Proportion of idiopathic pulmonary fibrosis risk explained by known genetic loci

Olivia C Leavy¹, Shwu-Fan Ma², Philip L Molyneaux^{3,4}, Toby M Maher^{3,4,5}, Justin M Oldham⁶, Carlos Flores^{7,8,9,10}, Imre Noth², R Gisli Jenkins^{11,12}, Frank Dudbridge¹, Louise V Wain^{1,13}, Richard J Allen¹

¹Department of Health Sciences, University of Leicester, Leicester, UK

²Division of Pulmonary & Critical Care Medicine, University of Virginia, Charlottesville, USA

³NIHR Respiratory Clinical Research Facility, Royal Brompton Hospital, London, UK

⁴National Heart and Lung Institute, Imperial College, London, UK

⁵Hastings Centre for Pulmonary Research and Division of Pulmonary, Critical Care and Sleep Medicine, Keck School of Medicine, University of Southern California, Los Angeles, USA

⁶Department of Internal Medicine, University of California Davis, Davis, USA

⁷Research Unit, Hospital Universitario N.S. de Candelaria, Universidad de La Laguna, Santa Cruz de Tenerife, Spain

⁸CIBER de Enfermedades Respiratorias, Instituto de Salud Carlos III, Spain

⁹Instituto Tecnológico y de Energías Renovables (ITER, S.A.), Santa Cruz de Tenerife, Spain

¹⁰Instituto de Tecnologías Biomédicas (ITB), Universidad de La Laguna, Santa Cruz de Tenerife, Spain

¹¹Division of Respiratory Medicine, University of Nottingham, Nottingham, UK

¹²National Institute for Health Research, Nottingham Biomedical Research Centre, Nottingham University Hospitals, Nottingham, UK

¹³National Institute for Health Research, Leicester Respiratory Biomedical Research Centre, Glenfield Hospital, Leicester, UK

Abstract

Genome-wide association studies have identified 14 genetic loci associated with susceptibility to idiopathic pulmonary fibrosis (IPF), a devastating lung disease with poor prognosis. Of these, the variant with the strongest association, rs35705950, is located in the promoter region of the *MUC5B* gene and has a risk allele (T) frequency of 30-35% in IPF cases. Here we present estimates of the proportion of disease liability explained by each of the 14 IPF risk variants as well as estimates of the proportion of cases that can be attributed to each variant. We estimate that rs35705950 explains 5.9-9.4% of disease liability, which is much lower than previously reported estimates. Of every 100,000 individuals with the rs35705950_GG genotype we estimate 30 will have IPF, whereas for every 100,000 individuals with the rs35705950_GT genotype 152 will have IPF. Quantifying the impact of genetic risk factors on disease liability improves our understanding of the underlying genetic architecture of IPF and provides insight into the impact of genetic factors in risk prediction modelling.

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

Introduction

Understanding how genetic factors contribute to disease risk improves our understanding of pathogenesis, supports drug development and aids risk prediction¹. Appropriate quantification and interpretation of this contribution is essential for measuring the impact of genetic variation and in motivating and informing future studies.

Idiopathic pulmonary fibrosis (IPF) is a chronic disease characterised by scarring of the lungs. Current therapies only slow disease progression and half of individuals die within 3-5 years of diagnosis. A genetic variant, rs35705950, in the *MUC5B* promoter region, is strongly associated with IPF susceptibility with the risk allele (T) associated with a five-fold increase in disease risk². Genome-wide association studies (GWAS) have identified 13 additional independent IPF susceptibility variants³.

The rs35705950_T allele frequency in IPF cases is 30-35%⁴ (compared to 11% in controls), but risk allele frequency does not reflect the disease risk accounted for by this variant. Explained risk can be measured in different ways, such as the proportion of risk explained in the general population, or alternatively the proportion of cases due to a specific variant.

Here we provide estimates of the proportion of IPF risk in the general population explained by known IPF susceptibility variants, and estimates of the proportion of cases attributable to each susceptibility variant. Our analyses focussed on non-familial IPF, therefore variants considered are just those evidenced by GWAS.

Methods

We used unrelated European IPF cases (diagnosed according to international guidelines⁵) and controls from a previously reported study³ with appropriate ethics approval.

To estimate the proportion of disease risk explained by each variant in the general population, we performed regression analyses including the susceptibility variant as the only covariate. R^2 is a measure of phenotypic variance explained by a model and, as our model only contains a single variant, the proportion of disease explained by that variant. R^2 cannot be directly calculated as the IPF phenotype is binary and the proportion of cases in our analysis is higher than observed in the general population. We therefore calculated a liability R^2 accounting for enrichment of cases⁶. The liability model assumes individuals have an unmeasured continuous trait, called the liability, and an individual develops IPF when the liability exceeds a critical value. We calculated the liability R^2 for IPF prevalence estimates of 1.25 and 63 cases per 100,000 people (the lowest and highest reported estimates of disease prevalence⁷). To estimate the variance in the liability explained by all variants, we fitted the model with the most significant variants from all 14 known IPF susceptibility loci and calculated the liability R^2 . Finally, we fitted the model with all susceptibility variants, minus rs35705950. We investigated whether results were biased by population stratification by repeating analyses including 10 genetic principal components to adjust for ancestry.

To estimate the proportion of cases attributable to each variant, we calculated the population attributable risk fraction⁸ (PARF). PARF is the proportion of cases that would be prevented if a risk factor were removed from the population. If any risk factors were removed, the PARFs of other risk factors would change. Therefore, PARFs cannot be summed to calculate the proportion of cases prevented if multiple risk factors were removed.

Results

A total of 792 IPF cases and 10,000 controls were included in the analysis. Variant rs35705950 alone explains 5.9-9.4% of disease liability in the general population. No other IPF susceptibility variant explained more than 0.6% and collectively the 13 non-*MUC5B* susceptibility variants explained 1.8-2.9% of variation in disease liability (**Figure 1**). The highest PARF was observed for rs35705950 (51%), however many of the susceptibility variants had PARF>10 (**Figure 2**). Effect sizes were similar after adjusting for principal components suggesting results are not biased by population stratification.

Discussion

The *MUC5B* promoter polymorphism explains three times more disease liability in the general population than the other 13 IPF susceptibility variants combined. In total, the 14 IPF susceptibility variants explain up to 12.4% of disease liability in the general population, which is smaller than previous reports that cited 30-35% of risk^{4,9}. Importantly, therapies that target variants that explain a small proportion of disease risk can still have a large clinical impact¹.

Our results suggest IPF cases could be halved if the *MUC5B* risk allele was removed from the population. Although the clinical relevance of PARF estimates may be limited as removing risk alleles from the general population is almost impossible, they do indicate the impact preventative interventions could have on disease incidence.

Different populations experience diverse environmental exposures and have varying allele frequencies, affecting the proportion of risk explained by these variants. We also only investigated known common IPF susceptibility variants, though previous studies suggest there could be many undiscovered genetic variants contributing to IPF risk³, and we have not investigated epistasis or gene-environment interactions. This means overall IPF risk explained by genetics will likely be much higher than the 12.4% explained by the known variants.

This study utilised an ascertained case-control study design and made assumptions about disease prevalence. Ideally, a general population cohort, such as UK Biobank¹⁰, would be used for these analyses. However, in UK Biobank there are few self-reported cases (n=104) and cases defined using hospital episode statistics do not genetically resemble clinically recruited cases (rs35705950_T allele frequency in these cases is 20%). Therefore, we restricted analyses to a study with clinically recruited cases. The study used was not used in the discovery of the IPF susceptibility variants³, meaning the estimates of risk explained should not be subject to winner's curse bias.

We could also consider absolute risk. Assuming disease prevalence is 63 cases per 100,000 and using the previously reported² effect size for the *MUC5B* risk allele (OR=4.99), for every 100,000 individuals with the rs35705950_GG genotype we would expect 30 to have IPF, whereas for every 100,000 individuals with the rs35705950_GT genotype we would expect 152 to have IPF. Therefore, although rs35705950 is strongly associated with disease risk, most individuals carrying the risk allele will not develop IPF.

Although risk allele frequencies in cases can be of interest, they are not a measure of explained risk. Many of the known IPF susceptibility variants have a high PARF, but individually explain a small overall proportion of the variation in risk. These results provide an important reference point to inform future genetic discoveries and for evaluation of the likely contribution of genetic factors in risk prediction models.

Figure 1: Proportion of liability explained in the general population.

Estimated proportion of variation explained is taken from the liability R^2 from the regression analyses with the lower bound given when assuming a disease prevalence of 1.25 cases per 100,000 and the upper bound given when assuming a disease prevalence of 63 cases per 100,000. The x-axis label “All 14 variants” refers to the model including all 14 sentinel IPF susceptibility variants, “13 non-MUC5B” refers to the model including all sentinel IPF susceptibility variants minus the *MUC5B* polymorphism rs35705950. Variants are ordered by the proportion of explained variation.

Figure 2: Population attributable risk fraction

Estimates of PARF for each variant with 95% confidence intervals. Variants are ordered by the proportion of explained variation in the general population (Figure 1).

Funding

R. Allen is an Action for Pulmonary Fibrosis Research Fellow. L. Wain holds a GSK/British Lung Foundation Chair in Respiratory Research. RG. Jenkins is supported by an NIHR Research Professorship (NIHR reference RP-2017-08-ST2-014). I. Noth: National Heart Lung and Blood Institute (R01HL130796). C. Flores: Spanish Ministry of Science and Innovation (grant RTC-2017-6471-1; AEI/FEDER, UE) co-financed by the European Regional Development Funds (ERDF) 'A way of making Europe' from the European Union, and by the agreement OA17/008 with Instituto Tecnológico y de Energías Renovables (ITER) to strengthen scientific and technological education, training, research, development and innovation in Genomics, Personalized Medicine and Biotechnology. P. Molyneaux is an Action for Pulmonary Fibrosis Research Fellow. T. Maher is supported by an NIHR Clinician Scientist Fellowship (NIHR Ref: CS-2013-13-017) and a British Lung Foundation Chair in Respiratory Research (C17-3). J. Oldham: National Heart Lung and Blood Institute (K23HL138190).

References

1. Zeggini E, Gloyn AL, Barton AC, Wain LV. Translational genomics and precision medicine: Moving from the lab to the clinic. *Science*. 2019;365(6460):1409-1413.
2. Zhu QQ, Zhang XL, Zhang SM, et al. Association between the MUC5B promoter polymorphism rs35705950 and idiopathic pulmonary fibrosis: A meta-analysis and trial sequential analysis in caucasian and asian populations. *Medicine (Baltimore)*. 2015;94(43):e1901.
3. Allen RJ, Guillen-Guio B, Oldham JM, et al. Genome-wide association study of susceptibility to idiopathic pulmonary fibrosis. *American journal of respiratory and critical care medicine*. 2020;201(5):564-574.
4. Seibold MA, Wise AL, Speer MC, et al. A common MUC5B promoter polymorphism and pulmonary fibrosis. *N Engl J Med*. 2011;364(16):1503-1512.
5. Raghu G, Remy-Jardin M, Myers JL, et al. Diagnosis of idiopathic pulmonary fibrosis. an official ATS/ERS/JRS/ALAT clinical practice guideline. *American journal of respiratory and critical care medicine*. 2018;198(5):e44-e68.
6. Lee SH, Goddard ME, Wray NR, Visscher PM. A better coefficient of determination for genetic profile analysis. *Genet Epidemiol*. 2012;36(3):214-224.
7. Nalysnyk L, Cid-Ruzafa J, Rotella P, Esser D. Incidence and prevalence of idiopathic pulmonary fibrosis: Review of the literature. *Eur Respir Rev*. 2012;21(126):355-361.
8. Witte JS, Visscher PM, Wray NR. The contribution of genetic variants to disease depends on the ruler. *Nature Reviews Genetics*. 2014;15(11):765-776.
9. 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(4):1567-1591.
10. Sudlow C, Gallacher J, Allen N, et al. UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015;12(3):e1001779.