective for severe COVID-19)6.

We aimed to further explore the shared genetic architecture and identify novel shared genetic loci between the two diseases, using new enlarged GWAS of IPF and COVID-19 risk.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Methods

Data
We used the largest GWAS of IPF risk which consisted of unrelated European individuals from across five studies. Cases were selected from centres in the USA, UK and Spain diagnosed using American Thoracic Society and European Respiratory Society guidelines.

For COVID-19, the summary statistics from version 6 of the COVID-19 Host Genetics Initiative (HGI_v6, https://www.COVID-19hg.org/results/r6/) were used. This analysis considered four different COVID-19 phenotypes according to the severity of the disease and the controls used; A2) Very severe respiratory confirmed COVID-19 vs. population, B1) Hospitalised COVID-19 vs. not hospitalised COVID-19, B2) Hospitalised COVID-19 vs. population and, C2) COVID-19 vs. population.

Genetic Overlap
Genome-wide genetic correlation analyses were performed using LD Score Regression. Twenty previously reported IPF genetic association signals were investigated for their association in the four COVID-19 GWAS, and 26 variants reaching genome-wide significance in the COVID-19 GWAS were tested for their association with IPF (proxy variants – \( r^2 > 0.8 \) in European population – were investigated if the top associated variant was not included). Genetic loci showing an association with both traits (after Bonferroni correction for multiple testing) were investigated to determine whether the same causal variant was driving both the IPF and COVID-19 associations using coloc. Regions with a posterior probability>80% of there being a shared causal variant (assuming up to one causal variant for each trait in the region and that variant has been measured) were deemed to have colocalised.

Identification of putative causal genes and pleiotropic effects
We investigated whether shared genetic signals were associated with gene expression in lung tissue (GTEx_lung, n=515) and whole blood (eQTLGen, n=31,684). Where the variant met a false discovery rate of 5%, colocalisation analyses were performed using coloc and deemed to be linked to gene expression if the posterior probability of a shared causal variant was greater than 80%.

We performed a phenome-wide association study (PheWAS) to identify if the overlapping genetic signals had been previously reported for association with other traits (\( p < 10^{-5} \)) using publicly available resources (PhenoScanner_v2, GWAS Catalog and Open Targets). Colocalisation analyses were performed to determine if the same causal variant was driving both traits. For overlapping signals for traits of relevance, we additionally assessed gene expression in all GTEx tissues, as described above.

Results
There was a significant weak positive genome-wide correlation between IPF and severe COVID-19 (A2 \( r^2 = 0.274 \), p=0.0045, B1 \( r^2 = 0.279 \), p=0.0093, and B2 \( r^2 = 0.261 \), p=0.0005) but not with COVID-19 infection (C2 \( r^2 = 0.066 \), p=0.433).

Four genetic association signals showed evidence of a shared causal variant between IPF and at least one COVID-19 phenotype (posterior probability>80%), namely loci at 7q22.1, near MUC5B, near ATP11A and near DPP9 (Table 1a). The 7q22.1 locus has not previously been reported for association with COVID-19. Three additional IPF genetic signals (at 17q21.31, DSP and DEPTOR) showed an association with COVID-19 but did not colocalise, suggesting there are different causal variants between the two traits at these loci. Visual inspection of the 17q21.31 locus revealed extended
linkage disequilibrium (due to the presence of a large inversion) meaning colocalisation analyses could not determine whether there were shared or distinct causal variants.

Three of the four shared signals colocalised with expression of the single nearest gene in blood or lung (MUC5B, ATP11A and DPP9) (Table 1b). The IPF and COVID-19 risk increasing alleles at the 7q22.1 signal colocalised with decreased expression of ZKSCAN1 and TRIM4 in blood.

The signal on chromosome 7 was associated with a number of blood traits and the signal near ATP11A was associated with blood traits and HbA1c (average blood glucose levels, used in diagnosing diabetes) (Table 1b). The IPF and HbA1c signals did not colocalise, however, as diabetes is a risk factor for COVID-1910, we further investigated the effects on gene expression for this signal. The allele (rs423117_T) associated with higher Hb1AC levels was associated with increased ATP11A expression in liver and decreased expression in cultured fibroblasts, but there was no association with ATP11A expression in blood.

**Discussion**

Genetic association signals near MUC5B, DPP9 and ATP11A have previously been reported for both COVID-19 severity and IPF risk; we show for the first time that these signals are likely due to the same underlying causal variant. In addition, we report a novel overlapping signal at 7q22.1 implicating TRIM4 and ZKSCAN1.

Despite a positive genome-wide genetic correlation between IPF risk and the COVID-19 severity phenotypes (A2, B1 and B2), we show that two of the four shared signals (at MUC5B and ATP11A) have opposite directions of effect on risk for the two diseases. The allele associated with increased risk of IPF and increased ATP11A expression in blood (rs9577395_C) was associated with decreased risk of severe COVID-19. Our PheWAS highlighted a potential link with HbA1c and diabetes risk at this locus via ATP11A expression, although effects were tissue dependent.

The rs2897075_T allele at 7q22.1, associated with increased IPF and COVID-19 risk, was linked to decreased TRIM4 and ZKSCAN1 expression. TRIM4 is an important regulator of virus-induced interferon induction pathways and a proteomic study identified significant adjacency between SARS-CoV-2 M protein and TRIM411. Viral infection-induced micro-injury to the alveolar epithelium is thought to be a trigger for development of IPF12, suggesting the interferon-mediated innate immune response could be central to both risk of chronic lung disease and worse outcomes due to SARS-CoV-2 infection.

Loci previously implicated by IPF GWAS relating to telomere dysfunction (TERT, TERC, RTEL1) and mitotic spindle assembly (KIF15, MAD1L1, SPDL1, KNL1) were not associated with COVID-19.

The colocalisation analyses assume a single measured causal variant. Although conditional analyses found no evidence of multiple independent association signals at the regions studied, we cannot guarantee all causal variants were measured. Furthermore, we utilised whole blood and lung tissue for gene expression so we cannot rule out cell-specific effects. Further analyses in non-European populations could help identify other ancestry specific overlapping variants and increase the generalisability of the results.

In conclusion, using the largest IPF and COVID-19 GWAS to date, we show there is a positive genetic correlation between IPF and severe COVID-19 risk. However, some IPF-related pathways may have an opposite (e.g. MUC5B and ATP11A pathways) effect on severe COVID-19 risk.
Funding and conflicts of interest

R Allen is an Action for Pulmonary Fibrosis Mike Bray Research Fellow. B Guillen-Guio is supported by Wellcome Trust grant 221680/Z/20/Z. G Jenkins is a trustee of Action for Pulmonary Fibrosis and reports personal fees from Astra Zeneca, Biogen, Boehringer Ingelheim, Bristol Myers Squibb, Chiesi, Daewoong, Galapagos, Galecto, GlaxoSmithKline, Heptares, NuMedii, PatientMPower, Pliant, Promedior, Redx, Resolution Therapeutics, Roche, Veracyte and Vicore. L Wain holds a GSK/British Lung Foundation Chair in Respiratory Research (C17-1) and reports research funding from GSK and Orion and consultancy for Galapagos, outside of the submitted work. The research was partially supported by the National Institute for Health Research (NIHR) Leicester Biomedical Research Centre; the views expressed are those of the author(s) and not necessarily those of the National Health Service (NHS), the NIHR or the Department of Health. This research used the SPECTRE High Performance Computing Facility at the University of Leicester.
Table 1: Variants reaching Bonferroni-corrected significance for both IPF and COVID-19.

All results are presented in terms of the allele which increases risk of IPF. Chr=chromosome. REF=reference allele. EFF=effect allele (i.e. the variant the effect estimates are in relation to. OR=odds ratio. CI=confidence interval. FEV1=forced expiratory volume in 1 second. FVC=forced vital capacity. HbA1c=Haemoglobin Type A1c.

**a) Genome-wide analysis results for IPF and COVID-19.** The coloc column gives the posterior probability there is a shared causal variant between IPF and that COVID-19 phenotype at that genetic loci. Colocalisation analyses were only performed on signals showing a possible association with both traits after correcting for multiple testing. Signals that colocalise (i.e. posterior probability>80%) between IPF and COVID-19 are highlighted in grey.

**Gene expression and PheWAS results.** Percentages shown for the eQTL columns show the posterior probability of colocalisation between the IPF risk signal and the gene expression eQTL signal (only genes with posterior probability>80% are presented in the table). eQTL results are presented in terms of the IPF risk allele. For the PheWAS results, phenotypes where the variant had p<10\(^{-5}\) and which colocalised with the IPF signal (posterior probability>80%) are presented. Only non-ILD and COVID-19 phenotypes were investigated for the PheWAS analysis. Proxy variants (with r\(^2\)>0.8) were also investigated in PhenoScanner. For Open Targets only traits with genome-wide summary statistics from GWAS Catalog were investigated.

<table>
<thead>
<tr>
<th>Chr</th>
<th>Position</th>
<th>rsid</th>
<th>Previously reported gene</th>
<th>REF</th>
<th>EFF</th>
<th>IPF 4,125 cases vs 20,464 controls</th>
<th>A2) Very severe respiratory confirmed COVID-19 vs. population 8,779 cases vs 1,001,875 controls</th>
<th>B1) Hospitalised COVID-19 vs. not hospitalised COVID-19 14,408 cases vs 73,191 controls</th>
<th>B2) Hospitalised COVID-19 vs. population 24,274 cases vs 2,061,529 controls</th>
<th>C2) COVID-19 vs. population 112,612 cases vs 2,474,079 controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>99630342</td>
<td>rs2897075</td>
<td>7q22.1</td>
<td>C</td>
<td>T</td>
<td>1.30 [1.23, 1.37] 1.77×10(^{-11})</td>
<td>1.07 [1.04, 1.12] 1.63×10(^{-4}) 88.2%</td>
<td>1.02 [0.99, 1.05] 0.238 -</td>
<td>1.04 [1.02, 1.06] 5.55×10(^{-6}) 48.1%</td>
<td>1.01 [1.00, 1.02] 0.014 -</td>
</tr>
<tr>
<td>11</td>
<td>1241221</td>
<td>rs35705950</td>
<td>MUC5B</td>
<td>G</td>
<td>T</td>
<td>5.06 [4.67, 5.47] 9.09×10(^{-12})</td>
<td>0.83 [0.77, 0.89] 1.17×10(^{-7}) 100%</td>
<td>0.89 [0.84, 0.94] 2.20×10(^{-5}) 98.5%</td>
<td>0.89 [0.86, 0.93] 1.22×10(^{-6}) 100%</td>
<td>0.99 [0.98, 1.01] 0.448 -</td>
</tr>
<tr>
<td>13</td>
<td>113534984</td>
<td>rs9577395</td>
<td>ATP11A</td>
<td>G</td>
<td>C</td>
<td>1.29 [1.21, 1.38] 4.78×10(^{-14})</td>
<td>0.90 [0.87, 0.94] 4.38×10(^{-4}) 99.0%</td>
<td>0.94 [0.90, 0.97] 8.76×10(^{-4}) 52.1%</td>
<td>0.94 [0.91, 0.96] 8.67×10(^{-7}) 99.5%</td>
<td>0.99 [0.98, 1.00] 0.037 -</td>
</tr>
<tr>
<td>19</td>
<td>4717672</td>
<td>rs12610495</td>
<td>DPP9</td>
<td>A</td>
<td>G</td>
<td>1.28 [1.21, 1.36] 2.56×10(^{-16})</td>
<td>1.20 [1.15, 1.26] 1.64×10(^{-15}) 97.9%</td>
<td>1.08 [1.04, 1.11] 1.73×10(^{-5}) 98.5%</td>
<td>1.11 [1.09, 1.14] 6.09×10(^{-18}) 97.9%</td>
<td>1.03 [1.02, 1.04] 5.10×10(^{-10}) 98.1%</td>
</tr>
</tbody>
</table>

---

The copyright holder for this version posted December 8, 2021. doi: medRxiv preprint CC-BY 4.0 International license It is made available under a perpetuity.
### b) Gene expression and PheWAS results

<table>
<thead>
<tr>
<th>Chr</th>
<th>Position</th>
<th>rsid</th>
<th>Gene expression (coloc posterior probability of shared causal variant)</th>
<th>PheWAS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>EFF GTeX (Lung, n=515) eQTLGen (whole blood, n=31,684) PhenoScanner v2</td>
<td>GWAS Catalog Open Targets</td>
</tr>
<tr>
<td>7</td>
<td>99630342</td>
<td>rs2897075</td>
<td>T -</td>
<td>Decreased <strong>ZKSCAN1</strong> (99.4%) Decreased <strong>TRIM4</strong> (85.4%) Blood traits&lt;br&gt;Mean corpuscular haemoglobin Red cell distribution width Impedance of leg right</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Blood traits&lt;br&gt;Mean corpuscular haemoglobin Red cell distribution width Impedance of leg right</td>
</tr>
<tr>
<td>11</td>
<td>1241221</td>
<td>rs35705950</td>
<td>T <strong>Increased MUC5B</strong> (100%) -</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Blood traits&lt;br&gt;Red cell distribution width HbA1c</td>
</tr>
<tr>
<td>13</td>
<td>113534984</td>
<td>rs9577395</td>
<td>C -</td>
<td><strong>Increased ATP11A</strong> (99.6%) Blood traits&lt;br&gt;Red cell distribution width HbA1c</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Blood traits&lt;br&gt;Mean corpuscular volume Mean corpuscular haemoglobin Red cell distribution width Platelet count Red blood cell count Lung function FEV₁/FVC</td>
</tr>
<tr>
<td>19</td>
<td>4717672</td>
<td>rs12610495</td>
<td>G -</td>
<td><strong>Decreased DPP9</strong> (88.3%) -</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Blood traits&lt;br&gt;Mean corpuscular volume Mean corpuscular haemoglobin Red cell distribution width Platelet count Red blood cell count Lung function FEV₁/FVC</td>
</tr>
</tbody>
</table>
References


