

Understanding Disease Relationships with Idiopathic Pulmonary Fibrosis through Genomic Methods

Thesis submitted for the degree of Doctor of Philosophy (PhD) at Imperial
College London

By Samuel Thomas Moss

The Margaret Turner-Warwick Centre for Fibrosing Lung Disease

National Heart and Lung Institute

Imperial College London

August 2024

Abstract

Idiopathic pulmonary fibrosis is a disease characterised by uncontrolled lung scarring, lung function decline, and a poor prognosis. Current treatments for IPF are limited and can only slow disease progression. To add to this problem, the biology underlying IPF disease is incompletely understood. This thesis aims to investigate how the genetics of IPF are linked to other human diseases, traits, and biological features to provide insights into disease mechanisms and potentially highlight opportunities for novel IPF treatments.

A phenome-wide association study (PheWAS) was conducted to investigate the spectrum of phenotypes associated with IPF genetic risk variants. This investigation highlighted shared genetic variants associated with IPF risk and multiple traits including lung function, telomere length, and total testosterone measurements.

Using genetic correlation, broad genetic overlap across IPF and related traits was also evaluated. This analysis identified the overlap of genetics associated with telomere length, lung function, and primary biliary cholangitis with genetics of IPF susceptibility.

Finally, a Mendelian randomisation approach was used to assess if the epidemiological association between IPF and diabetes mellitus is driven by direct causal effects between these diseases. Overall, no evidence of causal effects was suggested, indicating that other factors such as treatment effects may be driving this epidemiological relationship. In addition, a similar Mendelian randomisation approach was applied to investigate if a potential biological relationship between mosaic loss of Y chromosome (mLOY) and IPF is driven by direct causal effects. This investigation did not identify any evidence of causal effects between mLOY and IPF but highlights that shortened telomere length may be a central factor driving both traits.

Acknowledgments

I would firstly like to acknowledge my supervisors Professor Gisli Jenkins and Dr. Iain Stewart for going above and beyond to support me throughout my PhD. I can't thank them enough for their encouragement, guidance, and belief in me.

I am extremely grateful to Cosetta Minelli and Olivia Leavy for their time and incredible selflessness in guiding me through the methodological maze that is Mendelian randomisation. I would like to thank Professor Louise Wain and Richard Allen not only for their guidance, but also for the opportunities they have given me to integrate with the wider genomics and multimorbidity research community. I would also like to thank everyone at the Margaret Turner-Warwick Centre for Fibrosing Lung Disease for being such a friendly and motivated group.

A very special mention goes to my partner Vicky, who has been completing her own PhD at the same time. Her ability to find the space to be selfless despite experiencing the same stress has been nothing short of amazing. Finally, I would like to thank my family and friends, particularly my Mum and Dad who have been so kind and supportive throughout.

Statement of Originality

I confirm that the work presented in this thesis is my own and all other work has been appropriately referenced.

Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-NonCommercial 4.0 International Licence (CC BY-NC).

Under this license, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Table of Contents

List of Figures	8
List of Tables	15
1: Introduction	16
1.1 Introduction to idiopathic pulmonary fibrosis	16
1.1.1: Clinical features and current treatments	16
1.1.2 IPF pathogenesis and disease biology	20
1.1.3 IPF multimorbidity and fibrosis beyond the lung	21
1.2 Introduction to genetics and IPF genetic risk	26
1.2.1 Genetics, mutations, and phenotypes	26
1.2.2 Polygenic risk of human traits	29
1.2.3 Modern genomic approaches in disease research	33
1.2.4 IPF genetics and implicated disease biology	34
1.4 Outline and aims of thesis	38
2: Phenome-Wide Association Study of Common Idiopathic Pulmonary Fibrosis Genetic Risk Variants	42
2.1 Introduction	42
2.2 Methods	45
2.2.1 Methods for PheWAS of public GWAS data	45
2.2.2 Methods for genetic colocalisation	47
2.2.3 Methods for PheWAS using an independent cohort	52
2.3 Results	54
2.3.1 Results of PheWAS of public GWAS data	54
2.3.2 Conclusions from PheWAS of public GWAS data	83
2.3.3 Prioritisation of public PheWAS results for colocalisation analyses	83
2.3.4 Results from colocalisation analyses of prioritised traits	86
2.3.5 Conclusions from colocalisation analyses	92
2.3.6 Results of PheWAS using an independent cohort	93
2.3.7 Conclusions from PheWAS using an independent cohort	117

2.4 Discussion	118
3: Genetic Correlation Study of Idiopathic Pulmonary Fibrosis and Related Phenotypes.....	123
3.1 Introduction	123
3.2 Methods.....	124
3.2.1 Trait selection for genetic correlation analyses.....	124
3.2.3 Methods for genetic correlation analysis.....	125
3.3 Results.....	126
3.3.1 Traits selected for genetic correlation analysis with IPF risk GWAS.....	126
3.3.2 Genetic correlation analysis results.....	127
3.3.3 Conclusions from genetic correlation analysis of idiopathic pulmonary fibrosis and related phenotypes.....	130
3.4 Discussion	130
4: Investigating Causal Relationships between Diabetes Mellitus and Idiopathic Pulmonary Fibrosis	134
4.1 Introduction	134
4.2 Methods.....	139
4.2.1 Data retrieval and processing.....	139
4.2.2 Genetic instrument selection.....	140
4.2.3 Methods for statistical analyses	141
4.3 Results.....	143
4.3.1 Genetic instrument selection.....	143
4.3.2 Mendelian randomisation analysis results: Main analyses	143
4.3.3 Mendelian randomisation analysis results: Multivariable analyses.....	158
4.3.4 Mendelian randomisation analysis results: Secondary analyses.....	160
4.3.5 Conclusions from Mendelian randomisation study assessing causal relationships between diabetes mellitus and idiopathic pulmonary fibrosis	163
4.4 Discussion	164
5: Investigating a potential causal link between Mosaic Loss of Chromosome Y and Idiopathic Pulmonary Fibrosis.....	167
5.1 Introduction	167
5.2 Methods.....	169

5.2.1 Data retrieval and processing	169
5.2.2 Genetic instrument selection	171
5.2.3 Methods for statistical analyses	171
5.3 Results.....	173
5.3.1 Genetic instrument selection	173
5.3.2 Mendelian randomisation analysis results: Main analyses	174
4.3.4 Conclusions from Mendelian randomisation study investigating a potential causal link between mosaic loss of chromosome Y and idiopathic pulmonary fibrosis	188
5.4 Discussion	188
6: Discussion	191
6.1 The current understanding of IPF genetics and links with other human phenotypes	191
6.2 Summary of findings and biological insights	192
6.3 Clinical implications	196
6.4 Thesis strengths and limitations	197
6.5 Future work	201
6.5.1 Expansion of colocalisation analyses	201
6.5.2 Expanded investigation of shared genetic architecture across fibrotic disease traits.....	202
6.5.3 Mendelian randomisation study assessing if sex hormones are causally associated with IPF risk	202
6.6 Conclusion	203
Reference list	204

List of Figures

Figure 1.1.1: A diagram depicting how interstitial fibrosis leads to reduced gas exchange compared to healthy lung tissue	17
Figure 1.1.2: A diagrammatic representation of how comorbidities may be mechanistically linked to IPF and potential responses to different pathways of comorbidity	25
Figure 1.2.1: A simplified illustration of the structure of double stranded DNA	27
Figure 1.2.2: DNA replication of the leading strand	28
Figure 1.2.3: Examples of different types of genetic variation	29
Figure 1.2.4: Illustration of the basic methodology of genome-wide association studies (GWAS).....	30
Figure 2.2.1: An overview of the study design applied in a PheWAS of IPF genetic risk variants using public GWAS data	47
Figure 2.2.2: Prioritisation of PheWAS associations for follow-up genetic colocalisation analyses	48
Figure 2.2.3: An overview of methodology applied in a PheWAS of IPF genetic risk variants using the Genomics England 100,000 Genomes Project data resource	53
Figure 2.3.1: A volcano plot presenting phenotypes associated with rs35705950 (G > T) (<i>MUC5B</i>).....	55
Figure 2.3.2: A volcano plot presenting phenotypes associated with rs112087793 (T > C) (<i>STMN3</i>)	57
Figure 2.3.3: A volcano plot presenting phenotypes associated with rs2292181 (G > C) (<i>KIF15</i>).....	58

Figure 2.3.4: A volcano plot presenting phenotypes associated with rs2076295 (T > G) (<i>DSP</i>)	59
Figure 2.3.5: A volcano plot presenting phenotypes associated with rs9860874 (C > A) (<i>TERC</i>)	63
Figure 2.3.6: A volcano plot presenting phenotypes associated with rs41308092 (G > A) (<i>RTEL1</i>)	64
Figure 2.3.7: A volcano plot presenting phenotypes associated with rs7725218 (A > G) (<i>TERT</i>)	65
Figure 2.3.8: A volcano plot presenting phenotypes associated with rs12537430 (A > G) (<i>MAD1L1</i>)	66
Figure 2.3.9: A volcano plot presenting phenotypes associated with rs12912339 (G > A) (<i>KNL1</i>)	67
Figure 2.3.10: A volcano plot presenting phenotypes associated with rs74614704 (G > A) (<i>NPRL3</i>)	71
Figure 2.3.11: A volcano plot presenting phenotypes associated with rs10808505 (G > T) (<i>DEPTOR</i>)	72
Figure 2.3.12: A volcano plot presenting phenotypes associated with rs9577395 (G > C) (<i>ATP11A</i>)	73
Figure 2.3.13: A volcano plot presenting phenotypes associated with rs2304645 (G > C) (<i>IVD</i>)	74
Figure 2.3.14: A volcano plot presenting phenotypes associated with rs11073517 (C > T) (<i>AKAP13</i>)	75
Figure 2.3.15: A volcano plot presenting phenotypes associated with rs12610495 (A > G) (<i>DPP9</i>)	76

Figure 2.3.16: A volcano plot presenting phenotypes associated with rs2897075 (C > T) (ZKSCAN1)	77
Figure 2.3.17: A volcano plot presenting phenotypes associated with rs2077551 (C > T) (17q21.31).....	79
Figure 2.3.18: A volcano plot presenting phenotypes associated with rs79684490 (G > A) (10q25.1).....	80
Figure 2.3.19: A volcano plot presenting phenotypes associated with rs2609259 (C > A) (FAM13A)	81
Figure 2.3.20: Phenotypes associated with 19 IPF genetic risk variants	82
Figure 2.3.21: Colocalisation of association signals with IPF risk and total testosterone measurement.....	86
Figure 2.3.22: Non-colocalisation of association signals with IPF risk and telomere length measurement.....	87
Figure 2.3.23: Colocalisation of association signals with IPF risk and <i>DEPTOR</i> gene expression	88
Figure 2.3.24: Non-colocalisation of association signals with IPF risk and <i>ZGPAT</i> gene expression	89
Figure 2.3.25: Data processing and participant exclusion criteria to establish an appropriate cohort for PheWAS analyses within the GE 100,000 GP dataset	93
Figure 2.3.26: A volcano plot presenting phenotypes associated with rs35705950 (G > T) (<i>MUC5B</i>).....	94
Figure 2.3.27: A volcano plot presenting phenotypes associated with rs78238620 (T > A) (<i>KIF15</i>)	96

Figure 2.3.28: A volcano plot presenting phenotypes associated with rs2076295 (T > G) (<i>DSP</i>).....	97
Figure 2.3.29: A volcano plot presenting phenotypes associated with rs41308092 (G > A) (<i>RTEL1</i>).....	99
Figure 2.3.30: A volcano plot presenting phenotypes associated with rs7725218 (G > A) (<i>TERT</i>).....	100
Figure 2.3.31: A volcano plot presenting phenotypes associated with rs12699415 (G > A) (<i>MAD1L1</i>).....	101
Figure 2.3.32: A volcano plot presenting phenotypes associated with rs28513081 (A > G) (<i>DEPTOR</i>).....	104
Figure 2.3.33: A volcano plot presenting phenotypes associated with rs9577395 (C > G) (<i>ATP11A</i>)	105
Figure 2.3.34: A volcano plot presenting phenotypes associated with rs59424629 (T > G) (<i>IVD</i>)	106
Figure 2.3.35: A volcano plot presenting phenotypes associated with rs62023891 (G > A) (<i>AKAP13</i>).....	107
Figure 2.3.36: A volcano plot presenting phenotypes associated with rs12610495 (A > G) (<i>DPP9</i>)	108
Figure 2.3.76: A volcano plot presenting phenotypes associated with rs2897075 (C > T) (<i>ZKSCAN1/7q22.1</i>).....	109
Figure 2.3.38: A volcano plot presenting phenotypes associated with rs2013701 (G > T) (<i>FAM13A</i>)	111
Figure 2.3.39: A volcano plot presenting phenotypes associated with rs2077551 (T > C) (<i>MAPT</i>)	112

Figure 2.3.40 A volcano plot presenting phenotypes associated with r rs537322302 (C > G) (<i>HECTD2</i>).....	113
Figure 2.3.41: Phenotypes associated with 14 IPF genetic risk variants in the Genomics England 100,000 Genomes Project cohort.....	114
Figure 3.3.1: Results of genetic correlation analyses of IPF risk with all investigated phenotypes	129
Figure 4.1.1: Diagrams illustrating the three core assumptions required for robust causal estimation using MR	136
Figure 4.2.1 Overview of Mendelian randomisation study assessing if causal relationships exist between diabetes mellitus and idiopathic pulmonary fibrosis.....	142
Figure 4.3.1 Causal estimates from MR analyses of T1D on IPF using IVW-RE and pleiotropy-robust methods.....	144
Figure 4.3.2 Causal estimates from MR analyses of IPF on T1D using IVW-RE and pleiotropy-robust methods.....	145
Figure 4.3.3 Estimates from a SNP leave-one-out MR analysis of T1D on IPF	146
Figure 4.3.4 Estimates from a SNP leave-one-out MR analysis of IPF on T1D	147
Figure 4.3.5 Causal estimates from MR analyses of T2D on IPF using IVW-RE and pleiotropy-robust methods.....	149
Figure 4.3.6 Causal estimates from MR analyses of IPF on T2D using IVW-RE and pleiotropy-robust methods.....	149
Figure 4.3.7 Estimates from a SNP leave-one-out MR analysis of T2D on IPF	150
Figure 4.3.8 Estimates from a SNP leave-one-out MR analysis of IPF on T2D	151
Figure 4.3.9 Causal estimates from MR analyses of HbA1c on IPF using IVW-RE and pleiotropy-robust methods.....	152

Figure 4.3.10 Estimates from a SNP leave-one-out MR analysis of HbA1c on IPF.....	153
Figure 4.3.11 Causal estimates from MR analyses of fasting insulin level on IPF using IVW-RE and pleiotropy-robust methods	154
Figure 4.3.12 Estimates from a SNP leave-one-out MR analysis of fasting insulin level on IPF.....	155
Figure 4.3.13 Causal estimates from MR analyses of BMI on IPF using IVW-RE and pleiotropy-robust methods	156
Figure 4.3.14 Estimates from a SNP leave-one-out MR analysis of BMI on IPF	157
Figure 4.3.15 Causal estimates from multivariable MR analyses of HbA1c and T2D exposures on IPF using MVMR-IVW and pleiotropy-robust methods	158
Figure 4.3.16 Causal estimates from multivariable MR analyses of fasting insulin level and T2D exposures on IPF using MVMR-IVW and pleiotropy-robust methods.....	159
Figure 4.3.17 Causal estimates from multivariable MR analyses of BMI and T2D exposures on IPF using MVMR-IVW and pleiotropy-robust methods.....	160
Figure 4.3.18 Causal estimates from MR analyses of T2D (European ancestry GWAS) on IPF using IVW-RE and pleiotropy-robust methods.....	161
Figure 4.3.19 Causal estimates from MR analyses of IPF (all 19 susceptibility SNPs) on T2D (European ancestry GWAS) using IVW-RE and pleiotropy-robust methods.....	162
Figure 4.3.20 Causal estimates from MR analyses of HbA1c (glycemic SNPs only) on IPF using IVW-RE and pleiotropy-robust methods.....	163
Figure 5.2.1: Overview of Mendelian randomisation study investigating possible causal relationships between mLOY and IPF.....	173
Figure 5.3.1: Causal estimates from MR analyses of mLOY on IPF using IVW-RE and pleiotropy-robust methods.....	175

Figure 5.3.2: Estimates from a SNP leave-one-out MR analysis of mLOY on IPF	176
Figure 5.3.3: Causal estimates from MR analyses of IPF on mLOY using IVW-RE and pleiotropy-robust methods.....	177
Figure 5.3.4: Estimates from a SNP leave-one-out MR analysis of IPF on mLOY	178
Figure 5.3.5: Causal estimates from MR analyses of mLOY on prostate cancer risk using IVW-RE and pleiotropy-robust methods	179
Figure 5.3.6: Estimates from a SNP leave-one-out MR analysis of mLOY on prostate cancer risk	180
Figure 5.3.7: Causal estimates from MR analyses of mLOY on ovarian cancer risk using IVW-RE and pleiotropy-robust methods	181
Figure 5.3.8: Estimates from a SNP leave-one-out MR analysis of mLOY on ovarian cancer risk	182
Figure 5.3.9: Causal estimates from MR analyses of mLOY on IPF risk (female IPF cases only) using IVW-RE and pleiotropy-robust methods.....	183
Figure 5.3.10: Estimates from a SNP leave-one-out MR analysis of mLOY on IPF risk (female IPF cases only)	184
Figure 5.3.11: Causal estimates from MR analyses of telomere length (leukocyte) on mLOY using IVW-RE and pleiotropy-robust methods	186
Figure 5.3.12: Causal estimates from MR analyses of telomere length (leukocyte) on mLOY using IVW-RE and pleiotropy-robust methods	187

List of Tables

Table 1.2.1 Common genetic variants associated with IPF risk.....	35
Table 2.2.1: Genetic variants investigated in this PheWAS study and functional categorisation for each SNP	46
Table 2.3.1: Phenotypes prioritised for colocalisation analyses with IPF association signals	84
Table 2.3.2: Phenotypes prioritised for colocalisation analyses with IPF association signals	85
Table 2.3.3: A summary table of phenotypes with strong evidence of colocalisation with IPF risk association signals.....	91
Table 2.3.4: A summary table of eQTL association signals with strong evidence of colocalisation with IPF risk.....	92
Table 2.3.5: A summary table of phenotypes associated with IPF genetic risk variants in the Genomics England 100,000 Genomes project database.....	115
Table 3.3.1: Traits selected for genetic correlation analysis following identification of strong evidence of shared causal variants with IPF risk association signals	126
Table 3.3.2: Fibrotic disease traits selected for genetic correlation analysis	127
Table 3.3.3: Results of genetic correlation analyses of IPF risk with traits prioritised through colocalisation with IPF genetic risk variants.....	128
Table 3.3.4: Results of genetic correlation analyses of IPF risk with a sample of fibrotic disease phenotypes.....	129

1: Introduction

Idiopathic pulmonary fibrosis (IPF) is a disease of progressive and uncontrolled scarring of the lung. Our understanding of the biology of IPF is incomplete, and treatments for this devastating disease are limited. In this PhD thesis I aim to investigate how IPF biology relates to that of other human traits to develop understanding of disease mechanisms and potentially highlight opportunities for novel or repurposed treatments. In this chapter I will introduce genetics, how understanding our genomes can provide insights into human disease, and how we may be able to utilise this knowledge to improve treatments and outcomes for people with IPF.

1.1 Introduction to idiopathic pulmonary fibrosis

1.1.1: Clinical features and current treatments

Idiopathic pulmonary fibrosis (IPF) is characterised by progressive and uncontrolled scarring of lung tissue. Lungs are organs responsible for transporting oxygen from the air into the bloodstream and transporting carbon dioxide out. The development of fibrosis limits lung functionality, resulting in increasing respiratory difficulty. This decline in respiratory function over time is typically fatal within three-to-five years of diagnosis. IPF is considered to be a rare disease, but incidence has been shown to be comparable to many common cancers and 5-year survival rates are often worse^{1,2}.

IPF is the most common interstitial lung disease (ILD), which is a group of over 200 diseases which primarily affect the pulmonary interstitium. The interstitium constitutes of a group of tissues that structurally support lung alveoli and facilitate gas exchange between the alveolar space and capillary bloodstream. The characteristic features of ILDs are chronic inflammation and fibrosis of interstitial tissues, which physically restricts breathing and

reduces gas diffusion (Figure 1.1.1)³. While hypercapnia due to reduced CO₂ diffusion has been reported in patients with late-stage disease, fibrotic ILD symptoms primarily result from reduced oxygen diffusion into the bloodstream^{3,4}.

The development of fibrosis in ILDs is believed to be induced by damage to the alveolar epithelium, which can result from autoimmune injury or from damaging environmental exposure⁵ (as with asbestosis or silicosis). The factors contributing to the pathogenesis of IPF are less clear than for other ILDs, however current understanding suggests a complex interaction of chronic epithelial damage, accelerated lung ageing, and genetic susceptibility may underlie disease risk⁶. Common risk factors associated with IPF risk include male biological sex, increased age, and cigarette smoking⁶. Shortened telomere length is also a strong risk factor for IPF⁷ which may be further exacerbated by the effects of age and smoking^{8,9}. Increased IPF risk is also associated with occupational and environmental exposures, including exposure to wood dust, metal dust, and pesticides, and respiratory infection^{10,11}.

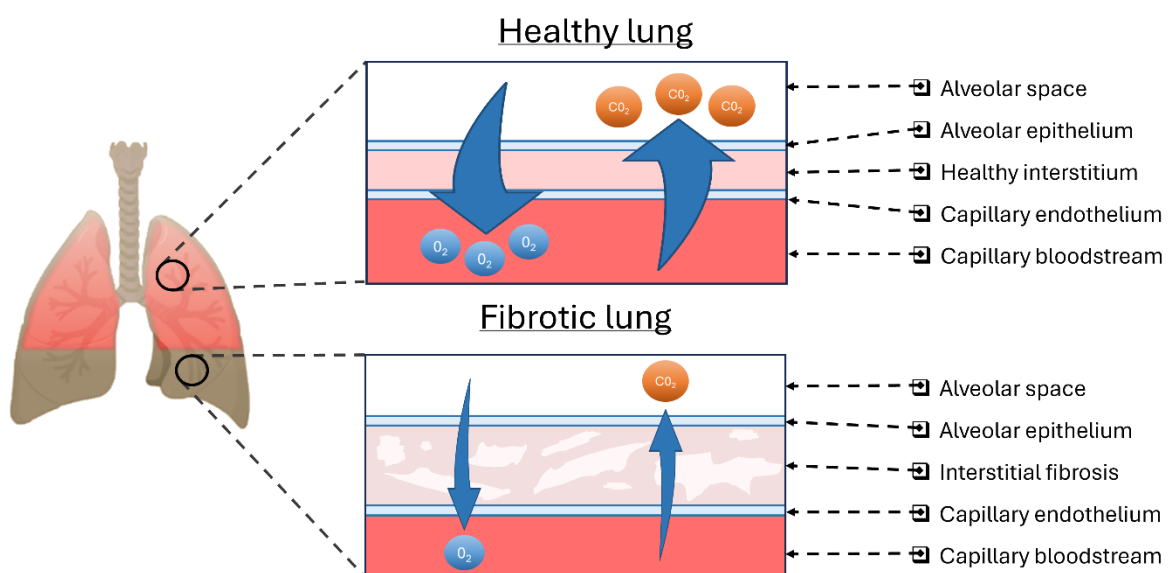


Figure 1.1.1: A diagram depicting how interstitial fibrosis leads to reduced gas exchange compared to healthy lung tissue. Reduced diffusion rate across thickened fibrotic interstitial tissue reduces oxygen (O₂; blue) diffusion from inhaled air into the bloodstream, and carbon dioxide (CO₂; orange) out.

IPF often presents in patients as shortness of breath (dyspnoea) and a dry cough. Typically, development of symptoms leads to identification of suspected ILD. To be diagnosed with IPF, the possibility of other ILDs or identifiable cause must be excluded⁶. Potential identifiable causes that would lead to a non-IPF diagnosis include connective tissue disease-related ILD (CTD-ILD), such as systemic sclerosis¹², or damaging exposures related to occupation, domestic environment, or drug-induced lung damage¹³. Under current guidelines¹⁴, if a multidisciplinary team find no identifiable cause of ILD (idiopathic) and a pattern of usual interstitial pneumonia (UIP) is identified through high-resolution computed tomography (HRCT) and/or lung biopsy, patients are diagnosed with IPF.

Characteristic IPF disease progression involves progressive decline of lung function, increased dyspnoea, and greatly reduced exercise tolerance. However, disease course can be variable across individuals, with a subgroup of patients experiencing a slow and relatively continuous decline and others possibly experiencing periods of rapid deterioration (acute exacerbations)¹⁵. In addition, by the time of symptom presentation and diagnosis, IPF disease progression can already be at a relatively advanced stage. In response to these disease features, many recent research efforts have focused on characterising the preclinical stages of IPF and predicting individual disease progression. As an example, interstitial lung abnormalities (ILAs) are abnormal findings during lung CT imaging that may be the key to characterising early stages of ILDs¹⁶. A high proportion of patients identified with ILAs have been shown to have a subsequent ILD diagnosis¹⁷ and studies have indicated considerable overlap of IPF and ILA genetic risk^{18,19}. In addition, the PROFILE (Prospective Observation of Fibrosis in the Lung Clinical Endpoints) study²⁰ has enabled the identification of distinct subgroups of lung function trajectories in IPF patients²¹, as well as serum biomarkers associated with disease progression^{22,23}.

Two antifibrotic treatments, pirfenidone and nintedanib, were approved by the U.S. Food and Drug Administration (FDA) in 2014. The approval of these two antifibrotic drugs

represented a major milestone in IPF treatment, as both were shown to reduce the rate of disease progression and lung function decline^{24,25}.

Pirfenidone is a synthetic derivative of pyridone that exhibits both anti-inflammatory and anti-fibrotic effects through multiple pathways. For example, pirfenidone has been shown to inhibit production of pro-inflammatory cytokines such as interferon gamma, $\text{TNF}\alpha$, and interleukin-6 in mouse models^{26,27}. The anti-fibrotic properties of pirfenidone are hypothesised to act through central fibrotic processes, such as suppression of fibroblast proliferation and collagen synthesis via inhibition of the $\text{TGF-}\beta 1$ signaling pathway²⁸⁻³⁰. While treatment with pirfenidone can slow lung function decline in IPF patients, it does not permanently stop or reverse disease progression. Pirfenidone use is also associated with adverse effects, with common side effects of pirfenidone treatment including nausea, diarrhoea, and fatigue (each occurring in roughly 15-20% of patients).

Nintedanib is an intracellular inhibitor of tyrosine kinases that, similarly to pirfenidone, has both anti-inflammatory and anti-fibrotic effects. The mechanisms of action of nintedanib include the inhibition of the platelet-derived growth factor (PDGF), fibroblast growth factor (FGF), and vascular endothelial growth factor (VEGF) signaling pathways. Through these pathways nintedanib is hypothesised to suppress fibroblast proliferation and extracellular matrix secretion, resulting in a slower rate of IPF disease progression³¹. Treatment with nintedanib is also associated with diarrhoea (60% of patients), nausea (24.5% of patients), vomiting (11.6% of patients), and reduced appetite (10.7% of patients), with between 5% and 10% of patients permanently discontinuing drug treatment due to these side effects^{32,33}.

Currently, the only other survival-improving intervention for IPF than pirfenidone and nintedanib is lung transplantation. While a lung transplant reduces mortality rate by roughly 75%, this intervention is not available to most IPF patients due to age over 70 and/or having serious comorbidities^{34,35}.

In summary, despite progressions in IPF treatment options over the past decade, IPF is still a fatal and life changing disease with no cure for most affected patients. Further, current treatments are associated with a range of life-limiting and unpleasant side effects. In particular, the adverse effects of current IPF treatments may limit use for early stage or less progressive disease. As research efforts continue to characterise pre-clinical disease stages and prognostic markers, earlier diagnosis and effective prediction of disease course may become possible through biomarkers, ILA screening, and genetic testing. This future potential only increases the current need for novel IPF treatments that are safe, have minimal side effects, and improve outcomes for IPF patients.

1.1.2 IPF pathogenesis and disease biology

The current paradigm of pathogenesis involves individuals that are genetically primed to developing IPF being exposed to damaging environmental stimuli, or effects of certain comorbidities. Environmental stimuli such as cigarette smoking, wood or metal dusts, or respiratory infections can cause microinjury to the alveolar epithelium, which can trigger a positive feedback loop of aberrant epithelial wound healing, pro-fibrotic signaling, and extracellular matrix deposition⁶.

Aberrant epithelial repair in IPF is hypothesised to be partly driven by an exaggerated or premature ageing phenotype. As an example, dysfunctional alveolar epithelial cells in IPF commonly show many of the hallmarks of ageing such as telomere attrition, epigenetic alterations, altered intercellular communications, and cellular senescence^{36,37}. In addition, IPF is increasingly common with age, which may be due to the interaction of the hallmarks of ageing with inherited genetics, or due to an extended period of interaction with potentially damaging environmental stimuli, or a combination of all three factors. Ultimately, the interaction of known risk factors in the pathogenesis of IPF is highly complex, with the relative proportion of risk contributed by each of these factors likely to vary across affected individuals.

1.1.3 IPF multimorbidity and fibrosis beyond the lung

An important feature of IPF disease is the high rate of comorbidity. A previous study of over 9,000 found that patients with an IPF diagnosis were more likely to additionally have other non-IPF conditions than age and sex-matched controls³⁸. Common comorbidities associated with IPF include gastro-oesophageal reflux disease (GORD), pulmonary hypertension, chronic obstructive pulmonary disease (COPD), lung cancer, diabetes mellitus (DM), and depression or anxiety^{6,39}.

The theorised factors that lead to the group of comorbidities associated with IPF are both diverse and potentially complex. In some cases, a comorbid relationship may be driven by shared environmental risk factors. As an example, the common environmental risk factor of smoking is one possible explanation for the comorbidity link between IPF¹⁰, COPD⁴⁰, and lung cancer⁴¹. Alternatively, a comorbidity may exist in some form prior to IPF that can trigger or exacerbate IPF disease. An example of comorbidity due to a risk increasing effect is GORD, which has been shown to causally affect IPF risk⁴². Comorbidities may also be driven by the mechanisms and symptomatic effects of IPF. Psychological effects of IPF diagnosis can be very impactful on patient quality of life and depressive symptoms are strongly correlated with the severity of respiratory symptoms⁴³. The mechanisms underlying the sphere of IPF comorbidities may be complex, with associations with each trait being driven by multiple pathways associated with IPF. The comorbid relationship of IPF with DM⁴⁴ is one such example. It has been proposed that hyperglycaemic effects of DM may exacerbate fibrosis through increased oxidative stress, DNA damage, and subsequently increased cellular senescence⁴⁵. The use of corticosteroids to treat IPF patients is also known to increase blood glucose levels, which can lead to treatment-induced DM⁴⁶. As with many IPF comorbidities, DM, particularly type 2 (T2D), is associated with older age⁴⁷ and therefore highlights the potential that premature ageing of lung tissues in IPF may be an early presentation of a more systemic accelerated ageing or senescence

phenotype. The positive epidemiological relationship between IPF and DM⁴⁴ could result from one or of these potential links between these diseases or may result from a combined effect of multiple pathways. The cellular mechanisms of ageing also overlap with the effects of damaging environmental factors, such as smoking, which may further contribute to comorbid relationships through this pathway of effect⁴⁸.

Fundamentally, the mechanisms or risk factors underlying the range of comorbidities associated with IPF are poorly characterised. This gap in understanding is directly relevant to clinical practice, as the number of comorbidities affecting IPF patients is negatively correlated with survival⁴⁹. Further, findings on the nature of comorbid associations with IPF may have different implications depending on the orientation of traits along causal pathways. Comorbidities that present upstream of IPF, therefore altering IPF risk, would indicate a greater requirement for management of the causal comorbidity to prevent the development or accelerated progression of IPF (Figure 1.1.2). Factors found to cause IPF may also provide insights into IPF biology and therapeutically targetable pathways, as comorbidities may induce or contribute to IPF disease mechanisms. As an example, if DM was found to cause IPF, this could implicate DM-associated factors such as chronic hyperglycaemia, dysregulated insulin signaling, or obesity in IPF pathogenesis. Alternatively, comorbidities that are found to be caused or exacerbated by IPF disease (presenting downstream of IPF) should be monitored more intensively in IPF patients to reduce impacts on survival and quality of life. Another possibility is that IPF may be linked to comorbidities through shared risk factors or biological processes acting independently on each trait, which can highlight the potential for shared management strategies or drug repurposing.

Fibrosis can affect most organs, typically as a late disease stage, resulting in decline of organ function as fibrotic tissue develops. As well as in IPF, excessive formation of fibrotic tissue through deposition of extracellular matrix is a common feature of serious liver, kidney, and

heart disease^{50,51}. Fibrosis is also associated with autoimmune conditions such as rheumatoid arthritis, Crohn's disease, systemic lupus erythematosus, and myelofibrosis⁵².

A common feature of many fibrotic diseases is the combination of a genetic risk background with the damaging effects of environmental or endogenous stimuli. Environmental factors that initiate or drive fibrosis may vary across tissues depending on the types of exposures relating to organ function, but tissue damage is a central theme in the development of fibrosis. As an example, cigarette smoking or pollution may cause or drive IPF due to lung tissue damage during and following inhalation, while excessive alcohol consumption causing liver fibrosis in alcoholic liver disease is related to the detoxifying function of this organ⁵³. In this example both cigarette smoking and alcohol may cause tissue damage, initiating a dysfunctional healing response and therefore causing fibrosis, but are risk factors for different fibrotic diseases due to varying exposure and cellular environments across organs. This paradigm implies that while fibrotic diseases may have some distinct pathogenic mechanisms, there is clear potential for shared pathways and overlapping risk factors. TGF- β signaling is one such example of a shared pathway, with TGF- β levels being positively associated with progression of fibrosis in lung, kidney, liver, and heart tissue⁵⁴. The TGF- β pathway may therefore appear to be a prime target for treating multiple fibrotic diseases, however the extensive regulatory function means that disruption of this pathway can promote cancer development⁵⁵. Some risk factors have also been identified in association with multiple fibrotic diseases, such as the association of loss of Y chromosome with cardiac fibrosis and primary biliary cirrhosis^{56,57}. In addition, it is plausible to consider that risk factors and biological mechanisms found to drive fibrosis in one disease may contribute to other fibrotic phenotypes. Such factors therefore present an important opportunity for targeted investigation across fibrotic conditions.

Fibrotic multimorbidity, where individuals are affected by more than one fibrotic condition, was shown to be present in 6% of patients in a recent study⁵⁸. The same study

also highlighted a positive epidemiological relationship between pulmonary fibrosis and multiple fibrotic conditions, including cardiac fibrosis, bile duct fibrosis, reproductive fibrosis, and fibrosis of the intestine and pancreas⁵⁸. Overall, current evidence highlights considerable potential that mechanistic similarities may exist across fibrotic conditions. The identification of common fibrotic pathways and disease similarities is therefore of high importance. In addition, identifying genetic and mechanistic similarities with related fibrotic conditions is a promising avenue for prioritising established therapeutic targets that could be repurposed for the treatment of IPF. However, in-depth understanding of IPF

biology is also required to facilitate identification and effective targeting of potential shared pathways.

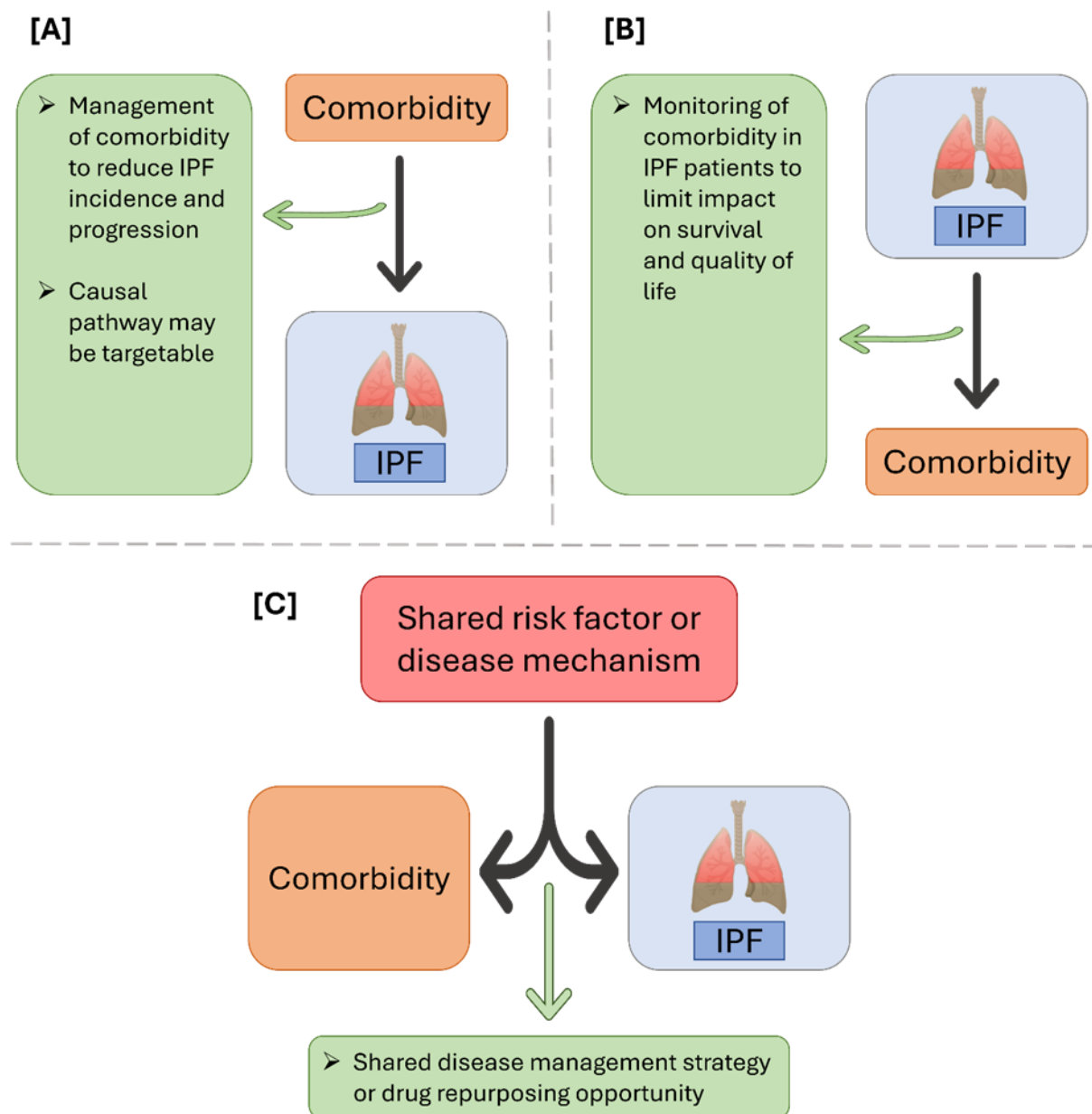


Figure 1.1.2: A diagrammatic representation of how comorbidities may be mechanistically linked to IPF and potential responses to different pathways of comorbidity. [A] Association driven by a causal effect of a comorbidity on IPF (IPF is downstream of comorbidity in mechanistic pathway). [B] Comorbidity resulting from a causal effect of IPF on a downstream trait. [C] An association between diseases driven by a shared risk factor or disease mechanism that is acting on IPF and the respective comorbidity independently. Green arrows and green boxes indicate potential responses once the direction of effect underlying associated traits is characterised.

1.2 Introduction to genetics and IPF genetic risk

1.2.1 Genetics, mutations, and phenotypes

Human genetics is the mechanism by which we inherit certain intrinsic traits such as eye colour, height, and how likely we are to develop certain diseases during our lifetime. Genetic information is contained within a molecule that is present in nearly every cell in our body called DNA (deoxyribonucleic acid). DNA is comprised of four different bases: adenine (A), cytosine (C), guanine (G), and thymine (T). These four bases combine with sugar-phosphate backbone to form a string of nucleotides to comprise the genetic code, which acts as an instruction manual for every biological process in the human body. Each DNA strand forms hydrogen bonds with another strand with a complementary sequence of nucleotides, as A specifically bonds with T and G specifically bonds with C (Figure 1.2.1). Complementary strands of DNA enable the formation of a more stable double-helix structure and provide a template for replication and repair. In human cells DNA is packaged into 23 pairs of chromosomes, with each pair containing a chromosome from each parent.

Genes are sequences of DNA that have a functional effect on biological mechanisms or traits when expressed. To influence a particular trait, the nucleotide sequence in a gene is transcribed to single-stranded messenger ribonucleic acid (mRNA), which is subsequently translated from nucleotide into a peptide sequence during protein synthesis. An individual's genetic make-up is referred to as a genotype, which mostly influences the corresponding range of observable traits (phenotype) through the transference of information from DNA to RNA to proteins.

The human genome contains roughly 20,000 protein coding genes, but this only constitutes 1.2% of the genome sequence^{59,60}. Initially, the remaining 98.8% of the genome

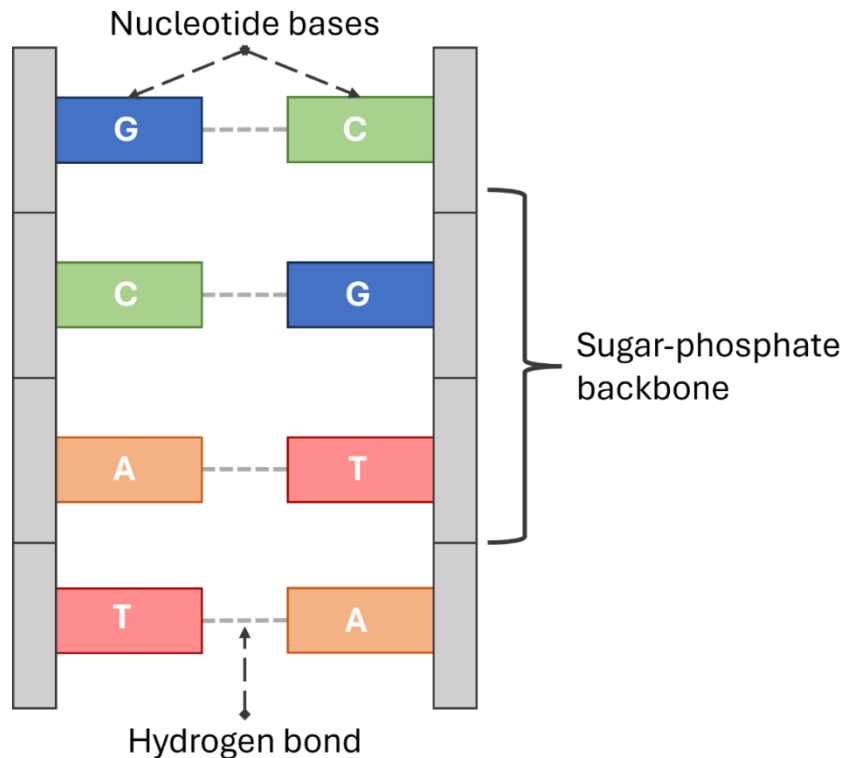


Figure 1.2.1: A simplified illustration of the structure of double stranded DNA. Nucleotide bases are indicated by G: Guanine (blue), C: Cytosine (green), A: Adenine (orange), T: Thymine (red).

was believed to be non-functional (commonly referred to as “junk” DNA). Over time, the crucial regulatory role of the noncoding genome has been uncovered. Mechanisms of gene regulation facilitate varying expression across different tissues and cell types, adaptation to a variable environment, and temporal control of gene expression.

In human cells, DNA is replicated during normal cell division. During replication, double stranded DNA is unzipped by DNA helicase enzymes and complementary pairing allows an exposed sequence (parental strand) to be used as a template for DNA polymerase to synthesise a new strand (daughter strand) (Figure 1.2.2). Despite the high efficacy of proofreading and sequence error repair mechanisms, DNA replication errors occur at a rate of roughly 1 in 10,000 to 1 in 100,000 nucleotides^{61,62}. If sequence errors persist through DNA replication during cell division, they become permanent mutations and are inherited through subsequent cell lineages. If mutations occur in gametes (sperm or egg cells) they

can be passed from parents to offspring. These germline mutations can result in multiple forms of a sequence at a given genomic position existing in a population (alleles). The genetic sequence changes that can vary across individuals are referred to as genetic variants (Figure 1.5). Some germline mutations increase the reproductive fitness of individuals that carry them, which often results in an increased frequency of that mutation in the population. Many germline mutations are deleterious and either exist at a lower frequency, maintained by factors such as *de novo* mutation rate and genetic drift, or are removed from the population due to selection. There are also situations where the fitness conferred by certain mutations is more complex.

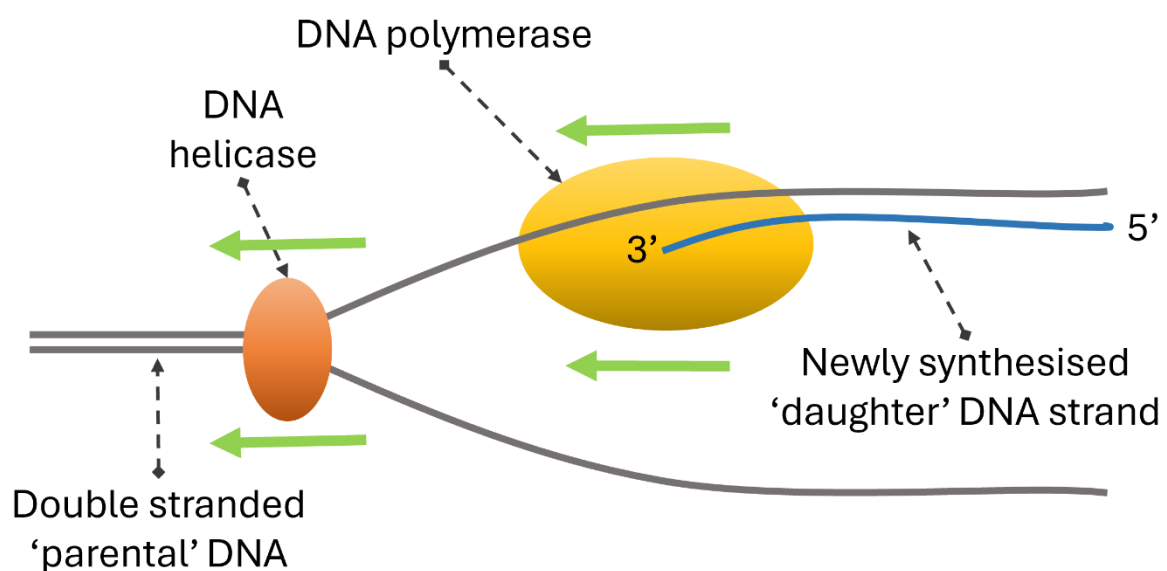


Figure 1.2.2: DNA replication of the leading strand. Double stranded 'parental' DNA is separated into individual strands by DNA helicase (orange). The exposed DNA sequence acts as a template for DNA polymerase (yellow) to synthesise a new 'daughter' strand (blue). Arrows (green) show the direction of DNA replication fork progression.

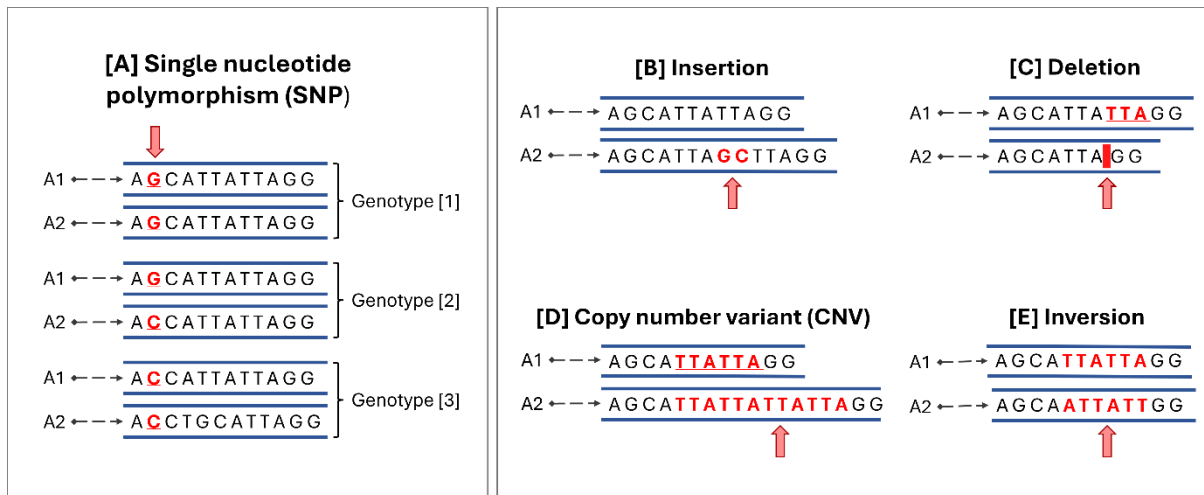


Figure 1.2.3: Examples of different types of genetic variation. [A] A single nucleotide polymorphism (SNP) at a single base position. Genotype [1] has a G nucleotide for both inherited alleles (A1 and A2). For this SNP, genotype [1] would be referred to as a GG genotype, where genotype [2] is a GC genotype, and genotype [3] is a CC genotype. [B] A two-base insertion of a GC nucleotide sequence in A2 compared to A1. [C] A three-base deletion of a TTA nucleotide sequence in A2 compared to A1. [D] A copy number variant (CNV), where a string of two repeats of a three-nucleotide sequence TTA in A1 has changed to four repeats of the same TTA sequence in A2. This could also be a duplication. [E] An inversion of the TTA repeat sequence present in A1, resulting in a reversed set of nucleotides in A2 compared to A1.

1.2.2 Polygenic risk of human traits

Many human traits are influenced by multiple genetic loci and do not follow Mendel's laws of inheritance. An individual's overall risk for a particular trait can be estimated by aggregating effects from often hundreds of genetic variants using a polygenic risk score (PRS)⁶³. To estimate individual polygenic risk, the genetic variants associated with a particular trait or disease must first be mapped out across the genome, typically using a genome-wide association study (GWAS). A typical GWAS study design involves identifying a large sample of individuals affected by a phenotype, a large sample of unaffected controls, and comparing the genetics of these two groups (Figure 1.2.4)⁶⁴. Statistical comparison of genetics across affected and control populations allows researchers to identify genetic variants that are strongly associated with the investigated phenotype. Results of GWAS

studies provide insights into the biology underlying human traits, which can contribute to understanding why diseases occur and how they can be treated.

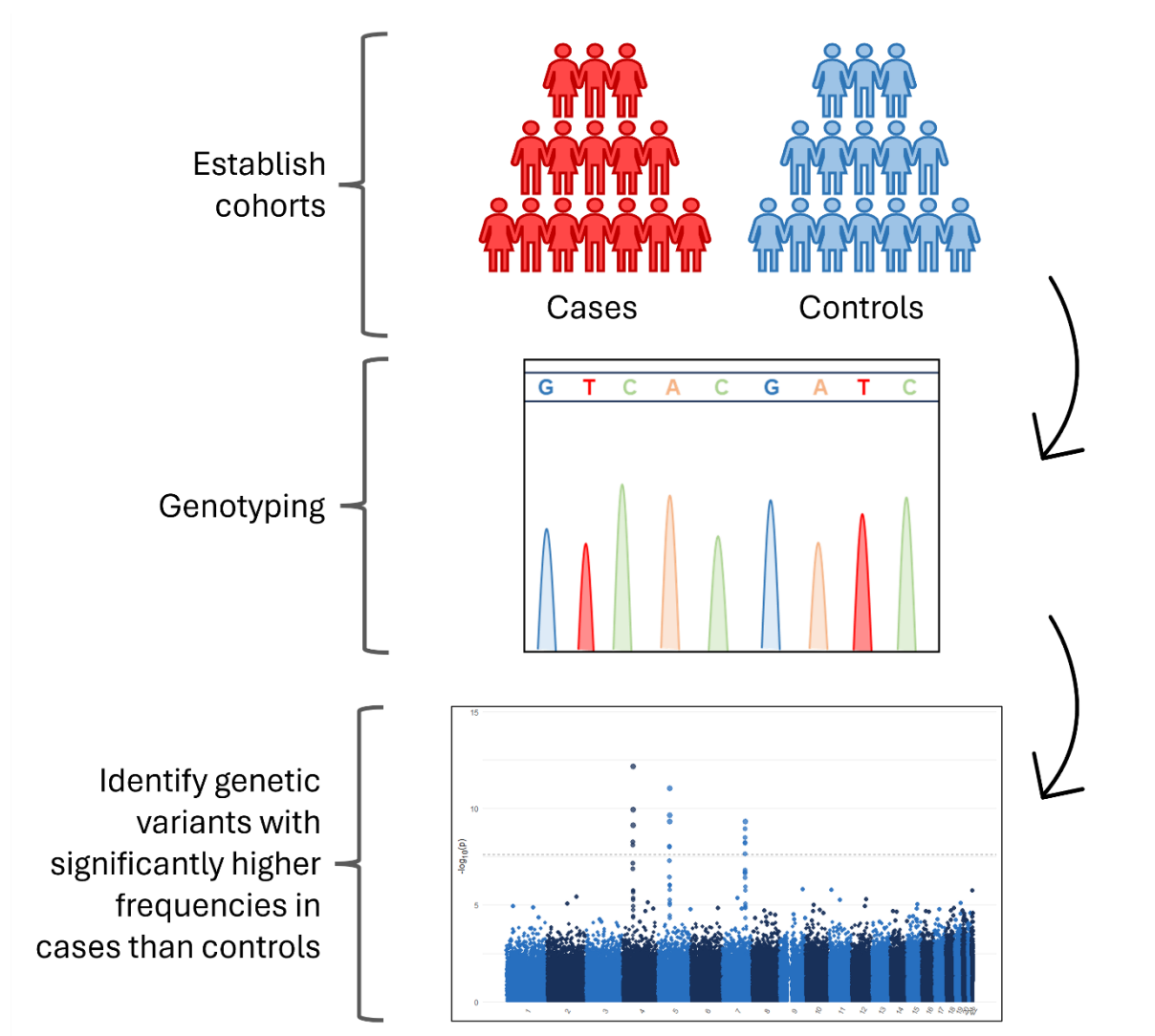


Figure 1.2.4: Illustration of the basic methodology of genome-wide association studies (GWAS). First, consenting participants are identified or recruited and phenotypic data are typically used to classify cases and controls for the trait of interest. The participants of each cohort are genotyped, often using targeted microarrays or next generation sequencing approaches. Retrieved data is processed, anonymised, and analysed to identify statistically significant associations between genetic variants and the trait of interest. Often, replication GWAS are subsequently performed using an independent cohort. Results are subsequently analysed further, for example by using fine-mapping approaches to identify likely causal variants from a genetically linked region.

If whole genome sequencing was required to genotype and subsequently compare the genetics of affected and control populations, historical GWAS studies would be very

computationally intensive and expensive to run. Fortunately, some sections of the genome are highly likely to be inherited together (linkage disequilibrium, or *LD*) and therefore more consistently contain the same genetic variation in a population. Other genomic sections are uncorrelated, either due to being on different chromosomes or being separated by recombination hotspots. Genetic recombination refers to the shuffling of genetic information, often between homologous chromosomes, during meiosis to promote genetic variation. This results in genetic variants that are closer together often being genetically linked, essentially subdividing the genome into highly correlated sections (LD blocks). This phenomenon means single variants can be used to effectively represent entire LD blocks, which enables researchers to capture a large amount of genetic variation using a relatively small number of SNPs. The International HapMap project⁶⁵ produced the first database of common genetic variation in 2005, showing that over 80% of common human genetic variation could be captured using roughly one million SNPs. The HapMap project was fundamental in enabling the first GWAS studies but has since been superseded by updated projects such as the 1000 Genomes project^{66,67}.

Between 2005 and 2018 3,639 GWAS studies were conducted across 3,508 traits⁶⁸, with the results of many GWAS being made publicly available through databases such as the EBI GWAS Catalog⁶⁹. GWAS results are typically available in a summary statistics format, containing data on the strength of association between a large number of SNPs and the trait of interest. Each row of the GWAS summary statistics data file will usually contain a variant identifier, such as an rsID (Reference SNP cluster ID), positional data describing genomic position (chromosome number and base position), and information on the altered allele that results in the effect on the trait of interest (effect allele and non-effect allele). The effect of the change in allele associated with each SNP is quantified using an effect estimate, often beta values or odds ratios, with a respective error statistic (such as standard error). A p-value statistic for each association quantifies how likely the observed

association is to occur under the null hypothesis that there are no true associations in the GWAS dataset. An extremely low p-value therefore provides greater confidence that an observed association is real. In most GWAS studies, a p-value threshold of $p < 5 \times 10^{-8}$ is typically used to filter for genome-wide significant associations. The use of this standardised p-value threshold is advantageous to minimise false positive results, but discussion of how to define such thresholds is continuously ongoing with new contextual challenges and as increasingly large datasets are developed^{70–72}. GWAS results can also include sample size for each variant (N) and the minor allele frequency (MAF) or effect allele frequency (EAF) within the study population.

A core motivation of performing GWAS is to identify the biological cause of heritable human traits. By understanding this underlying biology, this area of research aims to facilitate novel and effective therapeutic targeting, in the hope of leading to improved outcomes and quality of life for patients. In addition, prioritising genetically supported targets in clinical studies is estimated to double the likelihood of drug approval⁷³. The high volume of GWAS studies that have now been published, particularly with the development of public GWAS databases, has created an invaluable resource for further study. A finding of particular interest is that 67.2% of genes and 32.4% of common SNPs are associated with multiple traits⁷⁴. This phenomenon of single genetic loci, genes, or SNPs being associated with more than one trait is known as pleiotropy. When multiple traits are influenced by the same genetic change, this may be an indication of shared mechanisms underlying otherwise independent traits. Alternatively, pleiotropy can suggest the presence of interacting phenotypes where one trait is contributing to the aetiology of another. The existence of either form of pleiotropy can give insights into therapeutically targetable biology. The requirement for translating genetic association signals from GWAS into functional understanding and actionable targets has led to the development of a suite of innovative computational methods. Using these methods to study pleiotropy is a powerful

way to investigate disease risk factors, interacting disease mechanisms, and identify opportunities to discover novel uses for approved therapies. This use of a drug or therapy to treat a disease or condition different to that which it was originally developed is known as drug *repurposing* or drug *repositioning*.

1.2.3 Modern genomic approaches in disease research

Access to GWAS summary data has led to the development of many innovative *in silico* approaches that can provide powerful insights into the factors and mechanisms underlying human disease using open resources.

The finding that pleiotropy is highly prevalent across the human genome⁷⁴ has led to the development of study designs aiming to understand whether a particular genotype, such as a genetic risk variant, is associated with variation in any other phenotypes. This approach is referred to as a phenome-wide association study (PheWAS)⁷⁵ and can be performed using public GWAS data. This approach is complemented by methods that interrogate overlapping association signals such as colocalisation analysis, which aims to assess if phenotypes are associated with the same genetic variant^{76,77}. PheWAS are particularly powerful for understanding the biological effects of genetic variants and can potentially highlight shared or interacting biological pathways across human traits. Findings from PheWAS can also contribute to pharmaceutical research by identifying drug repurposing opportunities and highlighting potential adverse effects of therapeutic intervention⁷⁸.

Methods such as genetic correlation analyses utilise GWAS data to investigate broader genetic relationships. Genetic correlation can highlight shared genetic architecture and similar disease aetiology^{79,80}. In some cases, genetic correlation can also be an indicator of a causal relationship between traits. Causal relationships between traits can be investigated with a Mendelian randomisation (MR) analysis, which can also be performed using genetic variants identified in GWAS studies. MR studies typically use genetic variants as

instrumental variables to estimate causal effects of an exposure on an outcome. An MR study design has frequently been applied to focus on the effects of modifiable risk factors on disease^{81–84}. MR can be particularly powerful when the investigation of causal relationships between exposure and outcome variables would be unfeasible, costly, or unethical through other methods, such as through randomised clinical trials⁸⁵.

1.2.4 IPF genetics and implicated disease biology

Over the last two decades, multiple GWAS studies have identified 19 common genetic variants associated with IPF risk^{86–90} (Table 1.2.1). Inclusion of IPF cases in these studies required patient diagnosis to have followed official American Thoracic Society (ATS)/European Respiratory Society (ERS)/Japanese Respiratory Society (JRS)/Latin American Thoracic Society (ALAT) diagnostic guidelines^{14,90}. These studies currently represent the highest standard of IPF GWAS due to the rigorous IPF case inclusion criteria applied. A subset of 14 of these variants have been shown to explain 12.4% of overall IPF disease liability, increasing to 17.7% in people aged over 65⁹¹. A *MUC5B* promoter SNP, rs35705950, was shown to explain up to 9.4% of overall disease risk. The overall estimate of common genetic contribution to disease liability is likely to increase as larger sample sizes facilitate discovery of new IPF risk variants, such as the additional five SNPs identified in the most recent GWAS study published in 2022⁹⁰. The 19 common genetic variants associated with IPF were functionally annotated in recent studies using the Ensembl Variant Effect Predictor (VEP)^{89,90,92}.

The rs35705950 SNP (G>T), located within the promoter region of the mucin-5B gene (*MUC5B*), is strongly associated with increased risk of IPF^{87,89,93} and other interstitial lung diseases^{94–96}. The identification of a strong association between the *MUC5B* promoter SNP, rs3570950 (G>T), and IPF risk implicates a role of mucin biology in IPF pathobiology. This polymorphism is associated with increased *MUC5B* gene expression, with increased *MUC5B* protein expression shown to cause mucociliary dysfunction and lung fibrosis in

Functional category	SNP RSID	Chromosome	Base position	Gene / locus	Effect allele	Non-effect allele	Effect allele frequency (EAF)	OR (95% CI)	P-value
Mucin biology	rs35705950	11	1241221	MUC5B	T	G	14.50%	5.06 [4.69, 5.47]	9.09×10^{-418}
Cytoskeletal proteins	rs112087793	20	62284170	STMN3	C	T	91.50%	1.34 [1.21, 1.48]	1.09×10^{-8}
	rs2292181	3	44903434	KIF15	C	G	5.20%	1.52 [1.36, 1.70]	3.95×10^{-13}
	rs2076295	6	7563232	DSP	G	T	46.70%	1.49 [1.41, 1.57]	1.50×10^{-48}
Telomere biology	rs9860874	3	169486271	TERC	A	C	27.60%	1.29 [1.22, 1.37]	6.49×10^{-18}
	rs41308092	20	62324391	RTEL1	A	G	2.10%	1.75 [1.45, 2.10]	3.13×10^{-9}
	rs7725218	5	1282414	TERT	G	A	67.10%	1.41 [1.33, 1.50]	4.90×10^{-32}
Spindle assembly and cell cycle control	rs12537430	7	1868761	MAD1L1	G	A	62.50%	1.28 [1.21, 1.35]	4.20×10^{-18}
	rs12912339	15	40931708	KNL1	A	G	15.90%	1.30 [1.21, 1.39]	7.41×10^{-13}
MTOR pathway signaling	rs74614704	16	162240	NPRL3	A	G	5.60%	1.49 [1.33, 1.67]	2.57×10^{-12}
	rs10808505	8	120940206	DEPTOR	T	G	57.30%	1.20 [1.13, 1.26]	6.03×10^{-11}
Cell signaling and membrane trafficking	rs9577395	13	113534984	ATP11A	C	G	79.10%	1.29 [1.21, 1.38]	4.78×10^{-14}
	rs2304645	15	40716253	IVD	C	G	52.60%	1.28 [1.21, 1.35]	8.66×10^{-20}
	rs11073517	15	86287910	AKAP13	T	C	32.70%	1.19 [1.13, 1.26]	1.36×10^{-9}
	rs12610495	19	4717672	DPP9	G	A	30.60%	1.28 [1.21, 1.36]	2.58×10^{-16}
	rs2897075	7	99630342	ZKSCAN1	T	C	38.20%	1.30 [1.23, 1.37]	1.77×10^{-21}
Other or unclassified	rs2077551	17	44214888	17q21.31	T	C	80.70%	1.42 [1.32, 1.53]	1.92×10^{-20}
	rs79684490	10	111229861	10q25.1	A	G	4.60%	1.40 [1.24, 1.57]	3.52×10^{-8}
	rs2609259	4	89837808	FAM13A	A	C	22.40%	1.30 [1.22, 1.39]	6.47×10^{-17}

Table 1.2.1: Common genetic variants associated with IPF risk. Association statistics for each variant were derived from the most recent IPF risk GWAS⁹⁰. Functional categories were defined based on the predicted biological function^{102,103} of the gene linked to each SNP (VEP)^{90,92}. OR: odds ratio, 95% CI: 95% confidence interval.

murine models⁹⁷. The exact mechanisms by which altered MUC5B expression causes fibrosis are unclear. Altered lung microbiome and more persistent interaction between damaging factors and lung epithelium due to dysfunctional mucociliary clearance are possible pathways that may increase the likelihood of fibrosis development. While the T allele at this position may be expected to decrease fitness and therefore be at a lower population frequency, it is relatively common in populations of European ancestry with a

minor allele frequency (MAF) around 9%⁹⁸. This *MUC5B* SNP was recently found to be associated with a protective effect against COVID-19 infection and hospitalisation^{99–101}. This protective effect against viral infection is likely to confer an evolutionary advantage and may explain the relatively high population frequency of the rs35705950 polymorphism, particularly as the protective effects are likely to influence survival and fitness during reproductive years⁹⁸.

IPF risk variants linked to *STMN3*, *KIF15*, and *DSP* genes suggest that altered cytoskeletal proteins may also contribute to IPF. *STMN3* is a protein coding gene encoding stathmin-3, which has a predicted role in regulation of microtubule biology (Human Protein Atlas; *HPA*^{102,103}). Similarly, *KIF15* encodes the kinesin family member 15 protein, which is involved in microtubule force production and motility^{102,103}. Desmoplakin, encoded by *DSP*, is a desmosome protein predicted to be involved in cell-cell adhesion and wound healing responses^{102,103}. Desmosomes are also involved in tissue architecture, cell migration, proliferation, and differentiation. The routes by which these implicated proteins may lead to fibrosis are potentially multifaceted. Possible mechanisms include altered tissue structures conducive to scar tissue development or increased differentiation and proliferation of pro-fibrotic epithelial and immune cells.

Increased cellular senescence is believed to play a main role in the development of IPF and fibrosis generally. This is supported by the association of risk variants linked to *TERT*, *TERC*, and *RTEL1*, which are functionally involved in telomere maintenance^{102,103}. When expressed, *TERT* and *TERC* form a telomerase complex, with *TERT* encoding a reverse transcriptase and *TERC* encoding an RNA template to facilitate telomere repeat sequence elongation. Similarly, *RTEL1* encodes a DNA helicase that is required for normal telomere stability and elongation. These mutations affecting telomere maintenance are hypothesised to lead to shortened telomeres, subsequently increased cellular senescence, and accelerated fibrotic processes¹⁰⁴. The effect of disrupted telomere biology in causing IPF is supported

by *TERT* mutations being common in cases of familial pulmonary fibrosis¹⁰⁵, as well as the association of IPF with short telomere syndromes¹⁰⁶.

Other IPF risk-affecting genes may also result in increased cellular senescence, but through different pathways than telomere shortening. Kinetochore scaffold 1 (KNL1) protein, implicated by a common IPF risk variant affecting the *KNL1* gene, is responsible for spindle-chromosome binding during cell division^{102,103}. Dysregulated spindle assembly is likely to result in chromosomal instability and loss during cell division, which may trigger cellular senescence responses^{107,108}. The role of disrupted spindle assembly in the pathogenesis of IPF is also supported by the identification of a rare missense variant in *SPDL1*¹⁰⁹, encoding the Spindly protein, which also regulates spindle-chromosome binding and spindle orientation. Finally, IPF risk variants linked to *MAD1L1* suggest that normal *MAD1L1* function in spindle-assembly checkpoint regulation^{102,103} may be disrupted. Overall, lack of correct spindle assembly and spindle-assembly checkpoint regulation appears to be strongly linked to IPF risk¹⁰⁹. Failure to correctly regulate this cell cycle transition stage may result in IPF through increased chromosomal instability and cellular senescence.

The PI3K/mTOR signaling pathway plays a main role in regulation of cellular metabolism, proliferation, and survival. Biological research has shown that this pathway, particularly through mTORC1 (mammalian target of rapamycin complex 1) and subsequent 4E-BP1 (eukaryotic translation initiation factor 4E-binding protein 1) activation, regulates IPF fibroblast proliferation and extracellular matrix deposition^{110,111}.

The importance of the PI3K/mTOR pathway and mTORC1 activity is supported by the identification of IPF risk SNPs linked to *NPRL3* and *DEPTOR*, which both encode upstream inhibitors of mTORC1 complex activation^{102,103}. *DEPTOR* (DEP domain-containing mTOR-interacting protein) directly interacts with the mTORC1 complex, while *NPRL3* (nitrogen permease regulator 3) is a component of the GATOR1 complex, which

controls mTORC1 activity via upstream inhibition of RAG (ras-related GTP-binding protein complex).

Some IPF genetic risk variants are not yet functionally annotated or associated with expression of a particular gene, such as rs207551 (17q21.31) and rs79684490 (10q25.1). Many other IPF risk variants are linked to genes, but it is not clear how these SNPs or associated genes are functionally linked to IPF disease mechanisms. For example, the IPF risk variant rs9577395 implicates *ATP11A* in IPF disease, however it is not clear how the phospholipid transport function of this gene^{102,103} may contribute to disease pathogenesis. Similarly, *IVD* is predicted to be involved in leucine amino acid metabolism, *ZKSCAN1* is predicted to be a transcription factor, and *AKAP13* is involved in intracellular signaling pathways^{102,103}. Fundamentally, it is unclear how the downstream biology of many IPF risk variants or implicated genes contributes to disease risk. This gap in understanding limits the depth of biological insights that can be taken from IPF GWAS studies and subsequent therapeutic targeting. Investigation of the spectrum of human traits and diseases associated with each variant presents a powerful approach for building understanding of downstream biology of common IPF genetics.

1.4 Outline and aims of thesis

Fundamentally, the epidemiological factors and biology driving IPF pathogenesis are incompletely understood. By investigating genetic and epidemiological relationships with other human diseases and traits, aspects of shared biology may be highlighted that provide insights into the mechanisms driving IPF. Moreover, shared biology may highlight the potential for novel or repurposed treatments which may improve survival and quality of life for IPF patients.

This thesis aims to investigate if IPF shares genetic signatures with other human traits to highlight potential mechanistic similarities, identify possible risk factors, and build

understanding of the biology associated with IPF genetic risk variants. Secondly, this thesis aims to evaluate whether the relationship between IPF and DM is likely to be driven by direct causal effects between these diseases. Through these chapters, this work aims to contribute novel information towards the current understanding of IPF disease aetiology.

Chapter two of this thesis details a PheWAS of 19 common IPF risk variants (Table 1.2). This chapter describes the use of a public GWAS database to run a PheWAS study of common variants, including database selection and search criteria. Subsequently, this chapter describes an approach used to prioritise PheWAS results for follow up colocalisation analyses by categorising associations by variant function and by phenotype. Genetic colocalisation is then performed to estimate the probability that IPF susceptibility and PheWAS-identified phenotypes are associated with the same causal variant. Finally, this chapter details a secondary PheWAS investigation using data from the Genomics England 100,000 Genomes Project. This secondary PheWAS aimed to expand this overall investigation to include rare disease phenotypes that may not be represented in the previously used GWAS databases.

Chapter three details an investigation into shared genetic architecture across IPF and related traits using global genetic correlation analysis. This chapter first describes the use of global genetic correlation analysis to evaluate if traits with shared causal variants with IPF, prioritised by PheWAS and colocalisation analyses, share broad genetic similarity on a genome-wide scale. Secondly, genetic correlation analysis is applied to a sample of fibrotic conditions to investigate if diseases associated with fibrosis share broad genetic architecture with IPF.

Chapter four describes a Mendelian randomisation study investigating if the epidemiological association between IPF and DM is driven by a causal relationship. This chapter explains MR analysis and how current methods can be used to facilitate unconfounded estimation of causality. This chapter then details a bidirectional MR design

to assess for causal relationships between DM and IPF. Secondly, MR is used to evaluate if DM-associated factors (HbA1c, fasting insulin level, and BMI) are likely to be on the causal pathway for IPF. As these factors are also interlinked with the development of type 2 DM, a multivariable analysis to assess possible independent effects on IPF susceptibility is also included.

Chapter five of this thesis includes a second application of a Mendelian randomisation approach to investigate if causal effects underlie a potential biological relationship between mLOY and IPF. Overall, no evidence of a direct causal relationship between mLOY and IPF was identified. However, secondary analyses highlight a potential effect of shortened telomeres on mLOY, which suggests this potential relationship may result from shared pathways contributing to both mLOY and IPF in parallel.

The overall aims of this thesis are:

- 1) To investigate if IPF susceptibility variants are associated with other human traits**
 - 1.1) Run a PheWAS using public GWAS association data to investigate the spectrum of traits associated with IPF genetic risk variants
 - 1.2) Use colocalisation analyses to determine if prioritised overlapping association signals are likely to share the same causal variant
 - 1.3) Run a PheWAS using an independent cohort to extend phenotype discovery to include rare disease traits

- 2) To evaluate the extent of shared genetic architecture across IPF and related traits**
 - 2.1) Use genetic correlation to investigate if shared causal variants indicate broader shared genetic architecture across IPF and related traits

- 2.2) Use genetic correlation to explore if a sample of fibrotic diseases share broad genetic architecture with IPF

- 3) **Investigate if the epidemiological association of IPF with DM is driven by direct causal effects**
 - 3.1) Use a bidirectional MR approach to assess if DM and IPF are linked by direct causal effects
 - 3.2) Use MR to assess if DM-associated factors are likely to be causally associated with IPF susceptibility

- 4) **Investigate if a potential biological relationship between mLOY and IPF is driven by direct causal effects**
 - 4.1) Use a bidirectional MR approach to assess if mLOY and IPF are linked by direct causal effects

2: Phenome-Wide Association Study of Common Idiopathic Pulmonary Fibrosis Genetic Risk Variants

2.1 Introduction

This work aimed to explore the spectrum of traits associated with each IPF-associated SNP to build understanding of variant biology and potentially highlight shared mechanisms underlying different phenotypic traits.

Phenome-wide association studies were first conceptualised after it was discovered that many human genes and genetic variants affect multiple phenotypic traits (pleiotropy). A PheWAS can essentially be considered the reverse of a GWAS. In a GWAS, a variable phenotype is selected, genotyping is performed across different variations of that phenotype, and subsequent analyses aim to identify genotypes that are strongly associated with the phenotypic variation. In a PheWAS, a genotype is selected for study, a database with a broad range of phenotypes (often from electronic healthcare records; EHR) is developed, and analyses aim to identify phenotypes that are strongly associated with the studied genotype¹¹².

One previous PheWAS, published in 2016 by Claar *et al.*¹¹³, aimed to investigate the phenotypes associated with 15 IPF genetic risk variants. The aim of the Claar *et al.* PheWAS was to build understanding of how IPF variants confer disease risk by identifying the spectrum of diseases associated with each SNP. This study used linked genetic and EHR phenotype data (International Classification of Disease, Ninth Revision; ICD9) from the

BioVU (Vanderbilt University biobank) repository¹¹⁴. Of the 15 IPF risk variants originally identified for study, only one SNP (rs6793295) had enough genotyped participants with which to run a sufficiently powered PheWAS analysis.

Since the publishing of the previous IPF PheWAS in 2016, novel IPF risk variants have been identified and the data resources available for PheWAS analyses have improved. An important progression in this research area is the development of publicly available GWAS databases with associated PheWAS tools, such as Phenoscanner^{115,116} or IEU OpenGWAS¹¹⁷. Such databases typically integrate published summary statistics from thousands of GWAS studies. The current availability and depth of data available for PheWAS studies now presents an ideal opportunity to investigate the pleiotropic spectrum of IPF risk variants.

A key development in methods aimed towards understanding pleiotropy and genetic variant biology is the use of genetic colocalisation analyses. Colocalisation methods, such as *coloc*^{76,77}, aim to assess if two genetic association signals share the same causal variant within a given genomic region.

Colocalisation analyses can be applied to assess if two overlapping genetic signals, such as those identified through a PheWAS, are likely share the same causal variant. This approach can therefore be used to further understand downstream pathways and biological effects of genetic variants by identifying overlapping associations with multiple human traits, biomarkers, or gene expression signatures.

As well as public GWAS databases, other large-scale resources have been developed that contain sufficient data for a PheWAS. The Genomics England 100,000 Genomes Project contains genomic sequence and linked healthcare (Hospital Episode Statistics; HES) data from over 80,000 NHS patients, including 25,000 cancer patients and 17,000 rare disease patients^{118–120}. Using this data resource to run a PheWAS of IPF genetic risk variants therefore has the potential to expand the range of investigated phenotypes to include rare

diseases that may not be represented in GWAS databases. However, while a PheWAS using linked genomic and healthcare record data offers several advantages, routine follow-up analyses such as *coloc* require GWAS summary data and are therefore difficult to integrate into a methodological approach of this kind.

Characterising the biological effects and diseases associated with IPF SNPs has the potential to aid understanding of how these variants are linked to IPF pathogenesis. This would be of value as the downstream effects of many IPF SNPs are not fully understood, the function of many implicated genes is not known, and some IPF SNPs have not yet been functionally annotated (rs2077551: 17q21.31 and rs79684490: 10q25.1). Identifying overlapping association signals across diseases and traits may also highlight shared mechanisms that present drug repurposing opportunities¹²¹ for the treatment of IPF.

This chapter includes a PheWAS of 19 common IPF genetic variants. The first stage of this investigation applied a data mining PheWAS approach using a public database of GWAS results across many human traits. Following this, genetic colocalisation analyses were applied to assess if priority traits share causal variants with IPF. Finally, a PheWAS was performed in a non-public data resource (the Genomics England 100,000 genomes project) to investigate firstly for replication of trait associations in an independent resource, and secondly associations with rare disease traits that may not have been included in the public database PheWAS. While the Genomics England PheWAS study was not followed up with subsequent colocalisation analysis, any identification of novel or replicated phenotypes aimed to add depth to the overarching investigation.

2.2 Methods

2.2.1 Methods for PheWAS of public GWAS data

2.2.1.1 Database selection

To apply a data mining PheWAS approach (overview presented in Figure 2.1), a database containing a large volume of GWAS association data was required. The IEU *OpenGWAS* project¹¹⁷ database is regularly updated, has an associated PheWAS tool that can be used programmatically in R, and facilitates open access to the GWAS results incorporated into the database. *OpenGWAS* was therefore the database selected for this study.

2.2.1.2 Search criteria

A total of 19 IPF genetic risk variants (Table 2.2.1) were queried using the *phewas* function from the *ieugwasr* R package, with default batch ID search criteria applied (all available GWAS datasets included). Genetic associations with each IPF risk variant were filtered using a standard genome-wide significant p-value threshold of $p < 5 \times 10^{-8}$. Interpretation of effect sizes can be context dependent, for example small effect sizes for continuous traits, such blood biomarker levels, can still be biologically significant. It was therefore deemed appropriate to not use an effect size threshold for association filtering to allow objective prioritisation of both continuous and binary traits such as disease status.

Functional category	Subcategory	SNP RSID	Chromosome	Base position	Gene / locus	Effect allele	Non-effect allele
Mucin biology	Mucin biology	rs35705950	11	1241221	MUC5B	T	G
Cytoskeletal proteins	Cytoskeletal proteins	rs112087793	20	62284170	STMN3	C	T
		rs2292181	3	44903434	KIF15	C	G
		rs2076295	6	7563232	DSP	G	T
Chromosome maintenance and cell cycle control	Telomere biology	rs9860874	3	169486271	TERC	A	C
		rs41308092	20	62324391	RTEL1	A	G
		rs7725218	5	1282414	TERT	G	A
	Spindle assembly and cell cycle control	rs12537430	7	1868761	MAD1L1	G	A
		rs12912339	15	40931708	KNL1	A	G
Cellular transport, signaling, and metabolism	MTOR pathway signaling	rs74614704	16	162240	NPRL3	A	G
		rs10808505	8	120940206	DEPTOR	T	G
	Cell signaling and membrane trafficking	rs9577395	13	113534984	ATP11A	C	G
		rs2304645	15	40716253	IVD	C	G
		rs11073517	15	86287910	AKAP13	T	C
		rs12610495	19	4717672	DPP9	G	A
		rs2897075	7	99630342	ZKSCAN1	T	C
Other or unclassified	Other or unclassified	rs2077551	17	44214888	17q21.31	T	C
		rs79684490	10	111229861	10q25.1	A	G
		rs2609259	4	89837808	FAM13A	A	C

Table 2.2.1: Genetic variants investigated in this PheWAS study and functional categorisation for each SNP. Association statistics for each variant were derived from the most recent IPF risk GWAS⁹⁰. Functional categories were defined based on the predicted biological function^{102,103} of the gene linked to each SNP (VEP)^{90,92}. OR: odds ratio, 95% CI: 95% confidence interval.

2.2.1.3 Processing of PheWAS datasets

After PheWAS datasets for each IPF genetic risk variant (Table 2.2.1) has been retrieved from *OpenGWAS*, association data was aligned relative to effects on IPF risk. Datasets were aligned by firstly identifying variant positions where effect and non-effect alleles have different orientations in PheWAS results, for example, relative to the allele orientation associated with increased disease risk in the IPF GWAS dataset. Subsequently, non-matching alleles and corresponding effect sizes can be ‘flipped’ to ensure consistency when assessing shared or opposing effects of genetic variants across traits. When beta estimates are used to describe effect sizes, this would involve swapping the effect and non-effect allele and multiplying the effect estimate by -1.

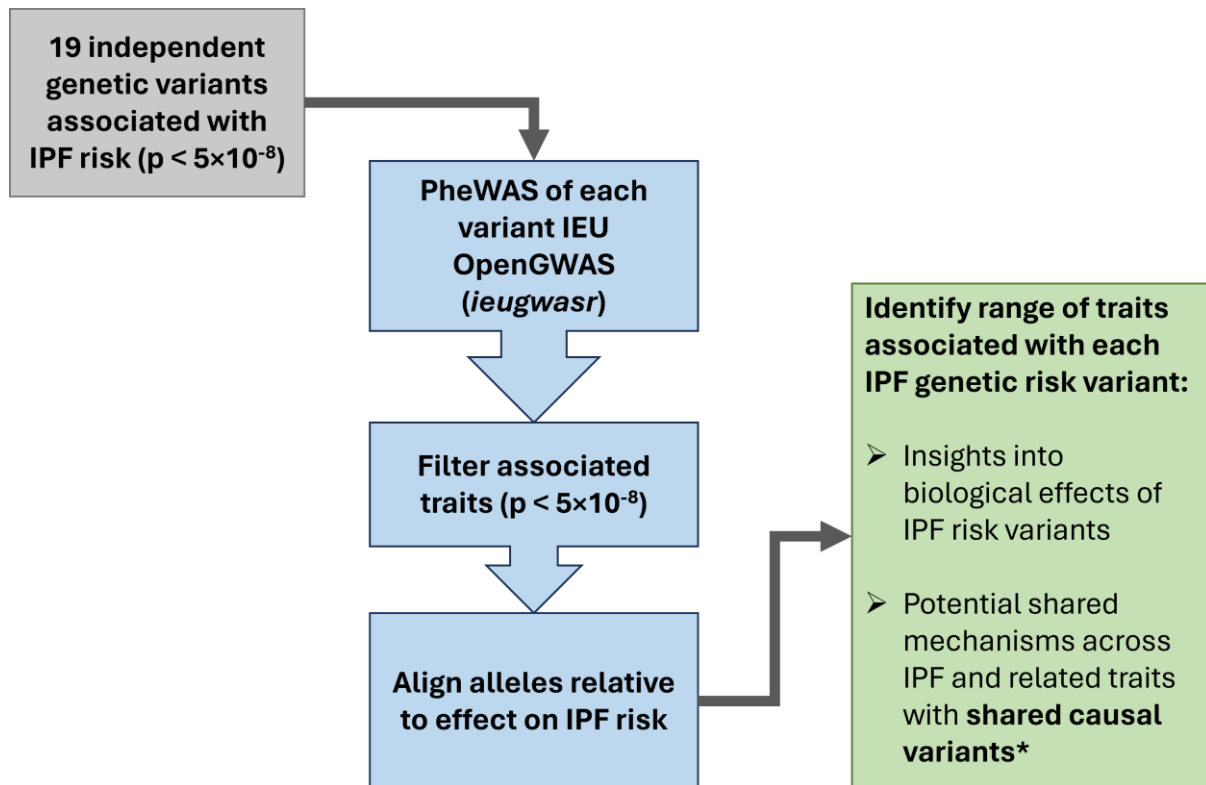


*Figure 2.2.1: An overview of the study design applied in a PheWAS of IPF genetic risk variants using public GWAS data. Input data (grey, left) consisted of 19 independent genetic risk variants associated ($p < 5 \times 10^{-8}$) with IPF risk. The analysis pipeline constructed (blue, middle) runs PheWAS searches of each specified genetic variant, before filtering and aligning output association data. This pipeline aims to output data that allows identification of the range of traits associated with each IPF variant and subsequent interpretation of biological effects (green, right). *Insights related to shared causal variants typically require colocalisation analyses to ensure informed interpretation.*

2.2.2 Methods for genetic colocalisation

2.2.2.1 Selection of overlapping signals for colocalisation analyses

Phenotypes associated with IPF risk variants (identified using public database PheWAS methodology: Section 2.2.1) were prioritised for follow-up colocalisation analyses using two approaches, each corresponding to a core aim of this PheWAS (Figure 2.2.2). Structural variation in the genomic region containing the rs2077551 IPF risk variant (17q21.31) is complex¹²². This structural complexity results in a high number of SNPs in high LD across

this region, which has previously led to inconclusive colocalisation analyses¹²³. The rs2077551 variant was therefore removed from signal selection for colocalisation analyses.

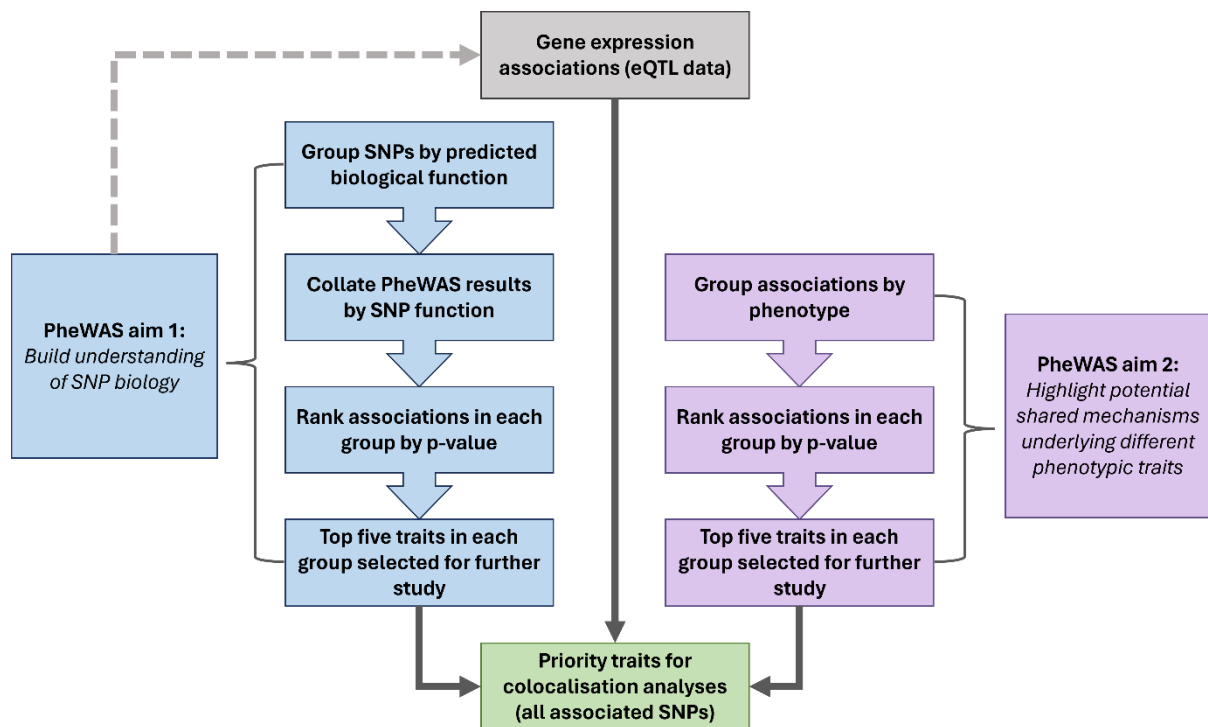


Figure 2.2.2: Prioritisation of PheWAS associations for follow-up genetic colocalisation analyses.

Towards the aim of building understanding of SNP biology, IPF genetic risk variants were grouped by the predicted biological function of linked genes. The variant-function groups and included variants were as follows:

1) Cytoskeletal proteins:

- *STMN3*: rs112087793
- *KIF15*: rs2292181
- *DSP*: rs2076295

2) Chromosome maintenance and cell cycle control:

- *KNL1*: rs12912339
- *MAD1L1*: rs12537430
- *TERC*: rs9860874

- *TERT*: rs7725218
- *RTEL1*: rs41308092

3) Cellular transport, signaling, and metabolism:

- *DEPTOR*: rs10808505
- *NPRL3*: rs74614704
- *DPP9*: rs12610495
- *IVD*: rs2304645
- *ATP11A*: rs9577395
- *AKAP13*: rs11073517
- *ZKSCAN1*: rs2897075

4) Mucin biology and immunity:

- *MUC5B*: rs35705950

5) Unknown function:

- *FAM13A*: rs2609259
- 10q25.1: rs79684490

Phenotypic associations were then ranked by p-value within each variant-function category. The top five associated traits in each group were selected for further investigation. To add depth to the investigation of SNP biology, all significant ($p < 5 \times 10^{-8}$) associations with eQTL data were also prioritised for further study.

Towards the aim of highlighting potential shared mechanisms underlying different traits, phenotypes were grouped into the following categories:

1) Pulmonary

2) Infectious

3) Neoplastic

4) Other disease traits:

- Psychiatric

- Neurologic
- Cardiovascular
- Digestive
- Genitourinary
- Dermatologic
- Musculoskeletal
- Ocular

5) Biomarkers:

- Any biomarkers not specifically linked to one disease category

Phenotypic associations were then ranked by p-value within each trait category. Once grouped by phenotype, the top five associated traits in each group were selected for further investigation.

2.2.2.2 Data retrieval and processing

To run colocalisation analyses, summary data from the most recent IPF GWAS⁹⁰ (4,125 European ancestry cases, 20,464 European ancestry controls) were retrieved by request from the authors. GWAS datasets for prioritised phenotypes were retrieved from the *OpenGWAS* database¹¹⁷. Downloaded GWAS datasets were subsequently processed using *tidyverse* (R). The *coloc*^{76,77} method requires GWAS data containing the following data fields:

- A SNP identifier (such as an RSID code)
- SNP base position
- Type (case-control: *cc*, or quantitative: *quant*)
- SNP effect size (*beta*)
- Variance of *beta* (squared standard error) (*varbeta*)

To estimate trait standard deviation (*sdY*), given *varbeta*, where not provided:

- SNP or GWAS sample size

- SNP minor allele frequency

Any phenotypes were removed from this analysis stage if required data fields were not available. For each colocalisation analysis, summary data were retrieved within a 500 kilobase (kb) region surrounding the relevant SNP (250kb upstream and 250kb downstream of the SNP base position).

2.2.2.3 Colocalisation analyses

Colocalisation analyses were performed using the Approximate Bayes Factor (*coloc.abf*) method. This approach involves fine mapping under the assumption of a single causal variant in the given region, followed by estimation of the posterior probability that fine mapped variants are shared⁷⁷. Assessment of likelihood of colocalisation using *coloc* is achieved by estimating the posterior probability of the following hypotheses:

- H_0 : neither trait has a genetic association in the specified genomic region.
- H_1 : only trait 1 has a genetic association in the specified genomic region.
- H_2 : only trait 2 has a genetic association in the specified genomic region.
- H_3 : both traits have genetic associations within the specific genomic region, but traits are associated with different, independent causal variants.
- H_4 : Association with trait 1 and trait 2 with one shared causal variant.

A high posterior probability of H_4 (PP.H4) is generally considered evidence of colocalisation between two independent association signals. In this analysis, a PP.H4 > 0.8 (80%) was considered sufficient evidence to infer colocalisation.

2.2.3 Methods for PheWAS using an independent cohort

2.3.3.1 *Data retrieval and processing*

This research was made possible through access to data in the National Genomic Research Library¹²⁴, which is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The National Genomic Research Library holds data provided by patients and collected by the NHS as part of their care and data collected as part of their participation in research. The National Genomic Research Library is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure.

The Genomics England (GE) 100,000 Genomes project (100,000 GP) resource contains de-identified hospital episode statistics (HES) data, relating to both admitted and outpatient care for over 80,000 NHS Digital patients. Corresponding genotyping data is available from the aggregate variant calls (AggV2) dataset. GE 100,000 GP HES and was retrieved using the *RLabKey* R package, and genotyping data using *bcftools*, within the Genomics England Research Environment (RE).

Any GE 100,000 GP participants with missing or invalid episode start dates, missing ICD10 codes, and non-concordant data for phenotypic and karyotypic sex were removed from the PheWAS cohort. All included participants were of European ancestry and unrelated. Only participants included in AggV2, therefore continuing to consent to data usage and with genotyping data available, were included. An overview of the PheWAS methodology applied is presented in Figure 2.2.3.

2.3.3.2 PheWAS analysis

The PheWAS methodology^{75,125} applied in this analysis involves converting ICD10 codes into PheWAS codes (phecodes) to categorise related phenotypes into hierarchical and more clinically relevant groups. This methodology also facilitates the exclusion of related phenotypes for each phecode to ensure only true controls are used in each association analysis.

PheWAS analyses were performed for the 14 IPF genetic risk variants identified in the 2020 GWAS⁸⁹. Linear regression analyses between each SNP and the range of included phenotypes were adjusted for the first ten genetic principal components to adjust for genetic ancestry, age, and sex. A minimum requirement of 20 combined cases and controls was set for each analysis. An exploratory p-value threshold of $p < 1 \times 10^{-5}$ was applied to filter significant results.

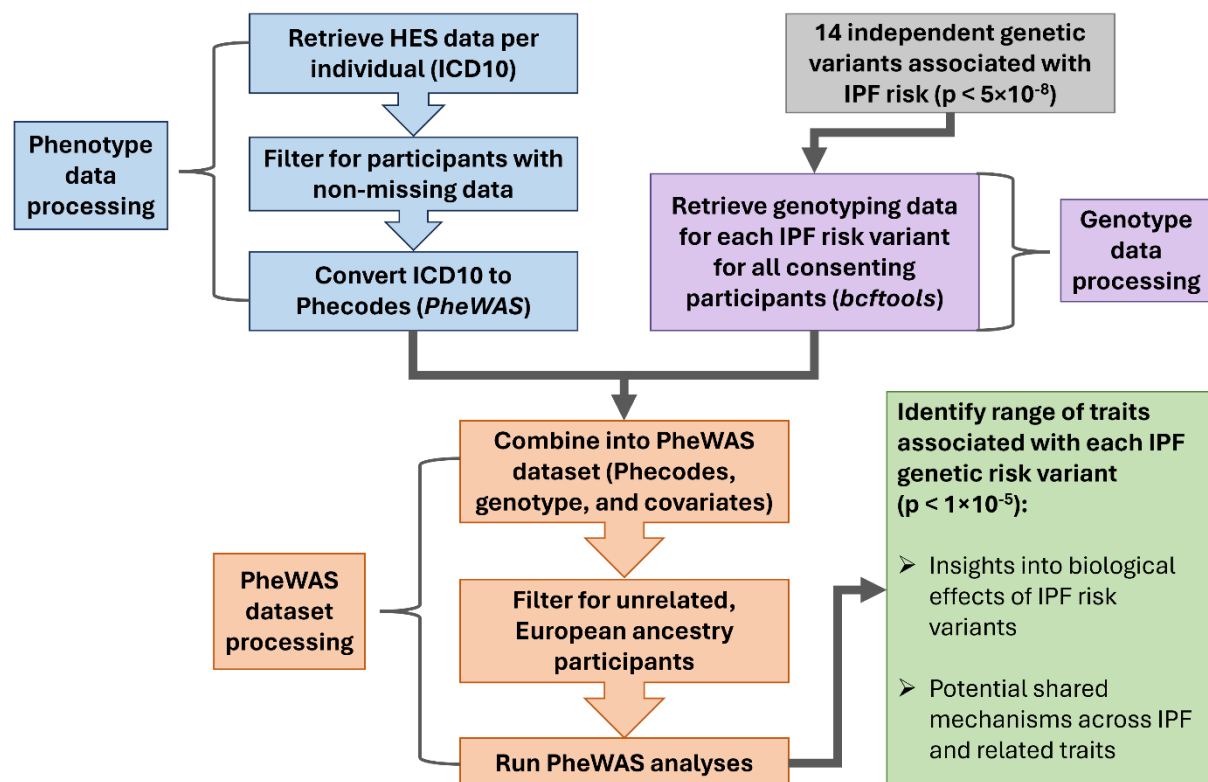


Figure 2.2.3: An overview of methodology applied in a PheWAS of IPF genetic risk variants using the Genomics England 100,000 Genomes Project data resource. Input data (grey, top) consisted of 14 independent genetic risk variants associated ($p < 5 \times 10^{-8}$) with IPF risk. In this pipeline, phenotype and genotype (blue, left, and purple, right, respectively) data are retrieved and processed into a combined dataset for PheWAS analyses (orange). This pipeline aims to output data that allows identification of the range of traits associated with each IPF variant and subsequent interpretation of biological effects (green, right).

2.3 Results

2.3.1 Results of PheWAS of public GWAS data

The data mining PheWAS approach using the IEU OpenGWAS database identified associated traits meeting inclusion criteria for 18 out of 19 IPF genetic risk variants.

2.2.2.1 Phenotypes associated with mucin biology genetic Variants

The *MUC5B* SNP, rs35705950 (G > T), is associated primarily with increased risk of lung disease phenotypes such as IPF ($\beta = 1.61, p = 3.88 \times 10^{-80}$), ILD ($\beta = 1.15, p = 1.65 \times 10^{-81}$), and ILD related to systemic autoimmune disease ($\beta = 0.80, p = 2.86 \times 10^{-10}$) (Figure 2.3.1). This *MUC5B* variant is also associated with decreased *CD151* gene expression ($\beta = -0.13, p = 2.86 \times 10^{-10}$), which is predicted to be involved in assembly of glomerular and tubular basement membranes in kidney tissue and host-virus interaction (Human Protein Atlas; *HPA*^{102,103}).

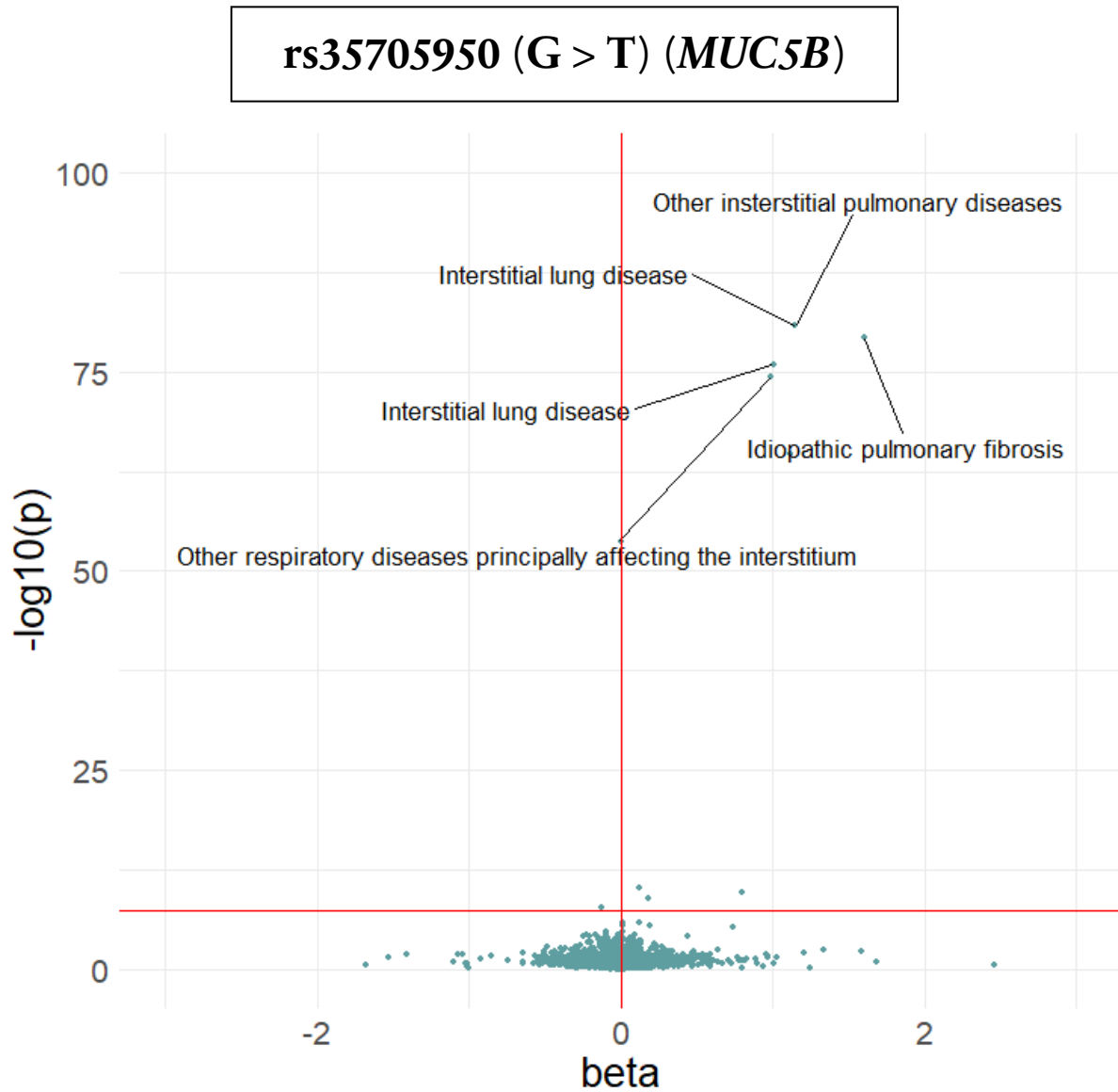


Figure 2.3.1: A volcano plot presenting phenotypes associated with rs35705950 (G > T) (*MUC5B*). The effect size (beta statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 5 \times 10^{-8}$ threshold applied in phenotypic association search criteria.

2.2.2.2 Phenotypes associated with cytoskeletal protein genetic variants

The variant rs112087793 (*STMN3*) is associated with serum protein biomarker levels, such as insulin-like growth factor (IGF-1) ($\beta = -0.03, p = 4.96 \times 10^{-13}$), C-reactive protein (CRP) ($\beta = 0.02, p = 8.18 \times 10^{-11}$), gamma glutamyl transferase (GGT) ($\beta = 0.02, p = 3.70 \times 10^{-10}$),

and tumour necrosis factor superfamily member 6B (TNFSF6B) ($\beta = -0.2557, p = 8.32 \times 10^{-9}$) (Figure 2.3.2). For context, IGF-1 is functionally associated with cell growth and proliferation, CRP regulates inflammation and host defense mechanisms, GGT is enriched in liver and kidney tissues and involved in lipid metabolism, and TNFSF6B is an apoptotic regulatory protein (HPA^{102,103}). This SNP is also associated with increased estimated glomerular filtration rate (a measure of kidney function) ($\beta = 0.003, p = 1.25 \times 10^{-10}$), decreased lung function (FVC) ($\beta = -0.02, p = 9.00 \times 10^{-10}$), and increased mean corpuscular haemoglobin levels ($\beta = 0.02, p = 4.42 \times 10^{-9}$). PheWAS results show that rs112087793 is an eQTL associated with increased expression of *ZGPAT* (a transcription factor involved in cell proliferation, survival, and migration via regulation of EGFR) ($\beta = 0.38, p = 1.24 \times 10^{-76}$), *UCKL1* (predicted to contribute to blast transformation and proliferation) ($\beta = 0.20, p = 1.34 \times 10^{-21}$), *PPDPF* (predicted to involved in cell differentiation and pancreatic development) ($\beta = 0.17, p = 6.37 \times 10^{-16}$), and *HELZ2* (transcription factor; downstream effects currently undefined) ($\beta = 0.15, p = 3.57 \times 10^{-13}$) (HPA^{102,103}).

The *KIF15* IPF risk variant (rs2292181) was not found to be associated with any observable phenotypic traits or diseases. This SNP is however an eQTL associated with decreased expression of *ZNF502* (transcription factor; downstream effects currently undefined) ($\beta = -0.34, p = 9.62 \times 10^{-39}$) and *TMEM42* (predicted to have an important role in membrane structure and function) ($\beta = -0.21, p = 4.62 \times 10^{-16}$) (Figure 2.3.3) (HPA^{102,103}).

The *DSP* IPF risk variant (rs2076295) is associated with increased FEV1/FVC ratio ($\beta = 0.02, p = 5.9 \times 10^{-27}$). Associations were also identified with decreased FVC ($\beta = -0.01, p = 1.60 \times 10^{-8}$) and increased risk of interstitial lung disease ($\beta = 0.15, p = 7.74 \times 10^{-10}$) which may underlie the previously described positive effect on FEV1/FVC ratio. In addition, rs2076295 is associated with increased receptor for advanced glycosylation end products (RAGE) levels (a proinflammatory protein enriched in type 1 alveolar cells and associated with fibrosis¹²⁶) ($\beta = 0.0611, p = 7.82 \times 10^{-9}$) (Figure 2.3.4). This SNP is also associated with

decreased expression of *SNRNP48* (mRNA processing protein; downstream effects currently undefined) ($\beta = -0.15, p = 2.07 \times 10^{-34}$) (HPA^{102,103}).

rs112087793 (T > C) (*STMN3*)

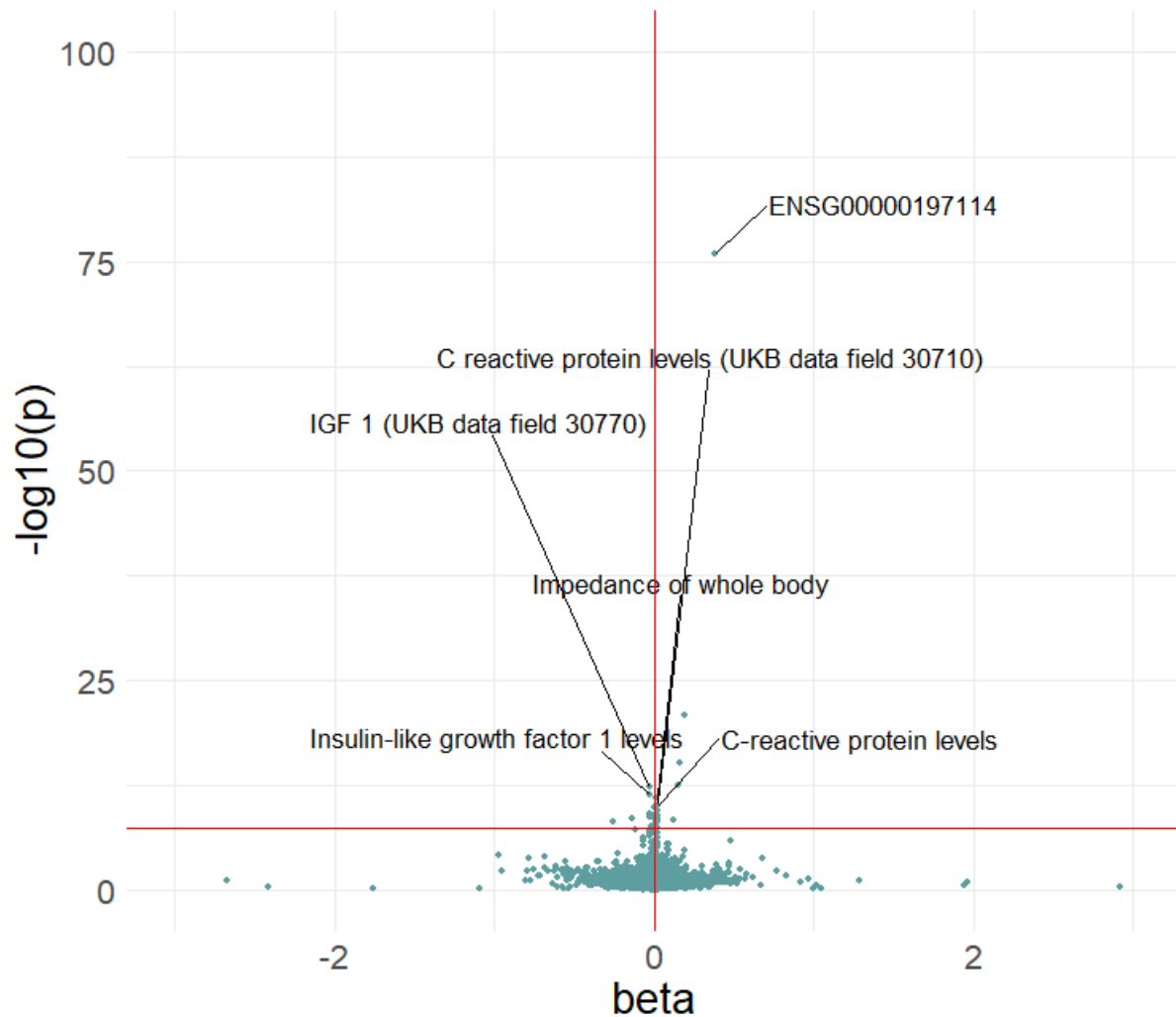


Figure 2.3.2: A volcano plot presenting phenotypes associated with rs112087793 (T > C) (*STMN3*). The effect size (β statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 5 \times 10^{-8}$ threshold applied in phenotypic association search criteria. IGF-1: insulin like growth factor 1. ENSG00000197114 is the Ensembl gene ID for ZGPAT.

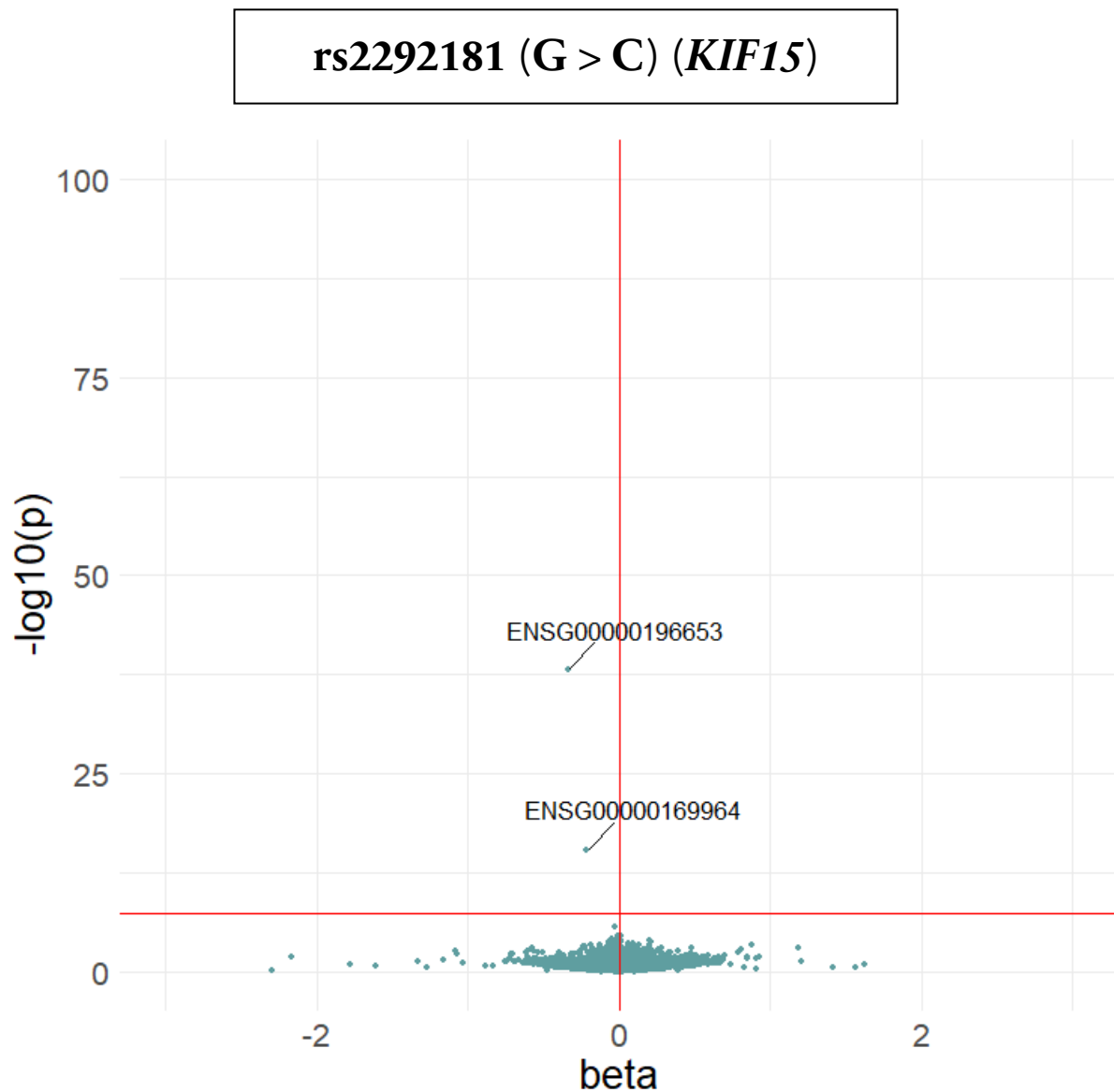


Figure 2.3.3: A volcano plot presenting phenotypes associated with rs2292181 (G > C) (*KIF15*). The effect size (beta statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance (-log₁₀(p)) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates beta = 0 (null effect), the horizontal red line shown indicates the $p = 5 \times 10^{-8}$ threshold applied in phenotypic association search criteria. ENSG00000196653 and ENSG00000169964 are the Ensembl gene IDs for ZNF502 and TMEM42, respectively.

rs2076295 (T > G) (DSP)

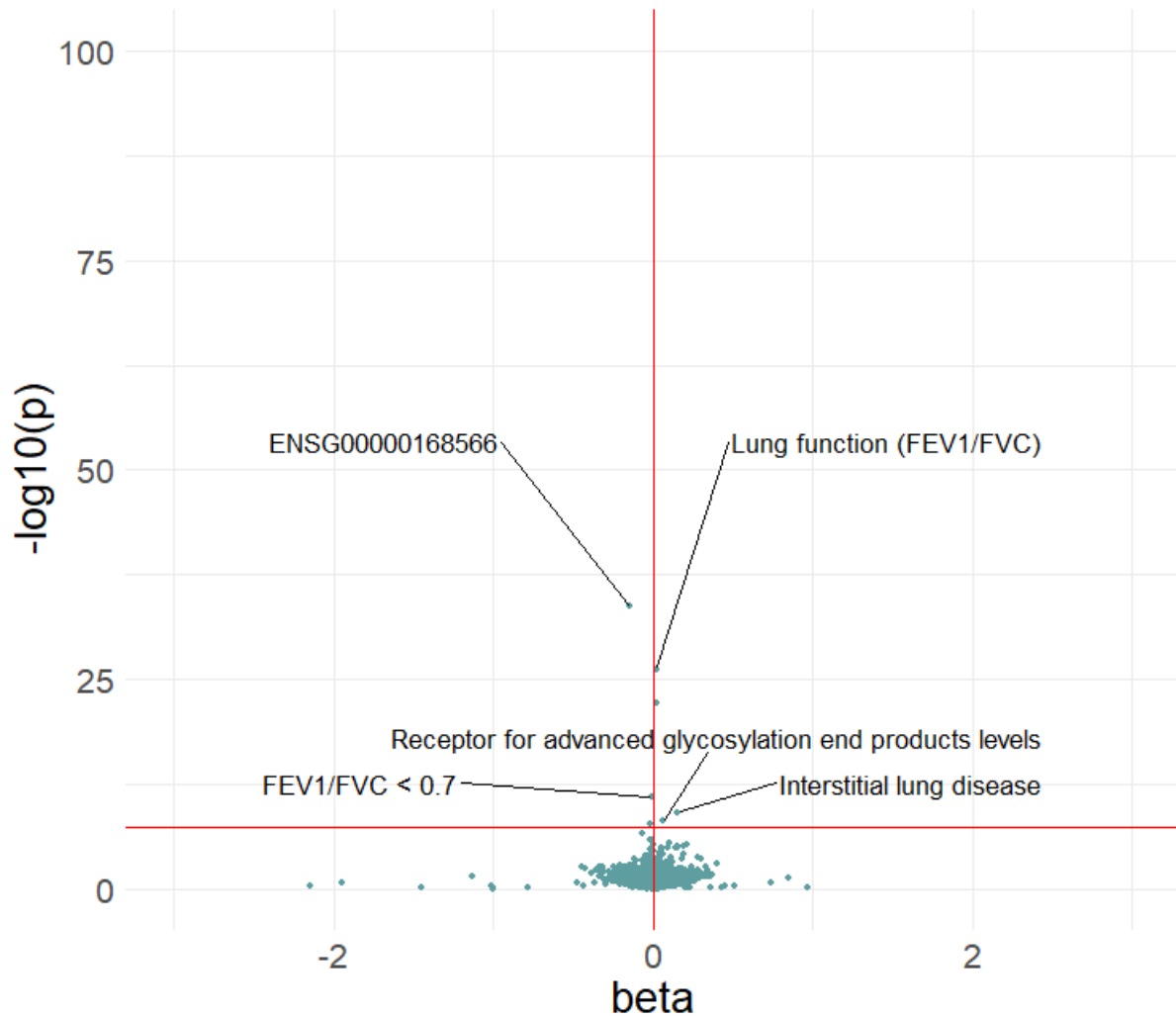


Figure 2.3.4: A volcano plot presenting phenotypes associated with rs2076295 (T > G) (DSP). The effect size (beta statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 5 \times 10^{-8}$ threshold applied in phenotypic association search criteria. ENSG00000168566 is the Ensembl gene ID for SNRNP48.

2.2.2.3 Phenotypes associated with chromosome maintenance and cell cycle control genetic variants

The *TERC* IPF risk variant (rs9860874) is strongly associated with shortened telomere length ($\beta = -0.08, p = 7.00 \times 10^{-294}$) (Figure 2.3.5). This SNP is associated with increased total white blood cell ($\beta = 0.02, p = 8.10 \times 10^{-21}$) and eosinophil ($\beta = 0.02, p = 8.1 \times 10^{-20}$), but decreased neutrophil cell counts ($\beta = -0.03, p = 1.10 \times 10^{-13}$). Results indicated negative effects of this SNP on red blood cell count ($\beta = -0.01, p = 5.00 \times 10^{-13}$), however positive effects were indicated on other red cell traits such as mean corpuscular haemoglobin ($\beta = 0.02, p = 1.03 \times 10^{-14}$) and mean corpuscular volume ($\beta = 0.02, p = 1.27 \times 10^{-14}$). Associations between rs9860874 and disease traits include negative effects on systolic blood pressure ($\beta = -0.28, p = 1.33 \times 10^{-15}$), uterine fibroids ($\beta = -0.0718, p = 5.61 \times 10^{-14}$), and pan-cancer risk ($\beta = -0.005, p = 5.50 \times 10^{-9}$). Finally, this IPF risk variant is also associated with increased sex hormone binding globulin (SHBG) levels in blood (protein regulator of steroid hormones, such as testosterone and oestrogen) ($\beta = 0.006, p = 6.90 \times 10^{-12}$) (HPA^{102,103}).

The rs41308092 IPF risk variant is associated with decreased telomere length ($\beta = -0.10, p = 3.30 \times 10^{-40}$) (Figure 2.3.6). In addition, this *RTEL1* SNP is associated with decreased expression of *STMN3* ($\beta = -0.1, p = 3.30 \times 10^{-40}$), another gene implicated in IPF disease by the rs112087793 SNP.

A strong association was identified between the rs7725218 IPF risk allele (A > G) (*TERT*) and decreased telomere length ($\beta = -0.07, p = 5.6 \times 10^{-220}$) (Figure 2.3.7). This *TERT* variant, similarly to the *TERC* IPF risk variant, is also associated with phenotypic variation in many blood cell traits. PheWAS results suggest that rs7725218 G allele is associated with decreased measurements of plateletcrit (blood platelet volume) ($\beta = -0.05, p = 4.40 \times 10^{-120}$), red blood cell count ($\beta = -0.04, p = 1.62 \times 10^{-171}$), overall white blood cell count ($\beta = -0.03, p = 2.53 \times 10^{-38}$), and neutrophil count ($\beta = -0.03, p = 1.46 \times 10^{-44}$), but increased eosinophil

count ($\beta = 0.02, p = 2.40 \times 10^{-16}$). This SNP is associated with increased risk of multiple disease or disease-associated traits, including interstitial lung disease ($\beta = 0.24, p = 5.70 \times 10^{-10}$), HDL cholesterol ($\beta = 0.01, p = 1.2 \times 10^{-9}$), and prostate cancer ($\beta = 0.11, p = 7.65 \times 10^{-39}$). Conversely, this variant is also associated with decreased risk of ovarian cysts ($\beta = -0.07, p = 4.33 \times 10^{-8}$), myeloproliferative disease ($\beta = -0.33, p = 4.83 \times 10^{-11}$), diastolic ($\beta = -0.12, p = 4.83 \times 10^{-11}$) and systolic ($\beta = -0.22, p = 8.82 \times 10^{-11}$) blood pressure, uterine fibroids ($\beta = -0.09, p = 5.6 \times 10^{-25}$), lung adenocarcinoma ($\beta = -0.16, p = 5.89 \times 10^{-18}$), serous ovarian cancer (low grade) ($\beta = -0.25, p = 7.25 \times 10^{-17}$), ER-negative breast cancer ($\beta = -0.09, p = 5.18 \times 10^{-13}$), and brain tumour diagnosis ($\beta = -0.28, p = 7.2 \times 10^{-11}$). Finally, rs7725218 is also associated with traits related to biological ageing, including intrinsic epigenetic age acceleration ($\beta = -0.23, p = 3.73 \times 10^{-10}$), DNA methylation Hannum age acceleration ($\beta = -0.22, p = 1.14 \times 10^{-8}$), and mitochondrial DNA copy number ($\beta = -0.016, p = 3.53 \times 10^{-13}$).

Associations identified with the *MAD1L1* IPF risk variant (rs12537430) included an increase in SHBG ($\beta = 0.02, p = 9.87 \times 10^{-12}$) and a decrease in IGF-1 ($\beta = -0.01, p = 3.32 \times 10^{-9}$) and plateletcrit levels ($\beta = -0.01, p = 1.6 \times 10^{-10}$) in blood (Figure 2.3.8). This SNP is also associated with increased expression of *MAD1L1* ($\beta = 0.23, p = 4.35 \times 10^{-79}$) and decreased expression of *MRM2* (Mitochondrial rRNA methyltransferase 2; predicted to be involved in cell cycle control and DNA repair) ($\beta = -0.12, p = 5.65 \times 10^{-24}$) (HPA^{102,103}).

The *KNL1* IPF risk variant (rs12912339) is associated with variation in white-grey contrast signatures identified through brain imaging. Effects of this SNP on white-grey contrast across different brain regions were consistently negative, such as left hand supramarginal gyrus white-grey intensity-contrast ($\beta = -0.09, p = 5.81 \times 10^{-17}$), right hand supramarginal gyrus white-grey intensity-contrast ($\beta = -0.08, p = 6.47 \times 10^{-14}$), and left hand rostral middle frontal gyrus white-grey intensity-contrast ($\beta = -0.08, p = 7.53 \times 10^{-15}$) (Figure 2.3.9). Additional associations with rs12912339 include increased mean spheroid cell volume (β

= 0.02, $p = 1.5 \times 10^{-14}$), mean reticulocyte volume ($\beta = 0.02$, $p = 2.32 \times 10^{-9}$), mean corpuscular hemoglobin ($\beta = 0.02$, $p = 6.8 \times 10^{-12}$), and SHBG levels in blood ($\beta = 0.02$, $p = 5.85 \times 10^{-12}$). This SNP is also associated with decreased prostate cancer risk ($\beta = -0.06$, $p = 1.40 \times 10^{-8}$) and decreased diastolic blood pressure ($\beta = -0.16$, $p = 4.85 \times 10^{-11}$). Associations identified with rs12912339 also included variation in expression of many genes, including *GCHFR* (predicted to be involved in phenylalanine metabolism in liver) ($\beta = -0.46$, $p = 1.44 \times 10^{-192}$), *RMDN3* (involved in microtubule binding and spindle biology) ($\beta = -0.39$, $p = 7.73 \times 10^{-136}$), and *RPAP1* (part of the RNA polymerase II enzyme complex) ($\beta = 0.12$, $p = 5.71 \times 10^{-13}$) (HPA^{102,103}).

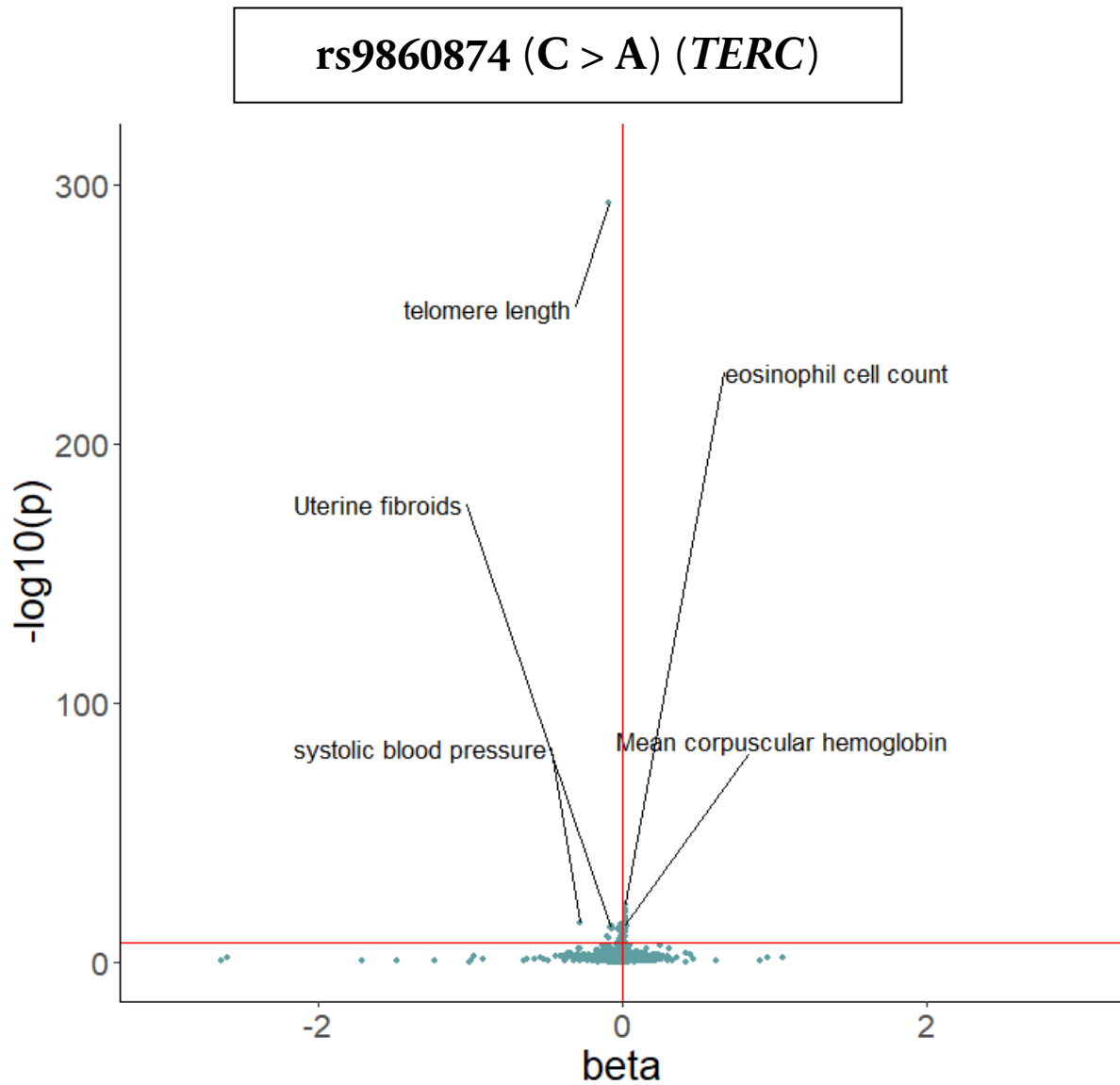


Figure 2.3.5: A volcano plot presenting phenotypes associated with rs9860874 (C > A) (TERC). The effect size (beta statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 5 \times 10^{-8}$ threshold applied in phenotypic association search criteria.

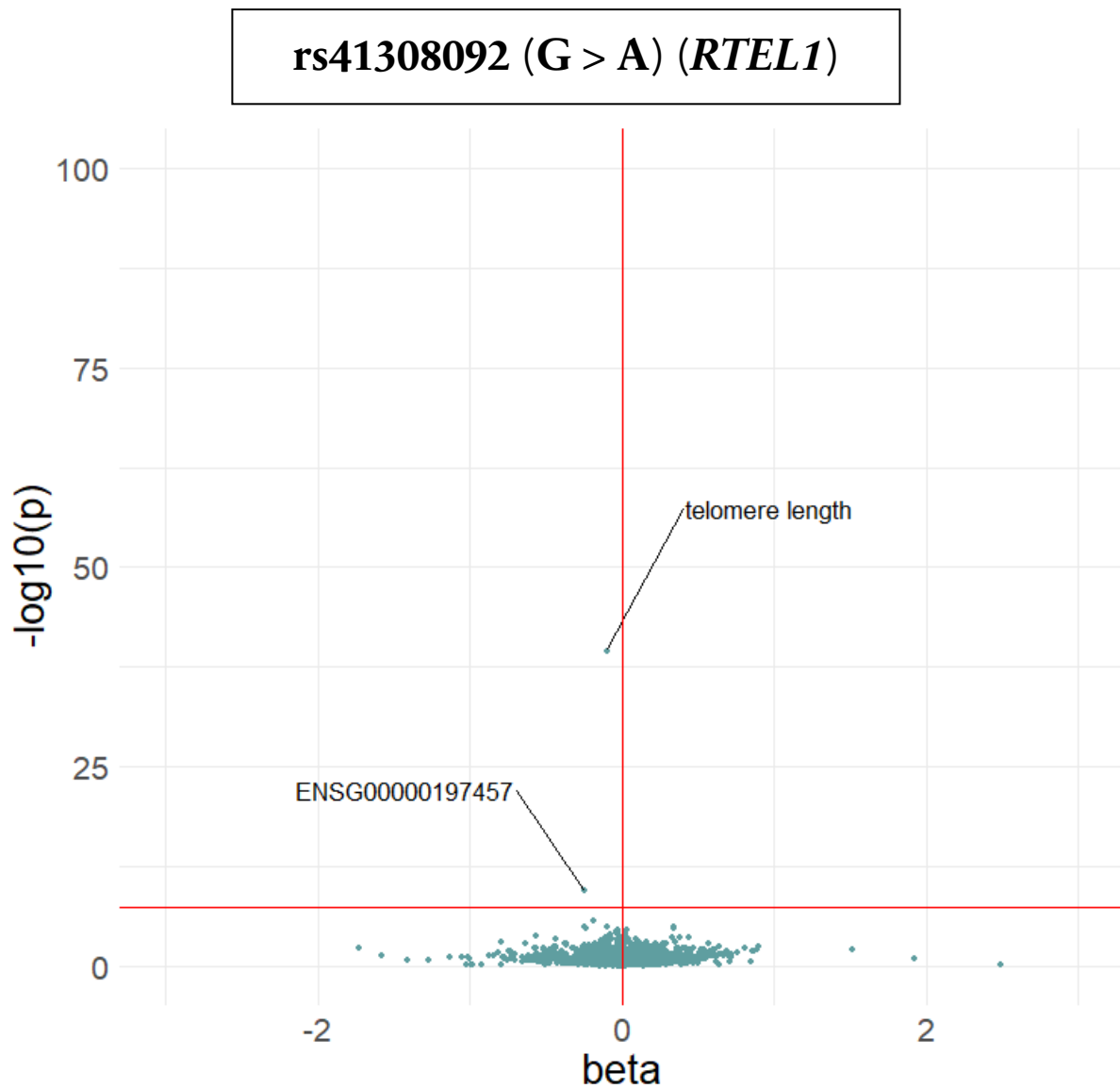


Figure 2.3.6: A volcano plot presenting phenotypes associated with rs41308092 (G > A) (*RTEL1*). The effect size (beta statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 5 \times 10^{-8}$ threshold applied in phenotypic association search criteria. ENSG00000197457 is the Ensembl gene ID for *STMN3*.

rs7725218 (A > G) (*TERT*)

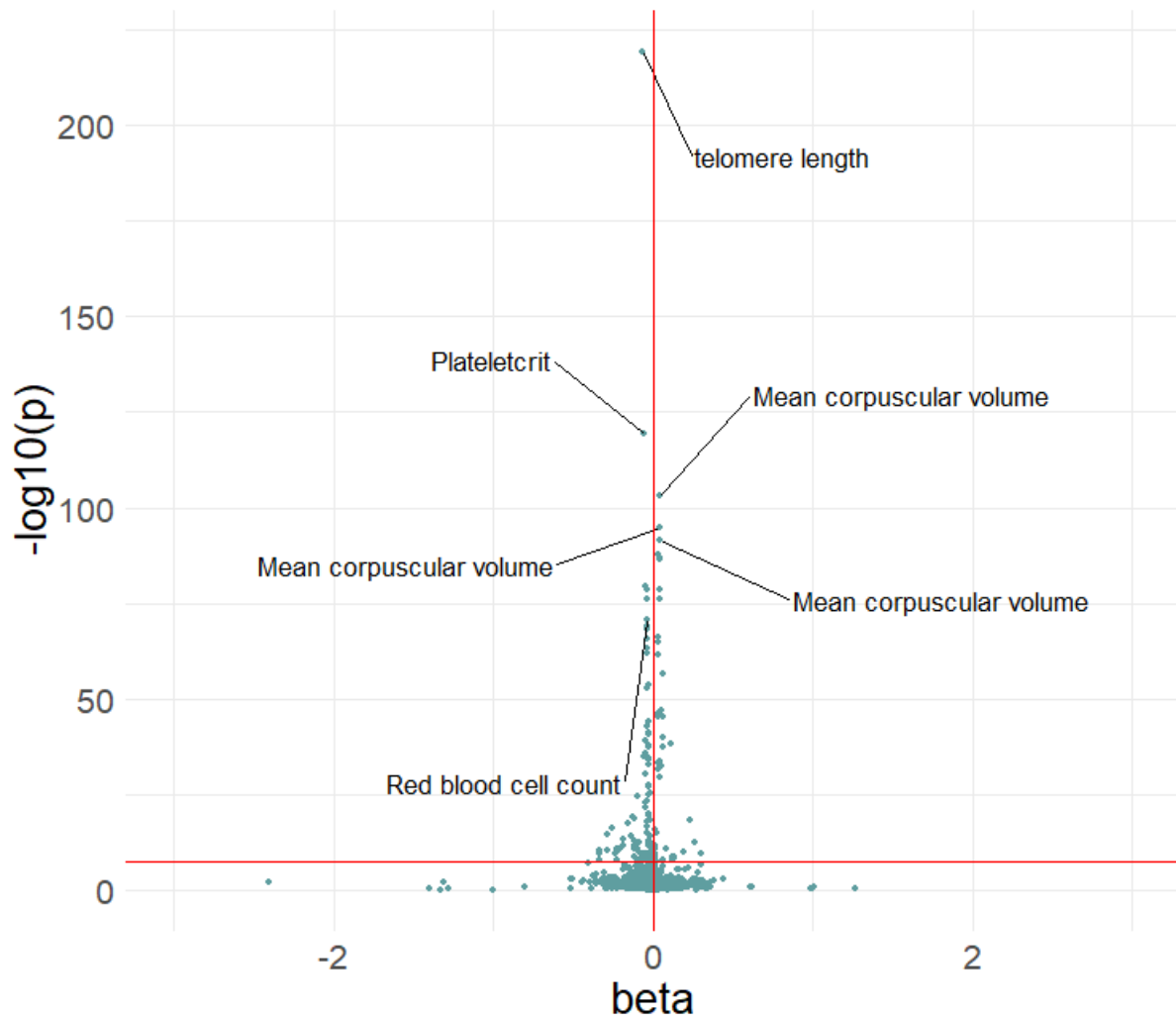


Figure 2.3.7: A volcano plot presenting phenotypes associated with rs7725218 (A > G) (*TERT*). The effect size (beta statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 5 \times 10^{-8}$ threshold applied in phenotypic association search criteria.

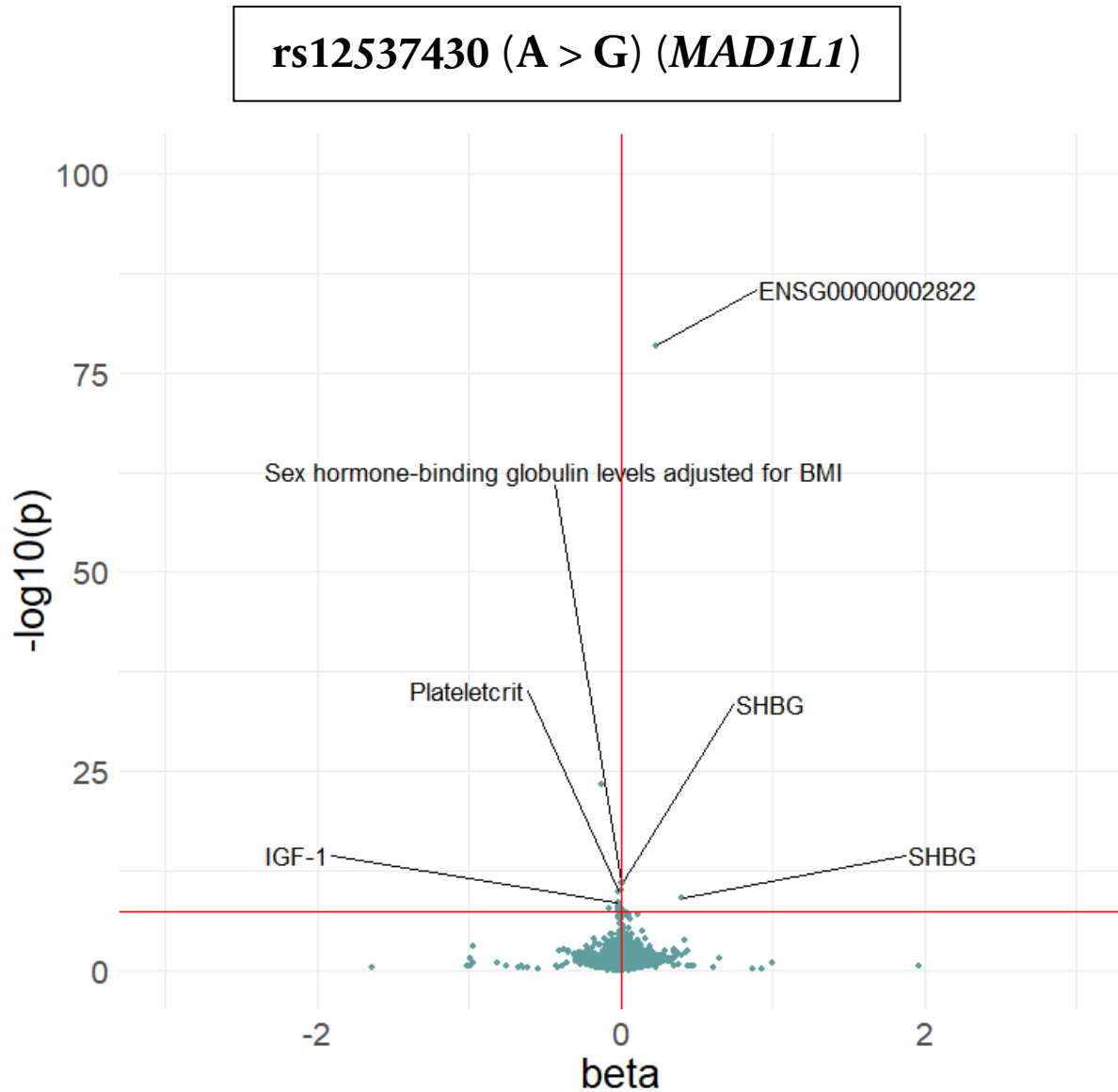


Figure 2.3.8: A volcano plot presenting phenotypes associated with rs12537430 (A > G) (*MAD1L1*). The effect size (β statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 5 \times 10^{-8}$ threshold applied in phenotypic association search criteria. SHBG: sex hormone binding globulin. IGF-1: insulin like growth factor 1. ENSG00000002822 is the Ensembl gene ID for *MAD1L1*.

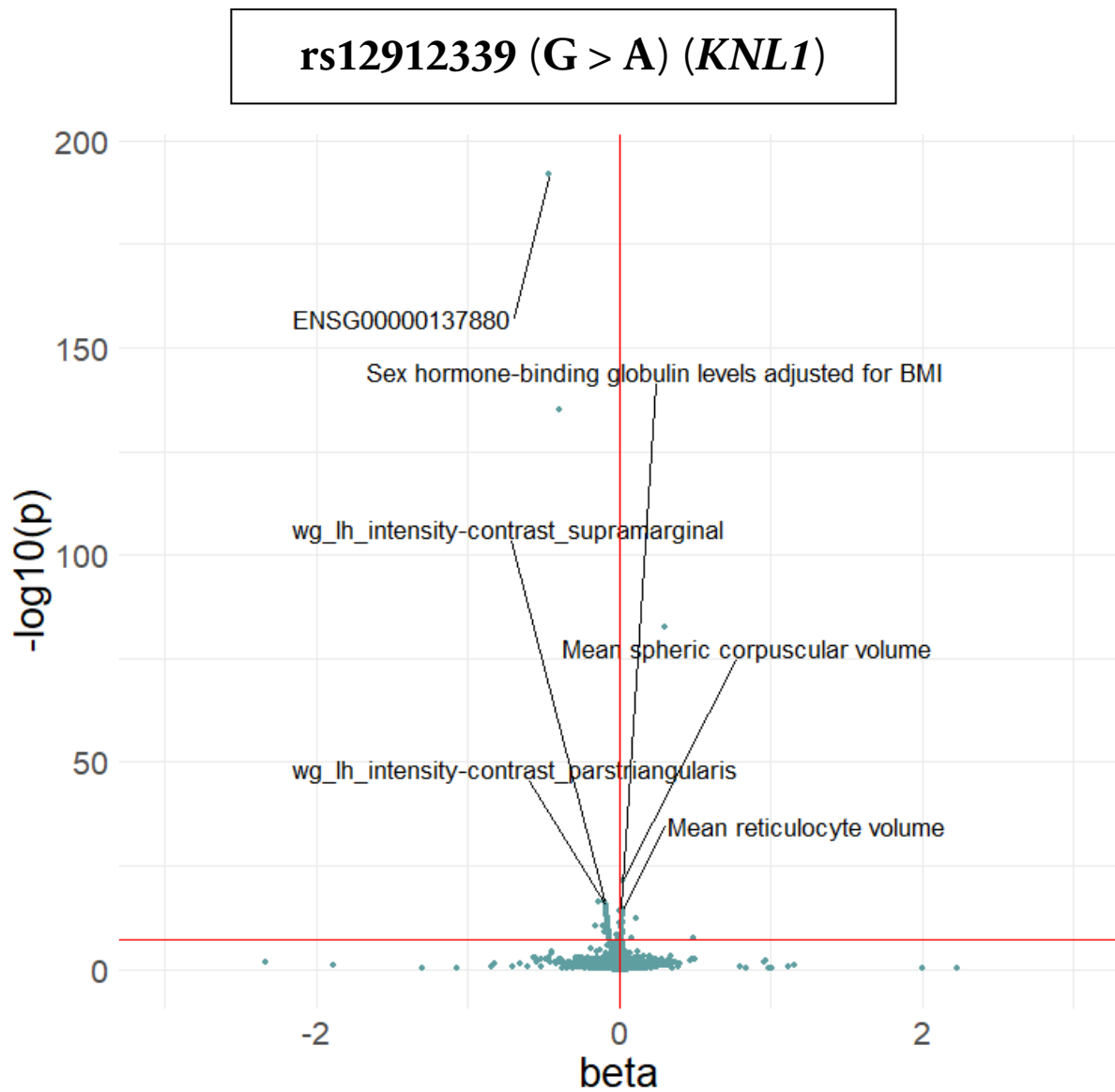


Figure 2.3.9: A volcano plot presenting phenotypes associated with rs12912339 (G > A) (KNL1). The effect size (beta statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 5 \times 10^{-8}$ threshold applied in phenotypic association search criteria. ENSG00000137880 is the Ensembl gene ID for GCHFR.

2.2.2.4 Phenotypes associated with cellular transport, signaling, and metabolism genetic variants

The *NPRL3* IPF risk SNP (rs74614704) was indicated to be associated with increased expression of *POLR3K* (RNA polymerase III subunit K) ($\beta = 0.39, p = 3.65 \times 10^{-35}$) (Figure 2.3.10). This variant is also associated with decreased expression of multiple genes including *IL9RP3* (interleukin 9 receptor pseudogene 3) ($\beta = -0.45, p = 1.22 \times 10^{-46}$), *NPRL3* ($\beta = -0.20, p = 2.17 \times 10^{-10}$), *HBZ* (haemoglobin subunit zeta) ($\beta = -0.19, p = 8.55 \times 10^{-10}$), and *DECR2* (involved in unsaturated fatty acid biosynthesis) ($\beta = -0.17, p = 4.18 \times 10^{-8}$).

Traits associated with rs10808505 (*DEPTOR*) include a decreasing effect on height ($\beta = -0.02, p = 1.00 \times 10^{-22}$), appendicular lean mass ($\beta = -0.01, p = 2.89 \times 10^{-12}$), estimated glomerular filtration rate ($\beta = -0.002, p = 3.4 \times 10^{-10}$), and brain imaging derived phenotypes (IDPs) such as mean diffusivity in the superior corona radiata (**left:** $\beta = -0.04, p = 3.44 \times 10^{-8}$, **right:** $\beta = -0.05, p = 5.36 \times 10^{-9}$) (Figure 2.3.11). This SNP is also associated with increased platelet count ($\beta = 0.01, p = 1.00 \times 10^{-8}$) and a range of brain IDPs, such as increased intracellular volume fraction (ICVF) in the superior longitudinal fasciculus ($\beta = 0.002, p = 1.67 \times 10^{-8}$), superior corona radiata ($\beta = 0.05, p = 9.00 \times 10^{-10}$), and the splenium of the corpus callosum ($\beta = 0.05, p = 1.19 \times 10^{-10}$). PheWAS results indicated further associations with decreased expression of *DSCC1* (predicted to be involved in regulation of DNA replication machinery) ($\beta = -0.14, p = 4.36 \times 10^{-33}$) and increased expression of *DEPTOR* ($\beta = 0.25, p = 8.64 \times 10^{-96}$), *ENSG00000254343* (long non-coding RNA of currently unknown function) ($\beta = 0.15, p = 6.34 \times 10^{-35}$), and *TAF2* (subunit of transcription factor IID; involved in initiation of RNA polymerase II-dependent transcription) ($\beta = 0.08, p = 5.64 \times 10^{-12}$) (HPA^{102,103}).

The *ATP11A* IPF risk variant (rs9577395) is associated with increased ease of skin tanning ($\beta = 0.03, p = 3.20 \times 10^{-34}$) and frequency of sunburns ($\beta = 0.01, p = 3.40 \times 10^{-10}$), red cell distribution width ($\beta = 0.03, p = 6.00 \times 10^{-26}$), and haemoglobin A1C (HbA1c) levels in blood ($\beta = 0.02, p = 8.8 \times 10^{-23}$) (Figure 2.3.12). This variant is also associated with

decreased mean corpuscular volume ($\beta = -0.01, p = 2.10 \times 10^{-12}$), and mean corpuscular haemoglobin ($\beta = -0.01, p = 1.8 \times 10^{-11}$). Finally, results show that rs9577395 is associated with increased expression of *ATP11A* ($\beta = 0.12, p = 2.48 \times 10^{-15}$).

The rs2304645 IPF risk variant (*IVD*) is associated with increased total testosterone ($\beta = 0.03, p = 7.00 \times 10^{-24}$) and bioavailable testosterone levels ($\beta = 0.02, p = 5.80 \times 10^{-13}$) (Figure 2.3.13). Identified phenotypic associations with rs2304645 also included increased standing height ($\beta = 0.008, p = 2.00 \times 10^{-9}$), increased platelet distribution width ($\beta = 0.01, p = 3.9 \times 10^{-11}$), and decreased lung function measures (FVC: $\beta = -0.02, p = 2.4 \times 10^{-12}$, FEV1: $\beta = -0.02, p = 2.93 \times 10^{-11}$). This SNP is also associated with expression of *KNSTRN* (essential function in correct mitotic spindle assembly) ($\beta = 0.36, p = 4.41 \times 10^{-209}$), *DISP2* (predicted to be required for normal Hedgehog protein signaling during embryonic development) ($\beta = 0.10, p = 6.96 \times 10^{-17}$), and *RMDN3* (involved in microtubule binding and spindle biology) ($\beta = -0.07, p = 1.43 \times 10^{-9}$) (*HPA*^{102,103}).

Phenotypes associated with the *AKAP13* IPF risk variant (rs11073517) include reduced basal metabolic rate ($\beta = -0.01, p = 2.80 \times 10^{-13}$), whole body fat-free mass ($\beta = -0.01, p = 7.40 \times 10^{-12}$), haemoglobin A1c levels ($\beta = -0.02, p = 7.30 \times 10^{-15}$), and aspartate aminotransferase levels ($\beta = -0.02, p = 1.20 \times 10^{-10}$) (Figure 2.3.14). This SNP is also associated with increased corneal resistance factor ($\beta = 0.07, p = 5.60 \times 10^{-14}$). Finally, results from PheWAS of eQTL data indicate associations with decreased expression of *AKAP13* ($\beta = -0.18, p = 4.18 \times 10^{-44}$) and increased expression of the ENSG00000259407 long non-coding RNA gene ($\beta = 0.12, p = 8.55 \times 10^{-21}$).

The *DPP9* IPF risk variant (rs12610495) is associated with increased severity of COVID-19 infection (very severe respiratory confirmed vs population: $\beta = 0.23, p = 3.75 \times 10^{-12}$, hospitalised vs population: $\beta = 0.16, p = 6.15 \times 10^{-11}$) (Figure 2.3.15). This SNP is also associated with reduced expression of *DPP9* ($\beta = -0.19, p = 3.05 \times 10^{-47}$) and increased

expression of *TNFAIP8L1* (a negative regulator of mTOR pathway signaling) ($\beta = 0.10, p = 1.43 \times 10^{-14}$) (*HPA*^{102,103}).

PheWAS results show the rs2897075 IPF risk (*ZKSCAN1*) variant is associated with increased FEV1/FVC ratio ($\beta = 0.02, p = 1.90 \times 10^{-25}$), platelet count ($\beta = 0.01, p = 1.90 \times 10^{-8}$), and mean corpuscular haemoglobin ($\beta = 0.01, p = 3.60 \times 10^{-17}$) (Figure 2.3.16). In addition, this SNP is associated with decreased total testosterone levels ($\beta = -0.03, p = 2.90 \times 10^{-33}$), red blood cell count ($\beta = -0.01, p = 5.74 \times 10^{-13}$), red cell distribution width ($\beta = -0.01, p = 1.83 \times 10^{-12}$), and diastolic blood pressure ($\beta = -0.10, p = 1.96 \times 10^{-8}$). Finally, this SNP is also associated with variation in expression of multiple genes, including *TRIM4* (E3 ubiquitin-protein ligase; biological function currently unknown) ($\beta = -0.48, p = 3.27 \times 10^{-310}$), *CNPY4* (regulates cell surface expression of TLR4) ($\beta = -0.36, p = 1.43 \times 10^{-201}$), *STAG3* (regulates sister chromatid cohesion during cell division) ($\beta = -0.24, p = 2.69 \times 10^{-84}$), and *PILRB* (involved in immune system regulation) ($\beta = 0.22, 1.87 \times 10^{-72}$).

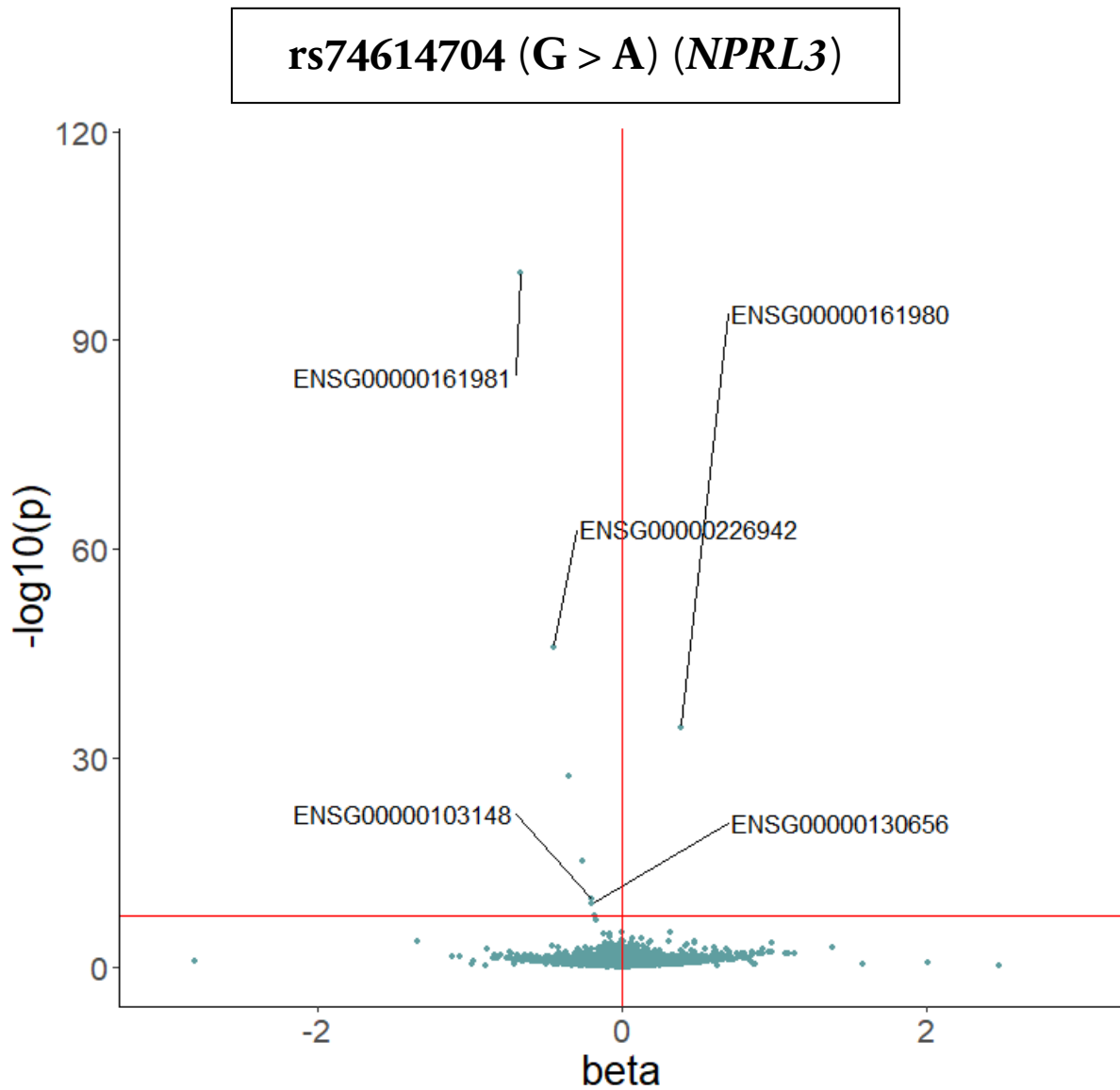


Figure 2.3.10: A volcano plot presenting phenotypes associated with rs74614704 (G > A) (NPRL3). The effect size (beta statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 5 \times 10^{-8}$ threshold applied in phenotypic association search criteria. ENSG000001378801, ENSG00000161980, ENSG00000226942, ENSG00000103148, and ENSG00000130656 are Ensembl gene IDs for SNRNP25, POLR3K, IL9RP3, NPRL3, and HBZ, respectively.

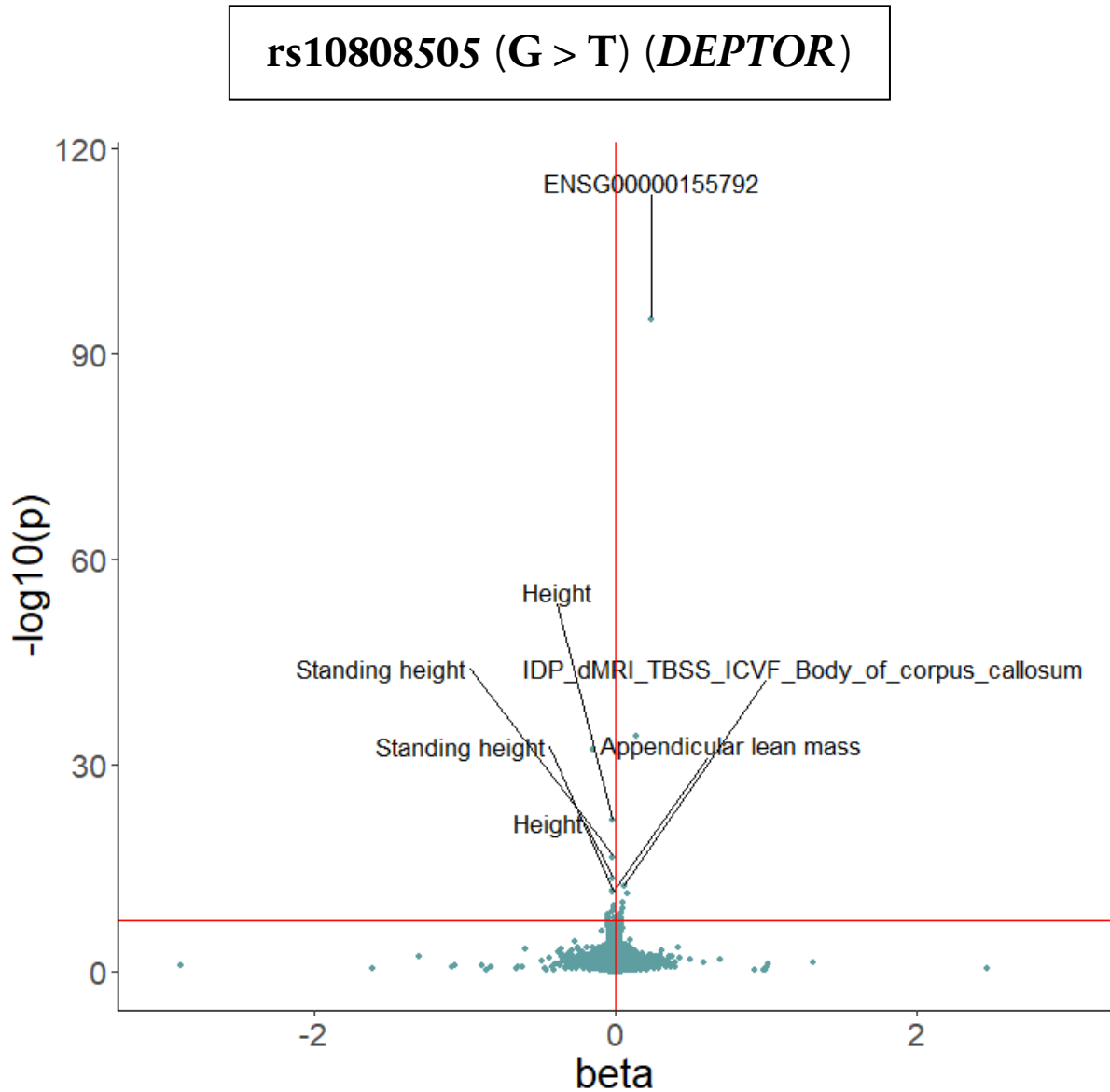


Figure 2.3.11: A volcano plot presenting phenotypes associated with rs10808505 (G > T) (DEPTOR). The effect size (beta statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 5 \times 10^{-8}$ threshold applied in phenotypic association search criteria. ENSG00000155792 is the Ensembl gene ID for DEPTOR.

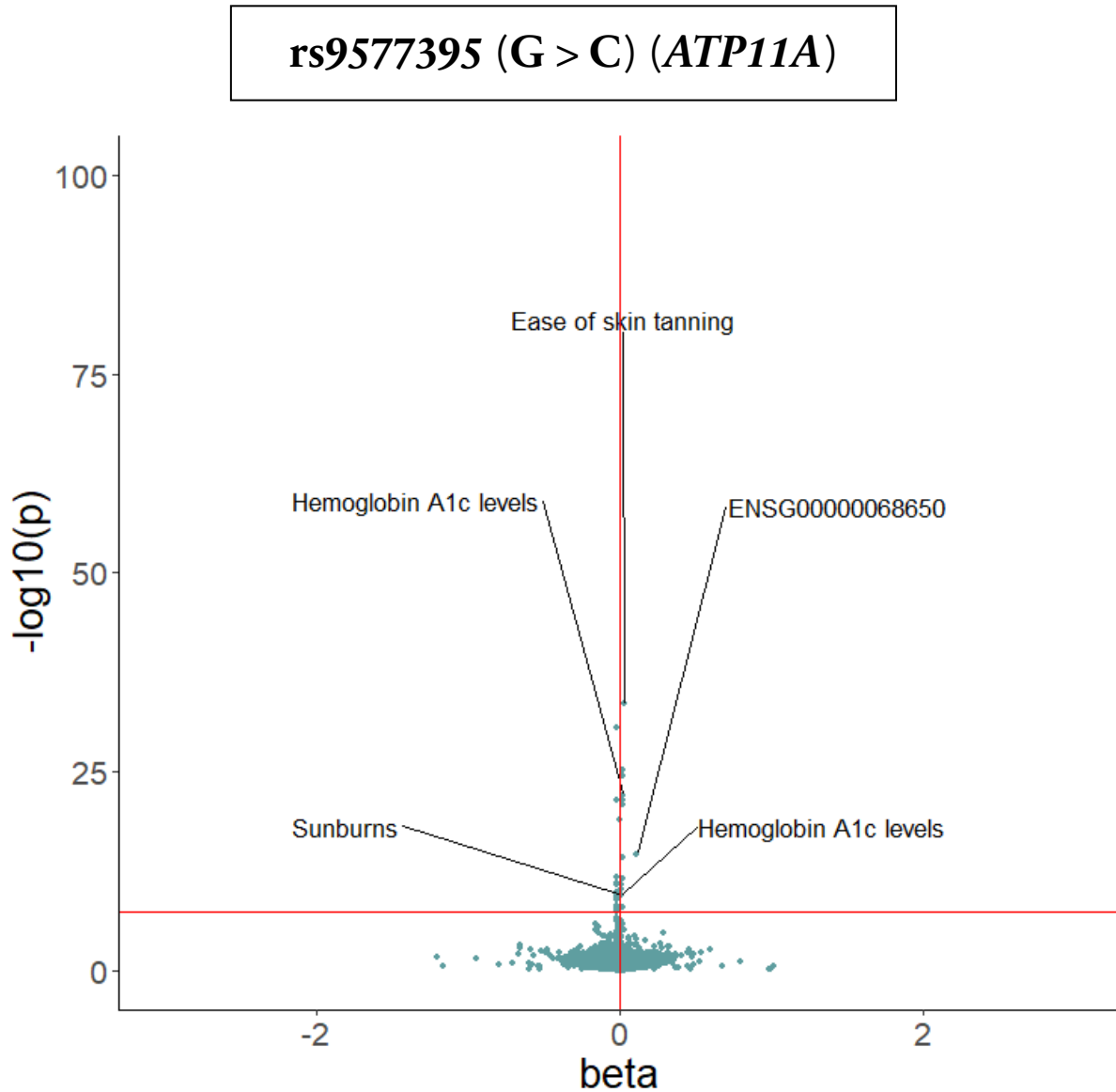


Figure 2.3.12: A volcano plot presenting phenotypes associated with rs9577395 (G > C) (ATP11A). The effect size (beta statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 5 \times 10^{-8}$ threshold applied in phenotypic association search criteria. ENSG00000068650 is the Ensembl gene ID for ATP11A.

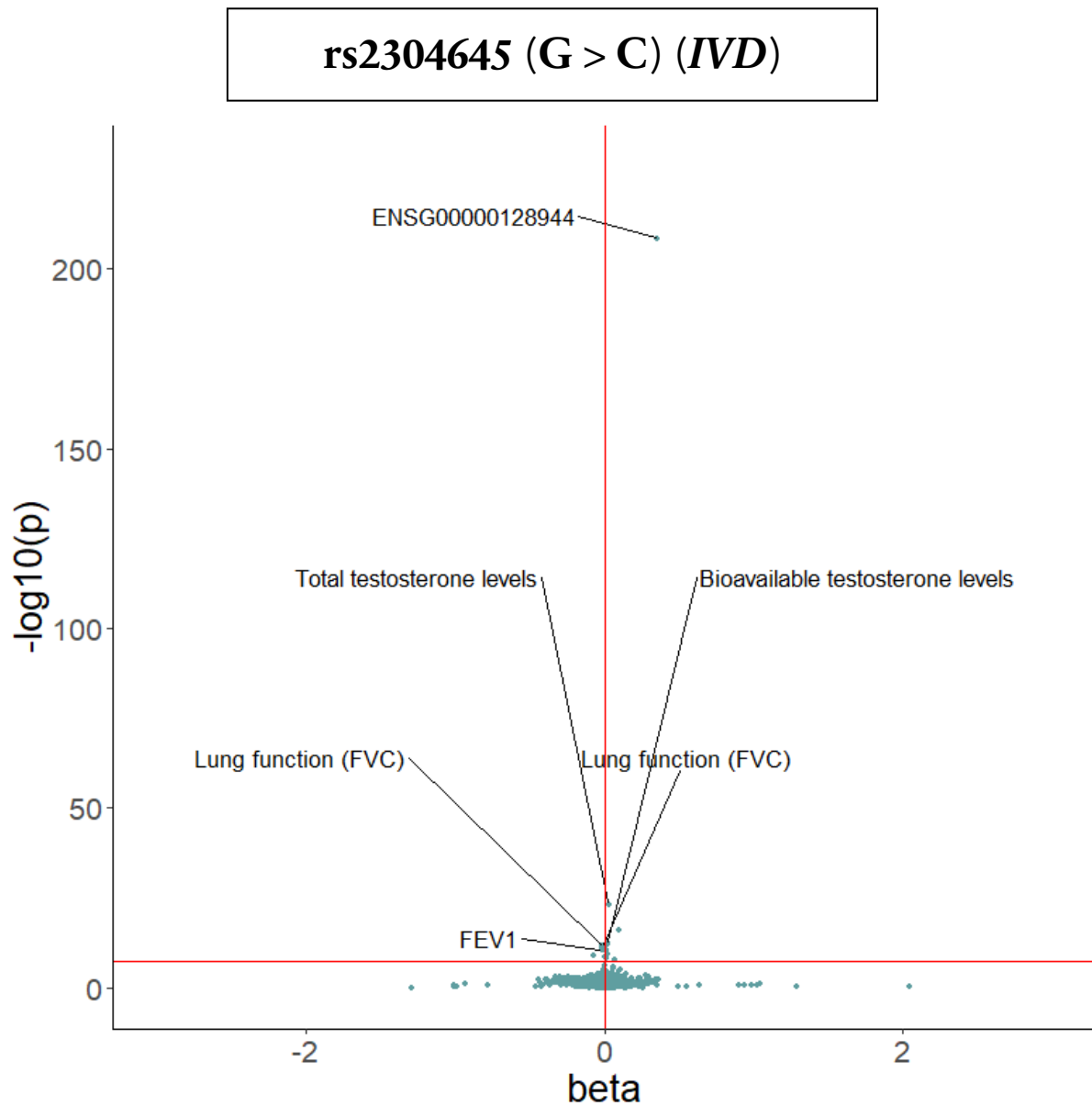


Figure 2.3.13: A volcano plot presenting phenotypes associated with rs2304645 (G > C) (IVD). The effect size (beta statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 5 \times 10^{-8}$ threshold applied in phenotypic association search criteria. ENSG00000128944 is the Ensembl gene ID for KNSTRN.

rs11073517 (C > T) (AKAP13)

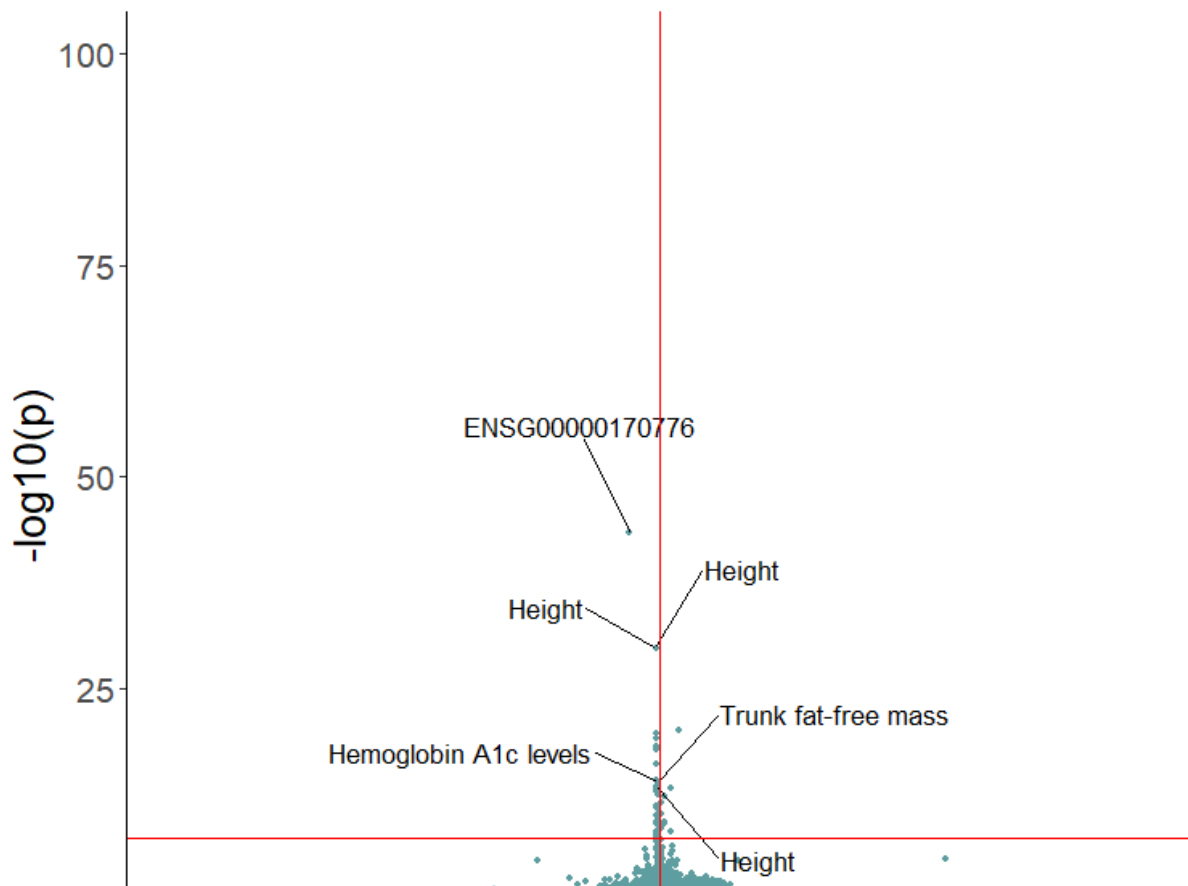


Figure 2.3.14: A volcano plot presenting phenotypes associated with rs11073517 (C > T) (AKAP13). The effect size (beta statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 5 \times 10^{-8}$ threshold applied in phenotypic association search criteria. ENSG00000170776 is the Ensembl gene ID for AKAP13.

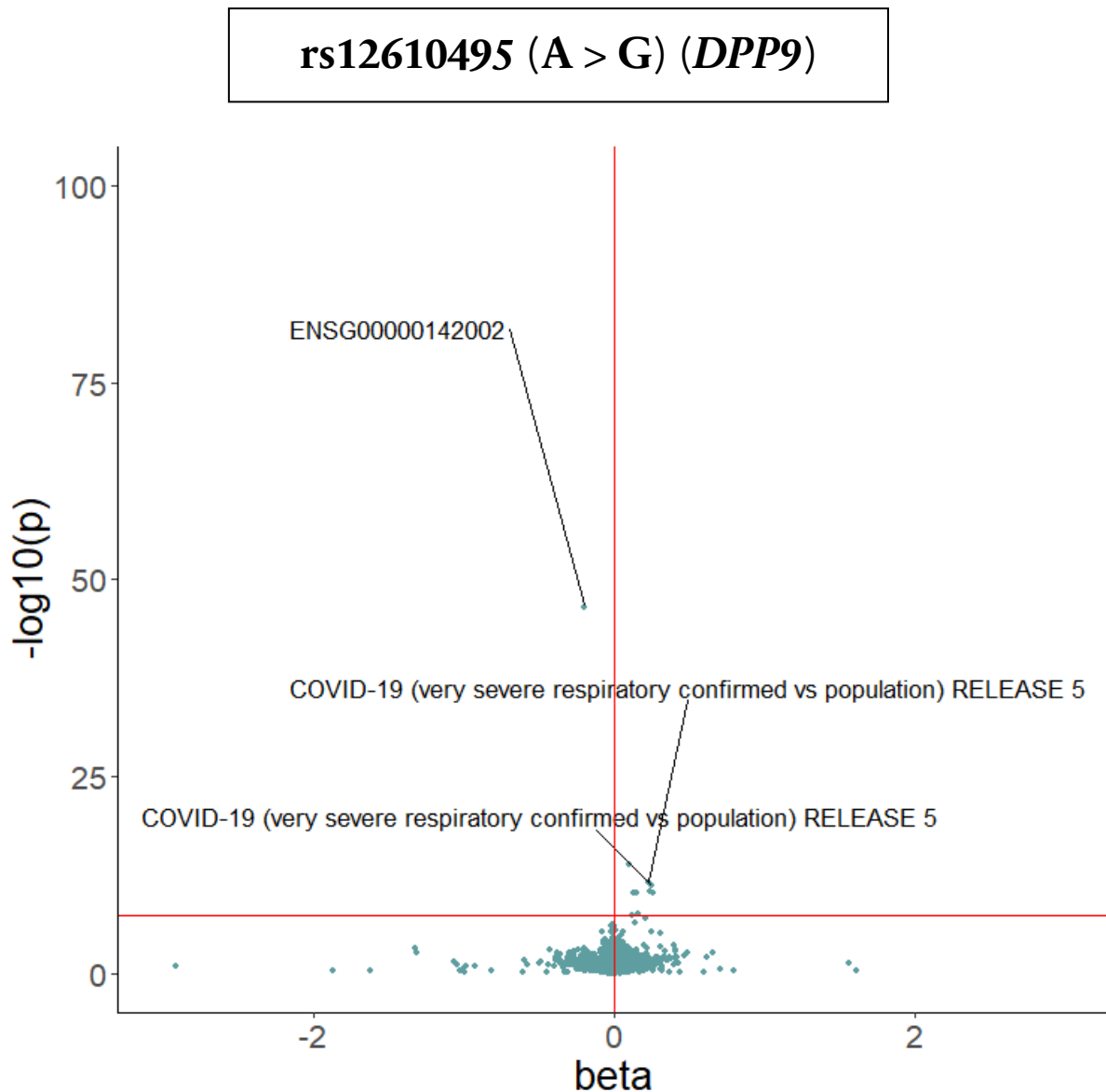


Figure 2.3.15: A volcano plot presenting phenotypes associated with rs12610495 (A > G) (DPP9). The effect size (beta statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 5 \times 10^{-8}$ threshold applied in phenotypic association search criteria. ENSG00000142002 is the Ensembl gene ID for DPP9.

rs2897075 (C > T) (ZKSCAN1)

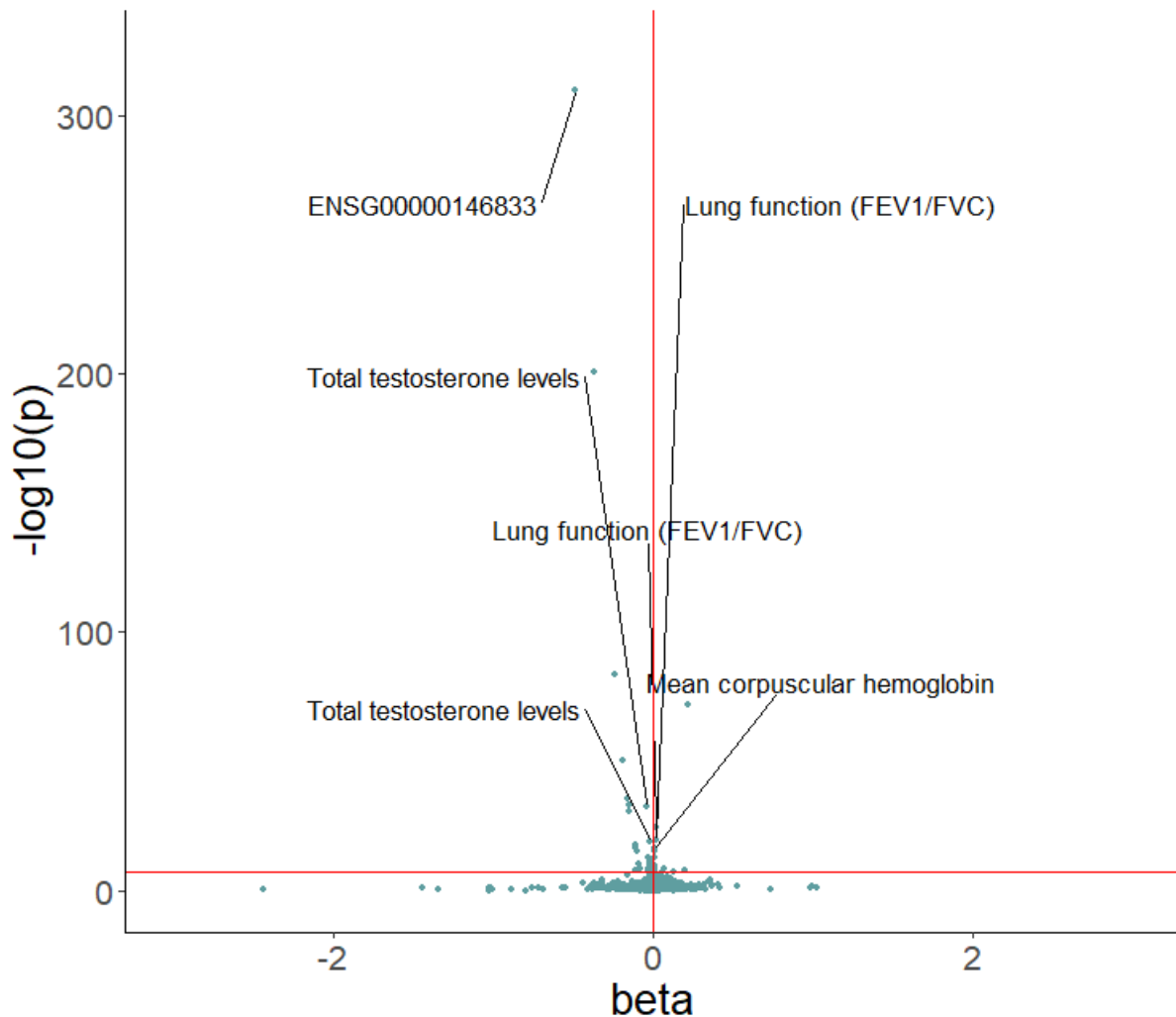


Figure 2.3.16: A volcano plot presenting phenotypes associated with rs2897075 (C > T) (ZKSCAN1). The effect size (beta statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 5 \times 10^{-8}$ threshold applied in phenotypic association search criteria. ENSG00000146833 is the Ensembl gene ID for TRIM4.

2.2.2.17 Phenotypes associated with genetic variants of other or unclassified function

The rs2077551 IPF risk variant was found to be highly pleiotropic. PheWAS results show this SNP has over 600 phenotypic associations after filtering with a $p < 5 \times 10^{-8}$ significance threshold. The high volume of varied phenotypic associations limits biological interpretation, however many traits associated with other IPF risk variants are also associated with rs2077551. Phenotypes associated with rs2077551 and other investigated SNPs include lung function (FVC: $\beta = 0.03, p = 1.40 \times 10^{-49}$, FEV1: $\beta = 0.03, p = 7.21 \times 10^{-27}$), HbA1c ($\beta = 0.02, p = 1.40 \times 10^{-37}$), neutrophil count ($\beta = -0.02, p = 1.63 \times 10^{-21}$), mean corpuscular haemoglobin ($\beta = 0.01, p = 3.50 \times 10^{-18}$), total testosterone levels ($\beta = -0.03, p = 4.70 \times 10^{-16}$), aspartate aminotransferase levels ($\beta = 0.02, p = 1.51 \times 10^{-15}$), and height ($\beta = 0.02, p = 3.65 \times 10^{-15}$) (Figure 2.3.17).

The IPF risk variant rs79684490 (10q25.1, G > A) is not associated with any traits within p-value threshold criteria ($p < 5 \times 10^{-8}$) (Figure 2.3.18).

PheWAS results show the rs2609259 (*FAM13A*) A allele is associated with decreased FVC ($\beta = -0.02, p = 1.50 \times 10^{-17}$) and increased FEV1/FVC ratio ($\beta = 0.05, p = 3.80 \times 10^{-94}$) (Figure 2.3.19). In addition, this SNP is associated with decreased white blood cell count ($\beta = -0.02, p = 3.46 \times 10^{-12}$), decreased monocyte count ($\beta = -0.01, p = 2.85 \times 10^{-9}$), decreased risk of chronic obstructive pulmonary disease ($\beta = -0.07, p = 6.90 \times 10^{-9}$), and increased mean platelet volume ($\beta = 0.01, p = 2.00 \times 10^{-8}$). This IPF risk variant is also associated with increased *FAM13A* gene expression ($\beta = 0.48, p = 3.97 \times 10^{-276}$).

rs2077551 (C > T) (17q21.31)

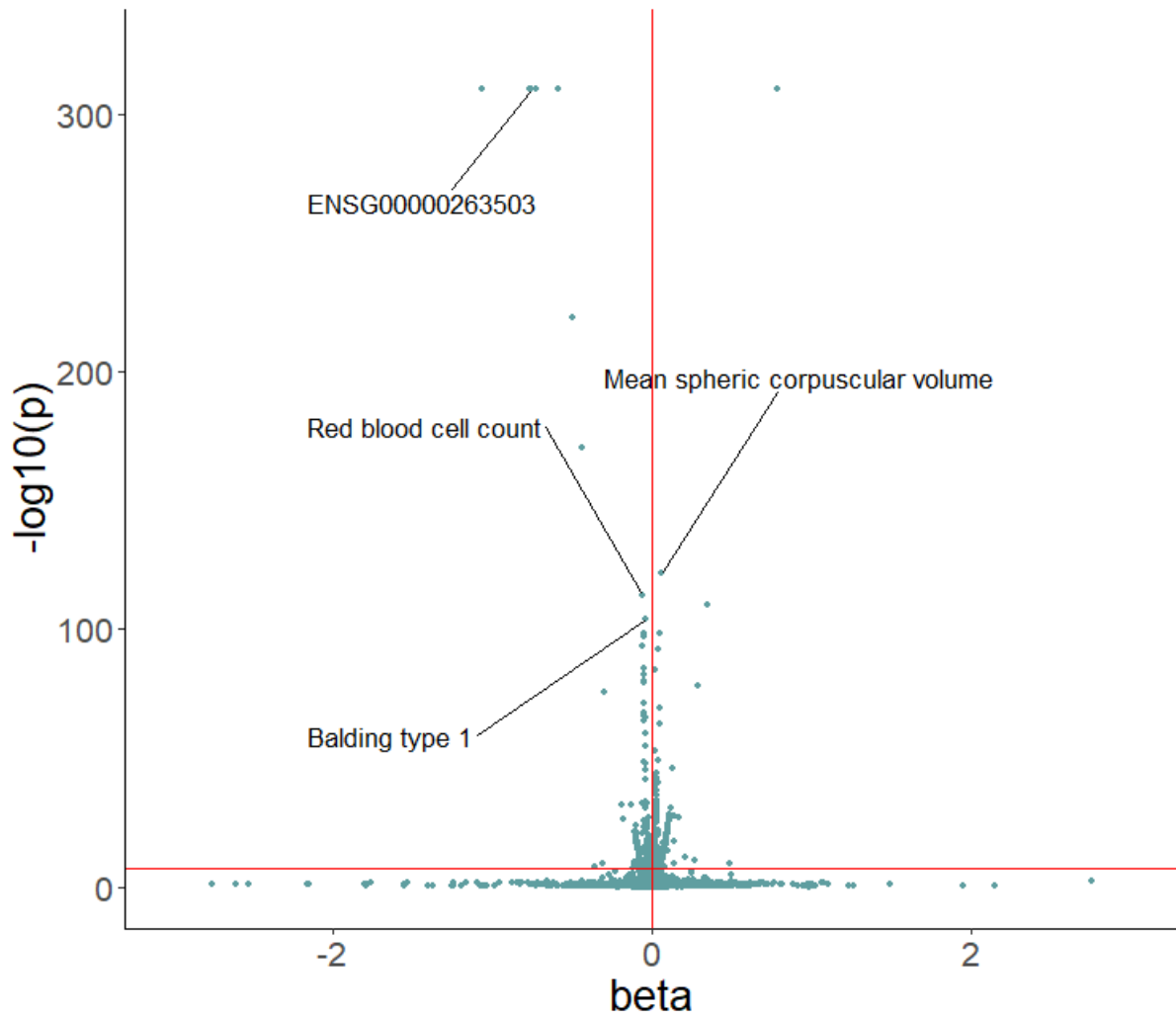


Figure 2.3.17: A volcano plot presenting phenotypes associated with rs2077551 (C > T) (17q21.31). The effect size (beta statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 5 \times 10^{-8}$ threshold applied in phenotypic association search criteria. ENSG00000146833 is the Ensembl gene ID for RP11-707O23.5 (pseudogene).

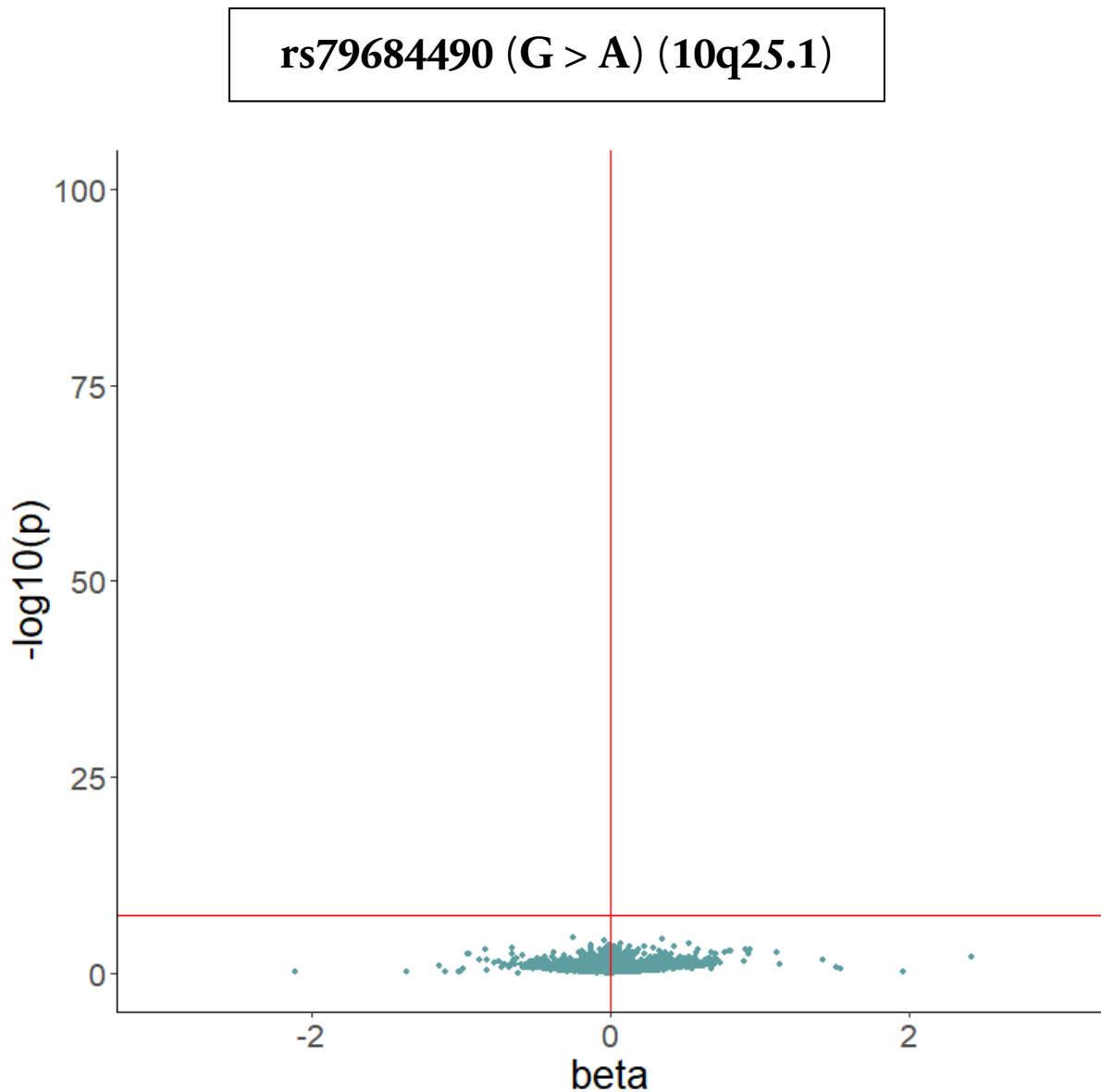


Figure 2.3.18: A volcano plot presenting phenotypes associated with rs79684490 (G > A) (10q25.1). The effect size (beta statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 5 \times 10^{-8}$ threshold applied in phenotypic association search criteria.

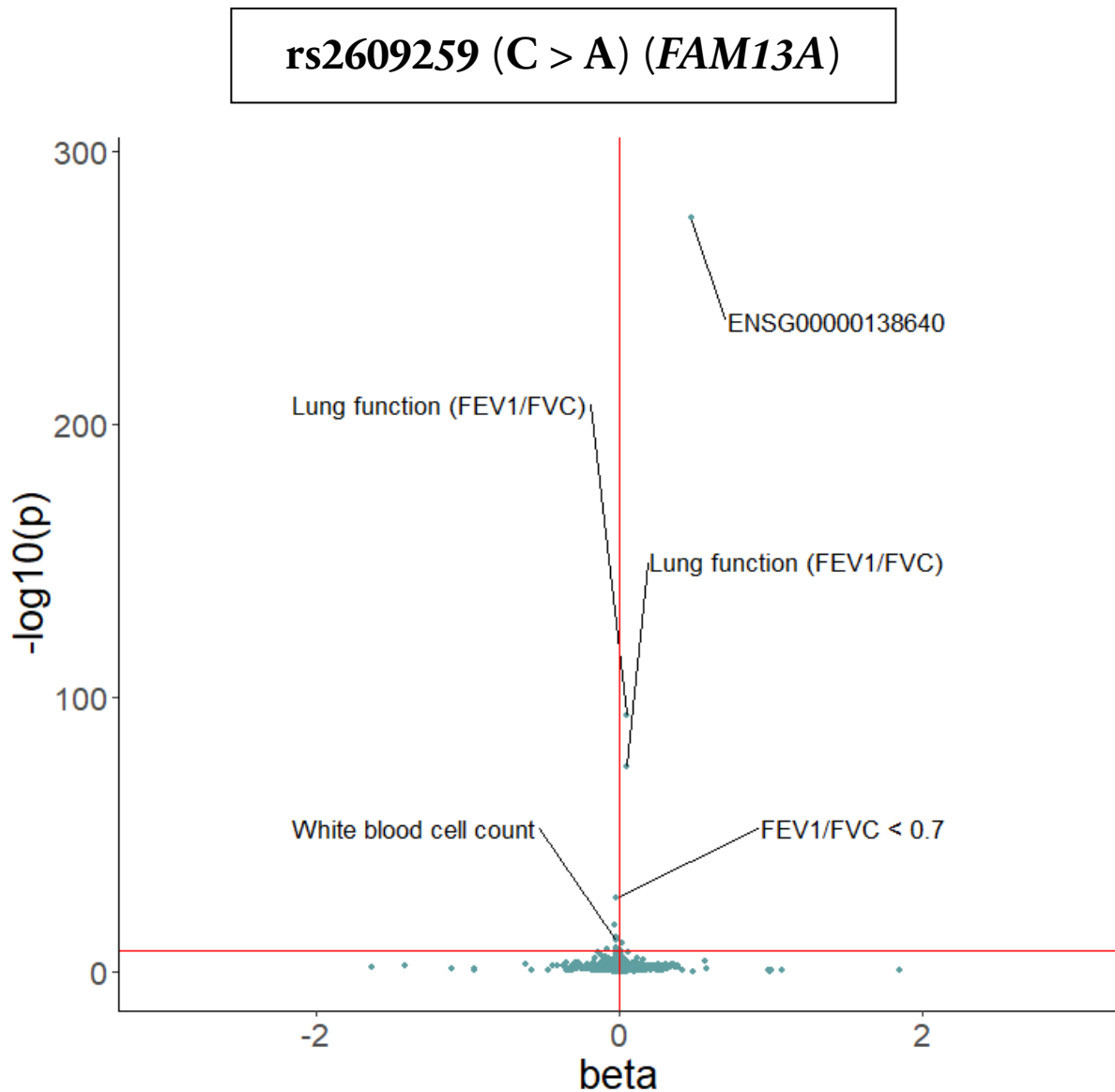


Figure 2.3.19: A volcano plot presenting phenotypes associated with rs2609259 (C > A) (*FAM13A*). The effect size (beta statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 5 \times 10^{-8}$ threshold applied in phenotypic association search criteria. ENSG00000138640 is the Ensembl gene ID for *FAM13A*.

Overall, the 19 IPF genetic risk variants tested are associated with a broad range of phenotypes. The most prominent associations are with blood traits (220 associations with $p < 5 \times 10^{-8}$) such as plateletcrit, mean corpuscular haemoglobin, red blood cell count, and white blood cell count (Figure 2.3.20). IPF genetic risk variants are also associated with 62 neurological traits, such as variation in intracellular volume fraction and grey-white contrast across multiple brain regions. Commonly associated phenotypes also include neoplastic traits (forty-seven associations within $p < 5 \times 10^{-8}$ threshold) such as prostate cancer, lung adenocarcinoma, and serous ovarian cancer. Associated pulmonary phenotypes include interstitial lung disease, lung function measures, and chronic obstructive pulmonary disease (36 total associations within $p < 5 \times 10^{-8}$ threshold). IPF risk variants are also associated with endocrine and metabolic phenotypes, such as HbA1c levels and basal metabolic rate, and a range of biomarkers including telomere length, total testosterone levels, sex hormone binding globulin levels, and height.

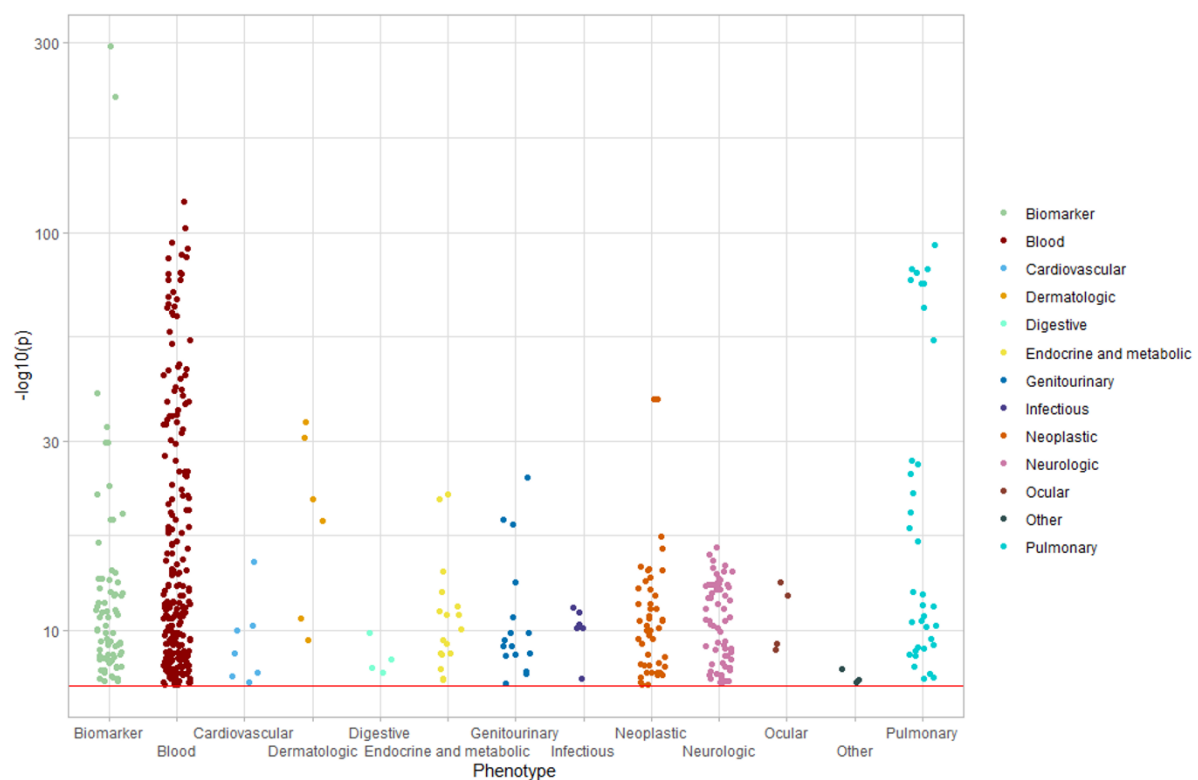


Figure 2.3.20: Phenotypes associated with 19 IPF genetic risk variants. Phenotype categories are presented across the x-axis and $-\log_{10}(p\text{-value})$ of SNP-phenotype associations are presented on the y-axis. Phenotype categories are ranked in alphabetical order. The red horizontal line indicates the $p = 5 \times 10^{-8}$ threshold applied in this analysis.

2.3.2 Conclusions from PheWAS of public GWAS data

The aim of this PheWAS approach was to investigate the spectrum of traits associated with 19 IPF genetic risk variants using large-scale data. Of the 19 IPF risk variants identified in currently published GWAS studies (Table 2.2.1), 18 were found to be associated with non-IPF traits. This method was therefore successful in achieving aim 1.1 of this thesis (Section 1.3). The 19 IPF genetic risk variants were generally found to be highly pleiotropic.

2.3.3 Prioritisation of public PheWAS results for colocalisation analyses

Traits associated with IPF genetic risk variants were prioritised for follow-up colocalisation analyses in line with two aims of the PheWAS investigation performed. This process aimed to firstly investigate the downstream biology of IPF genetic risk variants of similar function, by grouping SNPs by predicted function and selecting top associations by lowest p-value from each functional group. Secondly, overall associations with 19 IPF genetic risk were grouped by phenotype and top associations by lowest p-value were selected for follow-up colocalisation analyses.

Highly significant associations with functional SNP categories led to prioritisation of traits such as insulin-like growth factor 1 levels, impedance of whole body, c-reactive protein levels, and estimated glomerular filtration rate (creatinine) (associated with the *cytoskeletal proteins* SNP category) (Table 2.3.1). Further prioritised traits include mean corpuscular

SNP functional category	Associated trait prioritised for colocalisation analysis
Cytoskeletal proteins	Lung function (FEV1/FVC)
	Insulin-like growth factor 1 levels
	Impedance of whole body
	C-reactive protein levels
	Estimated glomerular filtration rate (creatinine)
Chromosome maintenance and cell cycle control	telomere length
	Plateletcrit
	Mean corpuscular hemoglobin
	Red blood cell count
	Neutrophil count
Cellular transport, signaling, and metabolism	Ease of skin tanning
	Total testosterone levels
	Height
	Red cell distribution width
	Lung function (FEV1/FVC)
Mucin biology	Interstitial lung disease
	Pulmonary fibrosis
Unclassified	Lung function (FEV1/FVC)
	FEV1/FVC < 0.7
	Lung function (FVC)
	Monocyte count
	White blood cell count

Table 2.3.1: Phenotypes prioritised for colocalisation analyses with IPF association signals. Phenotypes were prioritised by statistical significance (*p*-value) within SNP functional groups.

haemoglobin (*chromosome maintenance and cell cycle control*), total testosterone levels, height, red cell distribution width (*cellular transport, signaling, and metabolism*), monocyte count, and white blood cell count (*unclassified*).

Traits prioritised from phenotype categories include COVID-19 (very severe respiratory confirmed vs population) and COVID-19 (hospitalised vs population) (*infectious phenotype category*) (Table 2.3.2). Traits prioritised using this approach also include prostate cancer, lung adenocarcinoma, serous ovarian cancer: low grade and low malignant

potential, neoplasms, and ER- breast cancer (*neoplastic*). Further prioritised traits include ease of skin tanning, uterine fibroids, wg_lh-intensity-contrast_supramarginal (white-grey contrast measurement in left-hand supramarginal brain region), systolic blood pressure, corneal resistance factor (*other disease traits*), and mean corpuscular volume (*biomarkers*).

Phenotype category	Associated trait prioritised for colocalisation analysis
Pulmonary	Lung function (FEV1/FVC)
	Interstitial lung disease
	FEV1/FVC < 0.7
	Pulmonary fibrosis
	Lung function (FVC)
Infectious	COVID-19 (very severe respiratory confirmed vs population)
	COVID-19 (hospitalized vs population)
Neoplastic	Prostate cancer
	Lung adenocarcinoma
	Serous ovarian cancer: low grade and low malignant potential
	Neoplasms
	ER- Breast cancer
Other disease traits	Ease of skin tanning
	Uterine fibroids
	wg_lh_intensity-contrast_supramarginal
	systolic blood pressure
	Corneal resistance factor
Biomarkers	telomere length
	Plateletcrit
	Mean corpuscular volume
	Red blood cell count
	Neutrophil count

Table 2.3.2: Phenotypes prioritised for colocalisation analyses with IPF association signals. Phenotypes were prioritised by statistical significance (p-value) within phenotype groups.

Traits prioritised by both SNP functional grouping and phenotype grouping approaches include pulmonary traits such as lung function measures (FEV1/FVC, FEV1/FVC < 0.7, FVC), interstitial lung disease, and pulmonary fibrosis. Traits prioritised by both approaches also included telomere length, plateletcrit, red blood cell count, neutrophil count, and ease of skin tanning. Finally, 62 eQTL associations with IPF genetic risk variants were included for further investigation.

2.3.4 Results from colocalisation analyses of prioritised traits

2.3.4.1 Example 1: Colocalisation of associations with IPF risk and a non-IPF phenotype

Results of colocalisation analyses suggest associations signals of increased IPF risk ($\beta = 0.25$, $p = 8.66 \times 10^{-20}$) and increased total testosterone levels ($\beta = 0.03$, $p = 7.00 \times 10^{-24}$) have strong evidence of colocalisation (posterior probability of H4; $PP.H4 = 99.0\%$) (Figure 2.3.21).

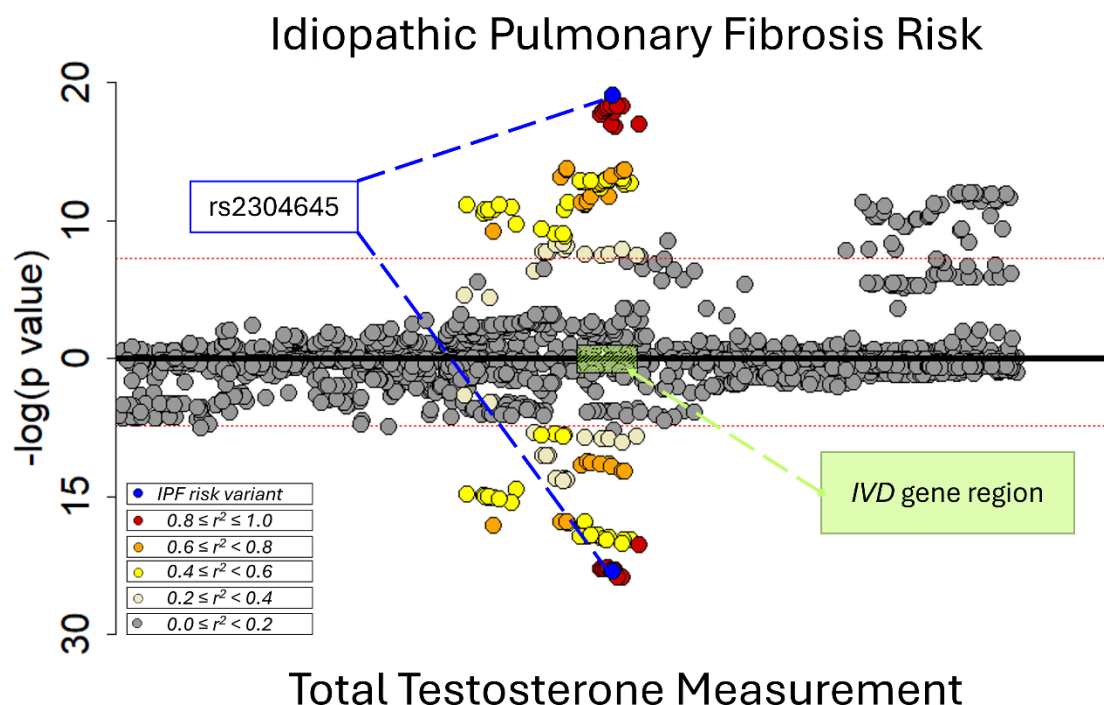


Figure 2.3.21: Colocalisation of association signals with IPF risk and total testosterone measurement. Each data point represents a SNP tested in each GWAS. The red dashed line indicates $p = 5 \times 10^{-8}$. Data point colour indicates LD (r^2) with the lead IPF risk variant.

2.3.4.2 Example 2: Non-colocalisation of associations with IPF risk and a non-IPF phenotype

Results did not support colocalisation of signals associated with increased IPF risk ($\beta = 0.35, p = 4.90 \times 10^{-32}$) and decreased telomere length ($\beta = -0.07, p = 5.6 \times 10^{-220}$) at rs7725218 (*TERT*) ($PP.H4 = 0.48\%$). Colocalisation analysis results suggested that these associations are instead likely to be driven by different causal SNPs ($PP.H3 = 99.50\%$) (Figure 2.3.22).

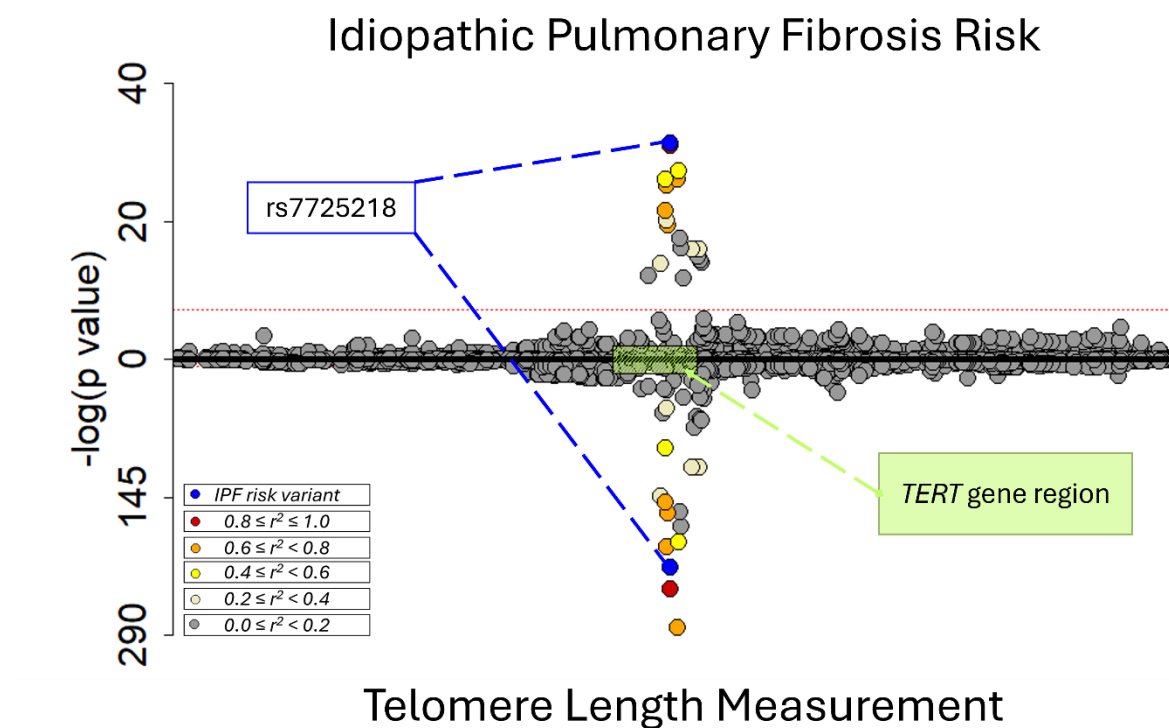


Figure 2.3.22: Non-colocalisation of association signals with IPF risk and telomere length measurement. Each data point represents a SNP tested in each GWAS. The red dashed line indicates $p = 5 \times 10^{-8}$. Data point colour indicates LD (r^2) with the lead IPF risk variant.

2.3.4.3 Example 3: Colocalisation of associations with IPF risk with an eQTL signal

Colocalisation analysis results suggest associations signals of increased IPF risk ($\beta = 0.18$, $p = 6.03 \times 10^{-11}$) and increased *DEPTOR* gene expression ($\beta = 0.25$, $p = 8.64 \times 10^{-96}$) are highly likely to share the same causal variant ($PP.H4 = 93.20\%$) (Figure 2.3.23).

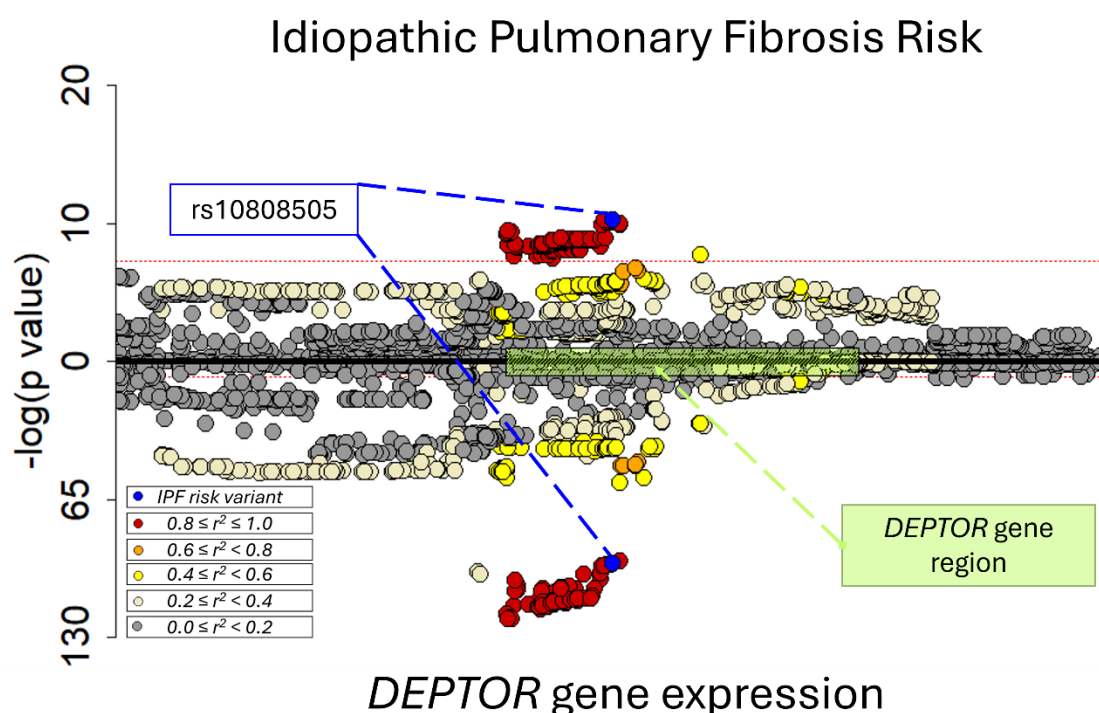


Figure 2.3.23: Colocalisation of association signals with IPF risk and *DEPTOR* gene expression. Each data point represents a SNP tested in each GWAS. The red dashed line indicates $p = 5 \times 10^{-8}$. Data point colour indicates LD (r^2) with the lead IPF risk variant.

2.3.4.4 Example 4: Non-colocalisation of associations with IPF risk with an eQTL signal

Results suggest that associations of rs112087793 with IPF risk ($\beta = 0.29$, $p = 1.09 \times 10^{-8}$) and potential trans-eQTL effect on *ZGPAT* gene expression ($\beta = 0.38$, $p = 1.24 \times 10^{-76}$) are unlikely to colocalise ($PP.H4 = 21.30\%$). These association signals are more likely to be driven by two independent causal variants ($PP.H3 = 78.60\%$) (Figure 2.3.24).

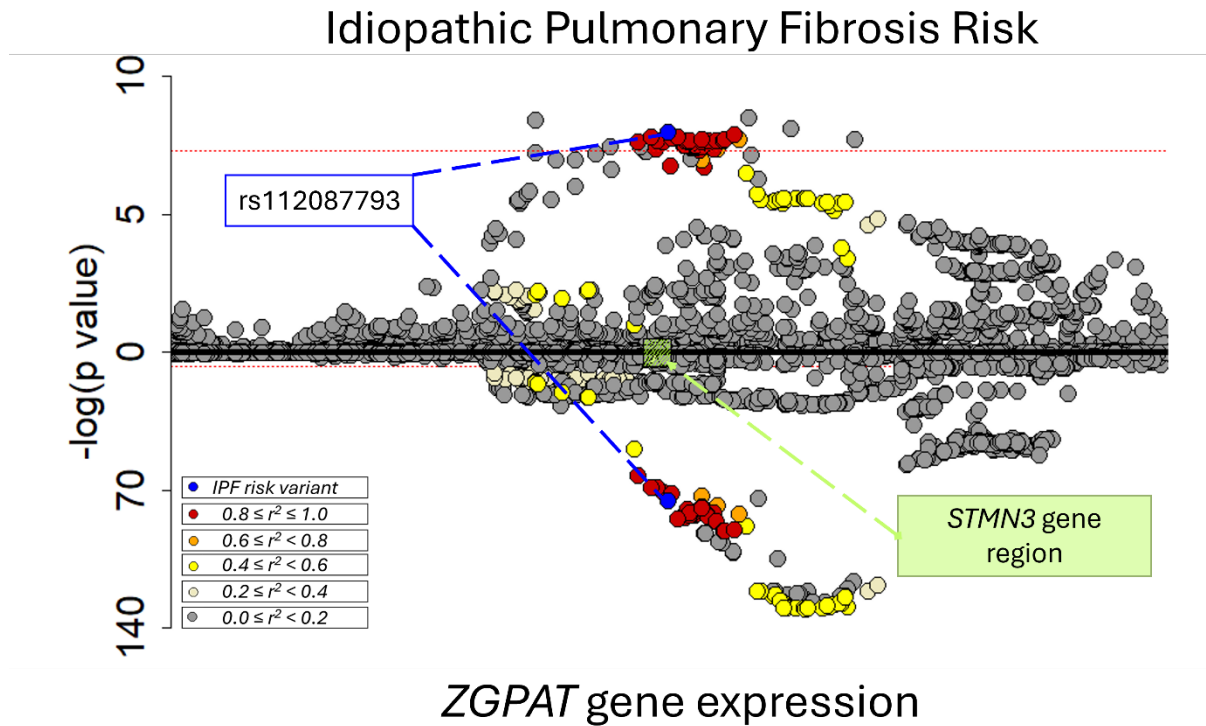


Figure 2.3.24: Non-colocalisation of association signals with IPF risk and ZGPAT gene expression. Each data point represents a SNP tested in each GWAS. The red dashed line indicates $p = 5 \times 10^{-8}$. Data point colour indicates LD (r^2) with the lead IPF risk variant.

2.3.4.5 Summary of traits with strong evidence of colocalisation with IPF risk association signals

SNP-phenotype associations with strong evidence of colocalisation with IPF risk (Table 2.3.3) include shared risk of COVID-19 infection severity, both severe vs population ($PP.H4 = 97.90\%$) and hospitalised vs population ($PP.H4 = 97.90\%$), with rs12610495 (*DPP9*).

Multiple associations with blood traits show strong evidence of colocalisation with IPF risk, including increased mean corpuscular haemoglobin (rs7725218; *TERT*, $PP.H4 = 86.80\%$), mean corpuscular volume (rs12912339; *KNL1*, $PP.H4 = 81.60\%$) (rs2897075; *ZKSCAN1*, $PP.H4 = 83.30\%$), red cell distribution width (rs9577395; *ATP11A*, $PP.H4 = 100.00\%$), and decreased plateletcrit (rs12537430; *MAD1L1*, $PP.H4 = 95.60\%$). Results also suggest colocalisation of associations with IPF risk and decreased monocyte count

(rs2609259; *FAM13A*, *PP.H4* = 97.30%) and overall white blood cell count (rs2609259; *FAM13A*, *PP.H4* = 96.90%).

IPF association signals are also highly likely to colocalise with associations with lung function measures, such as FEV1/FVC (rs2897075; *ZKSCAN1*, *PP.H4* = 99.40%) (rs2609259; *FAM13A*, *PP.H4* = 96.90%) (rs2076295; *DSP*, *PP.H4* = 100.00%) and FEV1/FVC < 0.7 (rs2609259; *FAM13A*, *PP.H4* = 96.90%) (rs2076295; *DSP*, *PP.H4* = 100.00%). It is unclear if colocalised associations with FEV1/FVC are independent of, or driven by, associations with decreased FVC (rs2609259; *FAM13A*, *PP.H4* = 97.20%) (rs2304645; *IVD*, *PP.H4* = 99.10%).

Additional associations with strong evidence of colocalisation with IPF risk include decreased risk of neoplasms (rs9860874; *TERC*, *PP.H4* = 81.40%) and uterine fibroids (rs9860874; *TERC*, *PP.H4* = 93.60%). Results suggest colocalisation of associations with IPF risk and decreased systolic blood pressure (rs9860874; *TERC*, *PP.H4* = 97.90%) and IGF-1 levels (rs12537430; *MAD1L1*, *PP.H4* = 99.20%) and telomere length (rs9860874; *TERC*, *PP.H4* = 89.60%), and increased corneal resistance factor (rs11073517; *AKAP13*, *PP.H4* = 96.40% (left), *PP.H4* = 83.90% (right)) and total testosterone levels (rs2304645; *IVD*, *PP.H4* = 99.00%).

Finally, results suggest strong evidence that associations with ILD (rs7725218; *TERT*, *PP.H4* = 95.50%) (rs35705950; *MUC5B*, *PP.H4* = 100.00%) (rs2076295; *DSP*, *PP.H4* = 100.00%) and pulmonary fibrosis risk (rs7725218; *TERT*, *PP.H4* = 96.70%) (rs35705950; *MUC5B*, *PP.H4* = 100.00%) colocalise with signals identified in the Allen et al (2022) GWAS⁹⁰ study used in PheWAS search criteria (Section 2.2.1). Prioritised associations with ease of skin tanning and *wg-lh-intensity-contrast_supramarginal* were removed from this investigation as neither dataset included sufficient data fields for colocalisation analyses, as minor allele frequency statistics were not present and *sdY* could therefore not be calculated.

IPF SNP	Gene	Phenotype	IEU GWAS ID	Effect relative to IPF risk (beta)	PP.H4
rs12610495	DPP9	COVID-19 - severe vs population	ebi-a-GCST011076	0.23	97.90%
		COVID-19 - hospitalised vs population	ebi-a-GCST011082	0.16	97.90%
rs7725218	TERT	Interstitial lung disease	ebi-a-GCST90018863	0.24	97.50%
		Pulmonary fibrosis	ebi-a-GCST90018908	0.27	96.70%
		Mean corpuscular haemoglobin	ebi-a-GCST004630	0.04	86.80%
rs9860874	TERC	Neoplasms	ukb-d-II_NEOPLASM	-0.01	81.40%
		Uterine fibroids	ebi-a-GCST90018934	-0.0718	93.60%
		Systolic blood pressure	ieu-b-38	-0.28	97.90%
		Telomere length	ieu-b-4879	-0.08	89.60%
rs9577395	ATP11A	Red cell distribution width	ebi-a-GCST90025988	0.03	100.00%
rs12912339	KNL1	Mean corpuscular volume	ebi-a-GCST90002334	0.02	81.60%
rs11073517	AKAP13	Corneal resistance factor (right)	ukb-b-15178	0.07	83.90%
		Corneal resistance factor (left)	ukb-b-4717	0.03	96.40%
rs12537430	MAD1L1	Plateletcrit	ebi-a-GCST90002400	-0.01	95.60%
		IGF-1	ebi-a-GCST90025989	-0.01	99.20%
rs2897075	ZKSCAN1	Mean corpuscular volume	ebi-a-GCST90002334	0.01	83.30%
		FEV1/FVC	ebi-a-GCST007431	0.02	99.40%
rs2609259	FAM13A	FVC	ebi-a-GCST007429	-0.02	97.20%
		FEV1/FVC	ebi-a-GCST007431	0.05	96.90%
		FEV1/FVC < 0.7	ebi-a-GCST007431	-0.01	96.90%
		Monocyte count	ieu-b-31	-0.01	97.30%
		White blood cell count	ieu-b-30	-0.02	96.90%
rs2304645	IVD	Total testosterone levels	ebi-a-GCST90012112	0.03	99.00%
		FVC	ebi-a-GCST007429	-0.02	99.10%
rs35705950	MUC5B	Interstitial lung disease	ebi-a-GCST90018863	1	100.00%
		Pulmonary fibrosis	ebi-a-GCST90018908	1.12	100.00%
rs2076295	DSP	FEV1/FVC < 0.7	ieu-b-106	-0.01	100.00%
		Interstitial lung disease	ebi-a-GCST90018863	0.15	100.00%
		FEV1/FVC	ebi-a-GCST007431	0.02	100.00%

Table 2.3.3: A summary table of phenotypes with strong evidence of colocalisation with IPF risk association signals. IEU GWAS ID: GWAS dataset code from the IEU OpenGWAS database. PP.H4: Posterior probability of H4 (colocalisation).

Most associations with gene expression variation that colocalise with IPF risk associations aligned with variant annotations, such as increased expression of *ATP11A* (*PP.H4* = 99.60%), *FAM13A* (*PP.H4* = 98.10%), and *DEPTOR* (*PP.H4* = 93.20%) (Table 2.3.4). Strong evidence of colocalisation was identified with IPF risk and increased expression of *RPAP1* (rs12912339; *KNL1*, *PP.H4* = 99.90%) and the RNA gene ENSG00000254343 (rs10808505; *DEPTOR*, *PP.H4* = 90.60%).

IPF SNP	Gene	Phenotype	IEU GWAS ID	Effect relative to IPF risk (beta)	PP.H4
rs9577395	ATP11A	ENSG00000068650 (<i>ATP11A</i>)	eqtl-a-ENSG00000068650	0.12	99.60%
rs12912339	KNL1	ENSG00000103932 (<i>RPAP1</i>)	eqtl-a-ENSG00000103932	0.12	99.90%
rs2609259	FAM13A	ENSG00000138640 (<i>FAM13A</i>)	eqtl-a-ENSG00000138640	0.48	98.10%
rs10808505	DEPTOR	ENSG00000155792 (<i>DEPTOR</i>)	eqtl-a-ENSG00000155792	0.25	93.20%
		ENSG00000254343	eqtl-a-ENSG00000254343	0.15	90.60%

Table 2.3.4: A summary table of eQTL association signals with strong evidence of colocalisation with IPF risk. IEU GWAS ID: GWAS dataset code from the IEU OpenGWAS database. PP.H4: Posterior probability of H4 (colocalisation).

2.3.5 Conclusions from colocalisation analyses

Colocalisation of GWAS signals associated with IPF risk, infection severity, multiple blood and disease phenotypes, and biomarkers highlight likely shared biology, possible systemic signatures of IPF disease, and potential risk factors. In addition, colocalised eQTL signals support current variant annotations and highlight potential trans-eQTL effects that may contribute to IPF pathogenesis.

2.3.6 Results of PheWAS using an independent cohort

Following retrieval and processing of phenotype and genotype data, a cohort of 38,264 GE 100,000 GP participants was established for this PheWAS study (Figure 2.3.25).

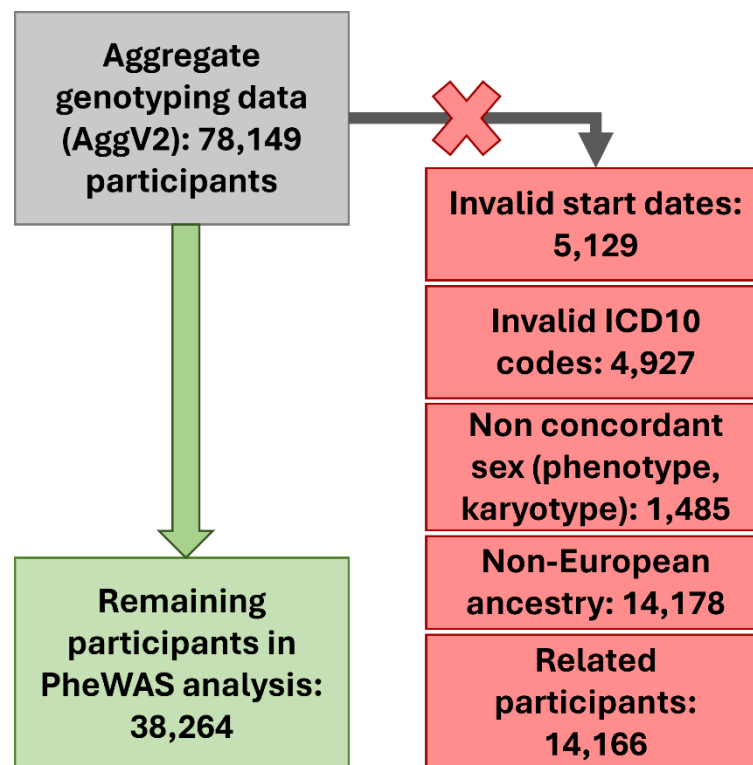


Figure 2.3.25: Data processing and participant exclusion criteria to establish an appropriate cohort for PheWAS analyses within the GE 100,000 GP dataset.

2.3.6.1 Phenotypes associated with mucin biology genetic variants

Only the *MUC5B* IPF risk variant (rs35705950) is significantly ($p < 1 \times 10^{-5}$) associated with any phenotypes in GE PheWAS analyses. Results suggest this SNP is associated with increased risk of postinflammatory pulmonary fibrosis ($\beta = 1.09$, $p = 8.30 \times 10^{-26}$) (Figure 2.3.38). Identified associations close to significance criteria (exploratory threshold: $p < 5 \times 10^{-3}$) include increased risk of pneumonia ($\beta = 0.15$, $p = 2.98 \times 10^{-3}$), other alveolar and parietoalveolar pneumonopathy ($\beta = 0.74$, $p = 1.90 \times 10^{-5}$), myopia ($\beta = 1.09$, $p = 1.22 \times 10^{-3}$), malignant neoplasm of gallbladder and extrahepatic bile duct ($\beta = 0.94$, $p = 5.93 \times 10^{-3}$).

⁴), and decreased risk of alteration of consciousness (somnolence) ($\beta = -1.80, p = 2.02 \times 10^{-3}$) (Figure 2.3.26).

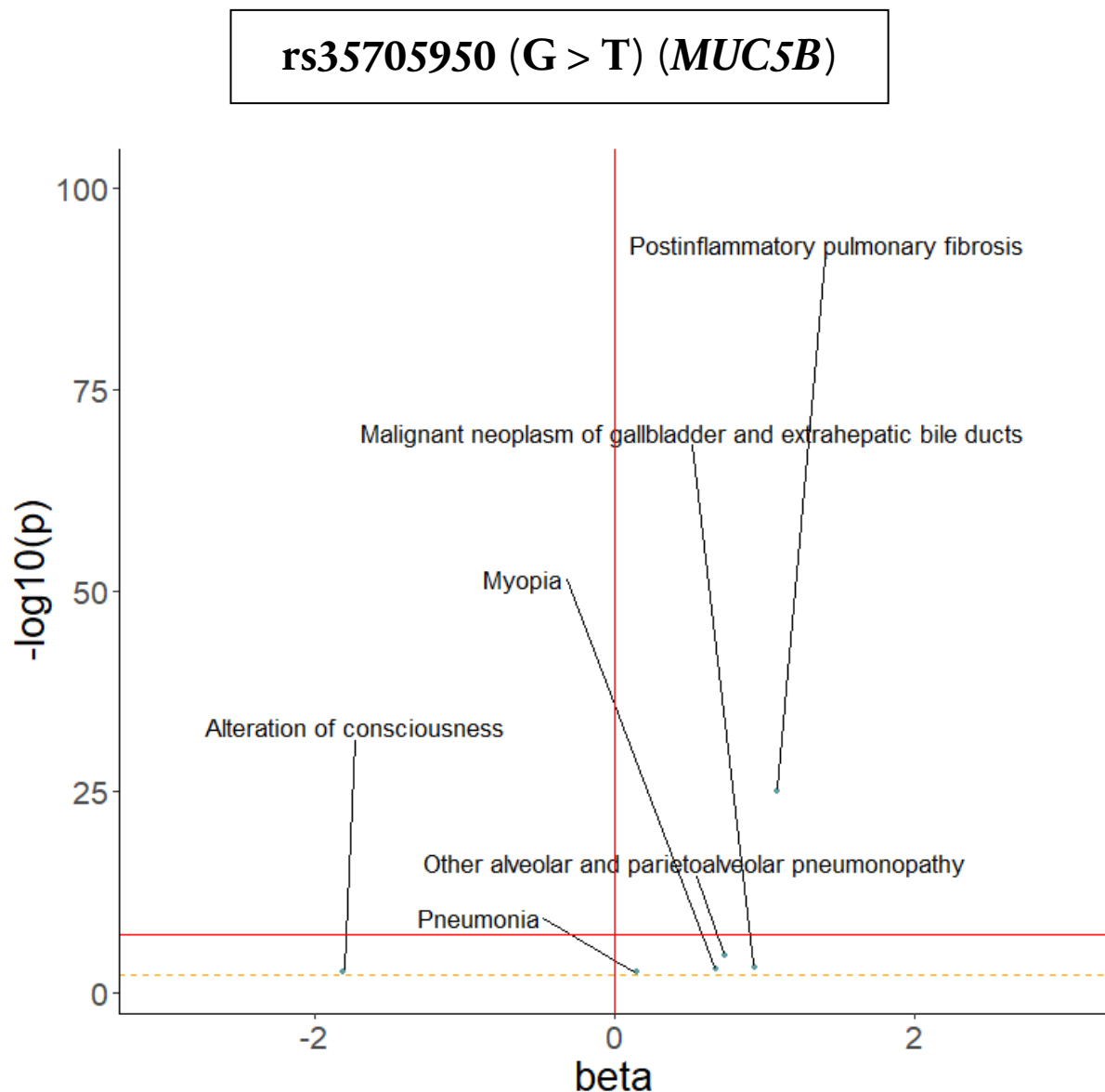


Figure 2.3.26: A volcano plot presenting phenotypes associated with rs35705950 (G > T) (*MUC5B*). The effect size (beta statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 1 \times 10^{-5}$ threshold applied in prioritising significant phenotypic associations. Phenotype associations were filtered at $p < 5 \times 10^{-3}$ (exploratory inclusion threshold), indicated by the orange dashed line.

2.3.6.2 Phenotypes associated with cytoskeletal protein genetic variants

The *KIF15* IPF genetic risk variant (rs78238620) is associated with increased risk of herpes zoster infection ($\beta = 0.93, p = 1.71 \times 10^{-3}$), personal history of diseases of blood or blood-forming organs ($\beta = 0.78, p = 1.67 \times 10^{-3}$), and increased risk of congenital musculoskeletal spine deformities ($\beta = 1.00, p = 2.61 \times 10^{-3}$) (Figure 2.3.27).

Results indicate that rs2076295 (*DSP*) is associated with decreased risk of skin cancer ($\beta = -0.16, p = 3.96 \times 10^{-3}$), left bundle branch block (heart) ($\beta = -0.29, p = 7.22 \times 10^{-4}$), and diabetes as complication of pregnancy ($\beta = -0.24, p = 4.21 \times 10^{-3}$) (Figure 2.3.28).

rs78238620 (T > A) (*KIF15*)

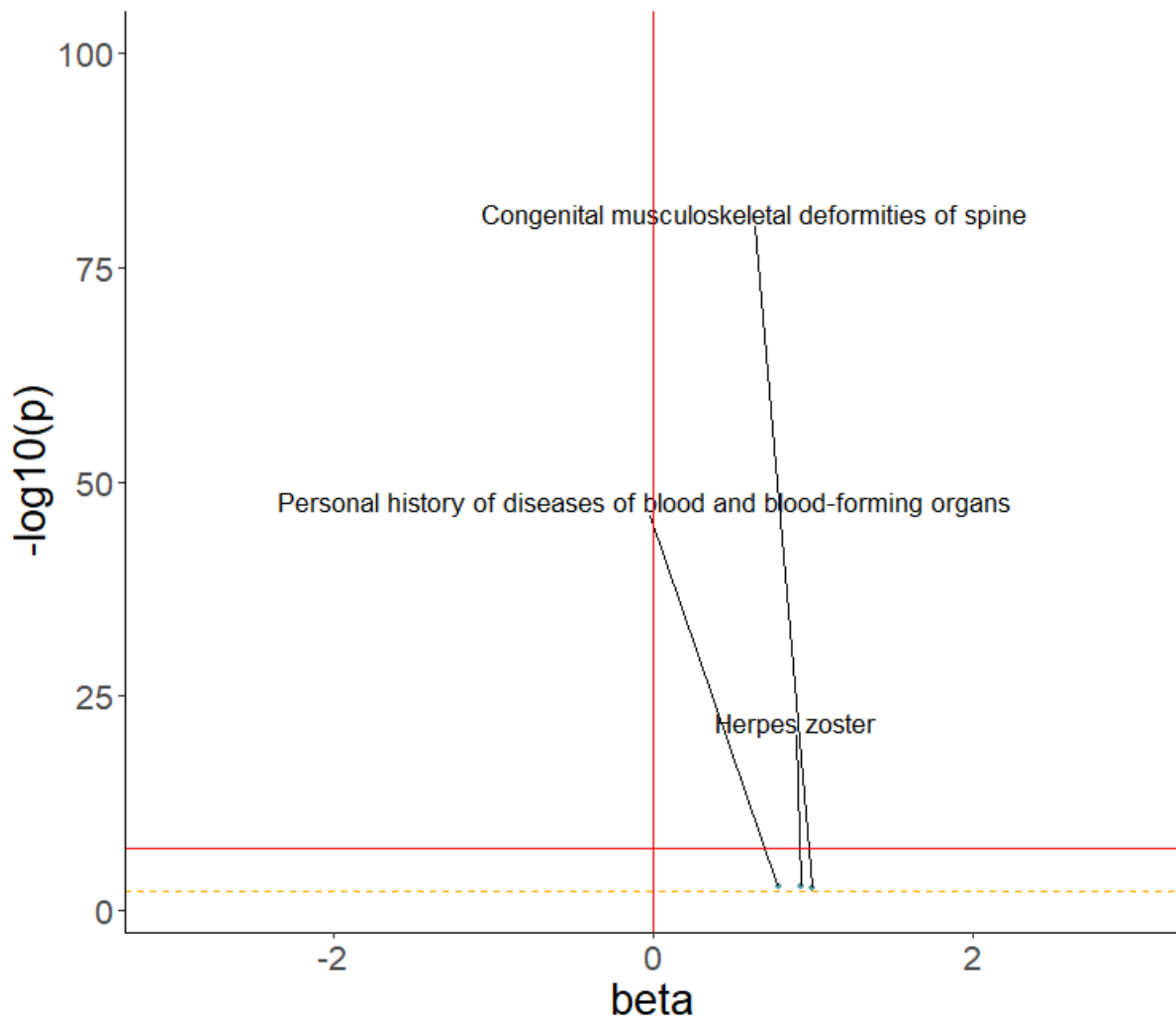


Figure 2.3.27: A volcano plot presenting phenotypes associated with rs78238620 (T > A) (*KIF15*). The effect size (beta statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 1 \times 10^{-5}$ threshold applied in prioritising significant phenotypic associations. Phenotype associations were filtered at $p < 5 \times 10^{-3}$ (exploratory inclusion threshold), indicated by the orange dashed line.

rs2076295 (T > G) (DSP)

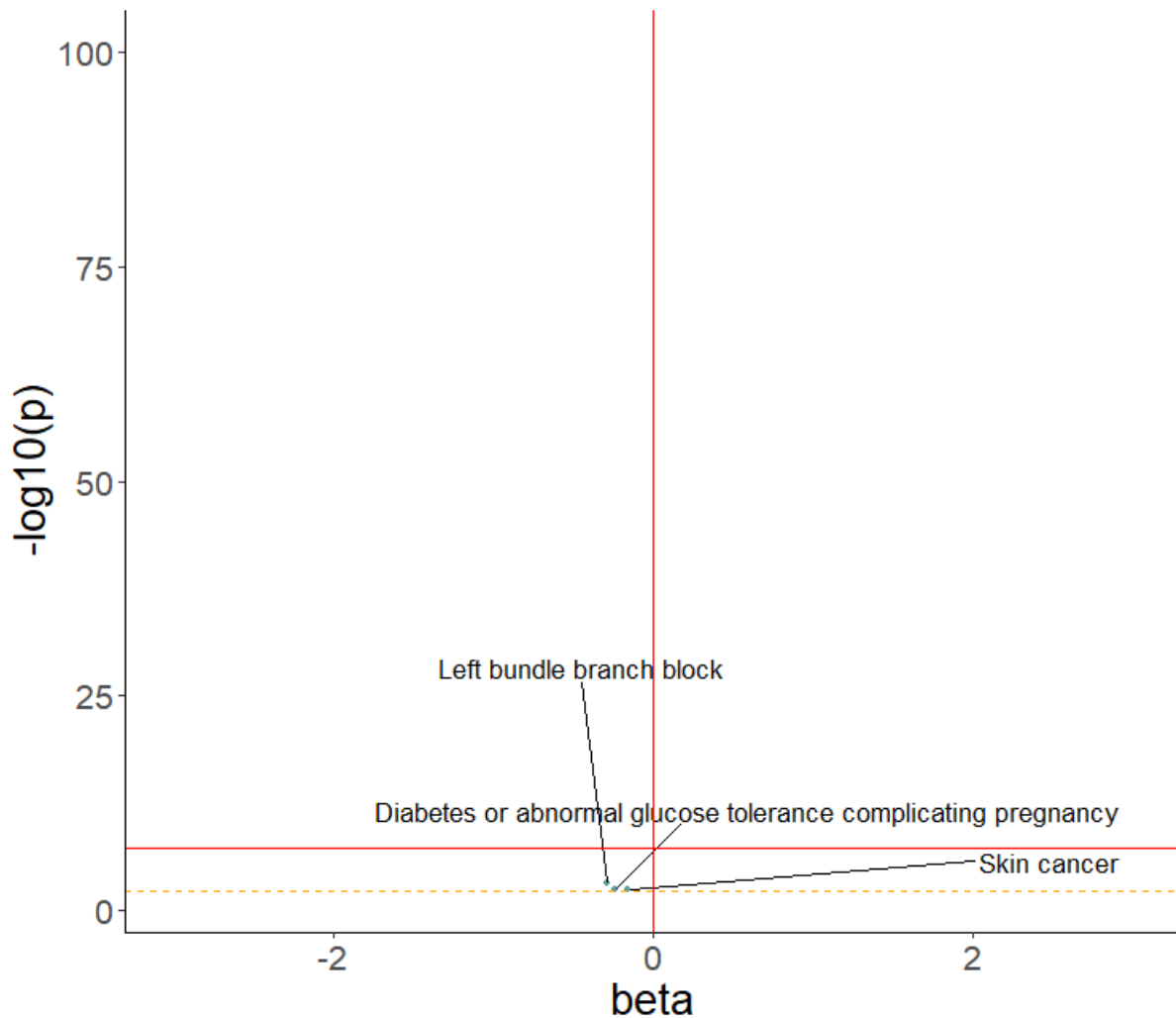


Figure 2.3.28: A volcano plot presenting phenotypes associated with rs2076295 (T > G) (DSP). The effect size (beta statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 1 \times 10^{-5}$ threshold applied in prioritising significant phenotypic associations. Phenotype associations were filtered at $p < 5 \times 10^{-3}$ (exploratory inclusion threshold), indicated by the orange dashed line.

2.3.6.3 Phenotypes associated with chromosome maintenance and cell cycle control genetic variants

No phenotype associations with rs12696304 (*TERC*) met significance criteria ($p < 1 \times 10^{-5}$) or were close to significance criteria (exploratory threshold: $p < 5 \times 10^{-3}$).

Results indicate that *RTEL1*-mapped SNP, rs41308092, is associated with increased risk of fluid overload ($\beta = 0.84, p = 4.27 \times 10^{-4}$), renal dialysis ($\beta = 0.59, p = 2.61 \times 10^{-3}$), glaucoma ($\beta = 0.57, p = 1.68 \times 10^{-3}$), and nonrheumatic tricuspid valve disorders ($\beta = 1.46, p = 1.90 \times 10^{-3}$). This genetic variant is also associated with disturbance of skin sensation ($\beta = 0.88, p = 1.66 \times 10^{-4}$), other disorders of cervical region ($\beta = 1.33, p = 4.45 \times 10^{-3}$), and other and unspecified back disorders ($\beta = 1.42, p = 2.27 \times 10^{-3}$).

Associations with rs7725218 (*TERT*) include decreased risk of cancer of tongue ($\beta = -0.72, p = 1.61 \times 10^{-4}$), cancer of brain and nervous system ($\beta = -0.26, p = 8.49 \times 10^{-5}$), and of secondary malignant neoplasms ($\beta = -0.08, p = 2.98 \times 10^{-3}$) (Figure 2.3.29). This SNP is also associated with decreased risk conduct disorders (behavioural) ($\beta = -0.62, p = 4.61 \times 10^{-3}$), disorders of tooth development ($\beta = -0.51, p = 5.96 \times 10^{-4}$), hydrocephalus ($\beta = -0.42, p = 8.83 \times 10^{-4}$), cerebral oedema and compression of brain ($\beta = -0.35, p = 1.10 \times 10^{-3}$), harmful effects of radiation ($\beta = -, p = -0.10, 2.71 \times 10^{-3}$), other disk (intervertebral) disorders ($\beta = -0.37, p = 2.46 \times 10^{-3}$), and increased risk of psoriasis vulgaris ($\beta = 0.25, p = 4.36 \times 10^{-3}$).

Results indicate that rs12699415 (*MAD1L1*) is associated with decreased risk of disorders of mineral metabolism ($\beta = -0.16, p = 2.20 \times 10^{-3}$) and increased risk of coagulation defects ($\beta = 0.29, p = 4.87 \times 10^{-3}$), stricture and stenosis of esophagus ($\beta = 0.38, p = 4.18 \times 10^{-3}$), and other disk (intervertebral) disorders ($\beta = 0.38, p = 1.66 \times 10^{-3}$) (Figure 2.3.30).

rs41308092 (G > A) (*RTEL1*)

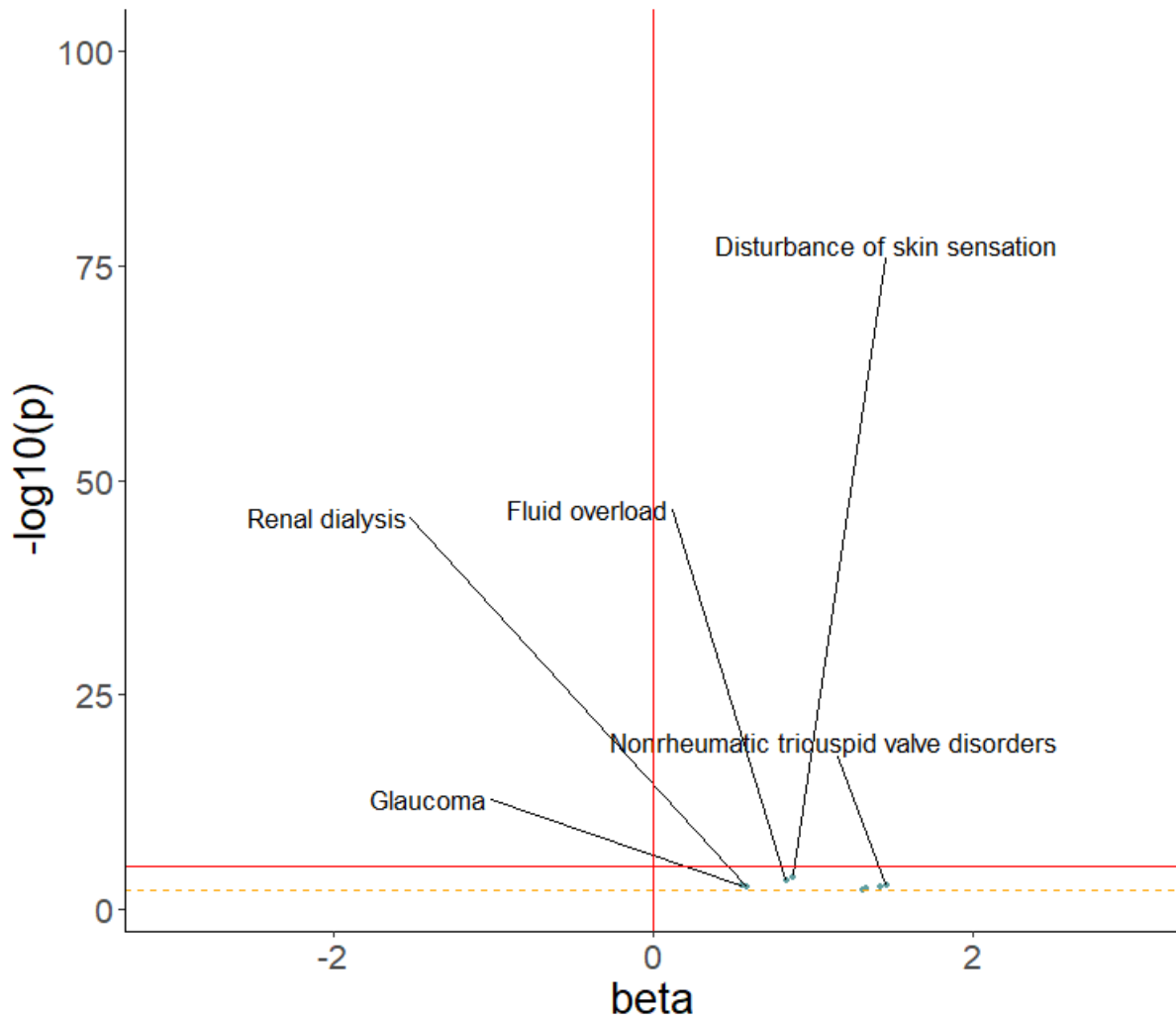


Figure 2.3.29: A volcano plot presenting phenotypes associated with rs41308092 (G > A) (*RTEL1*) The effect size (beta statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 1 \times 10^{-5}$ threshold applied in prioritising significant phenotypic associations. Phenotype associations were filtered at $p < 5 \times 10^{-3}$ (exploratory inclusion threshold), indicated by the orange dashed line.

rs7725218 (G > A) (*TERT*)

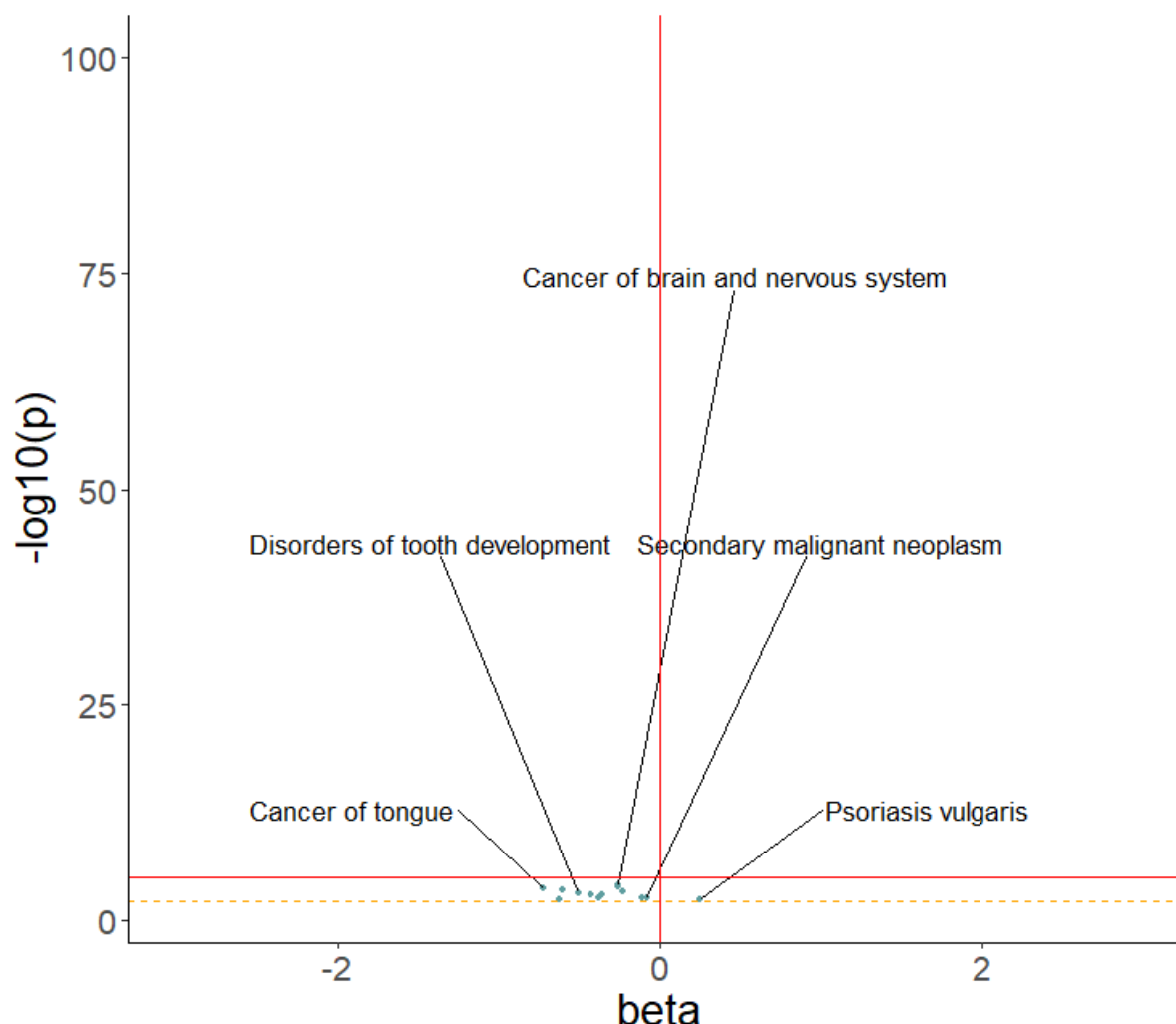


Figure 2.3.30: A volcano plot presenting phenotypes associated with rs7725218 (G > A) (*TERT*). The effect size (beta statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 1 \times 10^{-5}$ threshold applied in prioritising significant phenotypic associations. Phenotype associations were filtered at $p < 5 \times 10^{-3}$ (exploratory inclusion threshold), indicated by the orange dashed line.

rs12699415 (G > A) (*MAD1L1*)

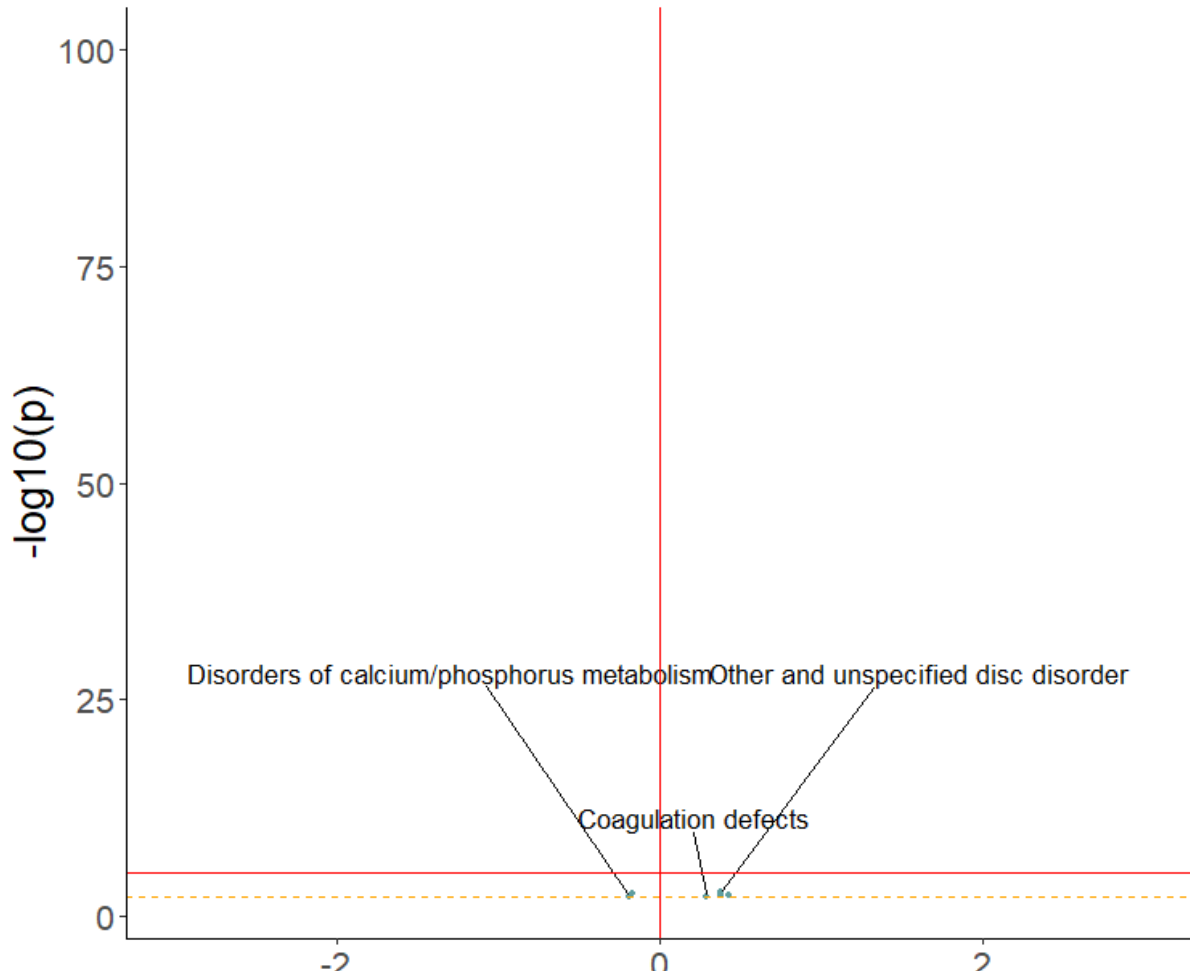


Figure 2.3.31: A volcano plot presenting phenotypes associated with rs12699415 (G > A) (*MAD1L1*). The effect size (beta statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 1 \times 10^{-5}$ threshold applied in prioritising significant phenotypic associations. Phenotype associations were filtered at $p < 5 \times 10^{-3}$ (exploratory inclusion threshold), indicated by the orange dashed line.

2.3.6.4 Phenotypes associated with cellular transport, signaling, and metabolism genetic variants

The rs28513081 genetic risk variant (*DEPTOR*) is associated with increased risk of spinal stenosis ($\beta = 0.55, p = 4.52 \times 10^{-3}$) and postmenopausal bleeding ($\beta = 0.35, p = 4.34 \times 10^{-3}$), and decreased risk of mitral valve disease ($\beta = -0.19, p = 3.90 \times 10^{-3}$) (Figure 2.3.31).

Associations with the rs9577395 IPF genetic risk variant (*ATP11A*) include increased risk of congestive heart failure ($\beta = 0.26, p = 3.77 \times 10^{-3}$), hemoptysis ($\beta = 0.51, p = 4.67 \times 10^{-3}$), and decreased risk of calculus of bile duct ($\beta = -0.42, p = 9.63 \times 10^{-4}$), and other complications of the puerperium ($\beta = -0.66, p = 9.33 \times 10^{-4}$) (Figure 2.3.32).

The rs59424629 genetic risk variant (*IVD*) is associated with increased risk of other disorders of liver ($\beta = 0.21, p = 2.81 \times 10^{-3}$) and secondary malignancy of bone ($\beta = 0.15, p = 3.31 \times 10^{-3}$). This SNP is also associated with decreased risk of polycystic ovaries ($\beta = -0.25, p = 4.06 \times 10^{-3}$), endometriosis ($\beta = -0.30, p = 1.96 \times 10^{-3}$), phobia ($\beta = -0.43, p = 3.17 \times 10^{-4}$), adrenal hyperfunction ($\beta = -0.72, p = 7.94 \times 10^{-4}$), disorders of penis ($\beta = -0.78, p = 4.61 \times 10^{-4}$), diabetes insipidus ($\beta = -0.81, p = 3.55 \times 10^{-3}$), and sinoatrial node dysfunction (bradycardia) ($\beta = -0.90, p = 1.32 \times 10^{-4}$) (Figure 2.3.33).

Results indicate that the rs62023891 IPF risk variant (*AKAP13*) is associated with increased risk of chronic thyroiditis ($\beta = 0.95, p = 4.23 \times 10^{-4}$), disorders of sweat glands ($\beta = 0.78, p = 4.94 \times 10^{-3}$), other congenital musculoskeletal ($\beta = 0.31, p = 4.63 \times 10^{-5}$), and complication of colostomy or enterostomy ($\beta = 0.38, p = 3.15 \times 10^{-3}$) (Figure 2.3.34).

The rs12610495 IPF genetic risk variant (*DPP9*) is associated with increased risk of alcohol-related disorders ($\beta = 0.15, p = 2.82 \times 10^{-3}$), complication of colostomy or enterostomy ($\beta = 0.38, p = 2.62 \times 10^{-3}$), and iatrogenic hypotension ($\beta = 1.02, p = 1.48 \times 10^{-3}$) (Figure 2.3.35).

The rs2897075 IPF risk variant (*ZKSCAN1/7q22.1*) is associated with increased risk of primary thrombocytopenia ($\beta = 0.40, p = 2.25 \times 10^{-3}$), malignant neoplasm of other and

ill-defined sites within the digestive organs and peritoneum ($\beta = 0.11, p = 8.82 \times 10^{-4}$). This SNP is also associated with decreased risk of fever of unknown origin ($\beta = -0.16, p = 2.92 \times 10^{-3}$), strabismus ($\beta = -0.51, p = 5.56 \times 10^{-4}$), and other disorders of neurohypophysis ($\beta = -0.91, p = 7.98 \times 10^{-4}$) (Figure 2.3.36).

rs28513081 (A > G) (DEPTOR)

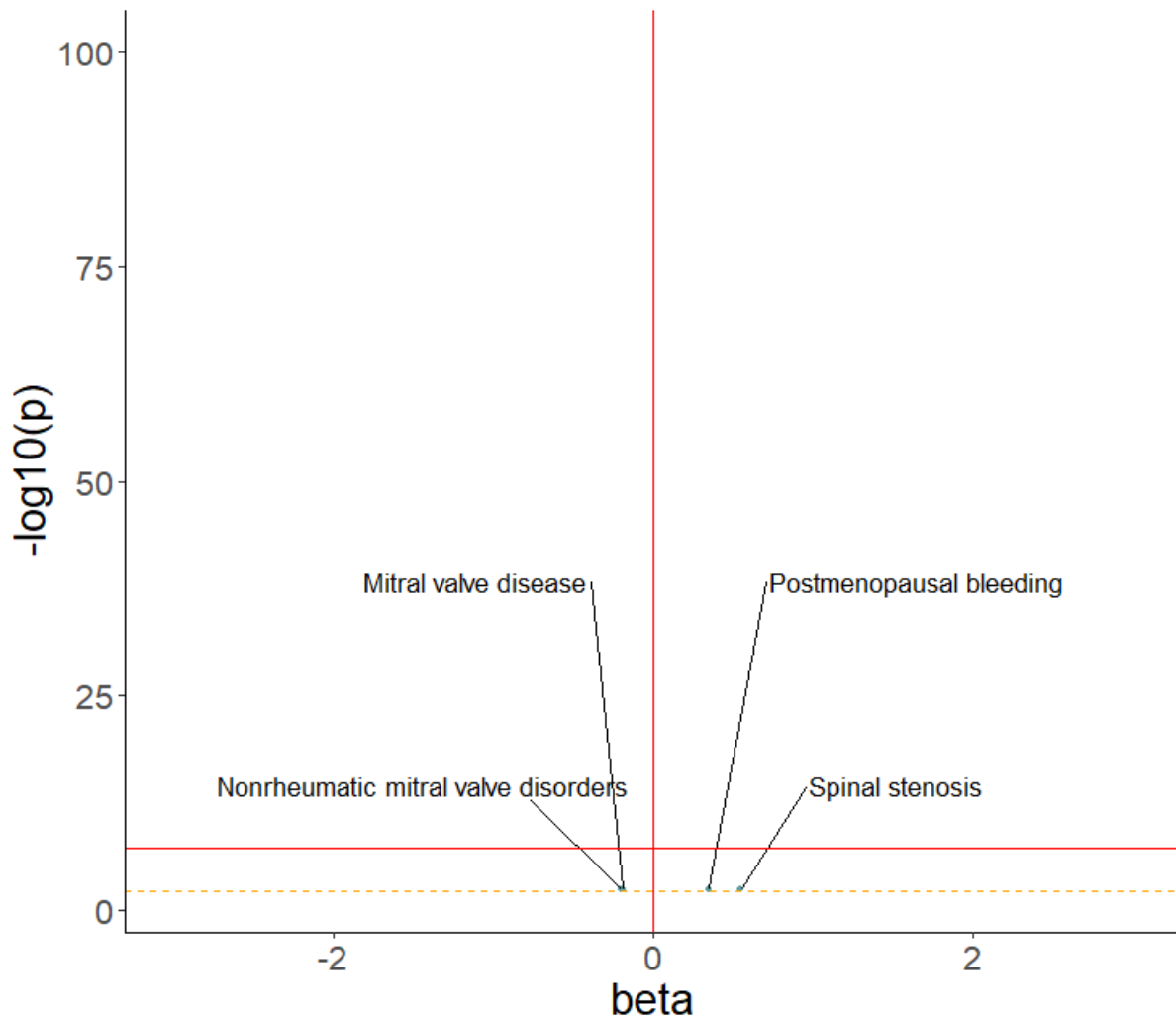


Figure 2.3.32: A volcano plot presenting phenotypes associated with rs28513081 (A > G) (DEPTOR). The effect size (beta statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 1 \times 10^{-5}$ threshold applied in prioritising significant phenotypic associations. Phenotype associations were filtered at $p < 5 \times 10^{-3}$ (exploratory inclusion threshold), indicated by the orange dashed line.

rs9577395 (C > G) (ATP11A)

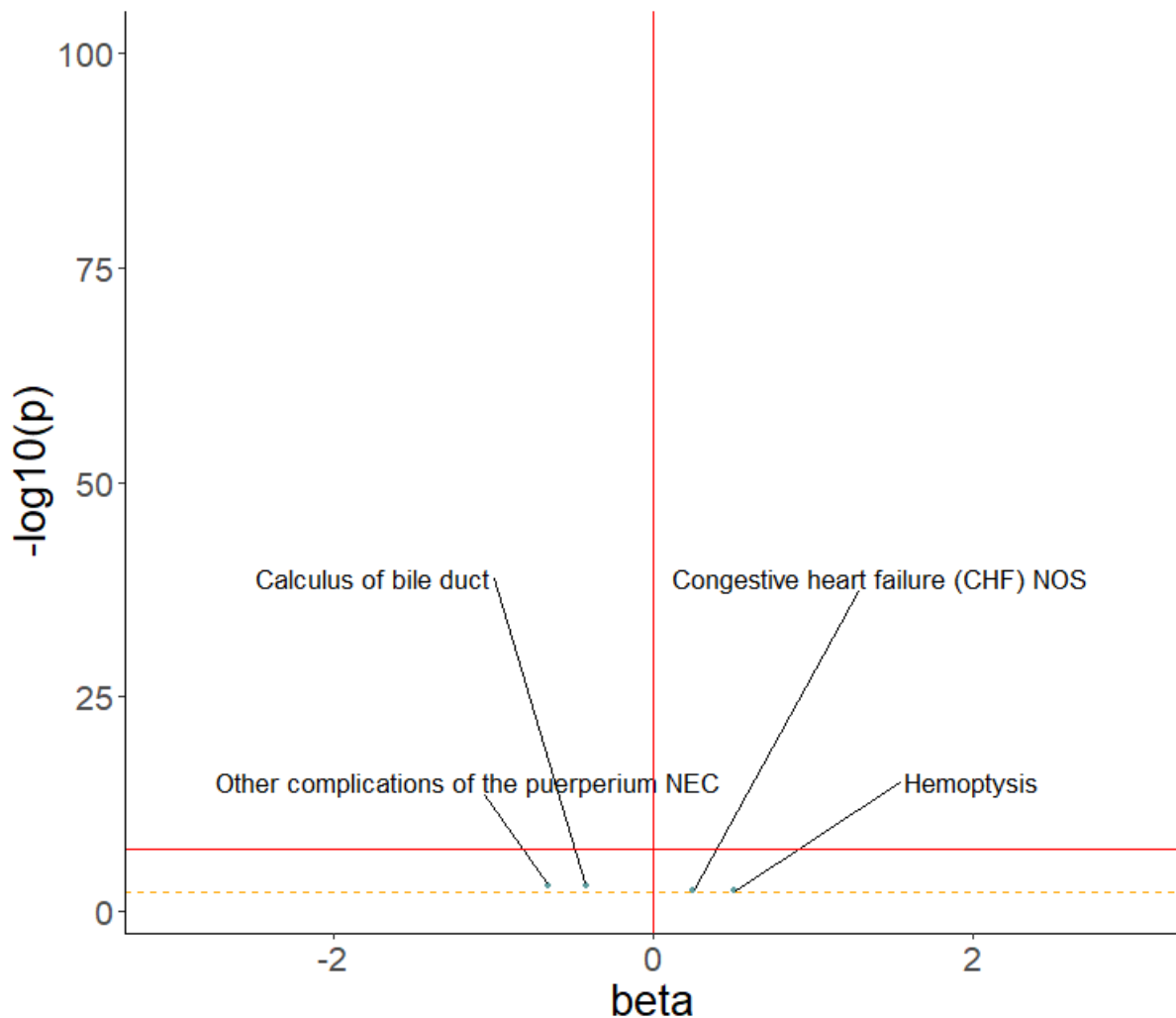


Figure 2.3.33: A volcano plot presenting phenotypes associated with rs9577395 (C > G) (ATP11A). The effect size (beta statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 1 \times 10^{-5}$ threshold applied in prioritising significant phenotypic associations. Phenotype associations were filtered at $p < 5 \times 10^{-3}$ (exploratory inclusion threshold), indicated by the orange dashed line.

rs59424629 (T > G) (IVD)

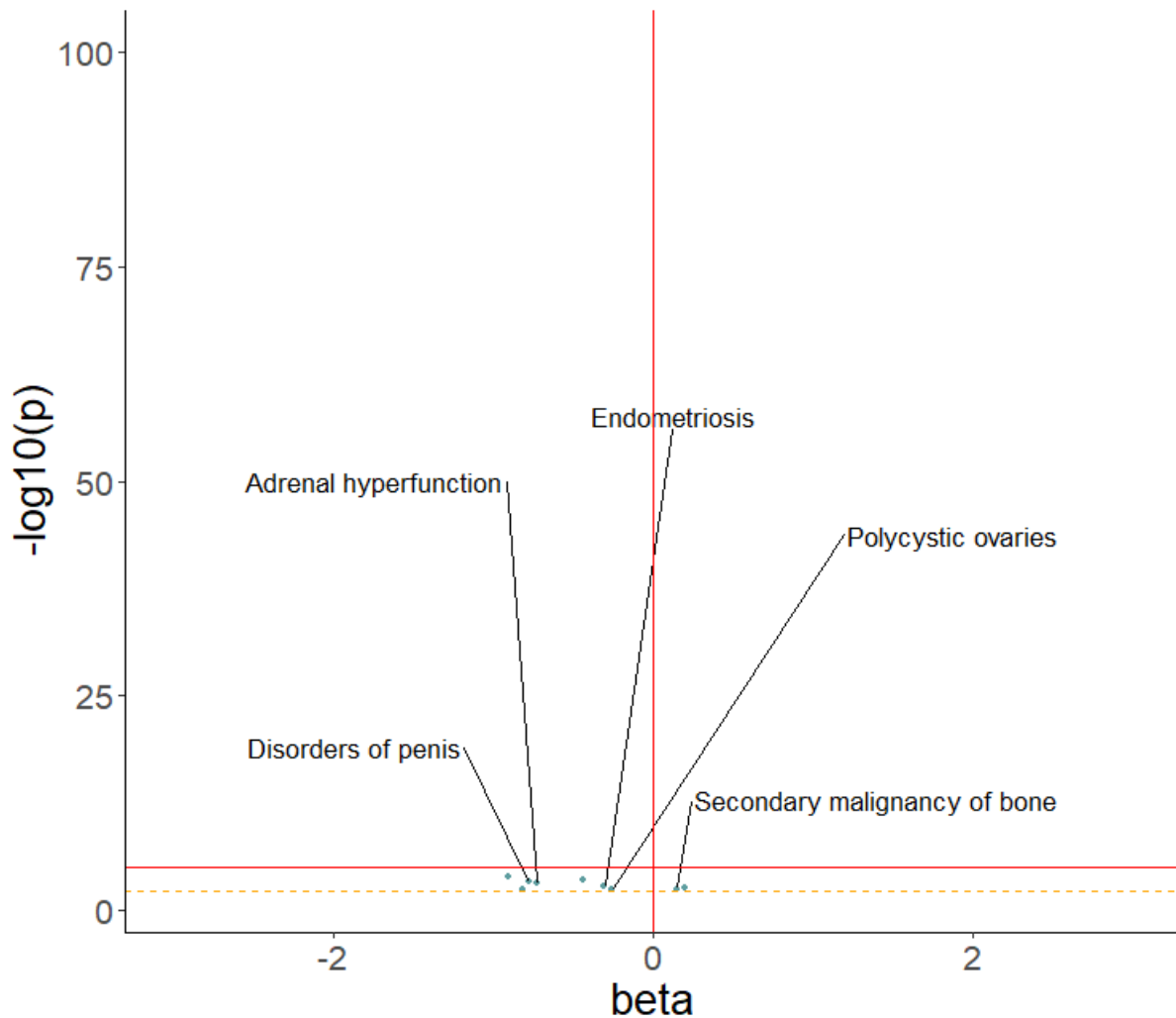


Figure 2.3.34: A volcano plot presenting phenotypes associated with rs59424629 (T > G) (IVD). The effect size (beta statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 1 \times 10^{-5}$ threshold applied in prioritising significant phenotypic associations. Phenotype associations were filtered at $p < 5 \times 10^{-3}$ (exploratory inclusion threshold), indicated by the orange dashed line.

rs62023891 (G > A) (AKAP13)

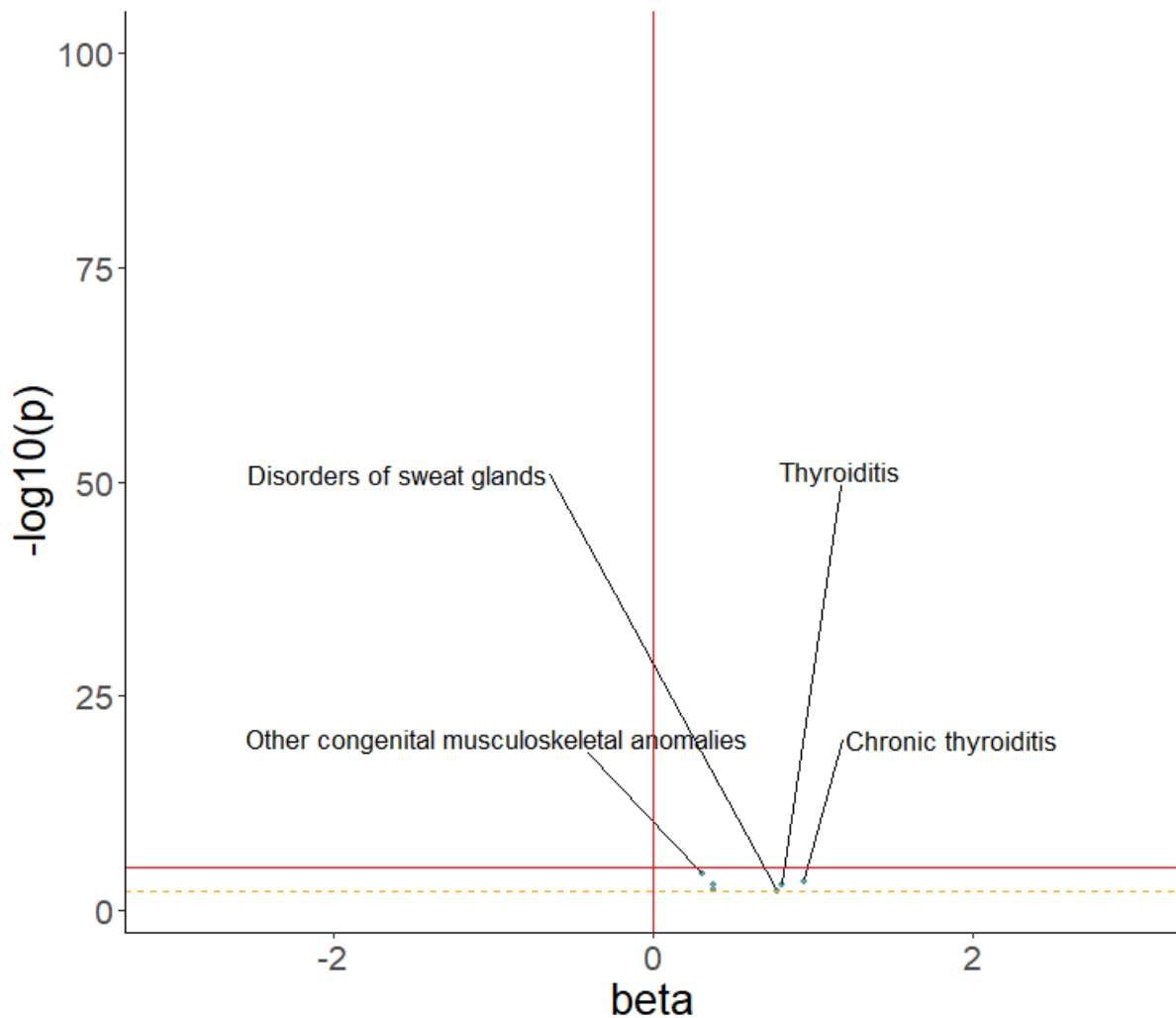


Figure 2.3.35: A volcano plot presenting phenotypes associated with rs62023891 (G > A) (AKAP13). The effect size (beta statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 1 \times 10^{-5}$ threshold applied in prioritising significant phenotypic associations. Phenotype associations were filtered at $p < 5 \times 10^{-3}$ (exploratory inclusion threshold), indicated by the orange dashed line.

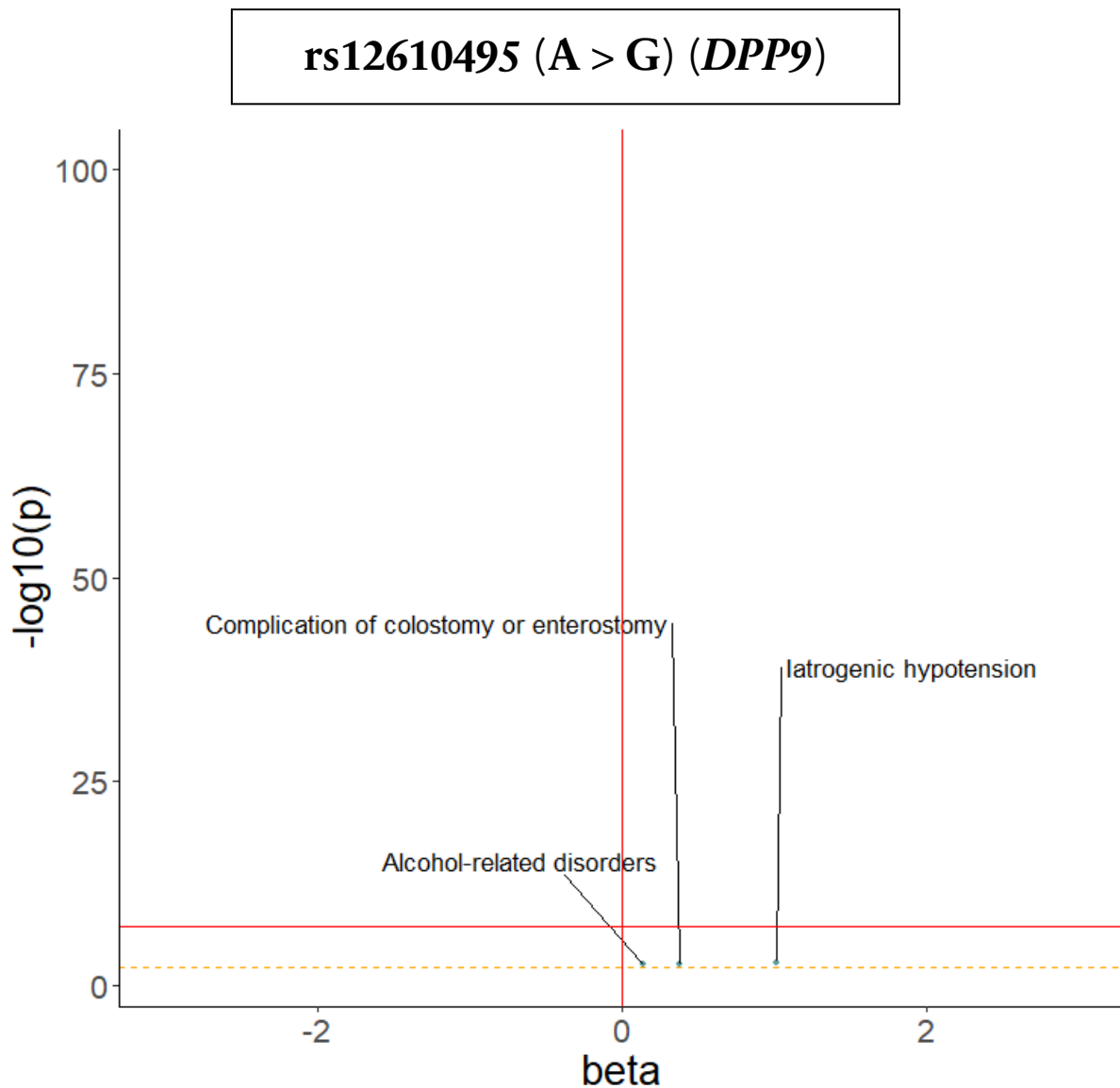


Figure 2.3.36: A volcano plot presenting phenotypes associated with rs12610495 (A > G) (DPP9). The effect size (beta statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 1 \times 10^{-5}$ threshold applied in prioritising significant phenotypic associations. Phenotype associations were filtered at $p < 5 \times 10^{-3}$ (exploratory inclusion threshold), indicated by the orange dashed line.

rs2897075 (C > T) (ZKSCAN1/7q22.1)

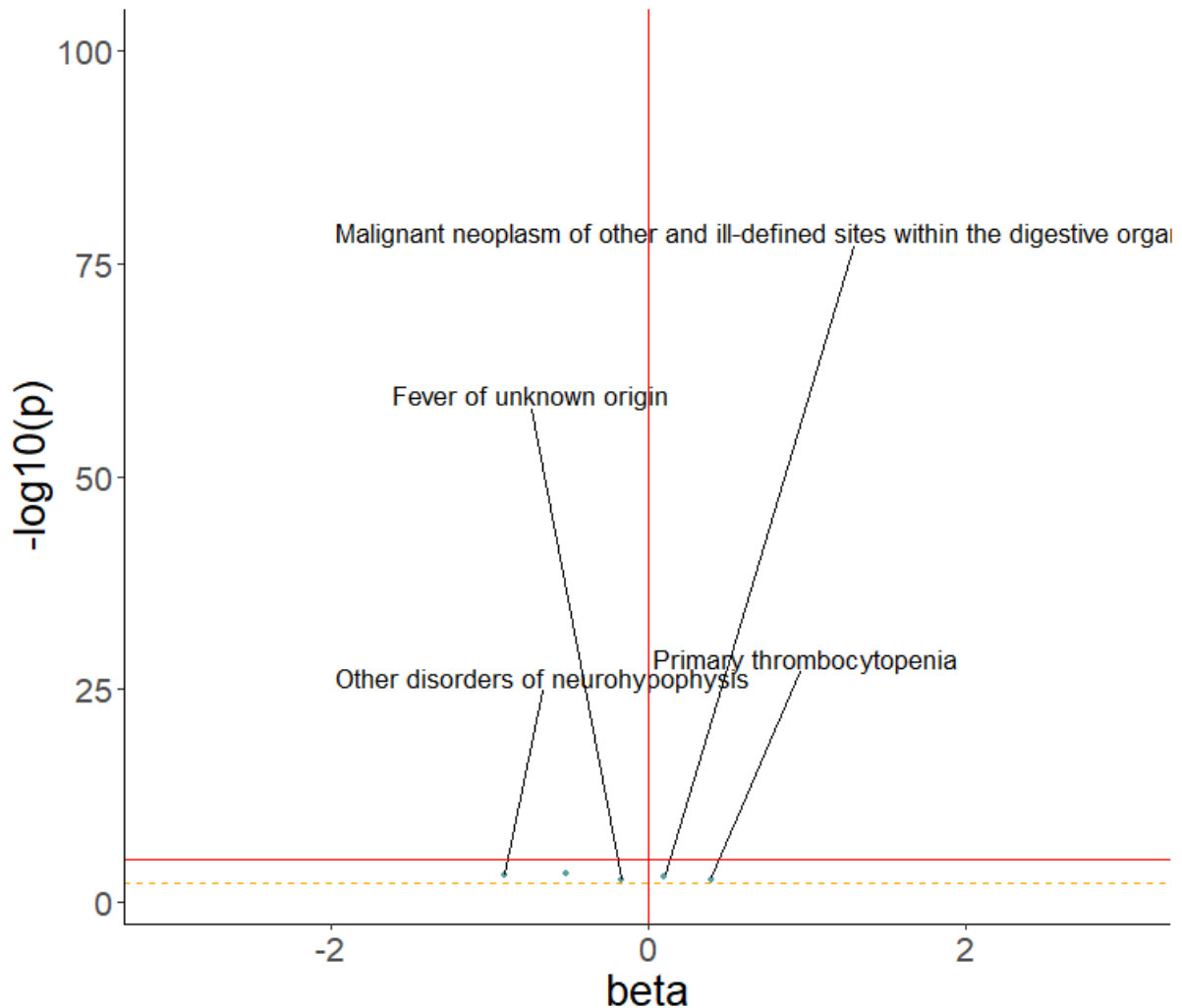


Figure 2.3.37: A volcano plot presenting phenotypes associated with rs2897075 (C > T) (ZKSCAN1/7q22.1). The effect size (beta statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 1 \times 10^{-5}$ threshold applied in prioritising significant phenotypic associations. Phenotype associations were filtered at $p < 5 \times 10^{-3}$ (exploratory inclusion threshold), indicated by the orange dashed line.

2.3.6.5 Phenotypes associated with genetic variants of other or unclassified function

Associations with the rs2013701 IPF genetic risk variant (*FAM13A*) include increased risk of adrenal hyperfunction ($\beta = 0.66, p = 2.38 \times 10^{-3}$) and Cushing's syndrome ($\beta = 0.76, p = 4.01 \times 10^{-3}$), and decreased risk of congenital anomalies of male genital organs ($\beta = -0.49, p = 2.70 \times 10^{-3}$) (Figure 2.3.37).

Results indicated that the *HECTD2*-mapped SNP, rs537322302, is associated with a range of diseases including chronic fatigue syndrome ($\beta = 1.66, p = 3.33 \times 10^{-4}$), celiac disease ($\beta = 1.42, p = 7.60 \times 10^{-4}$), and respiratory conditions of the fetus and newborn ($\beta = 2.81, p = 2.44 \times 10^{-3}$). This genetic variant is also putatively associated with polyneuropathy in diabetes ($\beta = 2.09, p = 4.51 \times 10^{-3}$), diseases of the larynx and vocal chords ($\beta = 1.30, p = 4.72 \times 10^{-3}$), disorders of the diaphragm ($\beta = 2.15, p = 3.75 \times 10^{-3}$), polyp of female genital organs ($\beta = 1.56, p = 2.61 \times 10^{-3}$), displacement of intervertebral disc ($\beta = 2.26, p = 2.27 \times 10^{-3}$), and other acquired deformities of limbs ($\beta = 2.69, p = 3.46 \times 10^{-4}$).

In this analysis, the rs2077551 IPF risk variant (*MAPT*) was found to be associated with decreased risk of pancreatic cancer ($\beta = -0.46, p = 2.19 \times 10^{-3}$), and increased risk of other demyelinating diseases of central nervous system ($\beta = 1.18, p = 4.67 \times 10^{-3}$) and postinflammatory pulmonary fibrosis ($\beta = 0.44, p = 1.21 \times 10^{-3}$).

A summary of all associations with the IPF risk variants investigated in this analysis is presented in Table 2.3.5.

rs2013701 (G > T) (*FAM13A*)

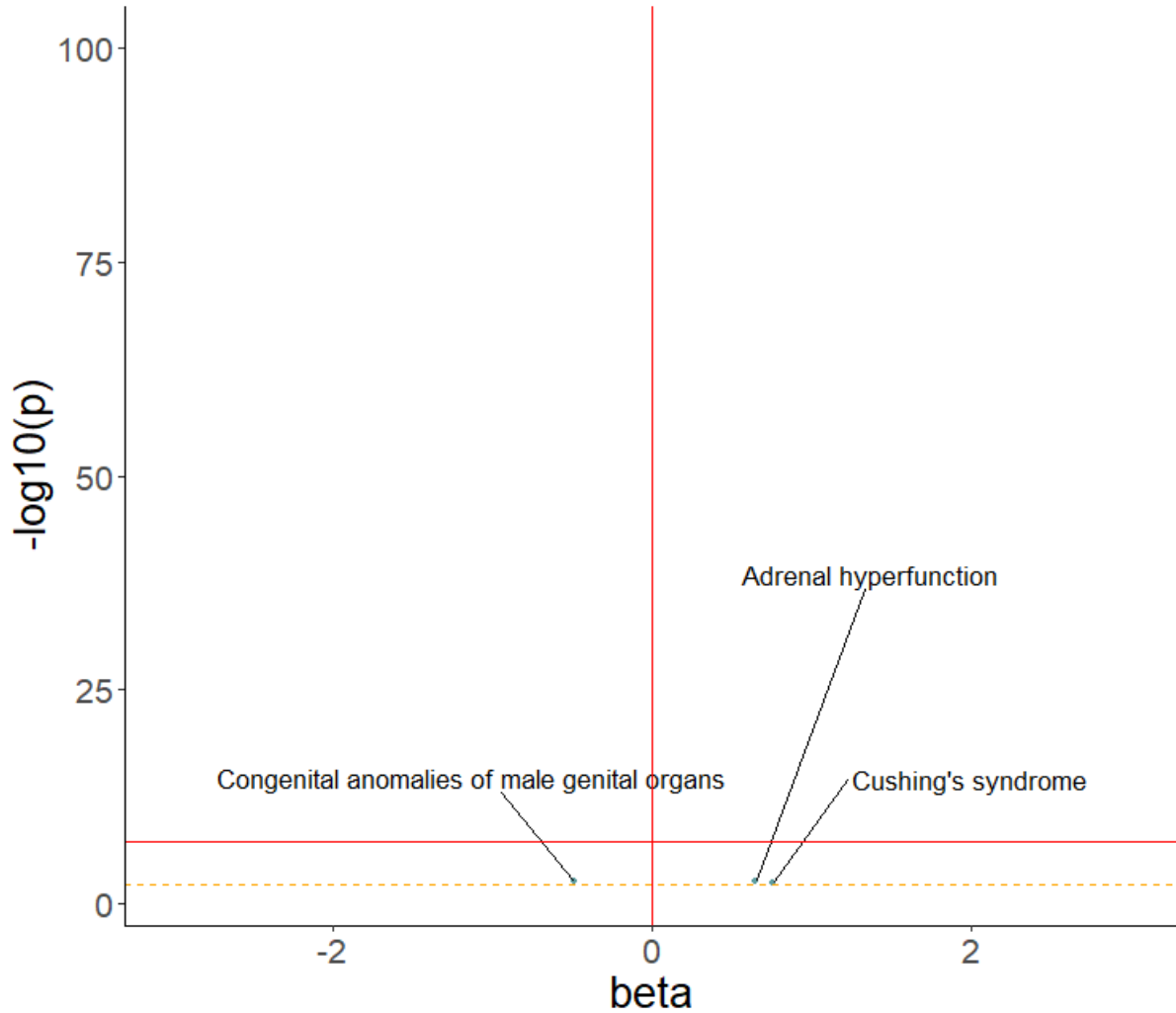


Figure 2.3.38: A volcano plot presenting phenotypes associated with rs2013701 (G > T) (*FAM13A*). The effect size (beta statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 1 \times 10^{-5}$ threshold applied in prioritising significant phenotypic associations. Phenotype associations were filtered at $p < 5 \times 10^{-3}$ (exploratory inclusion threshold), indicated by the orange dashed line.

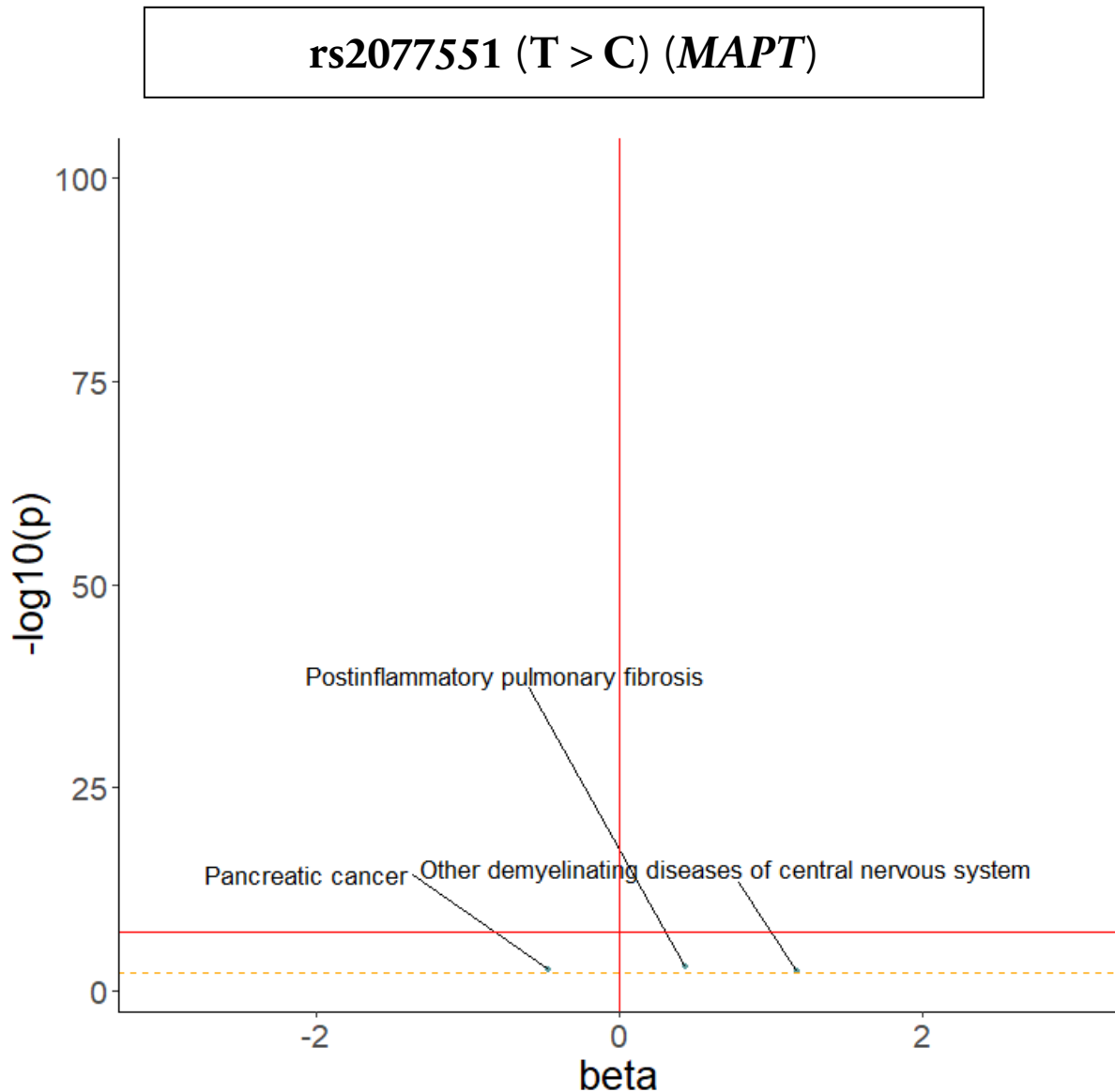


Figure 2.3.39: A volcano plot presenting phenotypes associated with rs2077551 (T > C) (MAPT). The effect size (beta statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 1 \times 10^{-5}$ threshold applied in prioritising significant phenotypic associations. Phenotype associations were filtered at $p < 5 \times 10^{-3}$ (exploratory inclusion threshold), indicated by the orange dashed line.

rs537322302 (C > G) (*HECTD2*)

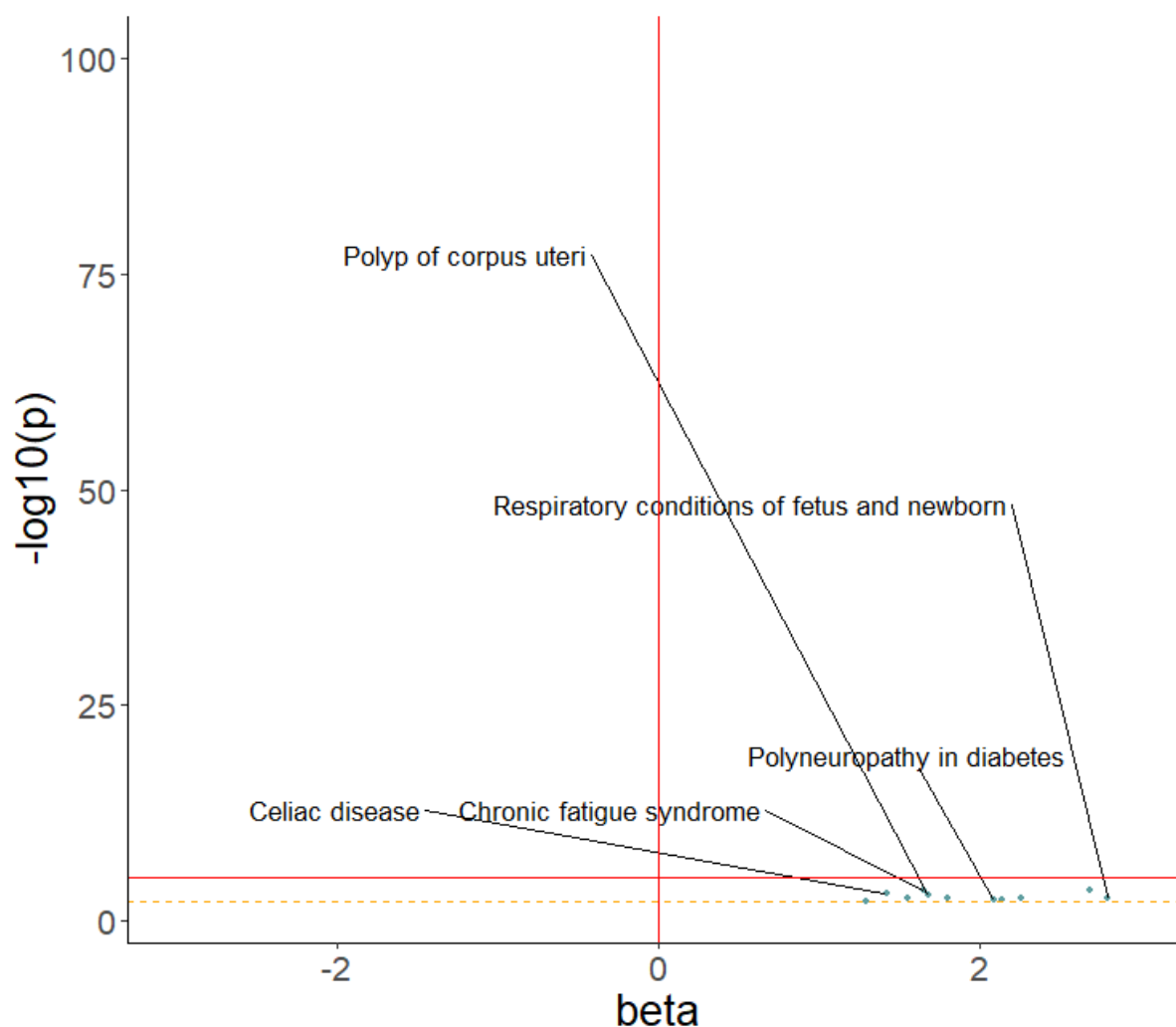


Figure 2.3.40 A volcano plot presenting phenotypes associated with *rs537322302* (C > G) (*HECTD2*). The effect size (beta statistic) of each data point (phenotype) is presented on the x-axis. Corresponding statistical significance ($-\log_{10}(p)$) is shown on the y-axis. Effect directions presented are relative to a positive effect of the investigated allele on IPF risk. The vertical red line shown indicates $\beta = 0$ (null effect), the horizontal red line shown indicates the $p = 1 \times 10^{-5}$ threshold applied in prioritising significant phenotypic associations. Phenotype associations were filtered at $p < 5 \times 10^{-3}$ (exploratory inclusion threshold), indicated by the orange dashed line.

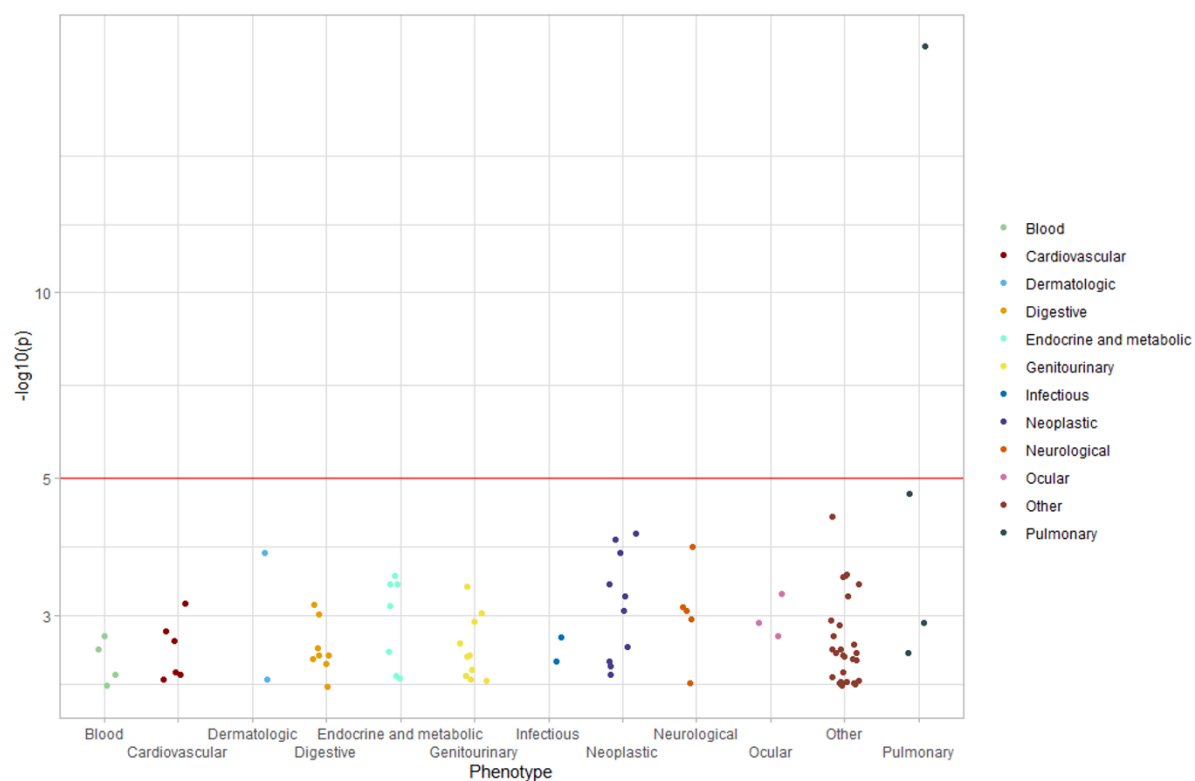


Figure 2.3.41: Phenotypes associated with 14 IPF genetic risk variants in the Genomics England 100,000 Genomes Project cohort. Phenotype categories are presented across the x-axis and $-\log_{10}(p\text{-value})$ of SNP-phenotype associations are presented on the y-axis. Phenotype categories are ranked in alphabetical order. The red horizontal line indicates the $p = 1 \times 10^{-5}$ threshold applied in this analysis to prioritise significant associations.

SNP	Gene	Phenotype	Category	No. cases	No. controls	Phecode	beta	p-value
rs35705950	MUC5B	Malignant neoplasm of gallbladder and extrahepatic bile ducts	Neoplastic	36	33413	159.3	0.94	5.93E-04
rs35705950	MUC5B	Alteration of consciousness	Other	72	35357	291.8	-1.80	2.02E-03
rs35705950	MUC5B	Myopia	Ocular	73	37459	367.1	0.68	1.22E-03
rs35705950	MUC5B	Pneumonia	Infectious	1971	33937	480	0.15	2.98E-03
rs35705950	MUC5B	Postinflammatory pulmonary fibrosis	Pulmonary	257	34950	502	1.09	8.30E-26
rs35705950	MUC5B	Other alveolar and parietoalveolar pneumonopathy	Pulmonary	104	34950	504	0.74	1.90E-05
rs2897075	ZKSCAN1	Malignant neoplasm of other and ill-defined sites within the digestive organs and peritoneum	Neoplastic	2350	33414	159	0.11	8.82E-04
rs2897075	ZKSCAN1	Other disorders of neurohypophysis	Neurological	44	36098	253.7	-0.91	7.98E-04
rs2897075	ZKSCAN1	Primary thrombocytopenia	Blood	119	36918	287.31	0.40	2.25E-03
rs2897075	ZKSCAN1	Strabismus (not specified as paralytic)	Ocular	119	36542	378.1	-0.51	5.55E-04
rs2897075	ZKSCAN1	Fever of unknown origin	Other	773	36007	783	-0.16	2.92E-03
rs62023891	AKAP13	Thyroiditis	Other	32	35202	245	0.81	1.28E-03
rs62023891	AKAP13	Chronic thyroiditis	Endocrine and metabolic	28	35202	245.2	0.95	4.23E-04
rs62023891	AKAP13	Chronic lymphocytic thyroiditis	Endocrine and metabolic	28	35202	245.21	0.95	4.23E-04
rs62023891	AKAP13	Disorders of sweat glands	Other	26	37164	705	0.78	4.94E-03
rs62023891	AKAP13	Other congenital musculoskeletal anomalies	Other	388	37305	756	0.31	4.63E-05
rs62023891	AKAP13	Congenital anomalies of muscle, tendon, fascia, and connective tissue	Other	154	37305	756.3	0.38	1.14E-03
rs62023891	AKAP13	Complication of colostomy or enterostomy	Digestive	130	33762	853	0.38	3.15E-03
rs12610495	DPP9	Alcohol-related disorders	Other	954	30062	317	0.15	2.83E-03
rs12610495	DPP9	Iatrogenic hypotension	Cardiovascular	20	31955	458.2	1.02	1.48E-03
rs12610495	DPP9	Complication of colostomy or enterostomy	Digestive	130	33763	853	0.38	2.62E-03
rs2076295	DSP	Skin cancer	Neoplastic	738	36483	172	-0.16	3.96E-03
rs2076295	DSP	Left bundle branch block	Cardiovascular	302	33074	426.32	-0.29	7.22E-04
rs2076295	DSP	Diabetes or abnormal glucose tolerance complicating pregnancy	Endocrine and metabolic	314	21023	649.1	-0.24	4.22E-03
rs537322302	HECTD2	Polyneuropathy in diabetes	Other	35	34696	250.6	2.09	4.51E-03
rs537322302	HECTD2	Diseases of the larynx and vocal cords	Other	173	33942	473	1.30	4.72E-03
rs537322302	HECTD2	Paralysis/spasm of vocal cords or larynx	Other	68	33942	473.3	1.81	2.47E-03
rs537322302	HECTD2	Disorders of diaphragm	Other	31	38034	513.8	2.15	3.75E-03
rs537322302	HECTD2	Celiac disease	Digestive	177	27095	557.1	1.42	7.60E-04
rs537322302	HECTD2	Polyp of female genital organs	Genitourinary	130	20979	622	1.56	2.61E-03
rs537322302	HECTD2	Polyp of corpus uteri	Genitourinary	119	20979	622.1	1.69	1.19E-03
rs537322302	HECTD2	Respiratory conditions of fetus and newborn	Pulmonary	58	37418	656.2	2.81	2.44E-03
rs537322302	HECTD2	Displacement of intervertebral disc	Other	33	36685	722.1	2.26	2.27E-03
rs537322302	HECTD2	Other acquired deformities of limbs	Other	21	37123	736	2.69	3.46E-04
rs537322302	HECTD2	Malaise and fatigue	Endocrine and metabolic	127	38055	798	1.66	3.33E-04
rs537322302	HECTD2	Chronic fatigue syndrome	Endocrine and metabolic	127	38055	798.1	1.66	3.33E-04

rs59424629	IVD	Secondary malignancy of bone	Neoplastic	758	27864	198.6	0.15	3.31E-03
rs59424629	IVD	Diabetes insipidus	Genitourinary	29	36098	253.3	-0.81	3.55E-03
rs59424629	IVD	Adrenal hyperfunction	Endocrine and metabolic	47	36098	255.1	-0.72	7.94E-04
rs59424629	IVD	Polycystic ovaries	Genitourinary	268	20684	256.4	-0.25	4.06E-03
rs59424629	IVD	Phobia	Other	142	25968	300.13	-0.43	3.17E-04
rs59424629	IVD	Sinoatrial node dysfunction (Bradycardia)	Neurological	42	33074	427.8	-0.90	1.32E-04
rs59424629	IVD	Other disorders of liver	Digestive	435	35155	573	0.21	2.81E-03
rs59424629	IVD	Disorders of penis	Genitourinary	46	14858	604	-0.78	4.61E-04
rs59424629	IVD	Endometriosis	Genitourinary	217	20255	615	-0.30	1.96E-03
rs78238620	KIF15	Herpes zoster	Infectious	61	36260	53	0.93	1.71E-03
rs78238620	KIF15	Personal history of diseases of blood and blood-forming organs	Blood	94	36183	289.3	0.79	1.67E-03
rs78238620	KIF15	Congenital musculoskeletal deformities of spine	Other	49	37306	754	1.00	2.61E-03
rs41308092	RTEL1	Fluid overload	Other	215	34948	276.6	0.84	4.27E-04
rs41308092	RTEL1	Glaucoma	Ocular	496	36040	365	0.57	1.68E-03
rs41308092	RTEL1	Nonrheumatic tricuspid valve disorders	Cardiovascular	31	36124	395.3	1.46	1.90E-03
rs41308092	RTEL1	Renal dialysis	Digestive	385	33172	585.31	0.59	2.61E-03
rs41308092	RTEL1	Disturbance of skin sensation	Dermatologic	209	36944	687.4	0.88	1.66E-04
rs41308092	RTEL1	Other disorders of cervical region	Genitourinary	35	36685	723	1.33	4.45E-03
rs41308092	RTEL1	Other and unspecified disorders of back	Other	35	36685	724	1.32	4.78E-03
rs41308092	RTEL1	Other unspecified back disorders	Other	32	36685	724.9	1.42	2.27E-03
rs9577395	ATP11A	Congestive heart failure (CHF) NOS	Cardiovascular	462	36435	428.1	0.26	3.77E-03
rs9577395	ATP11A	Hemoptysis	Other	131	37825	516.1	0.51	4.67E-03
rs9577395	ATP11A	Calculus of bile duct	Digestive	157	35047	574.2	-0.42	9.63E-04
rs9577395	ATP11A	Other complications of the puerperium NEC	Genitourinary	58	21366	674	-0.66	9.33E-04
rs28513081	DEPTOR	Mitral valve disease	Cardiovascular	485	36124	394.2	-0.19	4.37E-03
rs28513081	DEPTOR	Nonrheumatic mitral valve disorders	Cardiovascular	468	36124	395.1	-0.19	3.90E-03
rs28513081	DEPTOR	Postmenopausal bleeding	Genitourinary	146	18346	627.1	0.35	4.34E-03
rs28513081	DEPTOR	Spinal stenosis	Other	63	36685	720	0.55	4.53E-03
rs2013701	FAM13A	Adrenal hyperfunction	Endocrine and metabolic	47	36098	255.1	0.66	2.38E-03
rs2013701	FAM13A	Cushing's syndrome	Endocrine and metabolic	32	36098	255.11	0.76	4.01E-03
rs2013701	FAM13A	Congenital anomalies of male genital organs	Genitourinary	87	15325	751.12	-0.49	2.70E-03
rs12699415	MAD1L1	Disorders of mineral metabolism	Digestive	771	36825	275	-0.16	2.20E-03
rs12699415	MAD1L1	Disorders of calcium/phosphorus metabolism	Digestive	493	36825	275.5	-0.19	4.98E-03
rs12699415	MAD1L1	Coagulation defects	Blood	187	36918	286	0.29	4.87E-03
rs12699415	MAD1L1	Other and unspecified coagulation defects	Blood	91	36918	286.7	0.43	3.99E-03
rs12699415	MAD1L1	Stricture and stenosis of esophagus	Other	114	32956	530.3	0.38	4.18E-03
rs12699415	MAD1L1	Other and unspecified disc disorder	Other	136	36685	722.9	0.38	1.66E-03
rs2077551	MAPT	Pancreatic cancer	Neoplastic	177	33414	157	-0.46	2.19E-03
rs2077551	MAPT	Other demyelinating diseases of central nervous system	Neurological	40	32792	341	1.18	4.67E-03
rs2077551	MAPT	Postinflammatory pulmonary fibrosis	Pulmonary	257	34950	502	0.44	1.21E-03

rs7725218	TERT	Cancer of tongue	Neoplastic	55	37802	145.2	-0.72	1.61E-04
rs7725218	TERT	Malignant and unknown neoplasms of brain and nervous system	Neoplastic	537	37317	191	-0.22	4.28E-04
rs7725218	TERT	Cancer of brain and nervous system	Neoplastic	493	37317	191.1	-0.26	8.49E-05
rs7725218	TERT	Cancer of brain	Neoplastic	475	37317	191.11	-0.26	1.04E-04
rs7725218	TERT	Secondary malignant neoplasm	Neoplastic	4059	27863	198	-0.08	2.98E-03
rs7725218	TERT	Conduct disorders	Other	41	36816	312	-0.62	4.61E-03
rs7725218	TERT	Hydrocephalus	Neurological	126	32791	331.1	-0.42	8.83E-04
rs7725218	TERT	Cerebral edema and compression of brain	Neurological	178	32791	348.2	-0.35	1.10E-03
rs7725218	TERT	Disorders of tooth development	Other	92	34631	520	-0.51	5.96E-04
rs7725218	TERT	Disturbances in tooth eruption	Other	70	34631	520.2	-0.61	3.35E-04
rs7725218	TERT	Psoriasis vulgaris	Dermatologic	310	36095	696.41	0.25	4.36E-03
rs7725218	TERT	Other and unspecified disc disorder	Other	136	36684	722.9	-0.37	2.46E-03
rs7725218	TERT	Effects radiation NOS	Other	2069	33484	990	-0.10	2.72E-03

*Table 2.3.5: A summary table of phenotypes associated with IPF genetic risk variants in the Genomics England 100,000 Genomes project database. Associations with IPF risk variants were filtered for those within a p-value threshold of $p < 5 \times 10^{-3}$. Cases and controls for each phenotype were defined by the Denny *et al.* PheWAS pipeline¹²⁵ using the PheCode system^{75,78}. The direction of effect for each phenotype is presented relative to the IPF risk increasing allele. Where effect size is positive, this represents a shared direction of effect with IPF risk.*

2.3.7 Conclusions from PheWAS using an independent cohort

The only phenotype association meeting significance threshold criteria ($p < 1 \times 10^{-5}$) with any of the included 14 IPF genetic risk variants was postinflammatory pulmonary fibrosis with rs35705950 G>T (*MUC5B*). This phecode includes the ICD10 codes J84.1 (Other interstitial pulmonary diseases with fibrosis) and J84.8 (Other specified interstitial pulmonary diseases). However, other phenotypes associated with IPF genetic risk variants in this cohort may contribute to understanding of SNP biology when considered in combination with colocalised GWAS signals (Section 2.3.4).

2.4 Discussion

A PheWAS analysis was performed to investigate the range of phenotypes associated with IPF genetic risk variants, firstly using a data mining approach within a large-scale public database of GWAS results (OpenGWAS). This investigation was developed further by assessing probability of colocalisation of identified phenotypes with IPF risk to prioritise associations driven by the same causal variant. Finally, an independent PheWAS study was performed within the GE 100,000 GP cohort. A broad range of phenotypes were found to be associated with IPF genetic risk variants, and multiple phenotypic association signals from a prioritised subset were found to colocalise with IPF risk. While the PheWAS study performed in the GE 100,000 GP cohort appeared to lack sufficient statistical power, identified phenotypic associations may support interpretation of results from prior PheWAS and colocalisation analyses.

Results suggesting strong evidence of colocalisation of IPF risk associations with increased COVID-19 severity (rs12610495; *DPP9*) and decreased systolic blood pressure (rs9860874; *TERC*) align with the findings of previous studies by Allen *et al.* (2022)¹⁰¹ and Parcesepe *et al.* (2023)¹²⁷, respectively. The colocalisation of associations with IPF risk and COVID-19 severity potentially indicates either a fibrotic mechanism may be associated with more severe viral infection, or increased susceptibility to severe viral infection may subsequently contribute to IPF risk. The Allen *et al.* (2022) study also identified opposing effects of the *MUC5B* risk variant, rs35705950, on IPF risk and COVID-19 severity when a more lenient significance threshold of $p < 1 \times 10^{-5}$ was applied in selection of phenotypic associations. Colocalisation of opposing IPF risk and systolic blood pressure signals indicates that IPF patients may be at decreased risk of hypertension. Conversely, another signal that colocalised with IPF risk at this locus (rs9860874; *TERC*) is shortened telomere length, which has been shown to be associated with increased risk of both hypertension^{128,129} and IPF⁷. Overall, these results suggest there are possible mechanistic links between IPF,

telomere length, and blood pressure, but indicate that relationships between these traits are likely to be complex.

IPF risk associations with rs2076295 (*DSP*), rs2609259 (*FAM13A*), and rs2897075 (*ZKSCAN1*) had strong evidence of colocalisation with FEV1/FVC and $FEV1/FVC < 0.7$. Shared associations with pulmonary fibrosis and COPD ($FEV1 < 80\%$, $FEV1/FVC < 0.7$) with rs2076295 (*DSP*) and rs2609259 (*FAM13A*) have previously been identified¹³⁰ but shared causal variants were not confirmed through colocalisation. The results of this investigation therefore support an opposing effect of IPF genetics on COPD risk. More research is required to understand if this relationship is driven by biological factors, such as the negative association of rs2609259 (*FAM13A*) with monocyte count, or simply decreased FVC associated with IPF disease resulting in an increased FEV1/FVC ratio value.

This investigation identified novel colocalisation of IPF risk and total testosterone level associations with rs2304645 (*IVD*). This association with sex hormones is supported by associations between the rs59424629 (*IVD*) IPF risk variant and decreased risk of endometriosis and polycystic ovaries in the GE 100,000 GP PheWAS (Section 2.3.6). In addition, a positive association between IPF and total testosterone¹³¹ has been indicated. Intriguingly, a negative relationship with bioavailable testosterone has also been suggested^{131,132}, which contradicts the positive effect identified with this trait in the public data PheWAS analysis (Section 2.3.1). These results suggest a complex relationship between sex hormones and IPF which warrants further investigation.

Multiple IPF genetic risk signals are indicated to colocalise with associations with blood cell traits, including mean corpuscular haemoglobin (rs7725218; *TERT*), mean corpuscular volume (rs12912339; *KNL1*, rs2897075; *ZKSCAN1*), red cell distribution width (rs9577395; *ATP11A*), and plateletcrit (rs12537430; *MAD1L1*). Overall, these associations indicate that abnormal haematopoiesis may be linked to IPF disease mechanisms. Previous research has suggested that telomere length is inversely associated with mean corpuscular haemoglobin

and mean corpuscular volume^{133,134}. In addition, blood traits found to share causal variants with IPF risk have been indicated to vary with age and age-associated factors^{135–137}. A plausible interpretation of these findings is therefore that multiple IPF genetic risk variants may also be contributing to accelerated ageing of tissues outside of the lung.

Despite PheWAS results suggesting that IPF genetic risk variants, particularly rs7725218 (*TERT*), are associated with multiple cancers, no strong evidence of colocalisation was identified. Only associations with noncancerous tumours (neoplasms and uterine fibroids) were found to colocalise with IPF risk (rs9860874; *TERC*). These findings particularly contrast with the increased incidence of lung cancer in IPF patients⁴⁸. The results of PheWAS and colocalisation analyses therefore suggest that, at a single variant level, the comorbid relationship between IPF and lung cancer is unlikely to be driven by genetic factors.

Colocalisation of signals associated with IPF risk and gene expression of *ATP11A*, *FAM13A*, and *DEPTOR* aligned with previously published variant annotations using the Ensembl VEP^{89,92}. Strong evidence of colocalisation was also identified between IPF risk and *RPAP1* gene expression (rs12912339; *KNL1*). *RPAP1* (encoding RNA Polymerase II-associated protein 1) depletion has been linked to reduced epithelial to mesenchymal transition and downregulation of fibrosis-associated genes *TGFB3*, *VEGFA*, and *S100A4*^{138–141}. This finding highlights the potential that the rs12912339 A allele may contribute to IPF risk through an alternative pathway of increased *RPAP1* expression. However, individual regulatory variants can often influence the expression of multiple genes¹⁴², therefore this finding does not rule out a link between rs12912339 and *KNL1* gene expression. In addition, insights from are limited as the eQTL GWAS data used in this analysis were based on blood gene expression signatures only¹⁴³. More targeted investigation into the rs12912339 IPF risk variant, how it affects gene expression in lung tissues, and the involvement of *RPAP1* and *KNL1* in IPF biology are warranted to further understand the implications of this finding.

The association identified in the previous IPF PheWAS by Claar *et al.* (2016)¹¹³ of rs6793295 (*TERC*) with risk of asthma was not replicated in this investigation. However, this methodology offers some advantages over the methodological approach used in the previous IPF PheWAS study by Claar *et al.* (2016). Firstly, the association statistics used to perform PheWAS of GWAS databases typically represent large sample sizes. These data resources also typically represent a broader range of human traits, enabling a more comprehensive investigation of associated phenotypes. Using GWAS data also enabled direct follow-up colocalisation analyses to estimate the probability of shared causal variants.

The approach applied in this investigation also has several limitations. In particular, the combination of using GWAS databases and use of p-values to filter and prioritise trait associations could potentially induce bias towards common phenotypes with large sample sizes. As an example, this potential bias may underlie the high prevalence of associations identified with blood traits (Section 2.3.1). This potential bias is also indicated by many tested variants not being significantly associated with pulmonary fibrosis or ILD phenotypes in this study. The current largest IPF GWAS⁹⁰ is not included in public databases, therefore this example highlights potential de-prioritisation of important but rare phenotypes when using this approach.

The phenotype prioritisation method applied (Section 2.2.2.1) enabled objective filtering of priority phenotypes from a large number of associations. However, this approach may have de-prioritised phenotypes that may be relevant to IPF biology. As an example, a de-prioritised association between rs9577395 (*ATP11A*) and increased HbA1c may explain the increased incidence of diabetes mellitus observed in IPF patient cohorts⁴⁴. In addition, some prioritised traits did not have sufficient public data available to run follow-up colocalisation analyses, such as the association identified between rs10808505 (*DEPTOR*) and brain IDPs associated with brain ageing, such as white-grey contrast measurements¹⁴⁴. Ideally, when running colocalisation analyses researchers should also study the association

signals to define an appropriate genomic region for analysis. An arbitrary region of 500kb was used to enable analysis of a high volume of phenotypes, however specifically targeted definition of genomic regions would likely be a more optimised approach.

The GE 100,000 GP PheWAS aimed to facilitate inclusion of rare disease phenotypes, however most association analyses were limited by sample size and had low statistical power. Only 14 IPF genetic risk variants were investigated as this study was performed before the 2022 IPF GWAS⁹⁰, which identified five additional genetic risk signals. A re-analysis of this dataset was planned, but not completed due to time constraints. Associations with several disease phenotypes such as chronic fatigue syndrome, chronic thyroiditis, and endometriosis may indicate shared or interacting disease mechanisms if further evidence is found to support or explain relationships. For example, hypothyroidism has been shown to be more prevalent in IPF patients¹⁴⁵, however biological pathways linking these disease phenotypes have not clearly been defined. However, for many tested phenotypes, including hypothyroidism, PheWAS analyses performed in the GE 100,000 GP dataset may have been limited in statistical power due to a low number of cases (Table 2.3.5).

In summary, this investigation demonstrates that IPF genetic risk variants are highly pleiotropic. In particular, the findings of this investigation suggest that shortened telomere length, testosterone levels, viral infection, and blood pressure may affect risk of IPF or share underlying biological mechanisms with IPF.

3: Genetic Correlation Study of Idiopathic Pulmonary Fibrosis and Related Phenotypes

3.1 Introduction

Genetic correlation analyses aim to assess the extent of shared genetic architecture underlying heritable phenotypic variation in complex traits^{79,146}. Shared genetic architecture across traits can prioritise where shared factors or mechanisms may present opportunities for shared treatments. The identification of genetic correlation can also highlight potential causal relationships for further investigation.

Methods such as genetic colocalisation (Section 2.2.2) are powerful tools for understanding shared or causative disease aetiology when significantly associated genetic variants account for a large proportion of trait heritability. Many human traits are highly polygenic, with potentially thousands of genetic variants each contributing small effects to overall phenotypic variation¹⁴⁷. Genetic correlation approaches allow for genetic overlap to be assessed across all SNPs, not only those reaching genome-wide significance ($p < 5 \times 10^{-8}$), and may therefore enable a more comprehensive estimation of shared genetic architecture.

A commonly used method for estimating genetic correlation is linkage disequilibrium score regression (LDSC). Causal genetic variants are typically flanked by a region of correlated variants with inflated association statistics that are proportional to the LD with the causal variant. Alternatively, a similar inflation of test statistics can occur due to bias such as cryptic relatedness and population stratification due to genetic drift. LDSC primarily aims to distinguish inflation due to true genetic effects and from bias, therefore

allowing for unbiased estimation of heritability¹⁴⁸. Subsequently, LDSC can be used to estimate genetic correlation across traits without bias from sample overlap⁷⁹. This approach also only requires GWAS summary statistics and has a relatively low computational requirement, allowing time-efficient analysis of genetic correlation across multiple traits¹⁴⁹. LDSC has previously been successfully applied in IPF research to estimate genetic correlation between IPF and COVID-19 infection and severity¹⁰¹.

This chapter firstly includes a genetic correlation analysis to investigate whether shared causal variants with IPF at a genome-wide significant level ($p < 5 \times 10^{-8}$) are indicative of broader shared genetic architecture. Secondly, fibrosis is a common feature of many diseases and fibrotic disease incidence has been shown to be correlated^{52,58}. This chapter therefore includes an additional exploratory analysis investigating whether IPF shares genetic architecture with a sample of fibrotic disease traits.

3.2 Methods

3.2.1 Trait selection for genetic correlation analyses

3.2.1.1 Trait selection for follow-up investigation of colocalised phenotypes with IPF genetic risk variants

All phenotypic traits, excluding eQTL associations, with strong evidence of colocalisation with IPF risk ($PP.H4 > 80\%$) identified in prior analyses (Section 2.3.4) were prioritised for genetic correlation analyses.

3.2.1.2 Trait selection for investigation of shared genetic architecture across IPF and fibrotic disease traits

A sample of diseases classified as ‘fibrotic’ in a recent Delphi study⁵² with relevant publicly available GWAS summary data downloadable from the GWAS catalog⁶⁹ were selected for genetic correlation analyses.

3.2.3 Methods for genetic correlation analysis

To run genetic correlation analyses, summary data from the most recent IPF GWAS⁹⁰ (4,125 European ancestry cases, 20,464 European ancestry controls) were retrieved by request from the authors. GWAS summary dataset for each analysed phenotype were downloaded from either the GWAS Catalog database⁶⁹, or the IEU OpenGWAS database¹¹⁷. Downloaded GWAS datasets were subsequently processed using *tidyverse* (R). The LDSC genetic correlation method^{79,148} requires GWAS data containing the following data fields:

- A SNP identifier (such as an RSID code)
- Allele 1 (*A1*; effect allele)
- Allele 2 (*A2*; other allele)
- SNP effect size (*beta*)
- SNP p-value

To filter out rare variants or SNPs with a low sample size:

- SNP or GWAS sample size
- SNP minor allele frequency

All GWAS summary statistics datasets were converted to *sumstats* format using the *munge_sumstats.py* program included in the *ldsc* command line tool. Converting to *sumstats* format ensures that input data are correctly formatted and processed for heritability and genetic correlation estimation. To filter for well-imputed SNPs, datasets were filtered for

HapMap3 SNPs (downloaded from the Broad Institute public repository) using the `--merge-alleles` flag when running `munge_sumstats.py`.

Genetic correlations between IPF and related traits were estimated using `ldsc.py`. Pre-computed LD scores for LD score regression were downloaded from the Broad Institute public repository.

3.3 Results

3.3.1 Traits selected for genetic correlation analysis with IPF risk GWAS

Traits with strong evidence of genetic colocalisation (Section 2.3.4) each had appropriate data for LDSC, therefore all 17 traits were included in genetic correlation analyses (Table

Trait	Dataset ID	Database
COVID-19 - severe vs population	GCST011076	GWAS Catalog
COVID-19 - hospitalised vs population	GCST011082	GWAS Catalog
Neoplasms	ukb-d-II_NEOPLASM	IEU OpenGWAS
Uterine fibroids	GCST90018934	GWAS Catalog
Systolic blood pressure	ieu-b-38	IEU OpenGWAS
Telomere length	ieu-b-4879	IEU OpenGWAS
Red cell distribution width	GCST90025988	GWAS Catalog
Mean corpuscular volume	GCST90002334	GWAS Catalog
Mean corpuscular haemoglobin	GCST004630	GWAS Catalog
Corneal resistance factor	ukb-b-15178	GWAS Catalog
Plateletcrit	GCST90002400	GWAS Catalog
IGF-1	GCST90025989	GWAS Catalog
FEV1/FVC < 0.7	ieu-b-106	IEU OpenGWAS
FEV1/FVC	GCST007431	GWAS Catalog
FVC	GCST007429	GWAS Catalog
Monocyte count	ieu-b-31	GWAS Catalog
White blood cell count	ieu-b-30	GWAS Catalog
Total testosterone levels	GCST90012112	GWAS Catalog

Table 3.3.1: Traits selected for genetic correlation analysis following identification of strong evidence of shared causal variants with IPF risk association signals. IGF-1: Insulin-like growth factor 1.

Traits	Dataset ID	Database
Type 2 diabetes	GCST006867	GWAS Catalog
NAFLD	GCST90011885	GWAS Catalog
Primary biliary cholangitis	GCST90061440	GWAS Catalog
Chronic kidney disease	GCST008065	GWAS Catalog
Rheumatoid arthritis	GCST90132223	GWAS Catalog
Sarcoidosis	GCST005538	GWAS Catalog
Crohn's disease	GCST004132	GWAS Catalog
Endometriosis	GCST90269970	GWAS Catalog

Table 3.3.2: Fibrotic disease traits selected for genetic correlation analysis. NAFLD: Non-alcoholic fatty livers disease.

3.3.1). The sample of fibrotic diseases selected for genetic correlation analyses with IPF risk were type 2 diabetes (T2D), non-alcoholic fatty liver disease (NAFLD), primary biliary cholangitis (PBC), chronic kidney disease (CKD), rheumatoid arthritis (RA), Sarcoidosis, Crohn's disease (CD), and endometriosis (Table 3.3.2). In total, 26 traits or diseases were included in genetic correlation analyses (18 colocalised and eight fibrotic).

3.3.2 Genetic correlation analysis results

Results of genetic correlation analyses indicate a strong positive correlation between the genetics of IPF risk and genetics associated with increased risk of severe COVID-19 (genetic correlation score; $r_g = 0.24$, $p = 5.39 \times 10^{-3}$) and hospitalised COVID-19 ($r_g = 0.38$, $p = 3.19 \times 10^{-4}$) (Table 3.3.3).

IPF risk genetics are positively correlated with FEV1/FVC ($r_g = 0.19$, $p = 4.58 \times 10^{-6}$) and negatively correlated with FEV1/FVC < 0.7 ($r_g = -0.16$, $p = 8.06 \times 10^{-4}$). However, results suggest no overall correlation between IPF genetic risk and the genetics of FVC ($r_g = 0.01$, $p = 0.77$).

IPF risk genetics are negatively correlated with systolic blood pressure ($r_g = -0.07$, $p = 0.03$). Results also suggest a negative genetic correlation of IPF risk with red cell distribution

width ($r_g = -0.07, p = 0.04$) and a positive correlation with mean corpuscular haemoglobin ($r_g = 0.09, p = 0.02$) and mean corpuscular volume ($r_g = 0.08, p = 0.01$).

Results suggest a strong negative correlation between IPF risk and telomere length ($r_g = -0.31, p = 1.14 \times 10^{-7}$).

Trait	Category	Correlation score	Lower 95% CI	Upper 95% CI	P-value
COVID-19 - hospitalised vs population*	Infectious	0.38	0.17	0.59	3.19×10^{-4}
COVID-19 - severe vs population*	Infectious	0.24	0.07	0.40	5.39×10^{-3}
FVC	Pulmonary	0.01	-0.08	0.10	0.77
FEV1/FVC*	Pulmonary	0.19	0.11	0.28	4.58×10^{-6}
FEV1/FVC < 0.7*	Pulmonary	-0.16	-0.25	-0.07	8.06×10^{-4}
Neoplasms	Neoplastic	0.04	-0.27	0.34	0.82
Uterine fibroids	Genitourinary	-0.06	-0.19	0.07	0.39
Corneal resistance factor	Ocular	-0.01	-0.09	0.08	0.86
Systolic blood pressure*	Cardiovascular	-0.07	-0.14	-0.01	0.03
White blood cell count	Blood	-0.04	-0.11	0.03	0.23
Monocyte count	Blood	-0.05	-0.12	0.01	0.10
Plateletcrit	Blood	-0.04	-0.12	0.03	0.25
Mean corpuscular haemoglobin*	Blood	0.09	0.02	0.16	0.02
Mean corpuscular volume*	Blood	0.08	0.02	0.14	0.01
Red cell distribution width*	Blood	-0.07	-0.14	0.00	0.04
Total testosterone levels	Biomarker	0.01	-0.08	0.10	0.77
IGF-1	Biomarker	0.03	-0.05	0.10	0.45
Telomere length*	Biomarker	-0.31	-0.42	-0.19	1.14×10^{-7}

Table 3.3.3: Results of genetic correlation analyses of IPF risk with traits prioritised through colocalisation with IPF genetic risk variants. CI: confidence interval. Asterisks () indicate traits with statistically significant ($p < 0.05$) genetic correlations with IPF risk.*

Overall, genetic correlation scores of a sample of fibrotic diseases with IPF were either positive or null (Table 3.3.4). Results show a positive genetic correlation meeting a $p < 0.05$ threshold between IPF risk and risk of primary biliary cholangitis ($r_g: 0.14, p = 0.04$). A summary of genetic correlations of IPF with all phenotypes investigated in this analysis is presented in Figure 3.3.1.

Trait	Category	Correlation score	Lower 95% CI	Upper 95% CI	P-value
Endometriosis	Fibrotic	0.09	-0.05	0.23	0.20
Crohn's disease	Fibrotic	0.04	-0.06	0.14	0.45
Sarcoidosis	Fibrotic	0.31	-0.03	0.65	0.08
Rheumatoid arthritis	Fibrotic	-0.01	-0.15	0.12	0.87
Chronic kidney disease	Fibrotic	0.10	-0.04	0.23	0.18
NAFLD	Fibrotic	0.31	-0.07	0.68	0.11
Type 2 diabetes	Fibrotic	0.08	0.00	0.16	0.06
Primary biliary cholangitis*	Fibrotic	0.14	0.01	0.27	0.04

Table 3.3.4: Results of genetic correlation analyses of IPF risk with a sample of fibrotic disease phenotypes. CI: confidence interval. Asterisks (*) indicate traits with statistically significant ($p < 0.05$) genetic correlations with IPF risk.

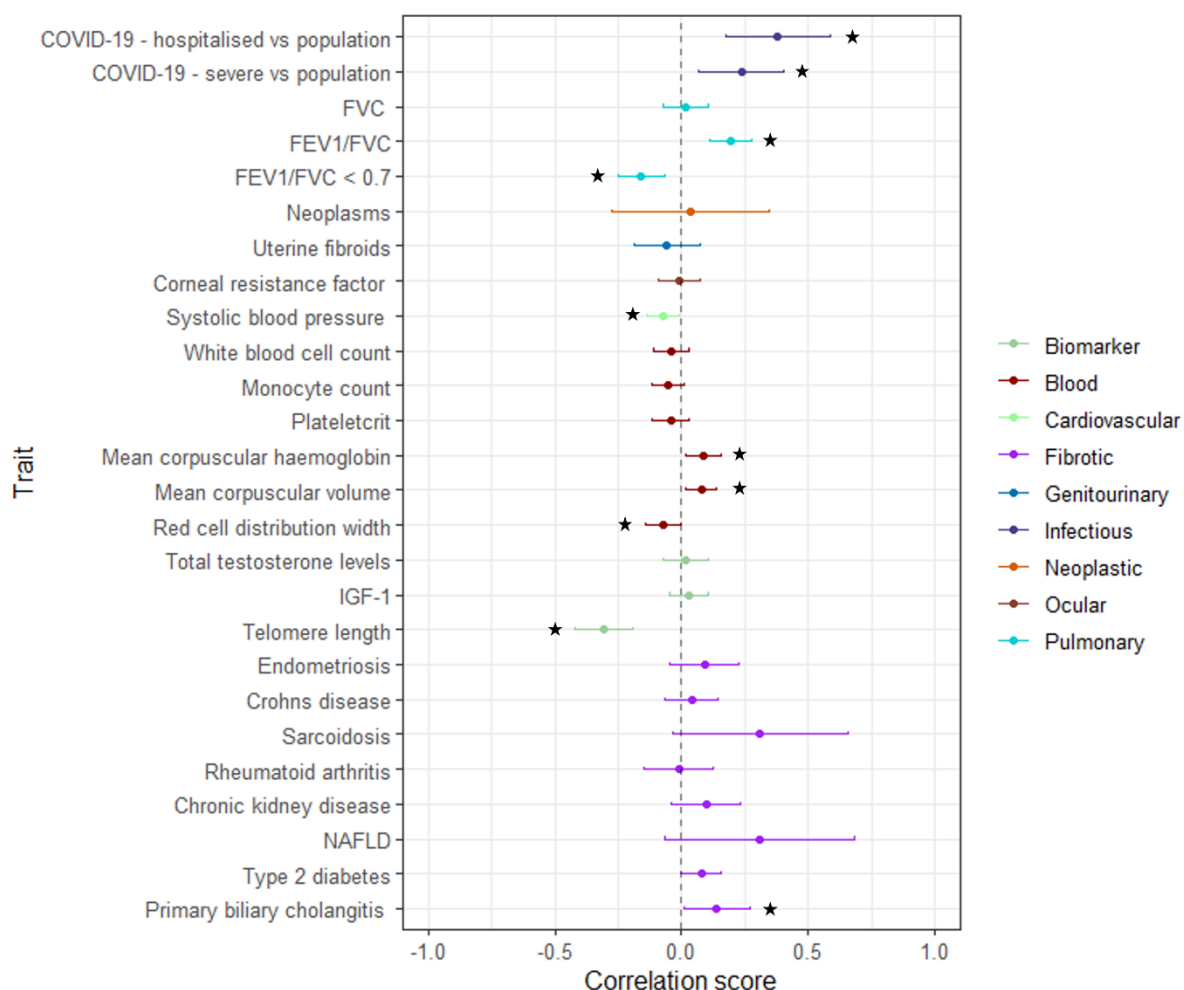


Figure 3.3.1: Results of genetic correlation analyses of IPF risk with all investigated phenotypes. Stars indicate statistically significant ($p < 0.05$) correlations with IPF risk. Error bars indicate 95% confidence intervals for each correlation score.

3.3.3 Conclusions from genetic correlation analysis of idiopathic pulmonary fibrosis and related phenotypes

The results of this investigation suggest that many phenotypes that share causal variants with IPF risk are not correlated with IPF risk on a genome-wide basis. IPF genetic risk is also not highly correlated with that of a sample of fibrotic disease traits. However, genetic correlations were found with IPF and severe COVID-19 disease, lung function measurements, blood pressure, blood cell traits, telomere length measurement, and primary biliary cholangitis risk.

3.4 Discussion

Genetic correlation analysis was applied to investigate whether phenotypes sharing causal genetic variants with IPF also share broader genetic architecture. An exploratory analysis also assessed genetic correlation of fibrotic traits with IPF risk to investigate whether fibrotic diseases are driven by similar genetics. Genetic correlations ($p < 0.05$) were identified with IPF and nine traits prioritised by prior colocalisation analyses (Section 2.3.4) and one fibrotic trait.

Genetic correlation analysis results suggesting a small overall negative correlation of systolic blood pressure with IPF risk align with previous research by Parcesepe *et al.* (2023)¹²⁷. This finding is unexpected as cardiovascular disease and IPF share multiple risk factors including male sex, age, smoking, and shortened telomere length^{128,150}. The correlation score of IPF with systolic blood pressure was negative but close to zero, which indicates that any relationship between IPF and systemic hypertension may be more driven by environmental factors than by genetic effects.

A positive genetic correlation between IPF risk and both severe and hospitalised COVID-19 disease has been previously reported^{101,151}. A possible interpretation of this finding is that increased susceptibility to severe viral infection may confer increased IPF risk. This concept would support the predicted role of viral respiratory infections in triggering or exacerbating IPF development^{11,152}. Alternatively, patients with IPF have been shown to be more susceptible to severe COVID-19 infection and more likely to be hospitalised¹⁵³. Increased IPF risk may therefore confer increased risk of severe COVID-19 and hospitalisation, resulting in positive genetic correlation with these phenotypes.

The results of this investigation indicate that IPF risk has a positive genetic correlation with FEV1/FVC ratio and, correspondingly, a negative genetic relationship with FEV1/FVC below 0.7. Alongside appropriate respiratory symptoms and prior exposure to harmful airborne stimuli, a postbronchodilator FEV1/FVC ratio below 0.7 is used to diagnose COPD¹⁵⁴. This negative genetic relationship potentially highlights opposing pathogenic mechanisms that increase IPF risk and decrease risk of COPD. COPD is a common comorbidity in IPF cohorts¹⁵⁵ that shares risk factors such as male sex, increased age, and smoking, and associated biological features including shorter telomeres, oxidative stress, and immune senescence¹⁵⁶. However, influential biological pathways such as WNT/ β -catenin, Notch, and Sonic hedgehog signaling have been shown to be upregulated in IPF and downregulated in COPD¹⁵⁷. A Mendelian randomisation study has also indicated that shortened telomere length, while being associated with both diseases, is likely to be on the causal pathway of IPF and not COPD⁷. It must also be considered that genetic correlation with FEV1/FVC may simply be driven by decreased FVC associated with IPF¹⁵⁸, despite the null genetic correlation observed between IPF risk and FVC in this analysis.

A negative genetic correlation with IPF risk and telomere length indicates shared genetic architecture of increased IPF risk with shortened telomeres. This finding aligns with many

previous studies showing that shortened telomere length is associated with IPF disease^{159,160} and is likely to be on the causal pathway for IPF⁷.

Mean corpuscular volume (MCV) and mean corpuscular haemoglobin (MCH) are both biomarkers used to detect macrocytic or microcytic anaemia. Red cell distribution width (RCDW) can be used to interpret possible causes underlying an identified anemia, with higher RCDW indicating greater cell size variation and a more complex underlying cause¹⁶¹. Increased MCV (macrocytosis) has previously been linked to lung disease associated with telomere disorders^{162–164}, which may explain the positive genetic correlation observed with IPF in this analysis. Increased MCV is also associated with vitamin deficiency, liver disease, and hypothyroidism¹⁶¹, and RCDW has been proposed to be a marker of disrupted metabolic and biological homeostasis associated with factor such as oxidative stress, telomere attrition, and inflammation¹⁶⁵. Anaemias have also previously been shown to be associated with morbidity in COPD patients¹⁶⁶. Overall, these correlations indicate that IPF genetic mechanisms are likely to affect blood cell biology. Blood cell biomarkers therefore have the potential to be markers of telomere biology in IPF, which is linked to patient survival¹⁶⁷, however this potential may be limited by the broad biology associated with these traits.

A positive genetic correlation of primary biliary cholangitis (PBC) with IPF indicates that these diseases may share some genetic architecture and pathogenic mechanisms. A recent study of electronic healthcare record data indicated that individuals with pulmonary fibrosis are more likely to also have biliary fibrosis⁵⁸. In addition, a small number of studies have shown a relatively high rate ILD and fibrotic ILD in PBC cohorts^{168,169}, although this comorbid relationship was attributed to the autoimmune effects of Sjögren's syndrome in one study¹⁷⁰. The pathogenesis of PBC is primarily driven by autoimmune mechanisms, but late-stage fibrosis is associated with increased TGF- β signaling^{171,172}. IPF and PBC are both diseases characterised by dysregulated epithelial repair and fibrosis, therefore shared

pathogenic mechanisms are plausible. However, more evidence is required to interpret the validity and possible biological implications of this genetic correlation.

A key limitation of global genetic correlation methods, such as LDSC, is that r_g indicates the overall average of shared genetic association. Varying positive and negative effects across the genome could therefore result in an overall null r_g of low statistical significance⁸⁰. A possible example of this effect is the low negative genetic correlation score of IPF risk with systolic blood pressure, where a previous study has identified four colocalised signals ($p < 1 \times 10^{-5}$) with opposite effects, and three with aligned risk effects¹²⁷. Similarly, no evidence of genetic correlation was observed between IPF and total testosterone levels. Intriguingly, total testosterone and bioavailable testosterone are positively genetically correlated¹⁷³ but have been shown to have opposite effects on IPF risk¹³¹, which may explain the null effect observed in this analysis. Overall, global correlation methods can provide insights into shared genetic aetiology but may have limited efficacy when applied to complex genetic relationships.

Use of the *munge sumstats* step in this methodology offers the advantage of helping to ensure reliable and robust genetic correlation analysis, but the disadvantage of filtering out rare variants ($MAF < 0.01$). As rare genetics represent a substantial proportion of IPF heritability^{174,175}, genetic correlation including rare variants may enable more comprehensive investigation of shared genetic architecture with related traits.

In summary, this genetic correlation analysis shows that multiple traits are genetically correlated with IPF risk, which may represent shared genetic architecture and trait mechanisms. Future research should aim to expand the range of investigated traits and include targeted follow-up analyses using, for example, local genetic correlation or Mendelian randomisation approaches.

4: Investigating Causal Relationships between Diabetes Mellitus and Idiopathic Pulmonary Fibrosis

4.1 Introduction

Diabetes Mellitus (DM) is a group of diseases that are characterised by dysregulation of blood glucose homeostasis and subsequent hyperglycaemia. DM is a common disease affecting over four hundred million people globally and disease incidence is predicted to increase¹⁷⁶.

Type 1 diabetes (T1D) is driven by an autoimmune response that targets pancreatic β -cells. As the primary role of β -cells is to produce and secrete insulin, autoimmune destruction of β -cells results in the characteristic dysregulation of glucose homeostasis¹⁷⁷.

Type 2 diabetes (T2D) is the most common form of diabetes, accounting for over 90% of diabetes cases¹⁷⁸. The current paradigm of T2D pathophysiology involves disruption, compensation, and subsequent failure of the biological mechanisms that normally maintain glucose homeostasis. A core process in the pathophysiology of T2D is progressive decline of systemic insulin sensitivity, leading to higher blood glucose levels due to decreased uptake by target tissues¹⁷⁹. To compensate for decreased insulin sensitivity, β -cells are driven to secrete more insulin to maintain normal blood glucose (hyperinsulinaemia). Over time, pancreatic tissues become unable to meet the increasing physiological insulin demand. This effect is also exacerbated by increased β -cell apoptosis due to oxidative stress¹⁷⁹. The interaction of diabetes-associated traits and T2D disease is made more complex by factors such as hyperglycaemia can precede disease development and are downstream consequences of disease biology¹⁸⁰.

There is a strong epidemiological association between IPF and DM⁵⁸. Studies have indicated increased incidence of DM in IPF cohorts when compared to controls, but it is unclear if there is increased incidence of IPF in DM cohorts¹⁸¹. Comorbid DM is also associated with increased mortality in IPF patient cohorts¹⁸². DM is also associated with cardiac fibrosis^{183,184}, significant liver fibrosis^{185–187}, retinal fibrosis, and kidney fibrosis associated with diabetic nephropathy¹⁸⁸, indicating a potential relationship with central fibrotic mechanisms¹⁸⁹.

It is unclear exactly how DM may be linked to IPF. One potential pathway of effect of DM on IPF is through chronic hyperglycaemia leading to increased oxidative stress, DNA damage, and subsequent senescence and fibrosis⁴⁵. Streptozotocin-induced diabetes has also been found to be associated with increased fibrotic signaling and epithelial to mesenchymal transition in rat lung tissue¹⁹⁰. DM has also been linked to systemic inflammation and increased rate of infection^{191–193}, which may contribute to the development of IPF¹¹.

Instrumental variable analysis is a method of estimating causality between two variables, an exposure variable and an outcome variable, by using a third (instrumental) variable that is strongly correlated with the exposure. Instrumental variable analyses are useful when direct assessment of a causal relationship, for example through a randomised trial, is unfeasible or unethical. Mendelian randomisation (MR) is a form of instrumental variable analysis that typically uses genetic variants that are strongly associated with an exposure to estimate causal relationships with a specific outcome¹⁹⁴. MR confers several advantages over observational studies, particularly in avoiding environmental confounding and reverse causation if performed correctly. Robust causal inference using MR requires genetic instruments that are¹⁹⁵:

- 1) Reliably associated with the investigated exposure (the *relevance* assumption)
(Figure 4.1.1A)

- 2) Only associated with the outcome through the effects of the exposure (vertical pleiotropy only) (the *exclusion restriction* assumption) (Figure 4.1.1B)
- 3) Independent of any confounders that influence both the exposure and the outcome (the *independence* assumption) (Figure 4.1.1C)

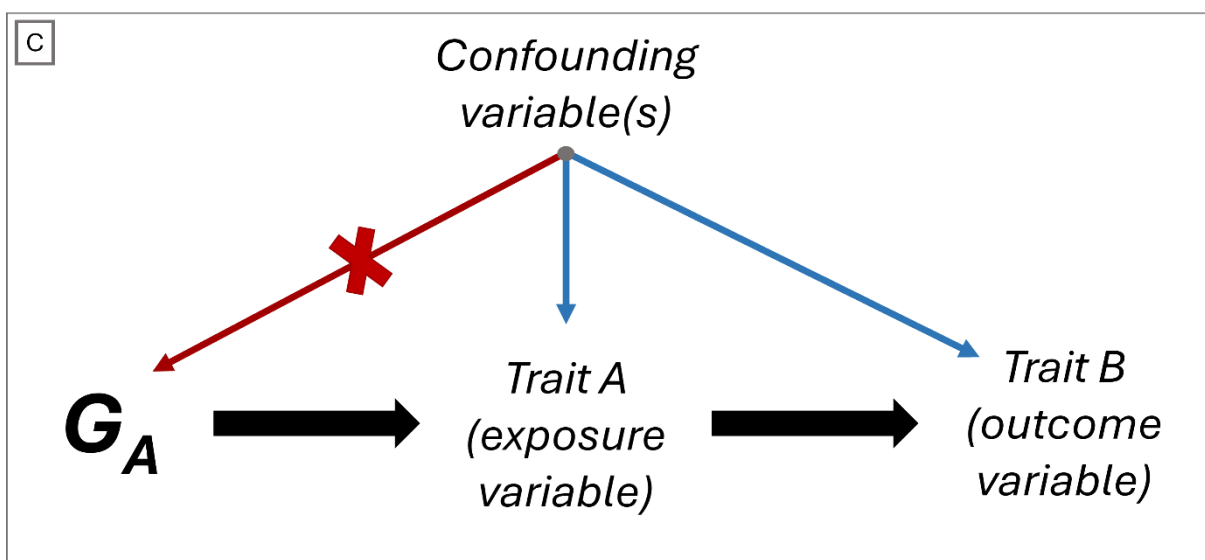
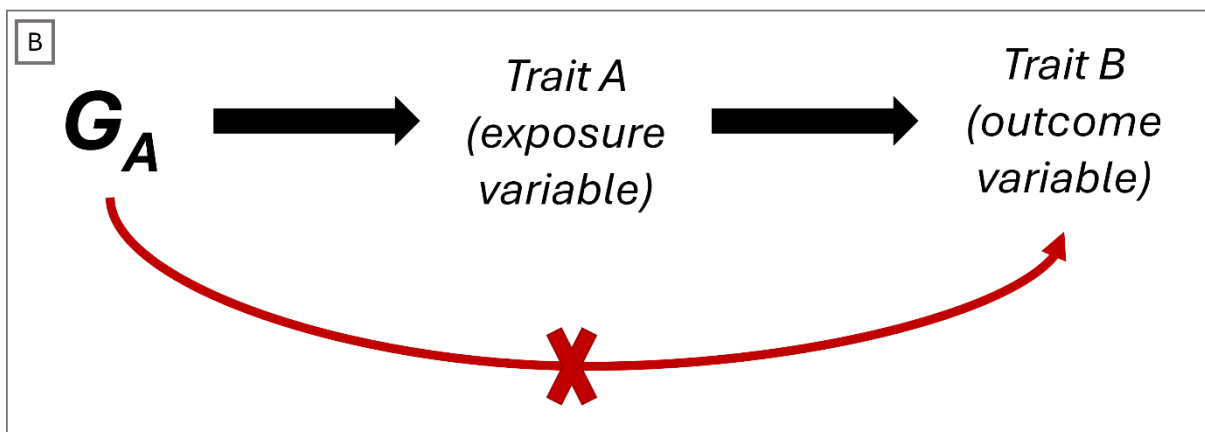
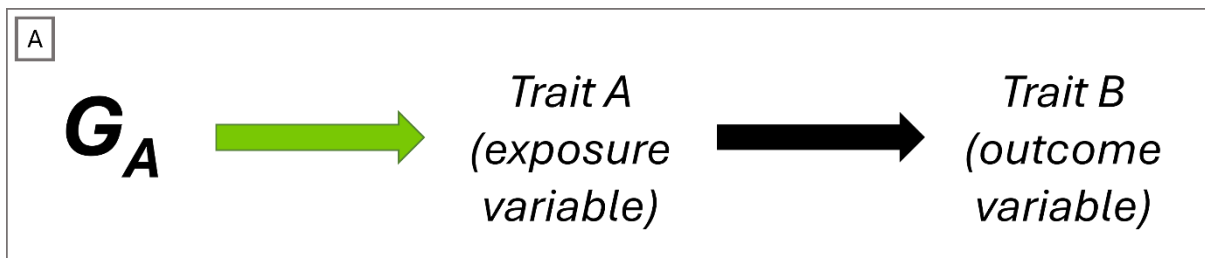


Figure 4.1.1: Diagrams illustrating the three core assumptions required for robust causal estimation using MR. [A] The genetic instrument for trait A (G_A) must be reliably associated with the exposure variable. [B] Any effects of G_A on the outcome variable should be mediated by the exposure variable and not act through any other pathways (horizontal pleiotropy). [C] G_A should not be related to any confounding factors that affect both the exposure variable and the outcome variable. The effects of any confounding factors should be minimally represented by G_A .

A common approach in modern MR studies is two-sample MR using genetic variants (SNPs) identified in GWAS to instrument an exposure. Reported GWAS SNPs are typically strongly associated with the investigated phenotype ($p < 5 \times 10^{-8}$), therefore the relevance assumption should be met by default. The genetic effects of the exposure instrument SNPs can subsequently be looked up in the outcome GWAS. If core assumptions (Figure 4.1.1) are met, SNPs are only associated with the outcome via the exposure, and there is no evidence of reverse causality, then significant effects can be considered evidence of a causal relationship between the exposure variable and the outcome variable¹⁹⁶.

The two-sample MR approach can use publicly available data to provide clinically relevant insights¹⁹⁵, but critical thought and appropriate use of sensitivity analyses are often required to avoid unintentional introduction of confounding effects. Genetic variants can influence the transcriptional regulation of multiple biological processes and effects of genotype on phenotype can act through multiple mediating or interacting pathways. These effects can lead to biologically complex genetic instruments that may invalidate the exclusion restriction assumption (Figure 4.1.1B) through horizontal pleiotropy, where selected SNPs are associated with the outcome variable via pathways that act independently of the exposure variable. Multiple methods have been developed that help to make MR analyses more robust to horizontal pleiotropy¹⁹⁷, however critical thought on overall study design and selection of genetic instruments is often the most important part of a robust MR methodology¹⁹⁸.

T2D is typically diagnosed with a repeated high blood glucose test result, such as HbA1c over 6.5%, with factors such as obesity (BMI over 25kg/m²) also recommended to be considered in prioritising patients for diagnostic testing¹⁹⁹. T2D can therefore be considered a dichotomisation of a continuous trait (blood glucose level) from a statistical perspective. MR with dichotomised or coarsened exposures based on an underlying continuous factor can lead to coarsening bias, which likely to result in violation of the exclusion restriction assumption²⁰⁰ (Figure 4.1.1B) and bias MR estimates due to potential paths of effect that are not mediated through the tested exposure^{201,202} (T2D). In other words, if raised blood glucose causes an outcome, genetic variants associated with blood glucose will still be associated with the same outcome in a population where no individuals reach the clinical threshold for T2D. It is therefore warranted that blood glucose should also be investigated as an exposure. In addition, a HbA1c (a measure of long-term blood glucose) genetic instrument has been shown to be stronger on average and explain more phenotypic variance than a genetic instrument for T2D²⁰³.

Multivariable MR is used in situations where genetic variants are associated with multiple related exposures to estimate the effect of each exposure on a single outcome¹⁹⁷. This approach can therefore distinguish between true exposures and mediators on a causal pathway²⁰⁴.

This chapter includes a two-sample bidirectional MR analysis that aimed to assess whether causal relationships exist between IPF susceptibility and both T1D and T2D. Following this, this chapter includes MR analyses assessing whether HbA1c level, BMI, and fasting insulin levels are likely to cause IPF. Multivariable MR analyses of T2D and continuous traits HbA1c, BMI, and fasting insulin levels were also performed to assess for potential independent causal effects on IPF risk. A manuscript describing this study is publicly available as a preprint through *MedRxiv*²⁰⁵ and has been submitted to the journal *Thorax*.

4.2 Methods

4.2.1 Data retrieval and processing

Summary data were retrieved from GWAS studies of T1D²⁰⁶ (6,683 European ancestry cases, 12,173 European ancestry controls, 2,601 European ancestry affected sibling pair families, 69 European ancestry trios), T2D²⁰⁷ (61,714 European ancestry cases, 1,178 Pakistani ancestry cases, 593,952 European ancestry controls, 2,472 Pakistani ancestry controls), and IPF susceptibility⁹⁰ (4,125 European ancestry cases, 20,464 European ancestry controls). A secondary analysis was also performed using summary data from a European-only T2D GWAS cohort²⁰⁸ (12,931 European ancestry cases, 57,196 European ancestry controls).

Summary data for MR analyses of continuous traits were retrieved from GWAS of HbA1c level²⁰⁹ (sample size: 88,355 of European ancestry), fasting insulin level²¹⁰ (sample size: 151,013 of European ancestry), and BMI²¹¹ (sample size: 681,275 of European ancestry). No participants in the HbA1c or fasting insulin GWAS cohorts had a diagnosis of diabetes or had blood glucose measurements meeting T2D diagnostic thresholds (fasting glucose over 7mmol/L, HbA1c over 6.5%, or 2hr glucose over 11.1 mmol/L). A secondary analysis was also conducted assessing if glycemic pathways contributing to HbA1c are likely to cause IPF using a subset of SNPs that were classified as ‘glycemic’ only, as defined by Wheeler *et al.* (2017)²⁰⁹. This secondary analysis aimed to ensure that any observed causal estimate was not being driven or altered by associations with haemoglobin or erythrocyte biology.

This MR methodology using *MendelianRandomization* in R with GWAS summary statistics required the following data fields in exposure and outcome datasets:

- A SNP identifier (such as an RSID code)
- Chromosome number
- Base position

- Allele 1 ($A1$; effect allele)
- Allele 2 ($A2$; other allele)
- SNP effect size (β)
- Effect standard error
- SNP p-value
- Effect allele frequency (EAF)
- GWAS sample size (N)

For ease of plotting and interpretation, alleles in the exposure instrument dataset were oriented to the positive effect direction on the exposure variable. Subsequently, alleles in the outcome dataset were harmonised to the same orientation as those in the exposure, with effect and allele frequency statistics ‘flipped’ accordingly (effect multiplied by -1, subtract allele frequency from 1)¹⁹⁵.

4.2.2 Genetic instrument selection

Genetic instruments were required to be associated with the relevant exposure in each MR analysis ($p < 5 \times 10^{-8}$). The IPF exposure instrument used the 19 SNPs reported in Allen et al. (2022) IPF GWAS⁹⁰. SNPs for all other exposure variables were clumped to ensure independence ($r^2 < 0.001$ in a 10,000kb window) using the 1000 Genomes Project reference panel (*ieugwasr*¹¹⁷). To avoid weak instrument bias, any SNPs with an F-statistic (instrument strength measurement) lower than ten were removed. If any SNPs used as genetic instruments for the exposure were not present in the outcome dataset, appropriate proxy SNPs were identified ($r^2 > 0.7$) or SNPs were removed. Proxy SNPs were identified using *LDLinkR*²¹².

4.2.3 Methods for statistical analyses

Main causal estimates were estimated using the random-effects inverse-variance weighted (IVW-RE) method (*MendelianRandomization*²¹³). If there is no significant heterogeneity between MR estimates across SNPs, then no precision is lost compared to a fixed-effects model, but if significant heterogeneity is present the fixed-effects IVW method is prone to bias¹⁹⁷. The IVW-RE method is therefore recommended over the fixed IVW method for most MR analyses.

MR-Egger²¹⁴, weighted median²¹⁵, MR-PRESSO²¹⁶ and MR-Lasso²¹⁷ methods were used to provide pleiotropy-robust causal estimates. The MR-Egger method incorporates average instrument pleiotropy to calculate pleiotropy-robust estimates, allowing all variants to be pleiotropic. The weighted median method allows up to 50% of SNPs to be pleiotropic while still providing pleiotropy robust estimates. MR-PRESSO is a development of the IVW method that removes (likely pleiotropic) outlier SNPs prior to analysis by calculating SNP contribution to overall heterogeneity. The MR-Lasso method identifies and removes any SNPs that have direct (pleiotropic) effects on the outcome variable²¹⁸. While considering the limitations of each approach, consensus across IVW-RE and pleiotropy-robust MR methods provides strong evidence of valid causal estimation. Leave-one-out MR analyses were also performed to assess whether any single variants are driving or distorting overall estimates within genetic instruments. Other than MR-PRESSO, all univariable MR methods have equivalent multivariable versions. MVMR analyses were performed with MVMR-IVW²¹⁹, MVMR-Median²²⁰, MVMR-Egger²²¹, and MVMR-Lasso²²⁰. An overview of this investigation is presented in Figure 4.2.1.

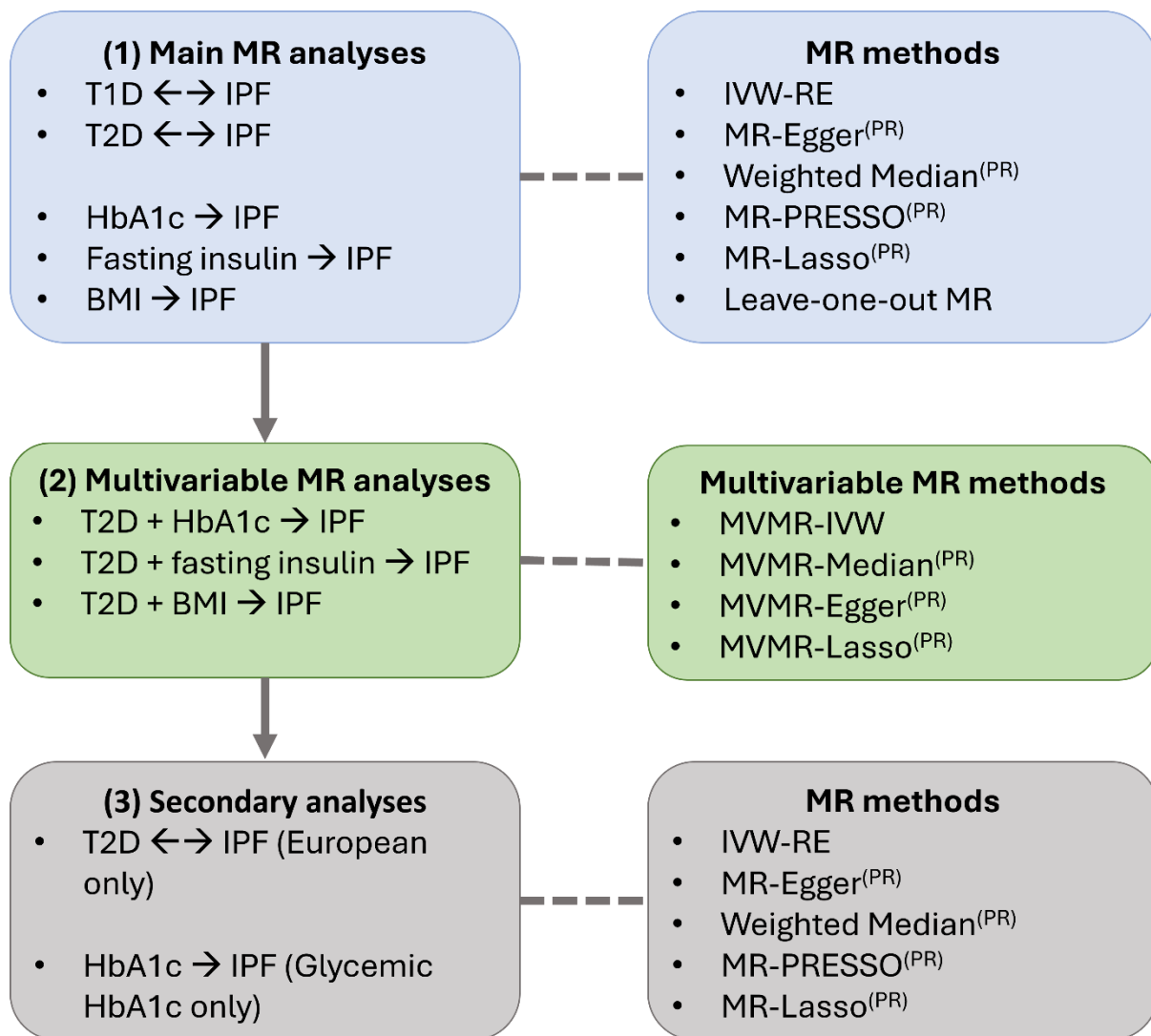


Figure 4.2.1: Overview of Mendelian randomisation study assessing if causal relationships exist between diabetes mellitus and idiopathic pulmonary fibrosis. (1) Main MR analyses included bidirectional univariable analyses of T1D and T2D on IPF susceptibility. Unidirectional univariable analyses of T2D-associated factors HbA1c, fasting insulin, and BMI on IPF susceptibility were also performed. (2) Multivariable MR analyses included analyses of T2D and associated factors HbA1c, fasting insulin, and BMI on IPF susceptibility to assess for possible independent effects. (3) Secondary analyses included a bidirectional univariable analysis of T2D on IPF susceptibility using a European-only T2D GWAS. A unidirectional analysis of HbA1c on IPF was also performed using only glycemic HbA1c SNPs as genetic instruments. (PR) indicates pleiotropy-robust MR methods.

4.3 Results

4.3.1 Genetic instrument selection

A total of 35 SNPs met inclusion criteria for use in the exposure instrument and were present in the outcome dataset (or had available proxies) for the MR analysis of T1D on IPF susceptibility. For other analyses with IPF as an outcome variable, genetic exposure instruments were selected for T2D (118 SNPs), HbA1c (37 SNPs), fasting insulin (36 SNPs), and BMI (131 SNPs).

For analyses with IPF as an exposure variable, of the 19 IPF risk SNPs five were available for T1D and 10 SNPs were available in T2D GWAS datasets.

In the European ancestry-only secondary analysis of T2D on IPF, 16 variants were available in both GWAS datasets for the T2D exposure instrument. All 19 IPF risk SNPs were present in the T2D GWAS dataset and were included in the IPF exposure instrument. For the secondary MR analysis assessing causal relationships with IPF risk using only glycemic HbA1c SNPs, 12 SNPs were selected.

For each tested exposure, all SNPs were sufficiently strong instruments ($F > 10$).

4.3.2 Mendelian randomisation analysis results: Main analyses

4.3.2.1 Bidirectional MR analysis of T1D and IPF susceptibility

Bidirectional MR analyses of T1D and IPF susceptibility did not suggest that T1D causes IPF (IVW-RE, odds ratio (OR) = 1.00, 95% confidence interval (95% CI): 0.93-1.07, p-value (p) = 0.902) (Figure 4.3.1) or that IPF is likely to cause T1D (IVW-RE, OR = 1.11, 95% CI: 0.95-1.28, p = 0.184) (Figure 4.3.2). Causal estimates from pleiotropy-robust analyses were consistent with those calculated using IVW-RE and similarly did not suggest any causal effects.

Results suggested significant heterogeneity was present in IVW-RE MR analyses of T1D on IPF (Cochran's Q (Q) = 65.5, Q -statistic p -value (p) = 9.00×10^{-4} , I^2 = 48.10%) and IPF on T1D (Q = 21.45, p = 3.00×10^{-4} , I^2 = 81.40%). Removal of likely pleiotropic SNPs using MR-Lasso reduced heterogeneity in MR analyses of T1D on IPF (Q = 33.22, p = 0.36, I^2 = 6.70%) and of IPF on T1D (Q = 2.82, p = 0.42, I^2 = 0.00%), however MR estimates remained null. Heterogeneity in IVW-RE analyses is therefore unlikely to be masking a true effect.

Leave-one-out MR results also did not suggest that any individual SNPs are driving or distorting a true causal effect between T1D and IPF (Figure 4.3.3), or between IPF and T1D (Figure 4.3.4).

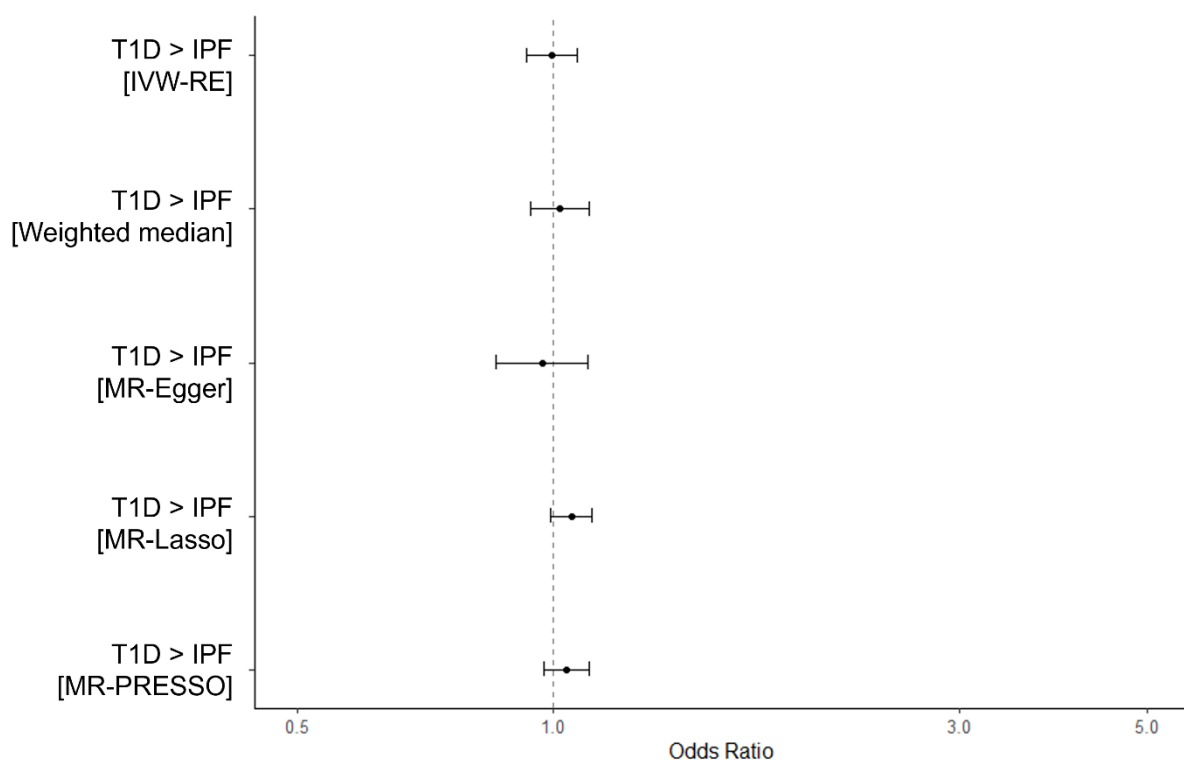


Figure 4.3.1: Causal estimates from MR analyses of T1D on IPF using IVW-RE and pleiotropy-robust methods. Error bars show 95% confidence intervals of MR estimates.

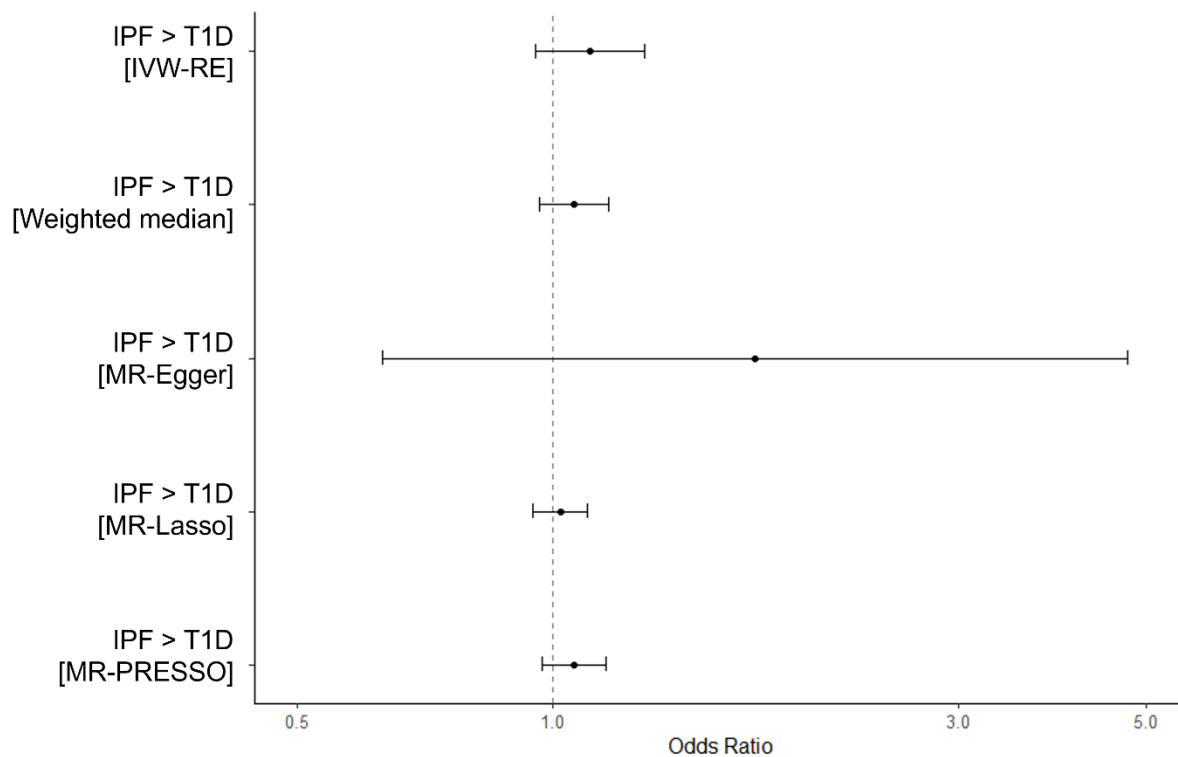


Figure 4.3.2: Causal estimates from MR analyses of IPF on T1D using IVW-RE and pleiotropy-robust methods. Error bars show 95% confidence intervals of MR estimates.

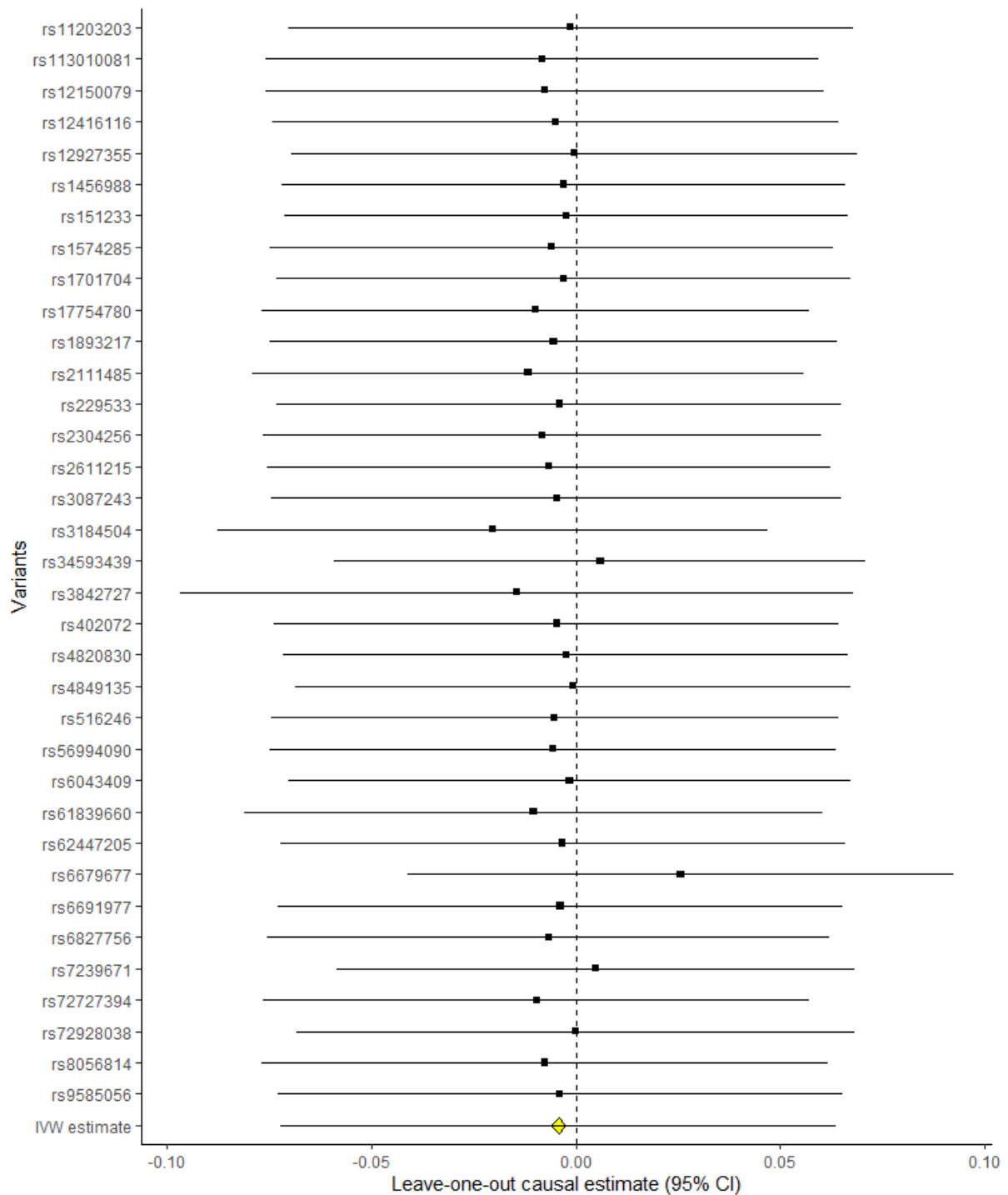


Figure 4.3.3: Estimates from a SNP leave-one-out MR analysis of T1D on IPF. The SNP removed from each MR analysis is indicated on the y-axis by the genetic variant RSID identifier. Effect estimates are presented as beta values ($\log(\text{OR})$) with error bars indicating 95% confidence intervals.

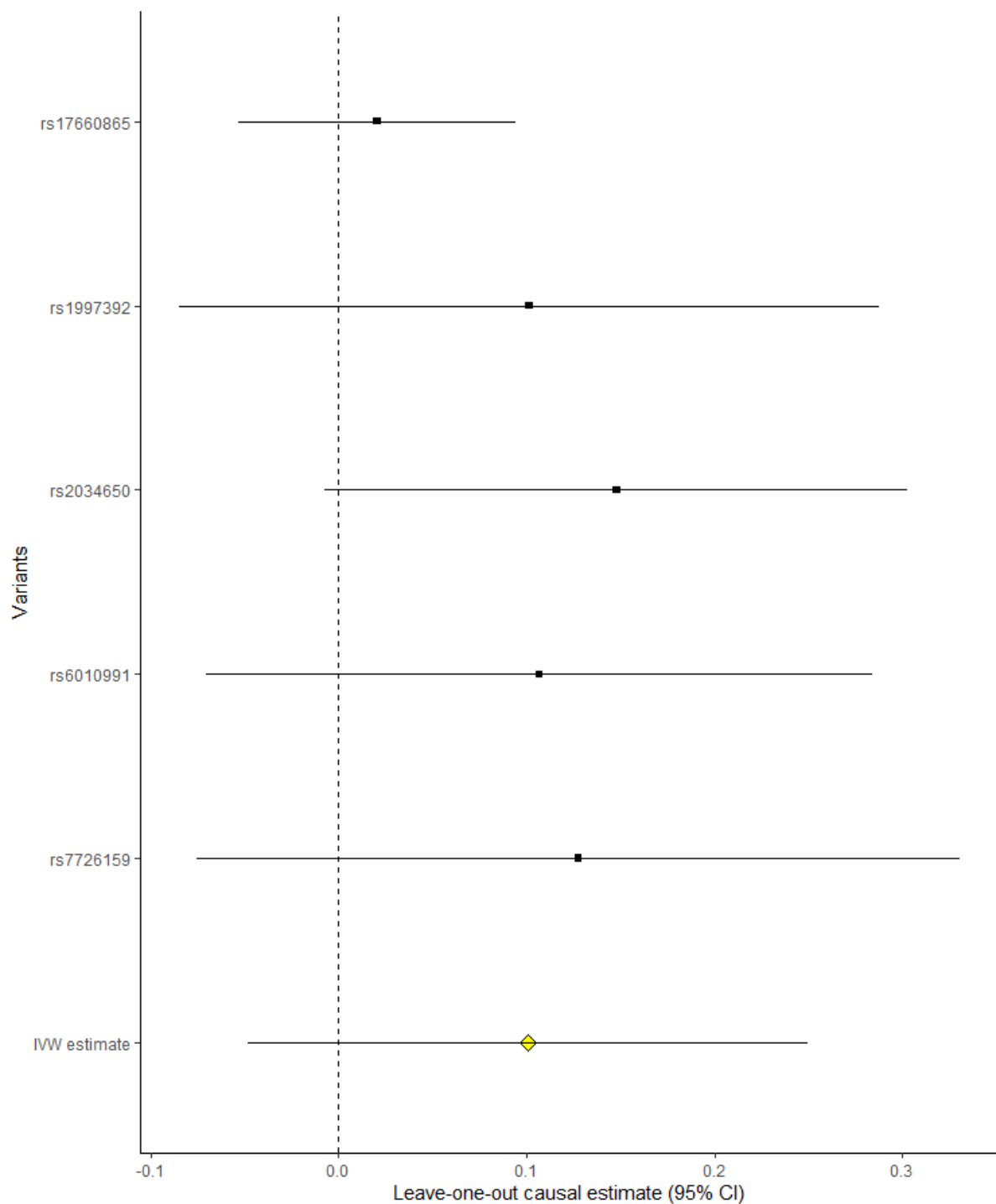


Figure 4.3.4: Estimates from a SNP leave-one-out MR analysis of IPF on T1D. The SNP removed from each MR analysis is indicated on the y-axis by the genetic variant RSID identifier. Effect estimates are presented as beta values ($\log(\text{OR})$) with error bars indicating 95% confidence intervals.

4.3.2.2 Bidirectional MR analysis of T2D and IPF susceptibility

Bidirectional MR analysis of T2D and IPF susceptibility suggested no evidence that T2D causes IPF (IVW-RE, $OR = 1.02$, 95% CI : 0.93-1.11, $p = 0.69$) (Figure 4.3.5) or that IPF causes T2D (IVW-RE, $OR = 1.00$, 95% CI : 0.97-1.03, $p = 0.80$) (Figure 4.3.6). Causal estimates from pleiotropy-robust analyses were consistent with those of IVW-RE.

Significant heterogeneity was present in IVW-RE analyses of T2D on IPF ($Q = 148.43$, $p = 0.03$, $I^2 = 21.2\%$) and IPF and T2D ($Q = 20.51$, $p = 0.02$, $I^2 = 56.1\%$). Removal of likely pleiotropic SNPs using MR-Lasso and MR-PRESSO resulted in reduced heterogeneity in analyses of T2D on IPF (MR-Lasso, $Q = 105.43$, $p = 0.61$, $I^2 = 0.00\%$) and IPF on T2D (MR-PRESSO, $Q = 5.89$, $p = 0.66$, $I^2 = 0.00\%$), however MR estimates remained null. Heterogeneity in IVW-RE analyses is therefore unlikely to be masking a true effect.

Results of leave-one-out MR analyses did not suggest that any individual SNPs are driving or distorting a true causal effect of T2D on IPF (Figure 4.3.7), or of IPF on T2D (Figure 4.3.8).

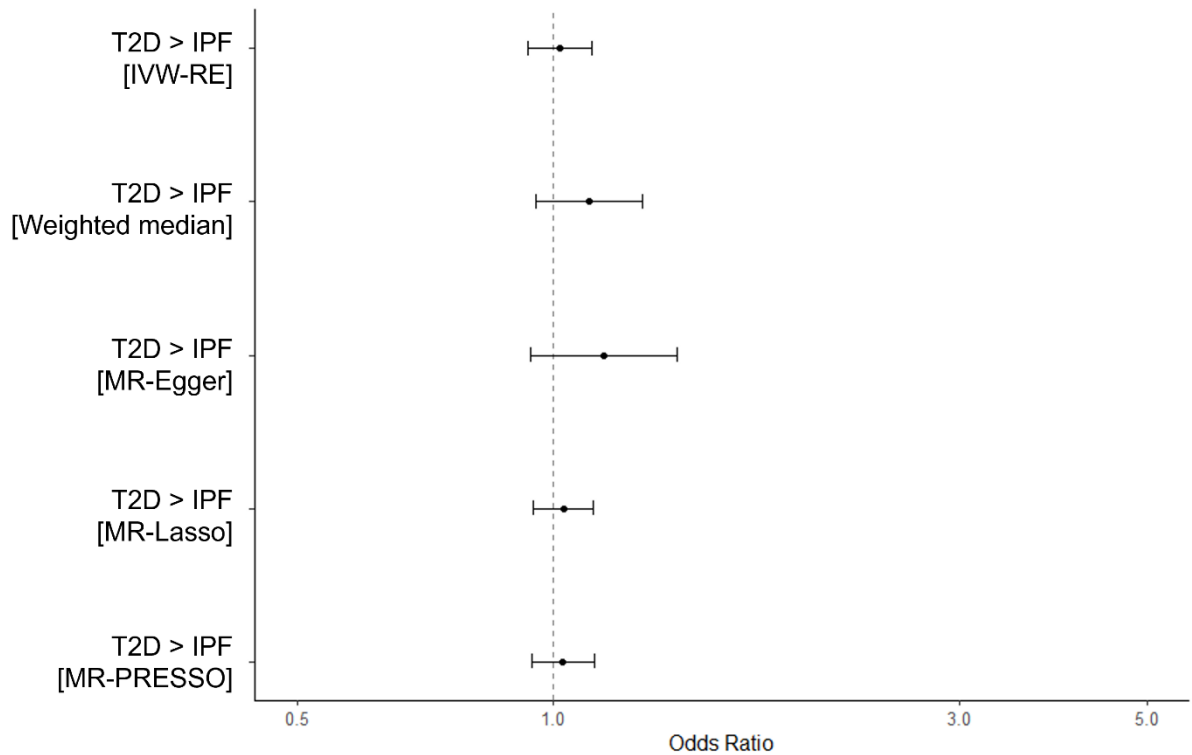


Figure 4.3.5: Causal estimates from MR analyses of T2D on IPF using IVW-RE and pleiotropy-robust methods. Error bars show 95% confidence intervals of MR estimates.

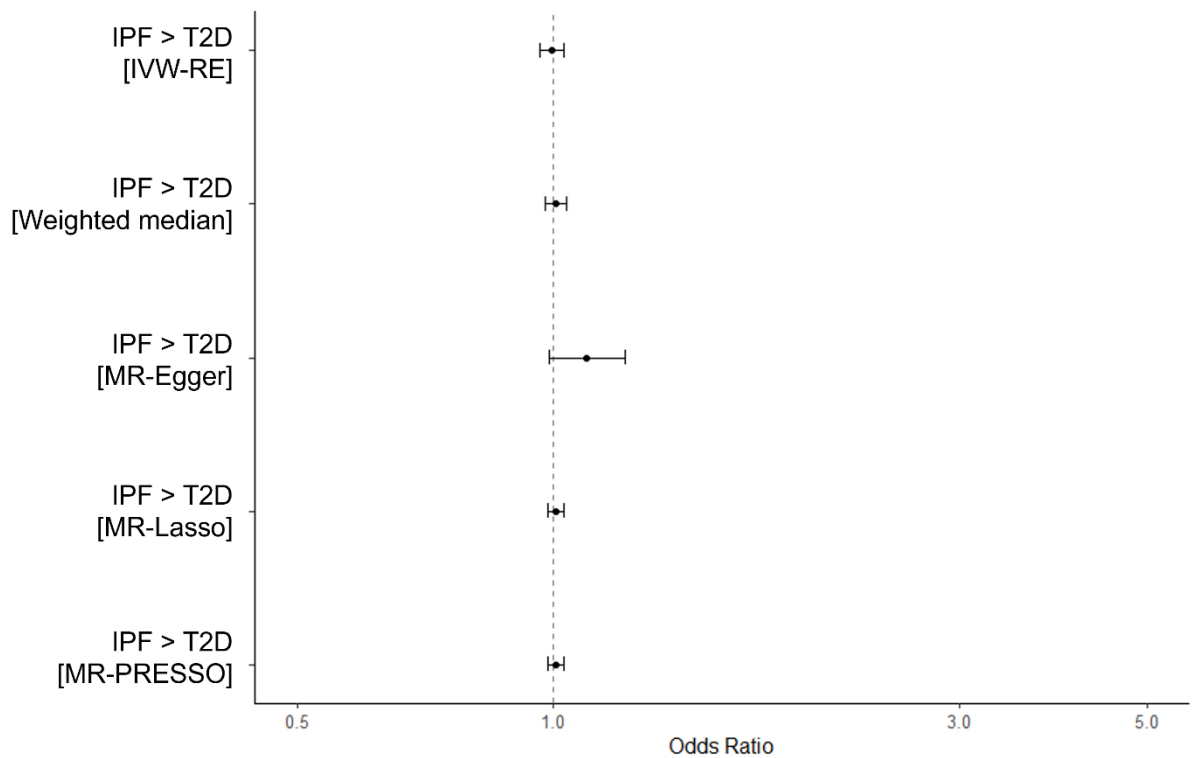


Figure 4.3.6: Causal estimates from MR analyses of IPF on T2D using IVW-RE and pleiotropy-robust methods. Error bars show 95% confidence intervals of MR estimates.

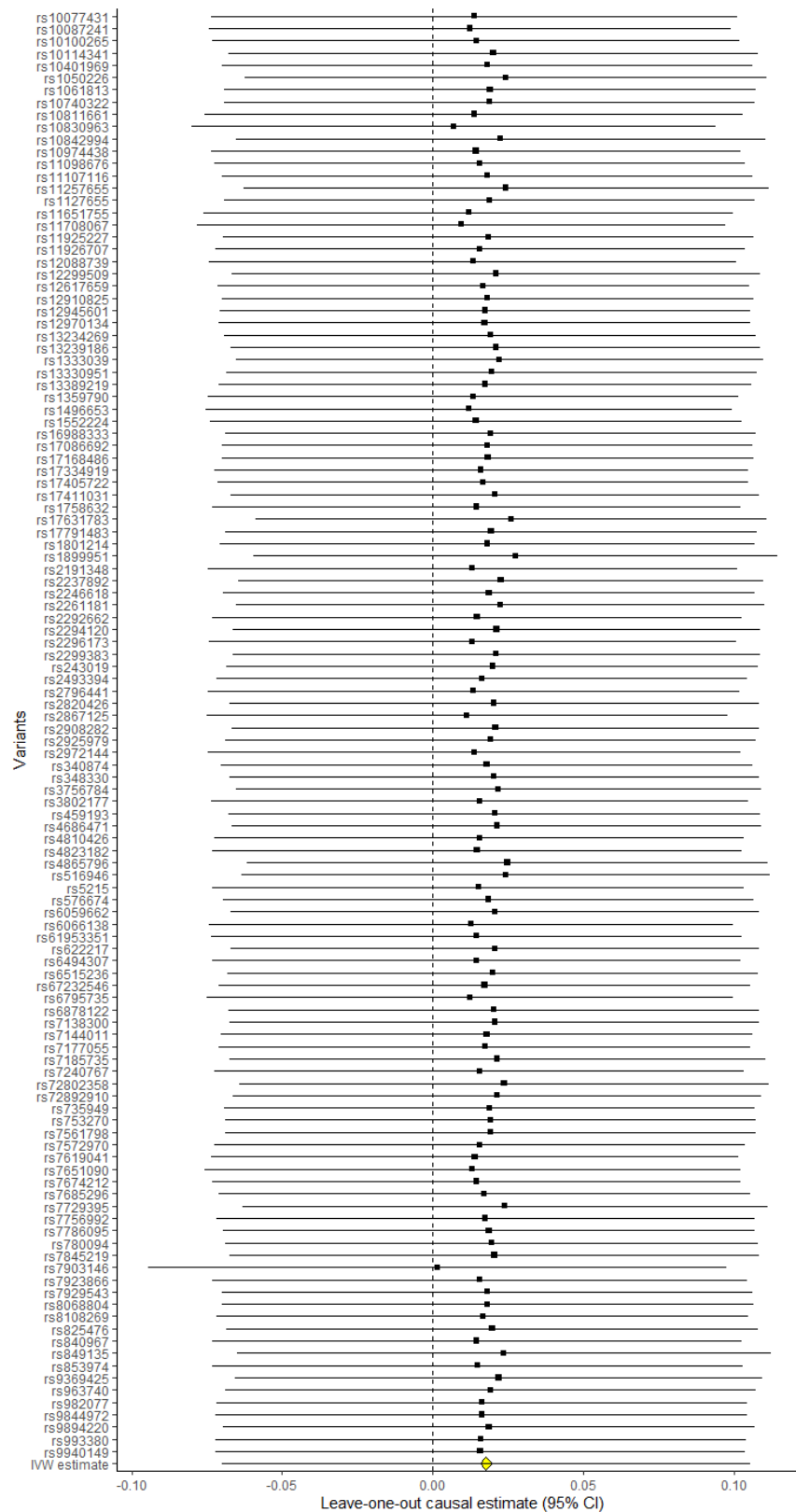


Figure 4.3.7: Estimates from a SNP leave-one-out MR analysis of T2D on IPF. The SNP removed from each MR analysis is indicated on the y-axis by the genetic variant RSID identifier. Effect estimates are presented as beta values ($\log(\text{OR})$) with error bars indicating 95% confidence intervals.

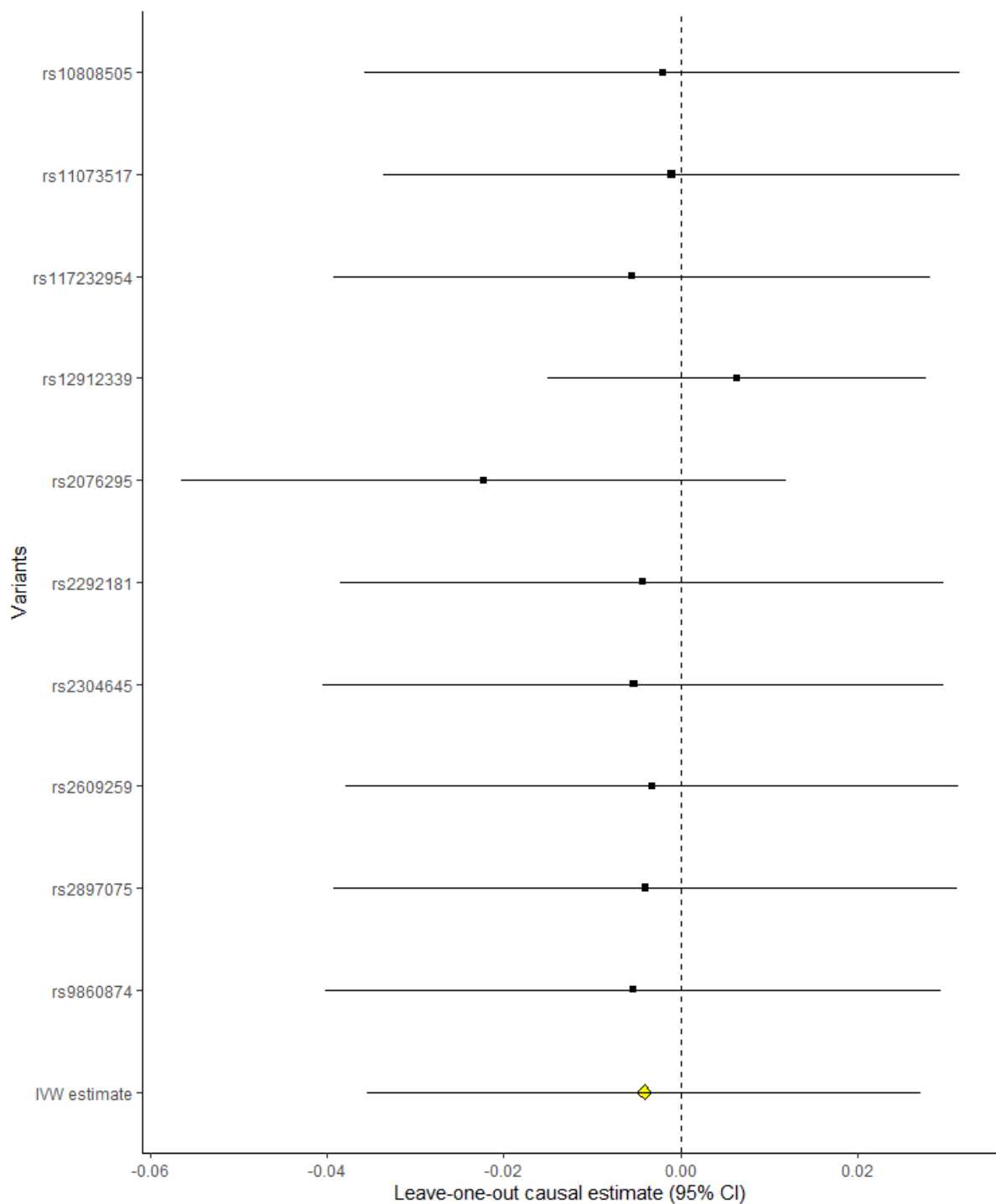


Figure 4.3.8: Estimates from a SNP leave-one-out MR analysis of IPF on T2D. The SNP removed from each MR analysis is indicated on the y-axis by the genetic variant RSID identifier. Effect estimates are presented as beta values ($\log(\text{OR})$) with error bars indicating 95% confidence intervals.

4.3.2.3 MR analysis of HbA1c on IPF susceptibility

Results from the MR analysis of HbA1c level on IPF susceptibility did not suggest that HbA1c is causally associated with IPF (IVW-RE, OR = 1.19, 95% CI: 0.63-2.22, $p = 0.592$). Causal estimates from pleiotropy robust analyses were concordant with those calculated using IVW-RE (Figure 4.3.9). Results did not indicate that significant heterogeneity was present in the IVW-RE analysis ($Q = 47.12, p = 0.10, I^2 = 23.60\%$). Leave-one-out MR analyses also did not suggest that any individual SNPs are likely to be driving or distorting a true causal effect of HbA1c on IPF susceptibility (Figure 4.3.10).

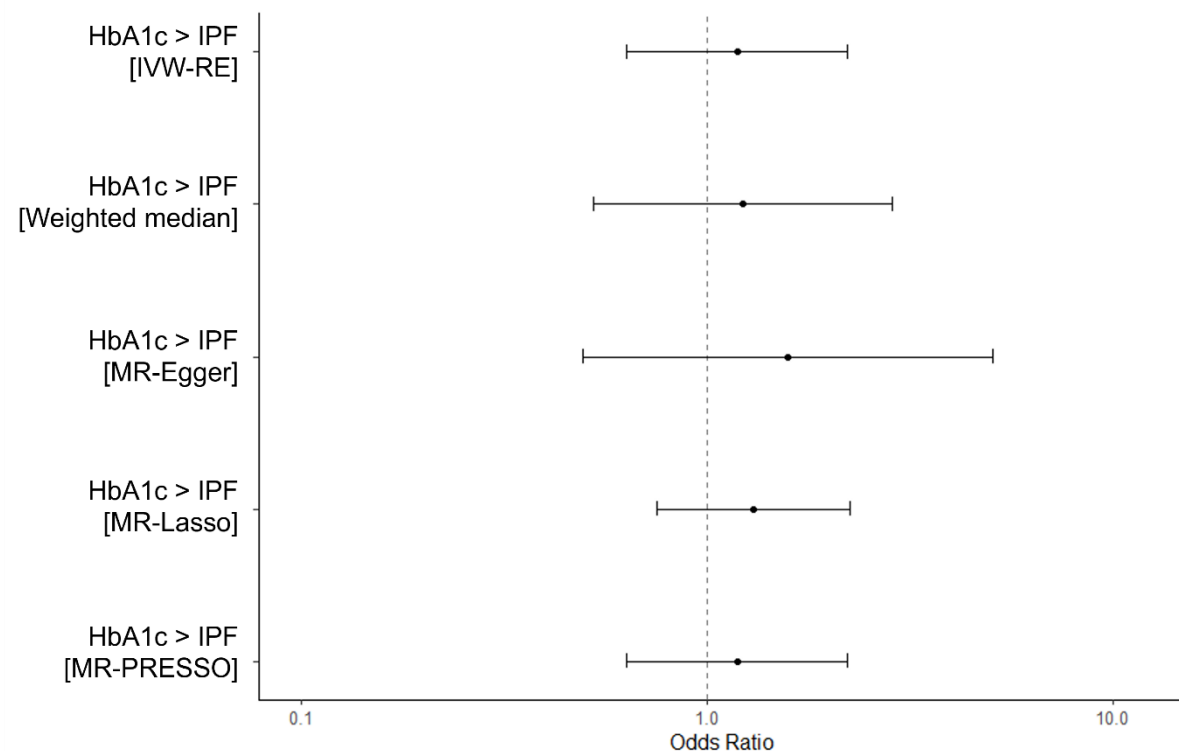


Figure 4.3.9: Causal estimates from MR analyses of HbA1c on IPF using IVW-RE and pleiotropy-robust methods. Error bars show 95% confidence intervals of MR estimates.

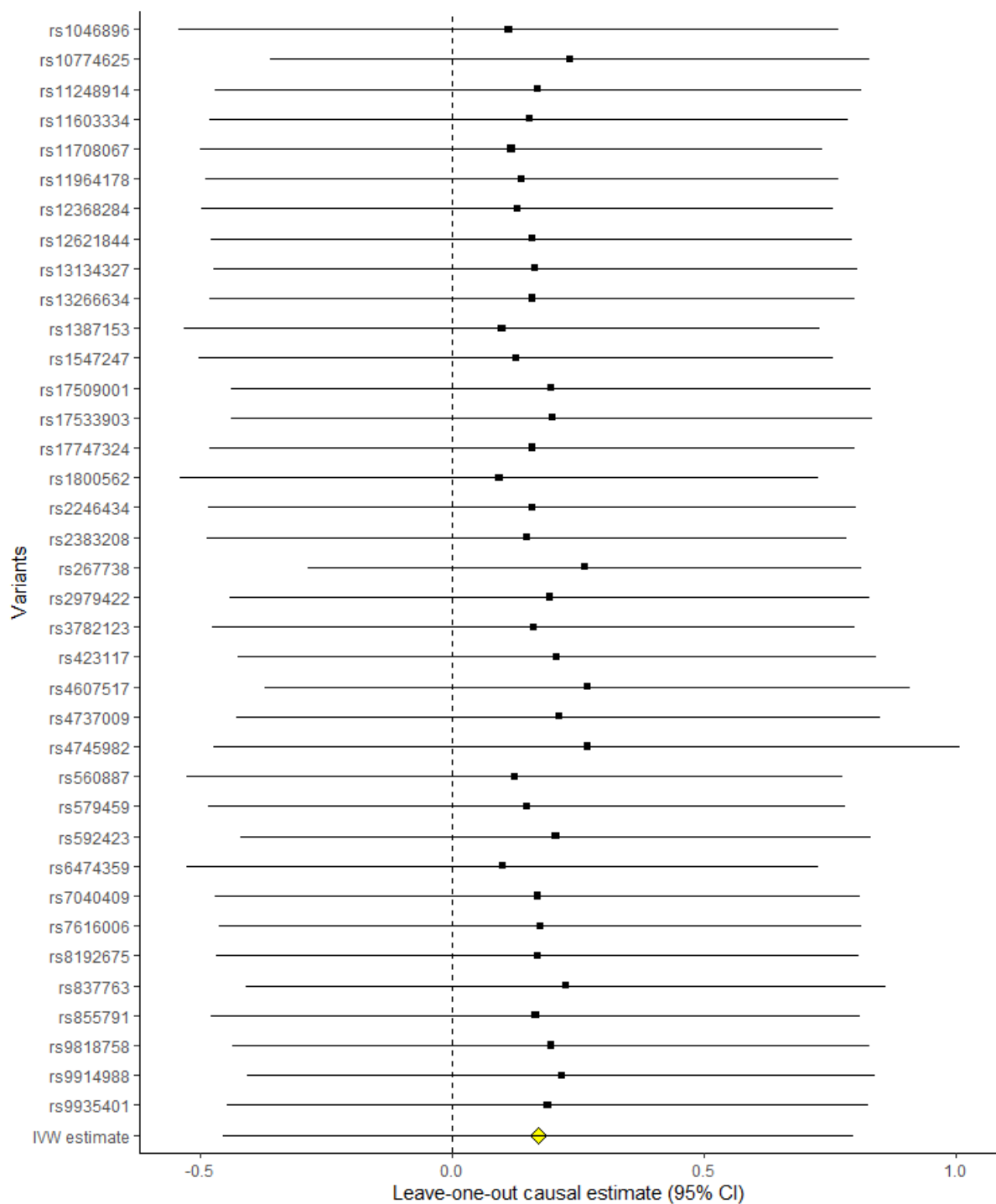


Figure 4.3.10: Estimates from a SNP leave-one-out MR analysis of HbA1c on IPF. The SNP removed from each MR analysis is indicated on the y-axis by the genetic variant RSID identifier. Effect estimates are presented as beta values ($\log(\text{OR})$) with error bars indicating 95% confidence intervals.

4.3.2.4 MR analysis of fasting insulin on IPF susceptibility

Results of the MR analysis of fasting insulin level on IPF susceptibility did not suggest that fasting insulin is causally associated with IPF (IVW-RE, $OR = 0.60$, 95% CI : 0.31-1.15, $p = 0.124$). Results from pleiotropy-robust MR analyses did not suggest any difference in outcomes compared to IVW-RE (Figure 4.3.11). Results indicate that heterogeneity was low in the main IVW-RE analysis ($Q = 35.72$, $p = 0.43$, $I^2 = 2.00\%$) and leave-one-out MR did not identify any individual SNPs that are likely to be distorting a true causal effect of fasting insulin on IPF (Figure 4.3.12).

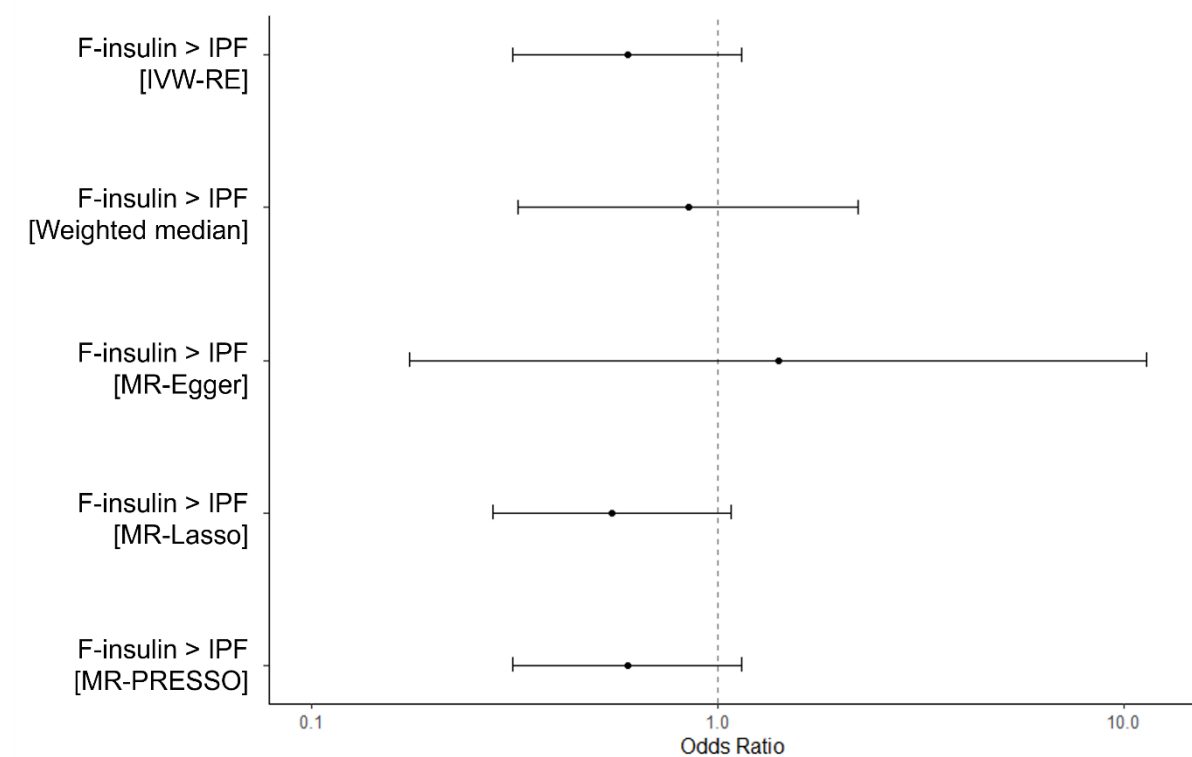


Figure 4.3.11: Causal estimates from MR analyses of fasting insulin level on IPF using IVW-RE and pleiotropy-robust methods. Error bars show 95% confidence intervals of MR estimates.

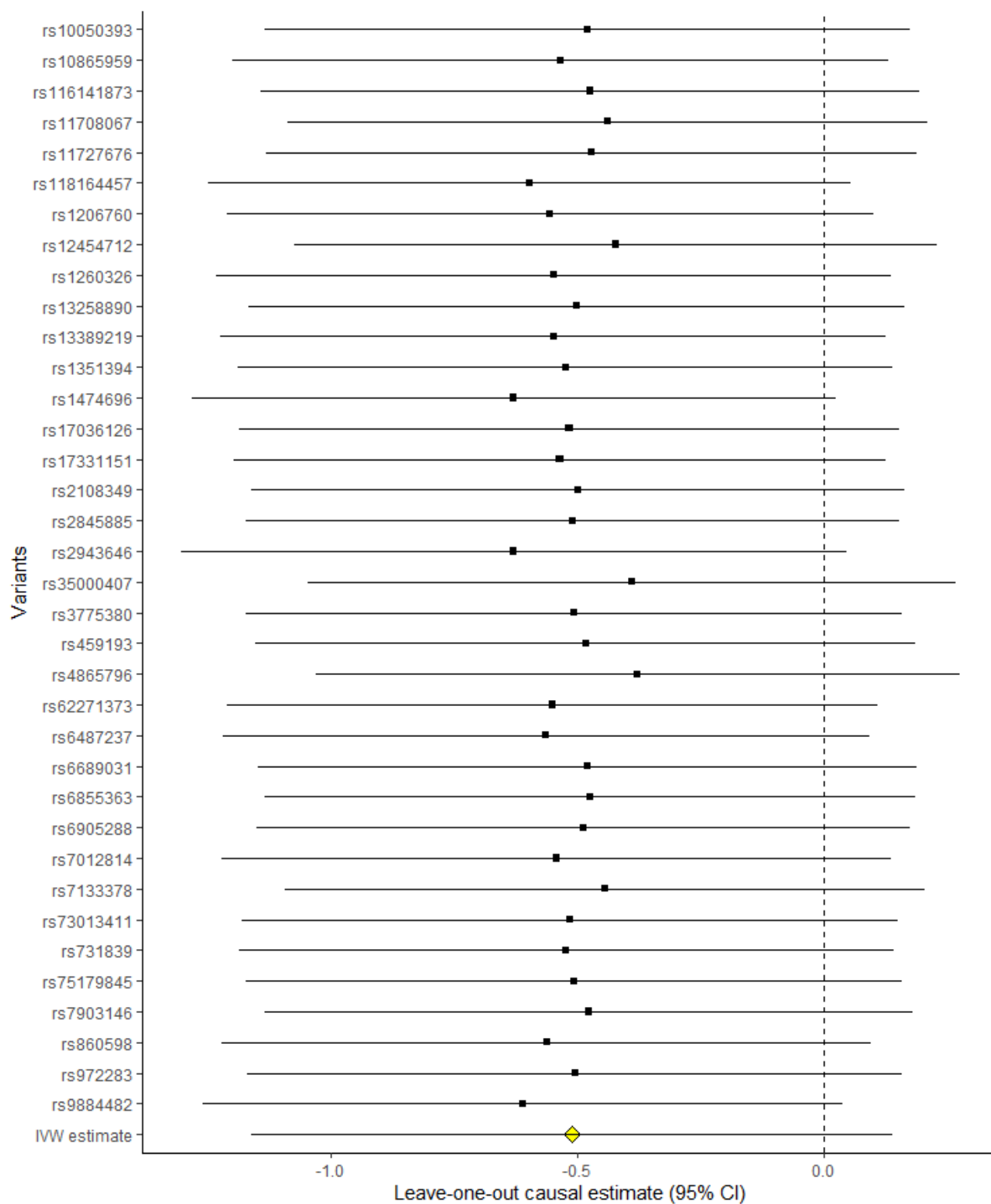


Figure 4.3.12: Estimates from a SNP leave-one-out MR analysis of fasting insulin level on IPF. The SNP removed from each MR analysis is indicated on the y-axis by the genetic variant RSID identifier. Effect estimates are presented as beta values (log(OR)) with error bars indicating 95% confidence intervals.

4.3.2.5 MR analysis of BMI on IPF susceptibility

MR analysis using the IVW-RE method indicated that BMI may be causally associated with IPF risk (IVW-RE, $OR = 1.44$, 95% CI : 1.12-1.85, $p = 4.00 \times 10^{-3}$). This outcome was supported by pleiotropy-robust analyses using MR-Lasso ($OR = 1.35$, 95% CI : 1.07-1.71, $p = 0.01$) and no invalid instruments were identified using MR-PRESSO. While MR estimates derived from weighted median and MR-Egger approaches were not statistically significant, effect directions were concordant with IVW-RE results (weighted median, $OR = 1.26$, 95% CI : 0.87-1.81, $p = 0.23$) (MR-Egger, $OR = 1.38$, 95% CI : 0.71-2.70, $p = 0.342$) (Figure 4.3.13). No significant heterogeneity was suggested by IVW-RE analysis results ($Q = 152.13$, $p = 0.09$, $I^2 = 14.50\%$). Leave-one-out MR also did not suggest that the positive effect indicated of BMI on IPF susceptibility is driven by any individual SNP (Figure 4.3.14).

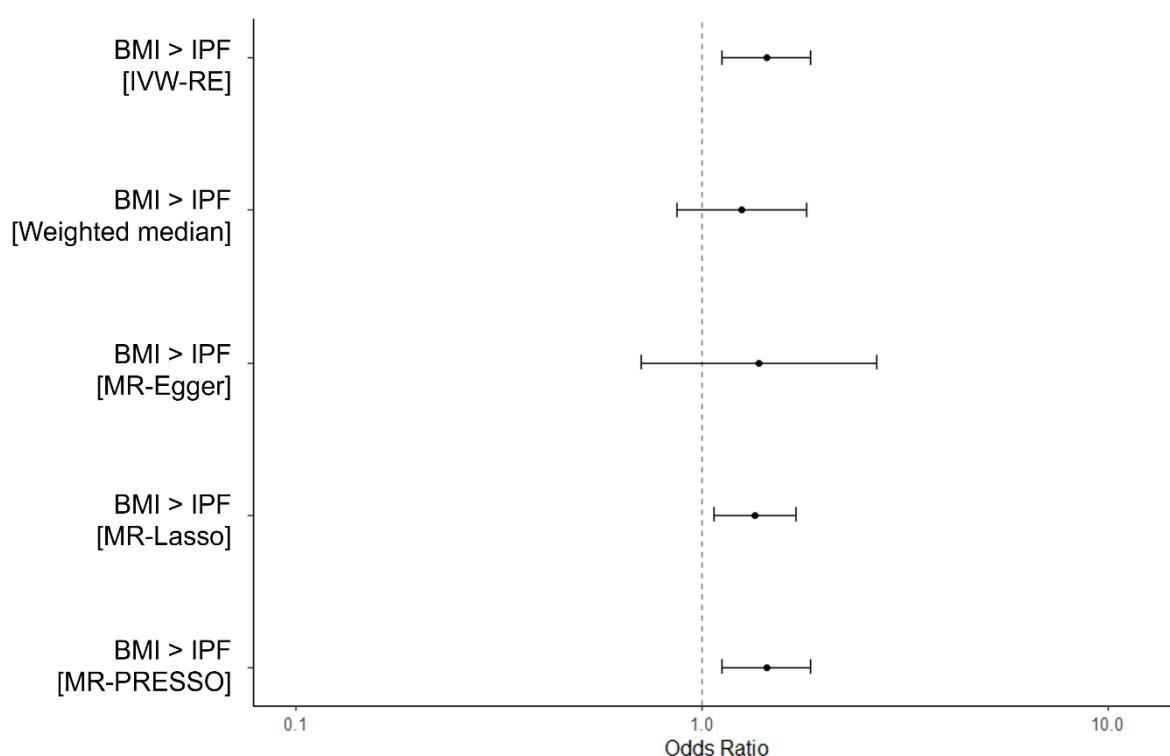


Figure 4.3.13: Causal estimates from MR analyses of BMI on IPF using IVW-RE and pleiotropy-robust methods. Error bars show 95% confidence intervals of MR estimates.

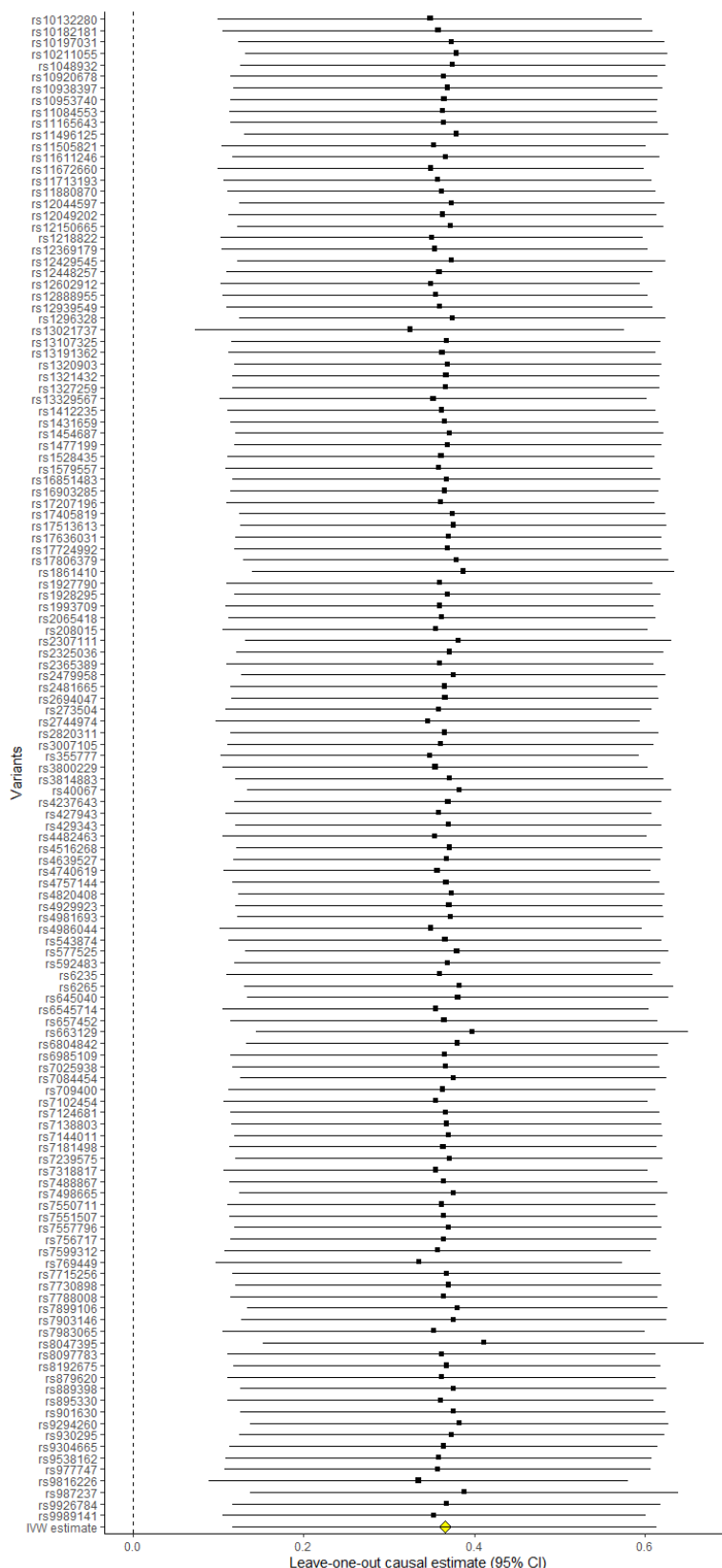


Figure 4.3.14: Estimates from a SNP leave-one-out MR analysis of BMI on IPF. The SNP removed from each MR analysis is indicated on the y-axis by the genetic variant RSID identifier. Effect estimates are presented as beta values ($\log(\text{OR})$) with error bars indicating 95% confidence intervals.

4.3.3 Mendelian randomisation analysis results: Multivariable analyses

4.3.3.1 Multivariable MR analysis of HbA1c and T2D on IPF susceptibility

Results of multivariable MR of HbA1c and T2D exposures on an IPF outcome did not suggest any independent causal effects of either HbA1c (MVMR-IVW, $OR = 1.14$, 95% CI : 0.52-2.52, $p = 0.75$) or T2D (MVMR-IVW, $OR = 1.03$, 95% CI : 0.92-1.15, $p = 0.60$) on IPF susceptibility. MR estimates calculated using pleiotropy-robust methods were concordant with those of MVMR-IVW (Figure 4.3.15).

Results suggest that significant heterogeneity was present in the MVMR-IVW analysis of T2D and HbA1c on IPF susceptibility ($Q = 154.31$, $p = 0.03$, $I^2 = 20.94\%$). Following removal of likely pleiotropic SNPs using MVMR-Lasso, heterogeneity statistics were reduced ($Q = 107.24$, $p = 0.71$, $I^2 = 0.00\%$) but MR estimates remained null. Heterogeneity in the MVMR-IVW analysis is therefore unlikely to be masking a true effect.

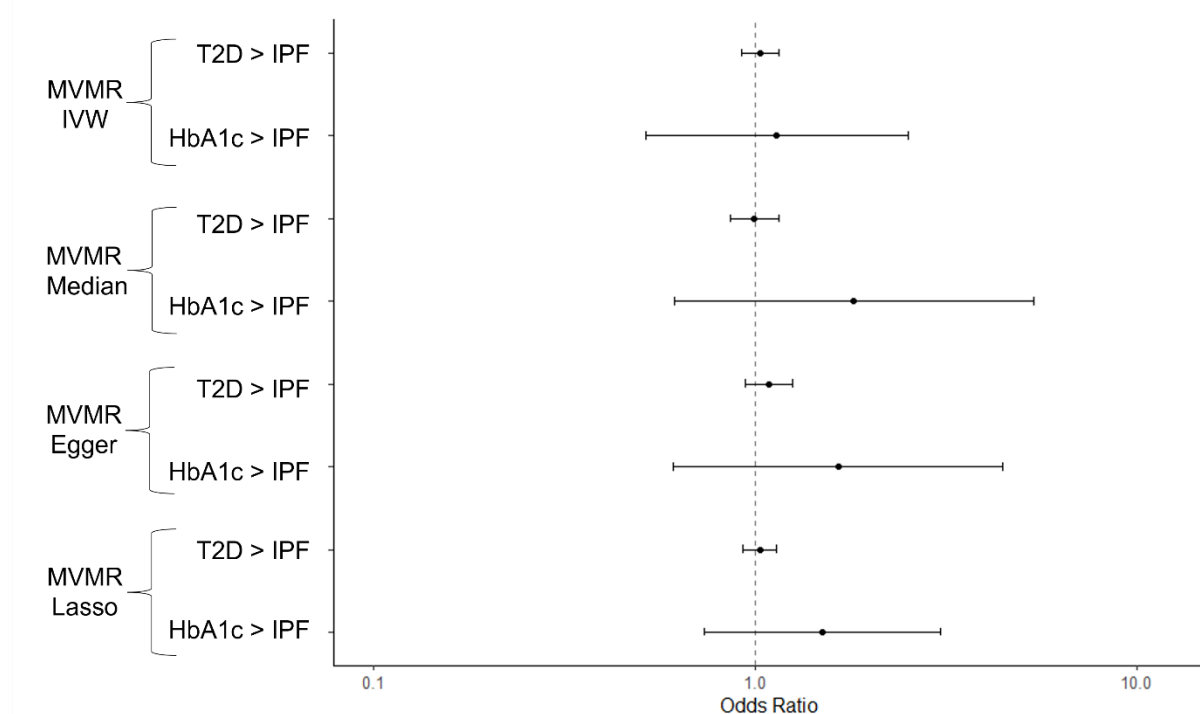


Figure 4.3.15: Causal estimates from multivariable MR analyses of HbA1c and T2D exposures on IPF using MVMR-IVW and pleiotropy-robust methods. Error bars show 95% confidence intervals of MR estimates.

4.3.3.2 Multivariable MR analysis of fasting insulin and T2D on IPF susceptibility

Results from multivariable MR of fasting insulin and T2D exposures on IPF did not indicate that fasting insulin (MVMR-IVW, $OR = 0.69$, 95% CI : 0.36-1.36, $p = 0.29$) or T2D (MVMR-IVW, $OR = 1.02$, 95% CI : 0.94-1.12, $p = 0.61$) have direct causal effects on IPF risk. Pleiotropy-robust methods calculated estimates that were consistent with outcomes from MVMR-IVW (Figure 4.3.16). Results suggest that no significant heterogeneity was present in the MVMR-IVW analysis ($Q = 142.70$, $p = 0.11$, $I^2 = 13.80\%$).

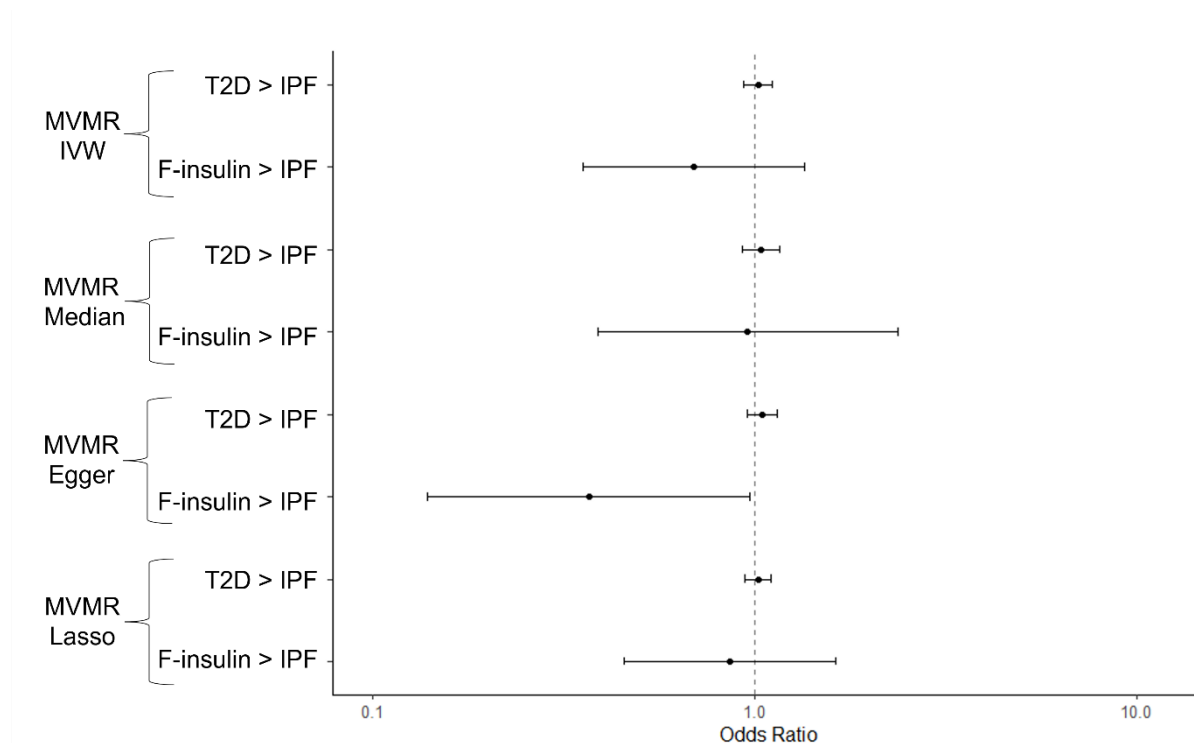


Figure 4.3.16: Causal estimates from multivariable MR analyses of fasting insulin level and T2D exposures on IPF using MVMR-IVW and pleiotropy-robust methods. Error bars show 95% confidence intervals of MR estimates.

4.3.3.3 Multivariable MR analysis of BMI and T2D on IPF susceptibility

MR results from the multivariable analysis of BMI and T2D exposures on IPF did not indicate that BMI (MVMR-IVW, $OR = 1.00$, 95% CI : 0.94-1.09, $p = 0.82$) or T2D (MVMR-IVW, $OR = 0.91$, 95% CI : 0.68-1.21, $p = 0.51$) are independently causally associated with IPF

susceptibility. Results of pleiotropy-robust analyses were concordant with MVMR-IVW in suggesting a null effect on IPF susceptibility (Figure 4.3.17). In addition, results did not suggest that significant heterogeneity was present in the MVMR-IVW analysis ($Q = 208.10$, $p = 0.09$, $I^2 = 12.54\%$).

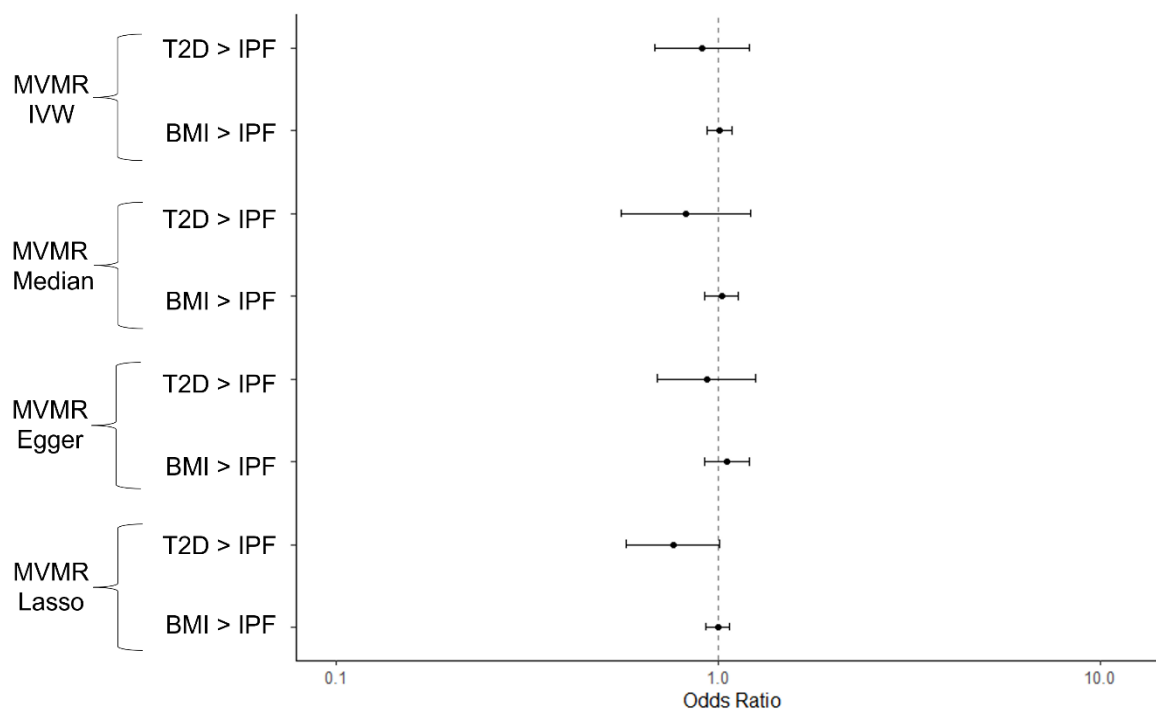


Figure 4.3.17: Causal estimates from multivariable MR analyses of BMI and T2D exposures on IPF using MVMR-IVW and pleiotropy-robust methods. Error bars show 95% confidence intervals of MR estimates.

4.3.4 Mendelian randomisation analysis results: Secondary analyses

4.3.4.1 Bidirectional MR analysis of T2D and IPF susceptibility with European ancestry GWAS

Bidirectional MR analysis of T2D and IPF susceptibility did not suggest that T2D causes IPF (IVW-RE, OR = 0.99, 95% CI: 0.89-1.11, $p = 0.92$) or that IPF causes T2D (IVW-RE, OR = 0.97, 95% CI: 0.93-1.01, $p = 0.176$). Results of pleiotropy-robust analyses also consistently suggested null effects of T2D on IPF (Figure 4.3.18) and T2D on IPF (Figure 4.3.19). Results did not suggest that significant heterogeneity was present in IVW-RE analyses assessing for

causal effects of T2D on IPF ($Q = 18.59, p = 0.23, I^2 = 19.30\%$) or IPF on T2D ($Q = 26.91, p = 0.08, I^2 = 33.10\%$). The results of these secondary analyses with European ancestry GWAS data were therefore concordant with main univariable MR analyses of T2D and IPF.

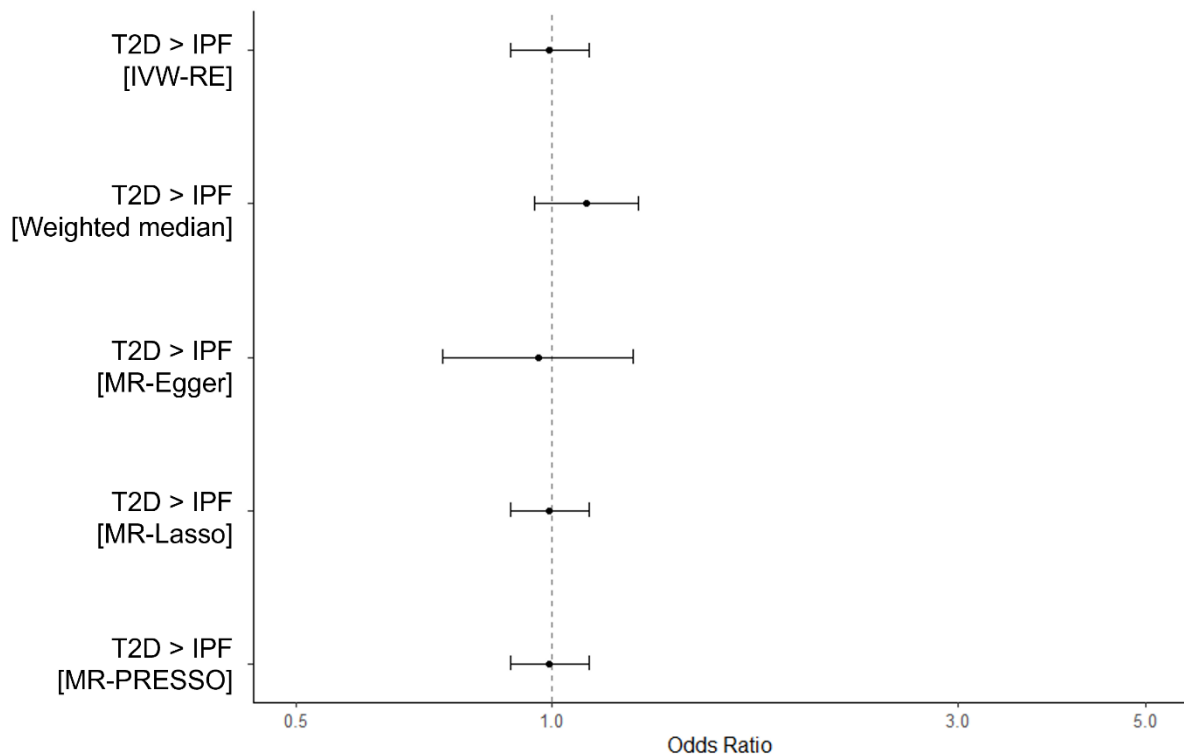


Figure 4.3.18: Causal estimates from MR analyses of T2D (European ancestry GWAS) on IPF using IVW-RE and pleiotropy-robust methods. Error bars show 95% confidence intervals of MR estimates.

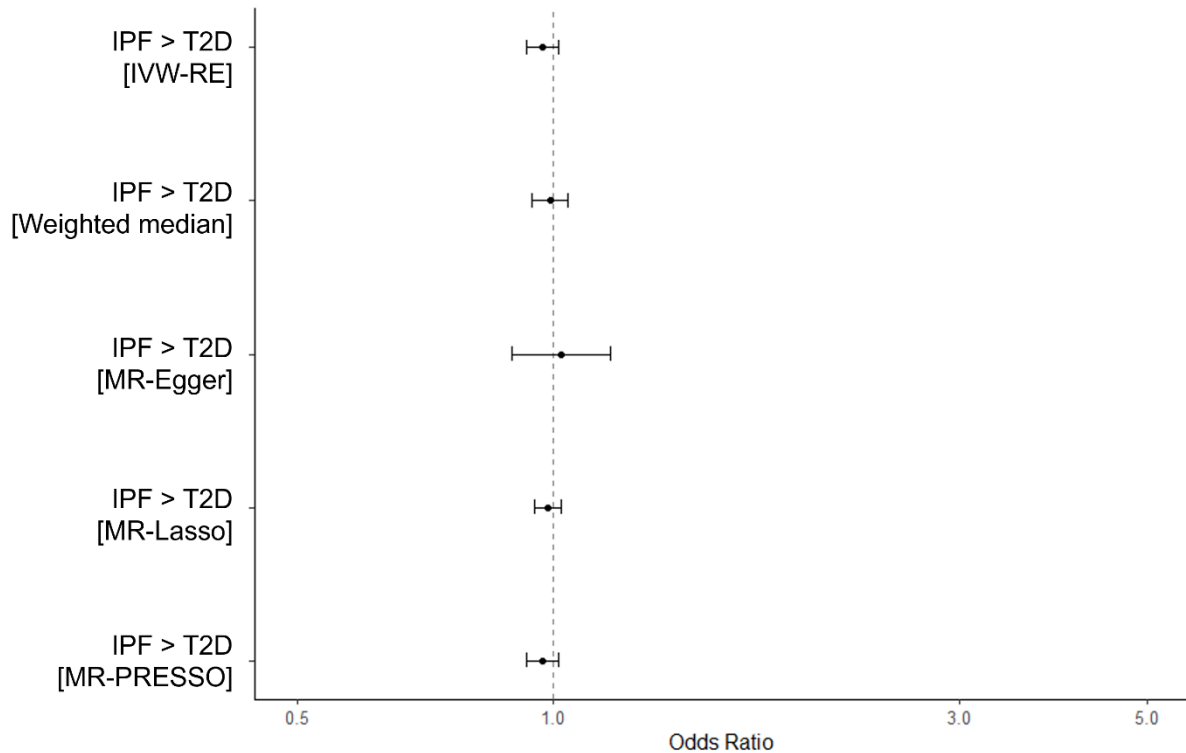


Figure 4.3.19: Causal estimates from MR analyses of IPF (all 19 susceptibility SNPs) on T2D (European ancestry GWAS) using IVW-RE and pleiotropy-robust methods. Error bars show 95% confidence intervals of MR estimates.

4.3.4.2 MR analysis of HbA1c on IPF susceptibility using only glycemic HbA1c SNPs

Results of an IVW-RE MR analysis assessing for causal effects of HbA1c on IPF susceptibility using only functionally glycemic SNPs did not suggest that HbA1c is causally associated with IPF ($OR = 2.78$, 95% CI : 0.97-8.02, $p = 0.06$). Results of pleiotropy-robust MR analyses were consistent in suggesting a null effect, but the MR estimate calculated using MR-Egger was negative ($OR = 0.37$, 95% CI : 0.02-7.34, $p = 0.517$) (Figure 4.3.20). No significant heterogeneity was indicated to be present in the IVW-RE analysis ($Q = 9.88$, $p = 0.54$, $I^2 = 0.00\%$). These results are generally concordant with MR analyses using the main HbA1c instrument (Section 4.3.2.3) in suggesting no causal relationship with IPF. Based on these results, it is therefore unlikely that effects of haemoglobin or red cell biology are greatly distorting main MR estimates of HbA1c on IPF susceptibility.

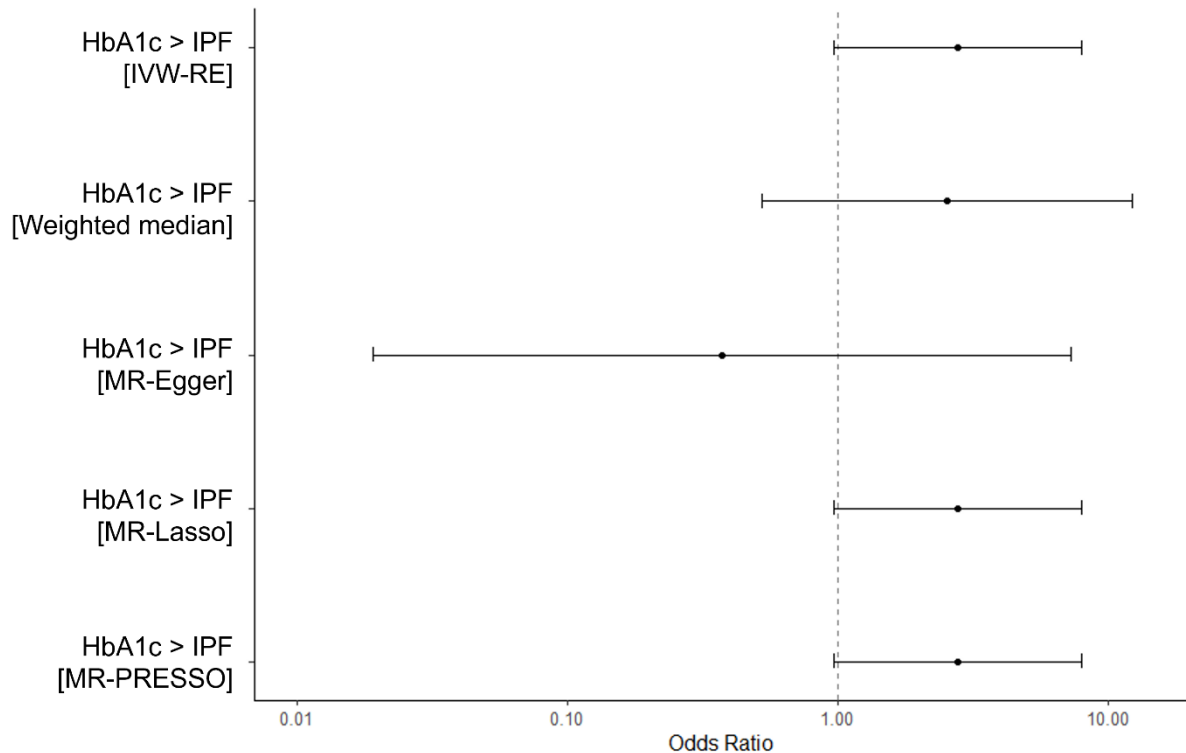


Figure 4.3.20: Causal estimates from MR analyses of HbA1c (glycemic SNPs only) on IPF using IVW-RE and pleiotropy-robust methods. Error bars show 95% confidence intervals of MR estimates.

4.3.5 Conclusions from Mendelian randomisation study assessing causal relationships between diabetes mellitus and idiopathic pulmonary fibrosis

Results indicate that DM and IPF are unlikely to be linked by direct causal effects. In addition, results suggest that HbA1c (a proxy for long term glucose levels) and fasting insulin levels are unlikely to be causally associated with IPF susceptibility. However, univariable analyses indicate a possible causal relationship between increased BMI and IPF risk.

4.4 Discussion

This research aimed to assess if the epidemiological association between DM and IPF is driven by causal effects. This was investigated using a bidirectional MR analysis of T1D and T2D with IPF susceptibility. Causal pathways between DM and IPF were investigated further through MR analyses of DM-associated continuous factors HbA1c, fasting insulin, and BMI. Finally, possible independent effects of T2D and associated factors on IPF susceptibility were investigated through multivariable MR. Overall, results did not suggest that DM is likely to cause IPF, or that IPF is likely to cause DM. Based on this study, DM-associated factors HbA1c and fasting insulin are also unlikely to cause IPF, however a positive effect of increased BMI on IPF susceptibility was indicated by univariable analyses.

It has been previously theorised that the association between DM and IPF is driven by DM-associated chronic hyperglycaemia contributing to fibrotic mechanisms⁴⁵. Null effects of T1D, T2D, and HbA1c on IPF susceptibility indicate that hyperglycaemia is unlikely to be on the causal pathway for IPF. Two recent MR studies^{222,223} have suggested that T1D may be causally linked to IPF. Despite using the same GWAS dataset for the T1D genetic instrument, MR results from this investigation do not support the positive effect on IPF risk suggested in these published studies. Neither published study used the 2022 IPF GWAS to instrument IPF risk, which is the largest IPF GWAS exclusively using cases diagnosed using ATS/ERS/JRS/ALAT guidelines^{89,90}. In addition, code-based IPF cases can be prone to misclassification and use in this context may result in a non-specific IPF phenotype²²⁴. This MR study is therefore more likely to represent a true result, suggesting that T1D does not cause IPF. However, results from published studies^{222,223,225} suggesting a null relationship between T2D and IPF aligns with the findings of this investigation.

Univariable MR results indicated that increased BMI (obesity) may be causally linked to increased IPF susceptibility, however this effect was not replicated in multivariable MR with T2D as a second exposure. IPF patients are typically normally nourished or obese, but

poorer disease outcomes are associated with lower BMI^{226,227}. BMI is a highly complex trait that is associated with a range of environmental factors²²⁸. While use of genetic instruments aims to evaluate causality without potential environmental confounding, obesity-associated gene-environment interactions with smoking behaviour, physical activity, and diet are well documented^{229,230}. The positive relationship indicated between BMI and IPF may therefore be confounded, particularly as there is a strong association between IPF and smoking²³¹. Overall, it remains unclear if obesity contributes to the association between IPF and DM, but results indicate a possible link between obesity and IPF that warrants further investigation.

As results of this investigation suggest that DM and IPF are unlikely to be directly causally linked, this comorbid relationship may instead be driven by upstream factors independently contributing to risk both traits. As an example, shortened telomeres are a strong risk factor for IPF⁷ and are associated with glycemic progression in T2D²³², increased HbA1c²³³, insulin resistance²³⁴, and obesity in elderly patients²³⁵. Pancreatic β -cells have been shown to have shorter telomeres in patients with T2D²³⁶. Dysregulated metabolic pathways may also contribute to both IPF and DM. In particular, the mTORC1 pathway is a key fibrogenic pathway¹¹⁰ that is upregulated in diabetic β -cells under increased metabolic stress^{237,238}.

Alternatively, in the absence of direct causal effects between DM and IPF, this comorbid relationship may be driven by environmental factors. T2D and IPF share risk factors, including smoking and increased age^{6,47}. In addition, Acute exacerbations in IPF are commonly treated with corticosteroids²³⁹, which may be contributing to treatment-induced hyperglycemia and DM²⁴⁰.

This research aimed to apply MR methods run an in-depth investigation into potential causal links between IPF and DM. Analyses using continuous DM-associated factors as exposures, subsequent multivariable analyses, and use of sensitivity analyses aimed to

enable robust assessment of any causal relationships. A main limitation of this overall approach is that DM and IPF are treated as single disease entities. It is possible that comorbid DM only exists when specific disease endotypes or pathobiological mechanisms interact. To investigate this possibility, the genetics of IPF patients with comorbid DM may need to be specifically studied.

Overall, the results of this investigation suggest that DM and IPF are unlikely to be directly linked by causality. Further research should focus on shared risk factors or treatment effects that may be driving the association of DM with IPF.

5: Investigating a potential causal link between Mosaic Loss of Chromosome Y and Idiopathic Pulmonary Fibrosis

5.1 Introduction

Somatic mutations can occur in human pluripotent and multipotent cell types and be passed along subsequent cell lineages. This can result in a “mosaic” pattern of cell populations that are genotypically different to the germline genome²⁴¹. Mosaic loss of the male-specific Y chromosome (mLOY) is a form of clonal mosaicism that can result due to the lower impact of Y chromosome loss on cell viability²⁴² relative to the loss of other chromosomes. It has also been hypothesised that mLOY may result in a selective advantage through loss of cell genes or variants involved in cell growth regulation, for example²⁴³. The rate of mLOY typically increases with age²⁴⁴, mLOY is associated with increased risk of multiple age-related diseases, including Alzheimer’s disease and non-haematological cancer²⁴², and mLOY is associated with all-cause mortality²⁴⁵. Overall, understanding relationships between mLOY and human disease has the potential to be highly relevant to biological research and clinical practice.

Two pseudoautosomal regions (PARs) in the human Y chromosome recombine with homologous regions on the X chromosome during cell division²⁴⁶. In population-level studies, mLOY is typically detected in blood leukocytes. Loss of Y chromosome can be measured by quantifying median genotyping intensities of the male-specific, non-PAR Y chromosomal regions^{247,248}. In more recent studies, mLOY has been estimated by identifying imbalance in genotyping intensities, or sequencing read depth, of heterozygous PAR regions between X (maternal) and Y (paternal) chromosomes^{243,249}. While it is possible to detect mLOY in blood leukocytes, mechanisms that are predicted to contribute to this

phenotype include disrupted cell cycle regulation, DNA damage response, and telomere biology^{242,243}. The extent of mLOY detected in blood is therefore likely to reflect a broader phenotype of genomic instability and is likely to be well correlated with mLOY in other human tissues.

Previous studies have indicated that mLOY may be a driver of, or linked to disease mechanisms in, both cardiac fibrosis²⁵⁰ and primary biliary cholangitis (previously named primary biliary cirrhosis)²⁵¹. Experimental investigation using single-cell RNA sequencing in Y chromosome-deficient mice has linked increased TGF β 1 and galectin-3 signaling to mLOY in cardiac fibrosis²⁵⁰. As both TGF β 1 and galectin-3 have also been shown to be potent drivers of fibrosis in IPF^{252,253}, this finding presents the potential that mLOY-driven mechanisms may also contribute to IPF pathogenesis. As IPF is also an age-related disease⁶, a potential causal link between mLOY and IPF is plausible given that mLOY increases with age²⁴⁵. The Y chromosome is generally specific to biological males, therefore a hypothetical link between mLOY and IPF is a possible explanation for the association between IPF risk and male sex⁶. Finally, cellular loss of Y chromosome (LOY) is predicted to result from segregation errors during mitosis^{242,243}. Genetic studies have indicated that disrupted spindle assembly is likely to be driving the mitotic segregation errors that result in mLOY, intriguingly implicating two IPF-linked genes - *SPDL1* and *MAD1L1*^{89,109,243}. Overall, the role of mLOY in other fibrotic disease and overlap of disease mechanisms with IPF supports the potential for a causal link between these traits.

This chapter describes a two-sample Mendelian randomisation approach that aimed to assess whether mLOY is likely to be on the causal pathway for IPF. This chapter also describes an MR analysis assessing for evidence of causal effects IPF on mLOY to investigate possible reverse causality. As an additional sensitivity analysis, the sex-specific nature of mLOY was considered in integrating positive and negative control analyses into the design of this study to evaluate gene-environment equivalence of the mLOY genetic instrument.

A positive control analysis assessed for causal effect of the mLOY instrument on prostate cancer risk, which has been suggested by previous research and published MR analyses^{243,254}. Further, two negative control analyses were performed to evaluate the robustness of the mLOY exposure instrument. Firstly, a negative control analysis assessed for any causal effect of the mLOY instrument on the biologically female-specific ovarian cancer risk outcome. Secondly, a negative control analysis assessed for any causal effects of the mLOY instrument on IPF risk using genetic variant effect data from a female-only IPF GWAS study. Finally, as shortened telomeres are a strong risk factor for IPF that may also lead to disrupted chromosomal maintenance and genetic mosaicism^{255,256}, a secondary analysis is described that assessed if telomere length is likely to be causally associated with mLOY. The work described in this chapter contributed to a wider investigation of mLOY in IPF by Wang *et al.* (2024). The study by Wang *et al.* (2024) identified a significantly lower copy number of chromosome Y in male IPF patients compared to controls. The identified relationship between IPF and mLOY was additionally supported by single-cell transcriptomic data, which also highlighted an increased rate of LOY in regions of the lung where more fibrosis was present. The study by Wang *et al.* is publicly available as a preprint on *BioRxiv*²⁵⁷ and has been submitted to the journal *American Journal of Respiratory and Critical Care Medicine*.

5.2 Methods

5.2.1 Data retrieval and processing

Summary data were retrieved from the most recent GWAS of IPF susceptibility⁹⁰ (4,125 European ancestry cases, 20,464 European ancestry controls) through formal request. In addition, summary data from a sex-stratified GWAS of IPF susceptibility on female-only IPF cases were retrieved from authors of a study that is currently under review for publication²⁵⁸. Summary data from an updated analysis of genetic predisposition to mLOY

were retrieved upon request from the authors of the most recent published GWAS²⁴³. In the updated analysis, loss of Y was detected in blood leukocytes and quantified using a combined approach in 204,770 male UK Biobank participants. The application of this combined approach has previously been applied in a published study of mosaic loss of chromosome X²⁵⁹. The combined measure of mLOY was calculated through a meta-analysis of three measures:

- 1) A dichotomous mLOY status variable derived from phased B-allele frequency, calculated using the MoChA pipeline²⁵⁹
- 2) A median log *R* ratio measure (mLRR), indicating average signal intensity²⁶⁰, derived from SNP array data
- 3) An allele dosage measure derived from whole exome sequencing (WES) data

Summary data for GWAS of prostate cancer²⁶¹ (79,148 cases, 61,106 controls), ovarian cancer²⁶² (25,509 cases, 40,941 controls), and leukocyte telomere length²⁶³ (sample size: 472,174) were retrieved from the public databases GWAS Catalog⁶⁹ and OpenGWAS¹¹⁷.

This MR methodology used the *MendelianRandomization* R package, requiring GWAS summary statistics with the following data fields in exposure and outcome datasets:

- A SNP identifier (such as an RSID code)
- Chromosome number
- Base position
- Allele 1 (*A1*; effect allele)
- Allele 2 (*A2*; other allele)
- SNP effect size (*beta*)
- Effect standard error
- SNP p-value
- Effect allele frequency (*EAF*)

- GWAS sample size (N)

5.2.2 Genetic instrument selection

Genetic instruments were required to be associated ($p < 5 \times 10^{-8}$) with the exposure in each MR analysis. The IPF exposure instrument used to investigate possible causal effects IPF on mLOY incorporated the 19 SNPs reported in Allen et al. (2022) IPF GWAS⁹⁰. SNPs for all other exposure variables were clumped to ensure independence ($r^2 < 0.001$ in a 10,000kb window) using the 1000 Genomes Project reference panel (*ieugwasr*¹¹⁷). To avoid weak instrument bias, any SNPs with an F-statistic (instrument strength measurement) lower than ten were removed from exposure instruments. If any SNPs used as genetic instruments for the exposure were not present in the outcome dataset, appropriate proxy variants were identified ($r^2 > 0.7$). If no appropriate proxy variants were available, SNPs were removed from the respective exposure instrument. Proxy SNPs were identified using *LDLinkR*²¹².

5.2.3 Methods for statistical analyses

A bidirectional MR analysis was performed as a primary analysis to assess for evidence of a causal relationship between mLOY and IPF. Main causal estimates of an mLOY exposure on an IPF outcome, and the reverse with IPF as the exposure variable, were calculated using the IVW-RE method (*MendelianRandomization*²¹³). Pleiotropy-robust causal estimates were calculated using MR-Egger²¹⁴, weighted median²¹⁵, MR-PRESSO²¹⁶ and MR-Lasso²¹⁷ MR methods. In addition, a SNP leave-one-out analysis was performed to identify any SNPs that may be independently driving or distorting causal effect estimates of mLOY on IPF, or of IPF on mLOY.

Inclusion of positive and negative control outcomes as sensitivity analyses can contribute to robust MR study design¹⁹⁷. In the context of MR, a positive control outcome should have an established causal relationship with the relevant exposure. Correspondingly, a negative

control outcome can be selected that should not be causally affected by the exposure¹⁹⁷. As the effects of mLOY should be sex-specific, a positive control outcome of prostate cancer and a negative control outcome of ovarian cancer were selected. To complement this negative control analysis using a female-specific outcome, an additional negative control analysis was performed using genetic instruments derived from a sex-stratified (female-only) GWAS of IPF susceptibility. Evidence of causal effects of mLOY on ovarian cancer risk, or IPF in a female-only cohort, would indicate that the mLOY instrument may influence IPF risk through alternative pathways and therefore highlight the potential for confounding through violation of the exclusion restriction assumption. Further, a secondary MR analysis was performed to investigate if telomere length as an exposure has causal effects on mLOY. Positive and negative control analyses, and secondary analyses of telomere length and mLOY, were performed using the IVW-RE, weighed median, MR-Egger, MR-Lasso, MR-PRESSO, and leave-one-out MR methods. An overview of this study is presented in Figure 5.2.1.

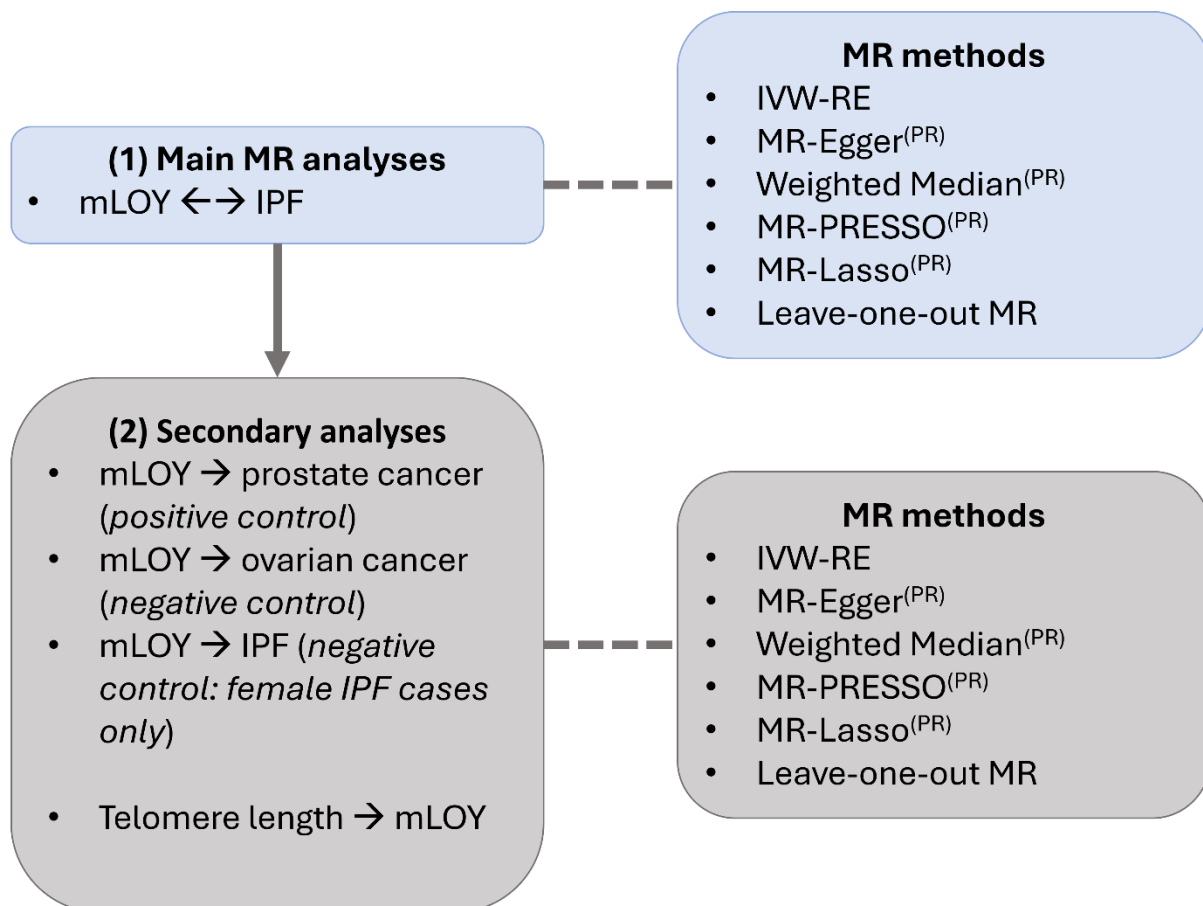


Figure 5.2.1: Overview of Mendelian randomisation study investigating possible causal relationships between mLOY and IPF. (1) Main MR analyses included bidirectional univariable analyses of mLOY on IPF susceptibility. (2) Secondary analyses included an MR analysis of mLOY on a positive control outcome of prostate cancer susceptibility and a negative control outcome of ovarian cancer susceptibility. An additional negative control analysis was performed of mLOY on female IPF susceptibility using genetic instruments derived from a sex-stratified IPF GWAS. To investigate possible effects of telomere attrition, an additional analysis was performed to assess if telomere length is causally associated with mLOY. (PR) indicates pleiotropy-robust MR methods.

5.3 Results

5.3.1 Genetic instrument selection

A total of 135 independent variants that are associated with mLOY met inclusion criteria and were present in the IPF GWAS dataset or had appropriate proxies available. For the MR

analysis assessing for causal effects of IPF on mLOY, 19 variants identified in the most recent susceptibility GWAS⁹⁰ were present in the mLOY GWAS dataset and were used for the IPF genetic instrument. For positive and negative control analyses, 117 and 123 mLOY-associated SNPs met inclusion criteria and were respectively present in the prostate cancer and ovarian cancer outcome datasets. Finally, 63 variants associated with leukocyte telomere length were selected to investigate possible causal effects of telomere length on mLOY.

For each tested exposure, all SNPs were sufficiently strong instruments ($F > 10$).

5.3.2 Mendelian randomisation analysis results: Main analyses

5.3.2.1 Bidirectional MR analysis of mLOY and IPF susceptibility

Initial estimates calculated using the IVW-RE MR method indicated that mLOY may have a protective effect on IPF susceptibility (IVW-RE, $OR = 0.53$, 95% CI : 0.33-0.84, $p = 0.008$) (Figure 5.3.1). However, results also indicated a high level of heterogeneity was present in the IVW-RE analysis of mLOY on IPF ($Q = 357.51$, $p < 1.00 \times 10^{-4}$, $I^2 = 62.50\%$), suggesting a high likelihood of horizontal pleiotropy. Removal of likely pleiotropic SNPs using MR-Lasso resulted in reduced heterogeneity ($Q = 109.02$, $p = 0.71$, $I^2 = 0.00\%$) and adjusted results suggested no causal effect of mLOY on IPF ($OR = 0.81$, 95% CI : 0.59-1.10, $p = 0.178$). In addition, results from other pleiotropy-robust methods consistently did not suggest any causal relationship between an mLOY exposure and an IPF risk outcome. Leave-one-out

MR analysis also did not suggest that any individual variants are likely to be driving or distorting a causal effect of mLOY on IPF (Figure 5.3.2).

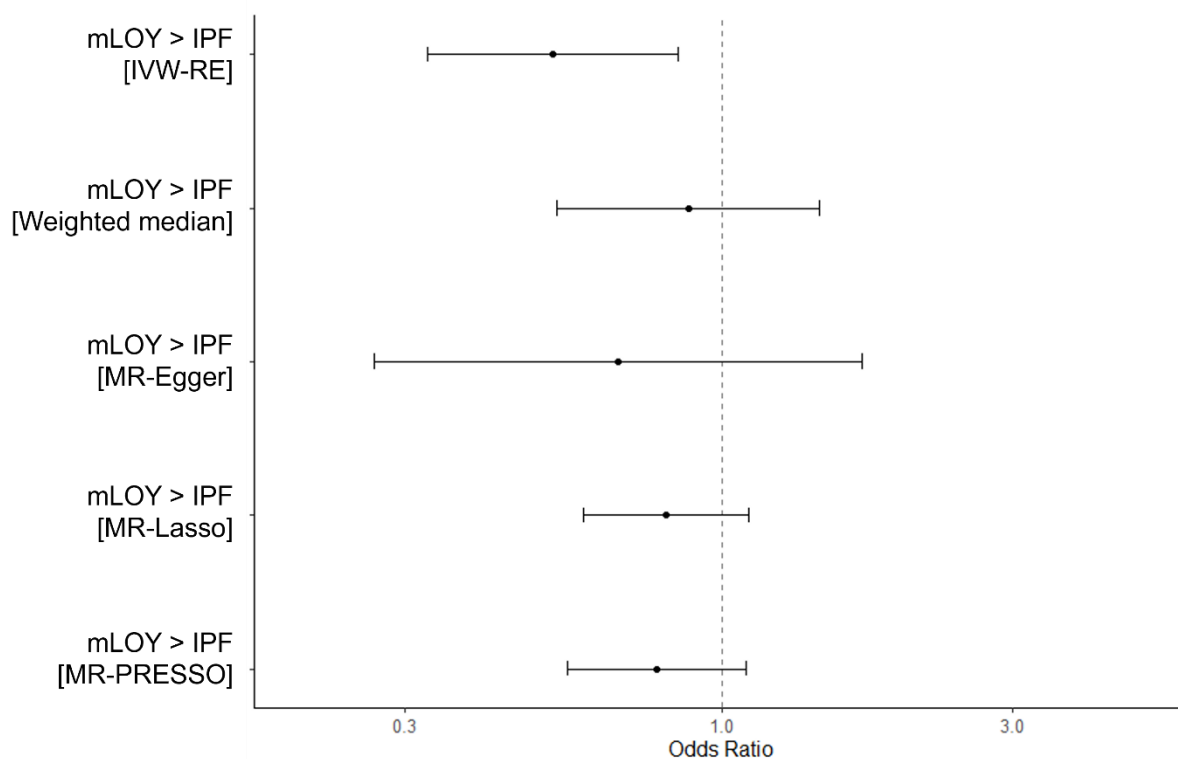


Figure 5.3.1: Causal estimates from MR analyses of mLOY on IPF using IVW-RE and pleiotropy-robust methods. Error bars show 95% confidence intervals of MR estimates.

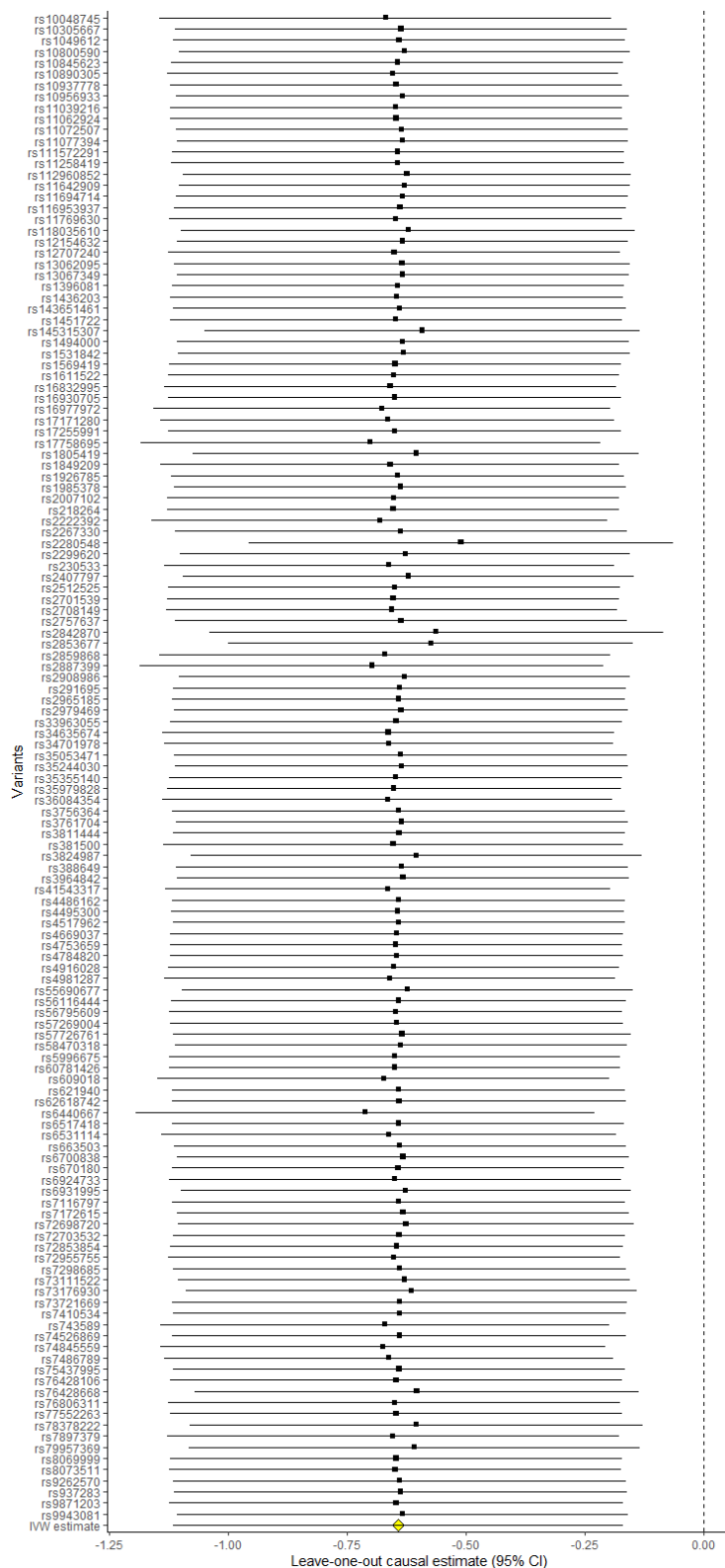


Figure 5.3.2: Estimates from a SNP leave-one-out MR analysis of mLOY on IPF. The SNP removed from each MR analysis is indicated on the y-axis by the genetic variant RSID identifier. Effect estimates are presented as beta values ($\log(\text{OR})$) with error bars indicating 95% confidence intervals.

Results from main MR analysis of an IPF exposure on an mLOY outcome did not suggest that IPF is likely to be causally associated with mLOY (IVW-RE, $OR = 1.00$, 95% CI : 0.99-1.00, $p = 0.38$) (Figure 5.3.3). Significant heterogeneity in the IVW-RE MR analysis suggested that horizontal pleiotropic effects may be present ($Q = 213.95$, $p < 1.00 \times 10^{-4}$, $I^2 = 91.60\%$), however causal estimates from pleiotropy-robust analyses were consistent with those of IVW-RE. Pleiotropy is therefore unlikely to be masking a true causal effect between IPF and mLOY. Leave-one-out MR analyses did not suggest that any individual variants are likely to be driving distorting a causal effect of IPF on mLOY (Figure 5.3.4).

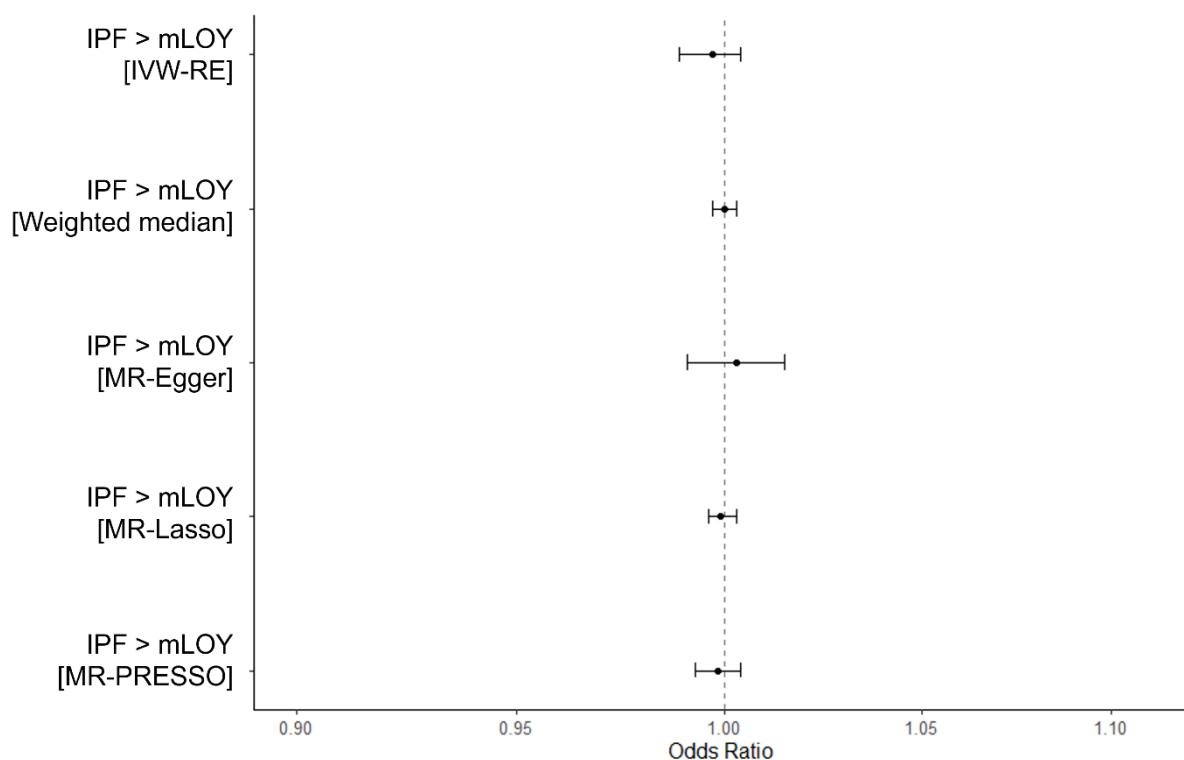


Figure 5.3.3: Causal estimates from MR analyses of IPF on mLOY using IVW-RE and pleiotropy-robust methods. Error bars show 95% confidence intervals of MR estimates. A narrow scale was used (0.90-1.10) on the x-axis for this plot to ensure error bars are clearly visible.

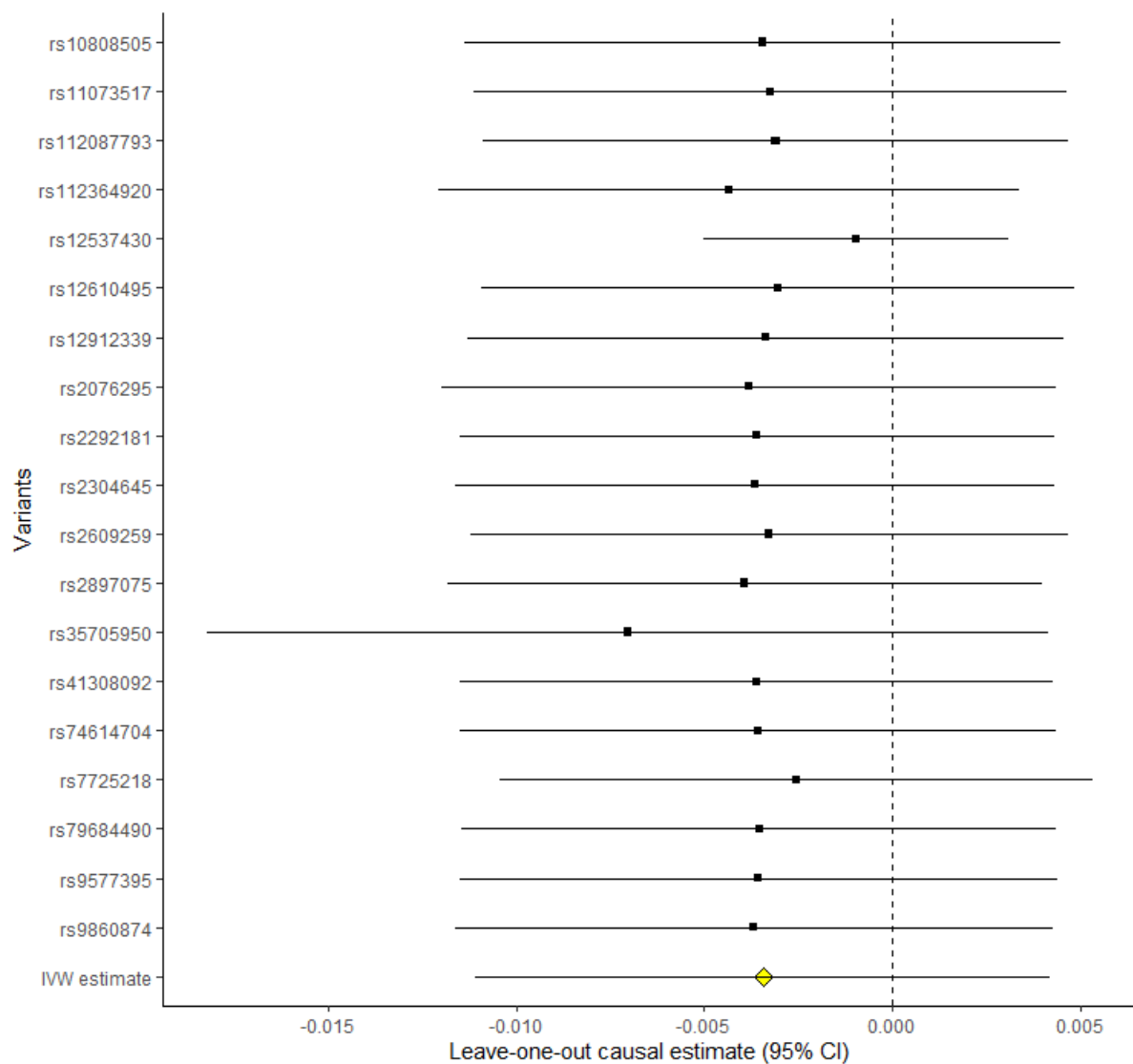


Figure 5.3.4: Estimates from a SNP leave-one-out MR analysis of IPF on mLOY. The SNP removed from each MR analysis is indicated on the y-axis by the genetic variant RSID identifier. Effect estimates are presented as beta values ($\log(\text{OR})$) with error bars indicating 95% confidence intervals.

5.3.3 Mendelian randomisation analysis results: Secondary analyses

Results from the positive control analysis supported a causal effect of mLOY on prostate cancer risk (IVW-RE, $\text{OR} = 1.46$, 95% CI: 1.25-1.71, $p = 2.56 \times 10^{-6}$) (Figure 5.3.5). Significant heterogeneity was present in the IVW-RE analysis ($Q = 338.23$, $p < 1.00 \times 10^{-4}$, $I^2 = 65.70\%$), indicating a high likelihood of horizontal pleiotropy. Removal of likely pleiotropic variants with the MR-Lasso method reduced overall heterogeneity statistics ($Q = 86.15$, $p = 0.65$, $I^2 =$

0.00%) but the adjusted causal estimate was concordant with results calculated using the IVW-RE method. These results suggest that causal effects of mLOY on prostate cancer risk are therefore unlikely to be driven only by horizontal pleiotropy. In addition, results from other pleiotropy-robust MR analyses were also concordant with the causal estimate calculated using the IVW-RE method. Leave-one-out MR analyses did not highlight any individual variants that may be independently driving this relationship between mLOY and prostate cancer (Figure 5.3.6).

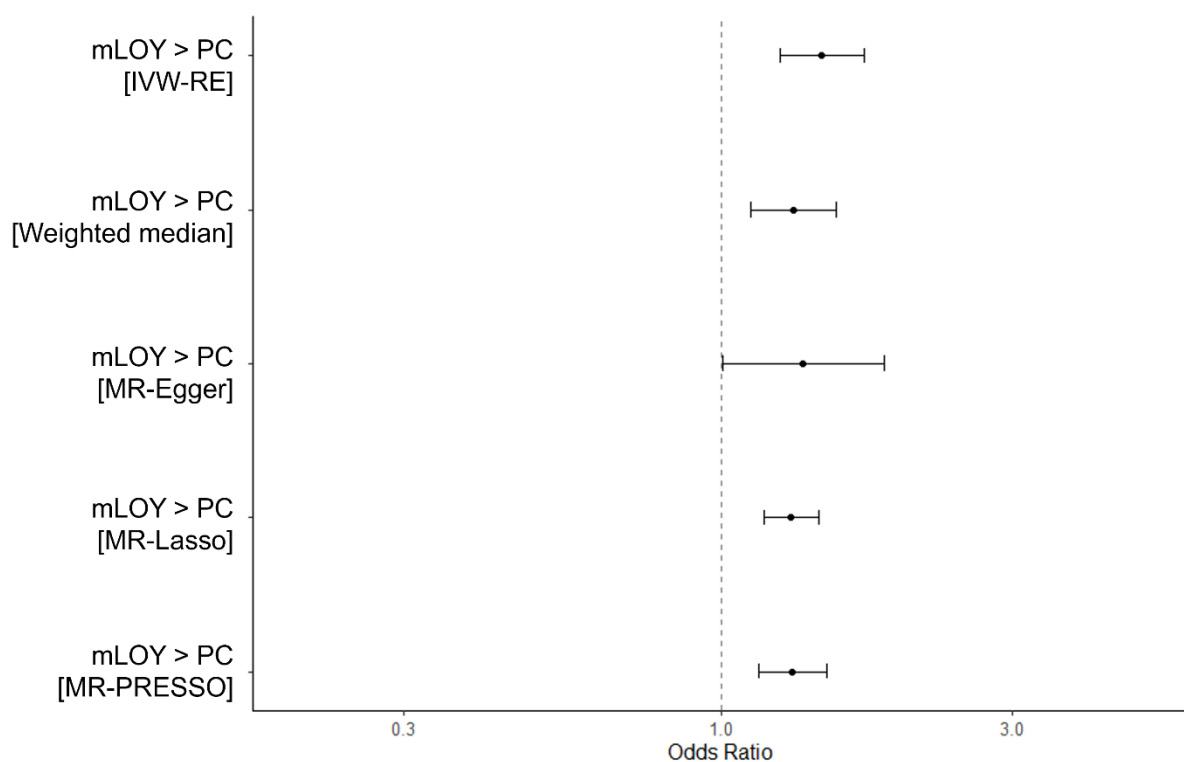


Figure 5.3.5: Causal estimates from MR analyses of mLOY on prostate cancer risk using IVW-RE and pleiotropy-robust methods. PC = prostate cancer. Error bars show 95% confidence intervals of MR estimates.

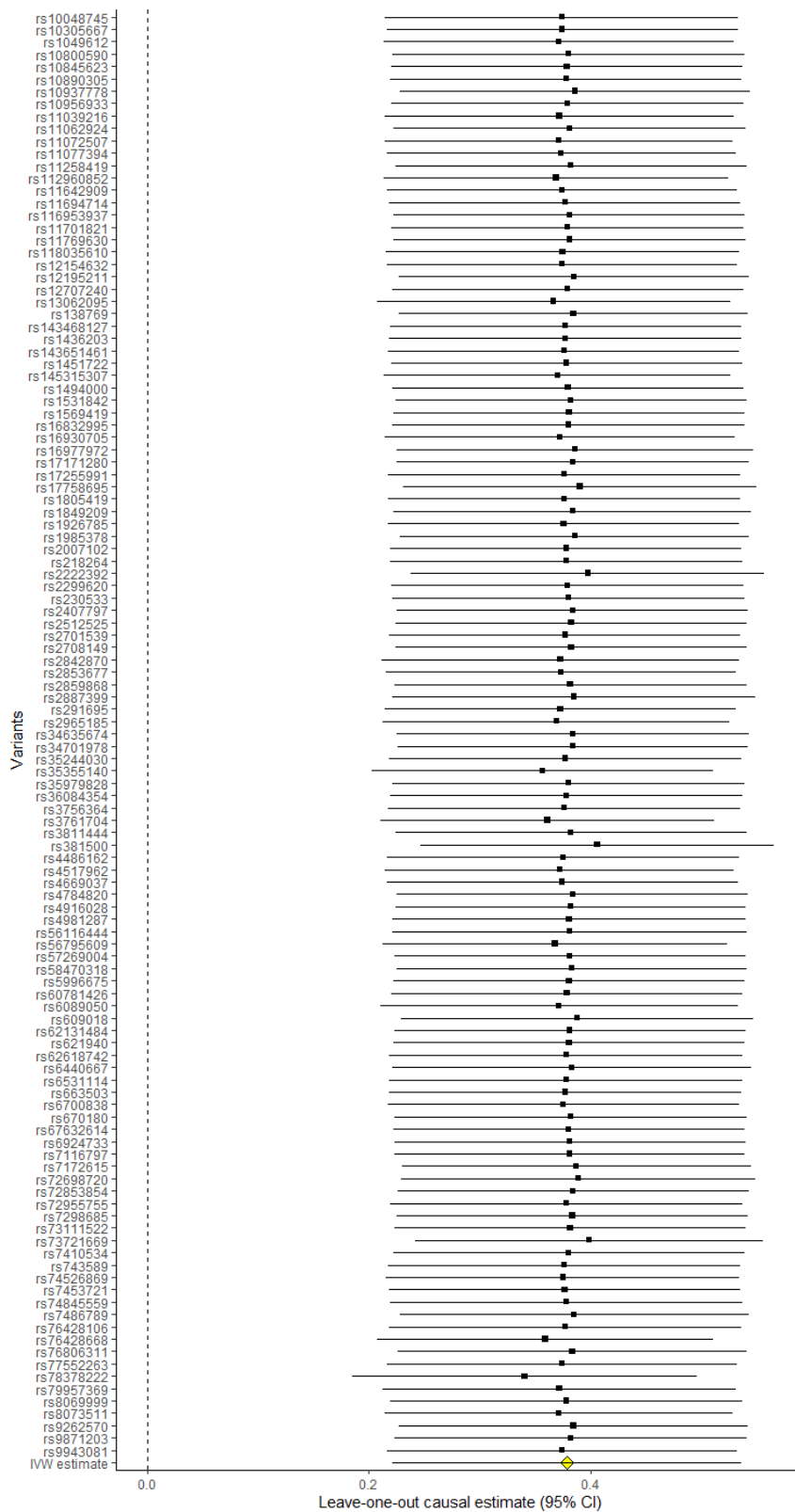


Figure 5.3.6: Estimates from a SNP leave-one-out MR analysis of mLOY on prostate cancer risk. The SNP removed from each MR analysis is indicated on the y-axis by the genetic variant RSID identifier. Effect estimates are presented as beta values ($\log(\text{OR})$) with error bars indicating 95% confidence intervals.

Results from negative control analyses did not suggest that mLOY causes ovarian cancer (IVW-RE, $OR = 1.05$, 95% CI : 0.87-1.27, $p = 0.63$) (Figure 5.3.7). Results also indicated significant heterogeneity was present in the IVW-RE analysis ($Q = 202.54$, $p < 1.00 \times 10^{-4}$, $I^2 = 39.80\%$). Results following removal of likely pleiotropic variants with MR-Lasso indicated reduced heterogeneity ($Q = 110.13$, $p = 0.59$, $I^2 = 0.00\%$) but also suggested no causal effect of mLOY on ovarian cancer risk. In addition, causal estimates across pleiotropy-robust MR methods were consistent with those of IVW-RE. Leave-one-out MR analyses also did not suggest that any individual variants in the mLOY instrument are distorting a true causal relationship between mLOY and ovarian cancer (Figure 5.3.8).

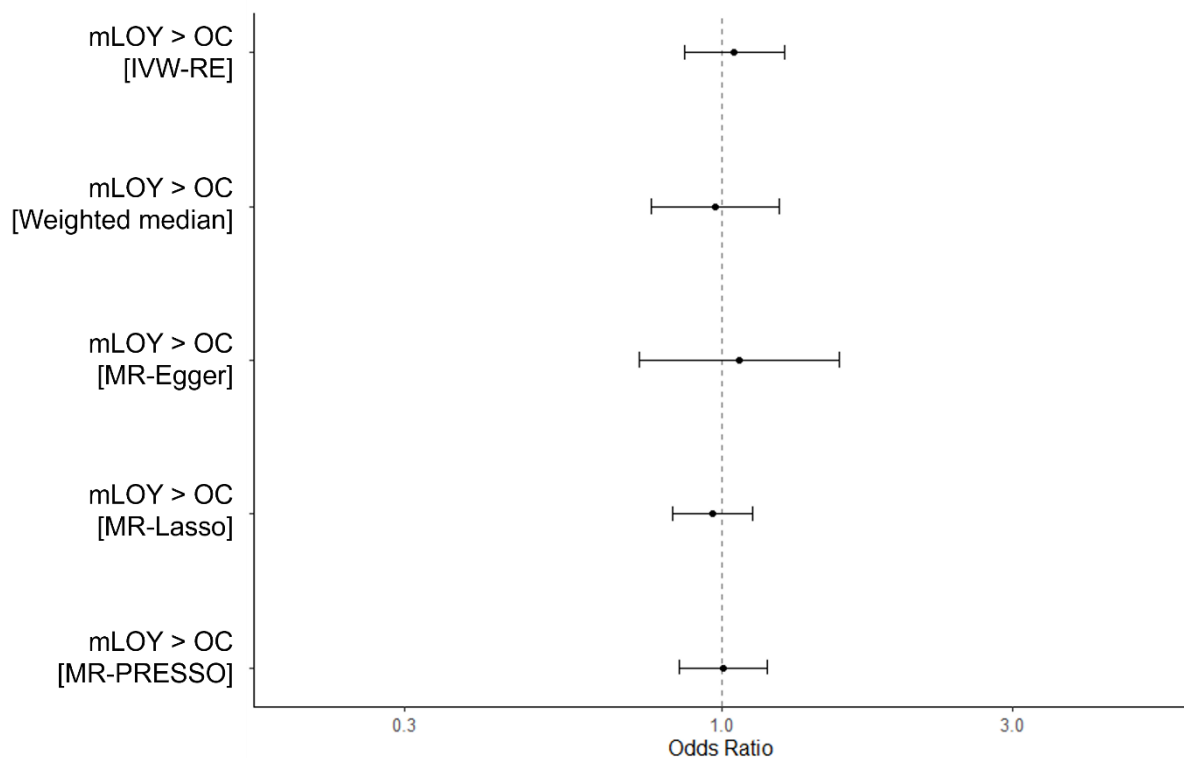


Figure 5.3.7: Causal estimates from MR analyses of mLOY on ovarian cancer risk using IVW-RE and pleiotropy-robust methods. OC = ovarian cancer. Error bars show 95% confidence intervals of MR estimates.



Figure 5.3.8: Estimates from a SNP leave-one-out MR analysis of mLOY on ovarian cancer risk. The SNP removed from each MR analysis is indicated on the y-axis by the genetic variant RSID identifier. Effect estimates are presented as beta values ($\log(\text{OR})$) with error bars indicating 95% confidence intervals.

Results from negative control analyses did not suggest that mLOY causes IPF in biological females (IVW-RE, $OR = 0.81$, 95% CI : 0.41-1.62, $p = 0.55$) (Figure 5.3.9). Results also indicated significant heterogeneity was present in the IVW-RE analysis ($Q = 259.66$, $p < 1.00 \times 10^{-4}$, $I^2 = 48.40\%$). Results following removal of likely pleiotropic variants with MR-Lasso indicated reduced heterogeneity ($Q = 112.88$, $p = 0.59$, $I^2 = 0.00\%$) but also suggested no causal effect of mLOY on IPF risk in biological females. In addition, causal estimates across pleiotropy-robust MR methods consistently suggested a null effect. Leave-one-out MR analyses also did not suggest that any individual variants in the mLOY instrument are distorting a true causal relationship between mLOY and IPF risk in biological females (Figure 5.3.10).

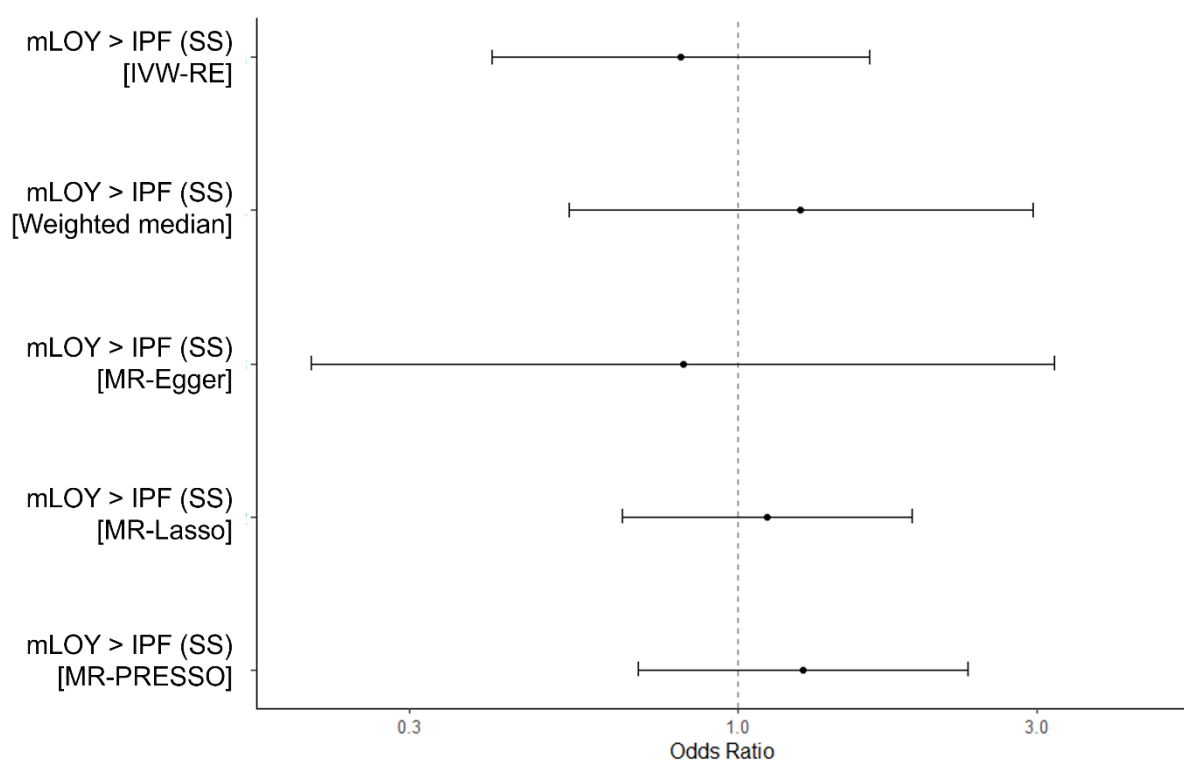


Figure 5.3.9: Causal estimates from MR analyses of mLOY on IPF risk (female IPF cases only) using IVW-RE and pleiotropy-robust methods. OC = ovarian cancer. Error bars show 95% confidence intervals of MR estimates.

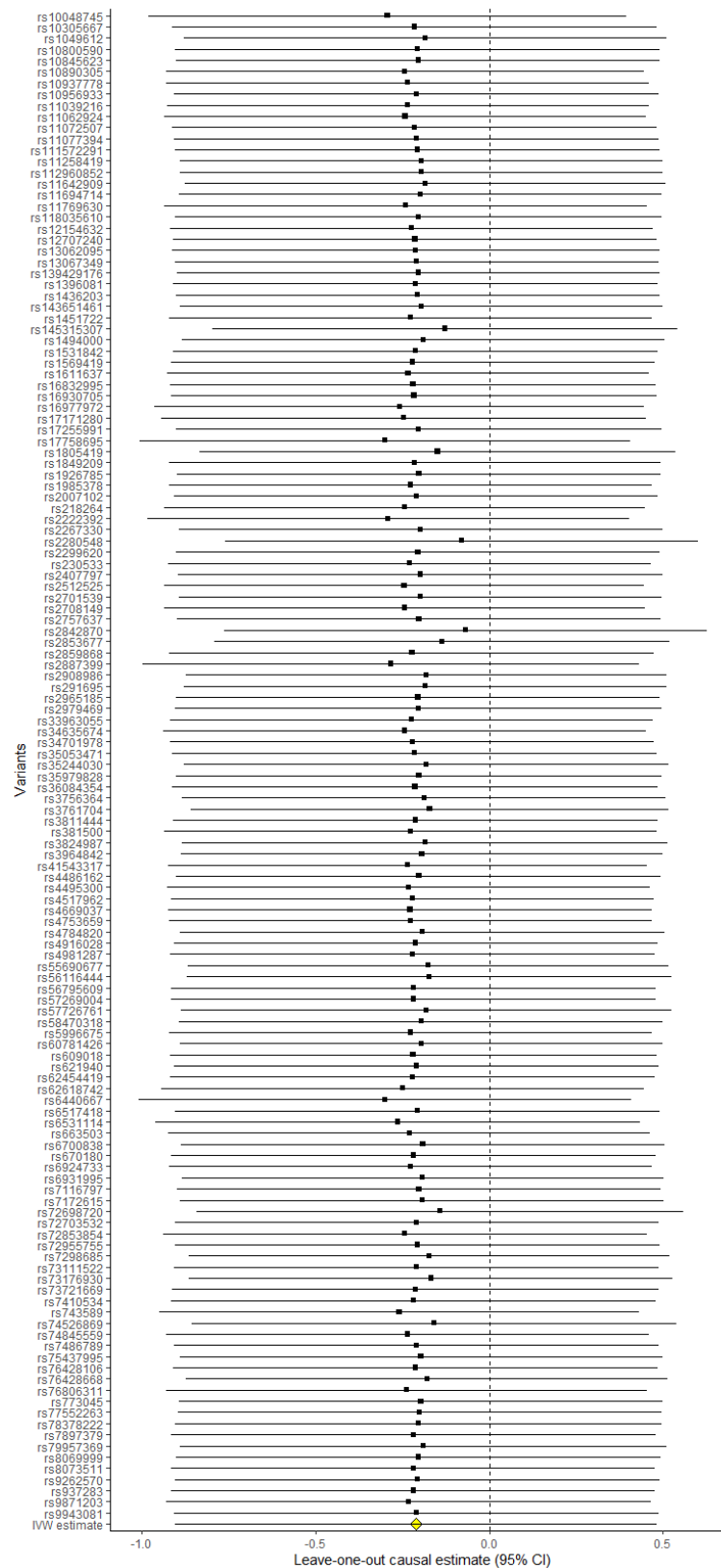


Figure 5.3.10: Estimates from a SNP leave-one-out MR analysis of mLOY on IPF risk (female IPF cases only). The SNP removed from each MR analysis is indicated on the y-axis by the genetic variant RSID identifier. Effect estimates are presented as beta values ($\log(\text{OR})$) with error bars indicating 95% confidence intervals.

Results initially did not indicate that telomere length is likely to be causally linked to mLOY (IVW-RE, $OR = 0.96$, $95\% CI = 0.91-1.01$, $p = 0.15$) (Figure 5.3.11). In addition, results indicated significant heterogeneity and potential horizontal pleiotropy in this analysis ($Q = 924.80$, $p < 1.00 \times 10^{-4}$, $I^2 = 93.30\%$). However, the causal estimates of three out of four pleiotropy-robust MR approaches (Weighted median, MR-Lasso, and MR-PRESSO) indicated a negative causal relationship between telomere length and mLOY (Figure 5.3.9). Removal of likely pleiotropic variants by MR-Lasso also resulted in greatly reduced heterogeneity ($Q = 41.28$, $p = 0.50$, $I^2 = 0.00\%$). In addition, leave-one-out analyses identified two single variants that, when removed from the telomere length instrument, resulted in causal estimates suggesting a negative relationship between telomere length and mLOY (rs16978028, $OR = 0.95$, $95\% CI = 0.91-0.99$, $p = 0.02$) (rs7705526, $OR = 0.95$, $95\% CI: 0.90-1.00$, $p = 0.04$) (Figure 5.3.12).

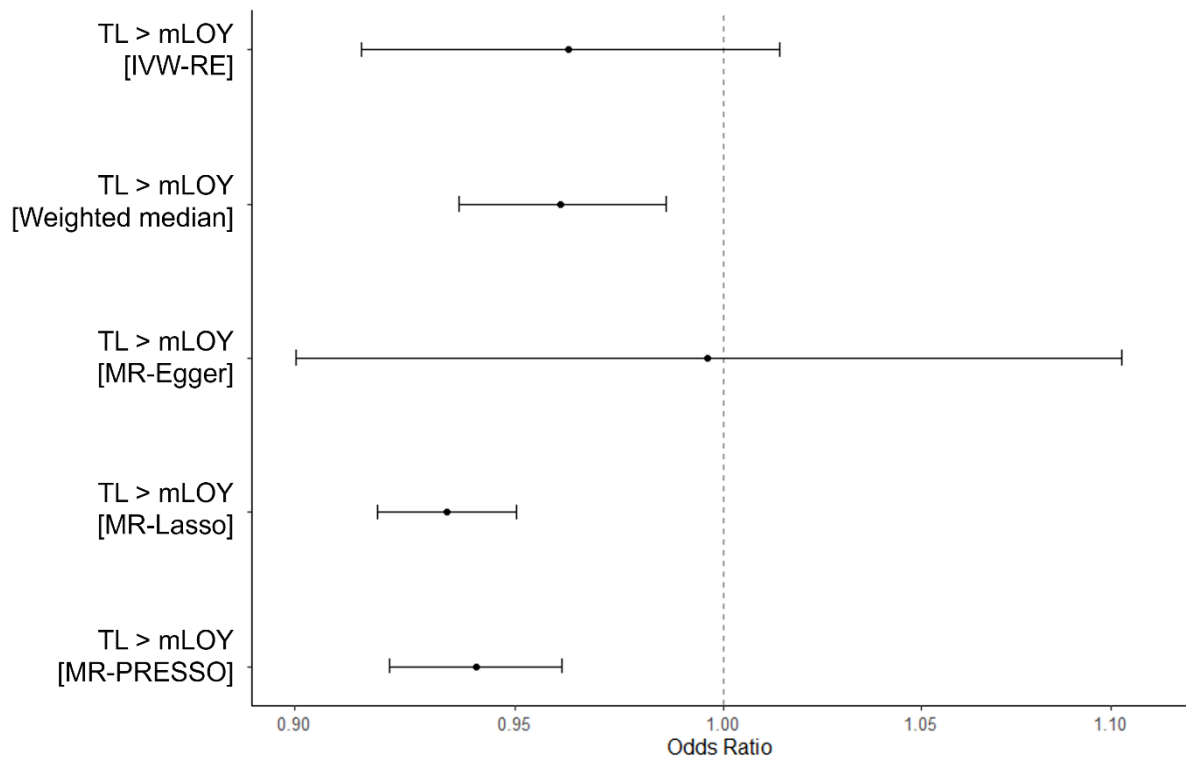


Figure 5.3.11: Causal estimates from MR analyses of telomere length (leukocyte) on mLOY using IVW-RE and pleiotropy-robust methods. TL = telomere length. Error bars show 95% confidence intervals of MR estimates. A narrow scale was used (0.90-1.10) on the x-axis for this plot to ensure error bars are clearly visible.

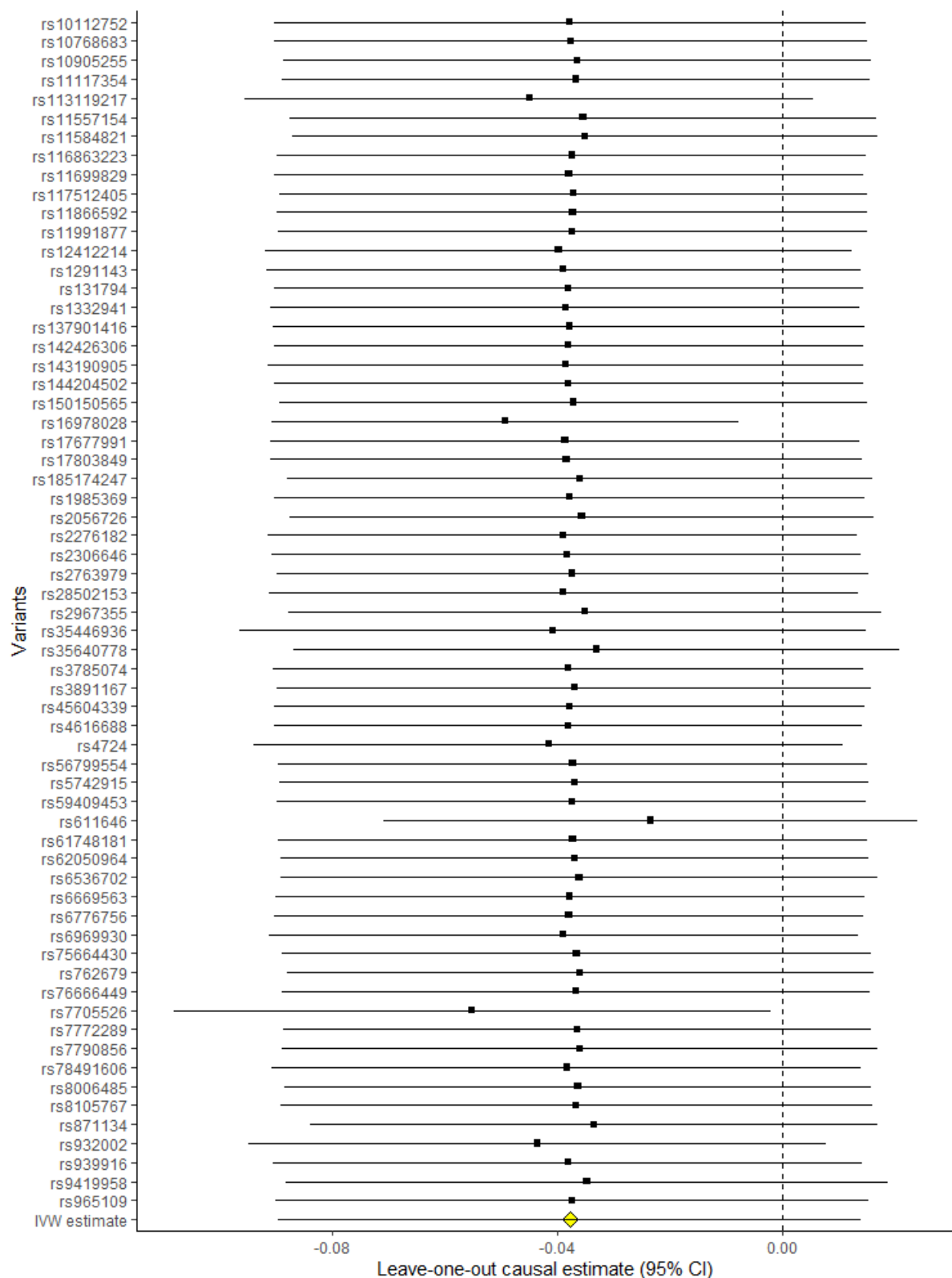


Figure 5.3.12: Estimates from a SNP leave-one-out MR analysis of telomere length (leukocyte) on mLOY. The SNP removed from each MR analysis is indicated on the y-axis by the genetic variant RSID identifier. Effect estimates are presented as beta values ($\log(\text{OR})$) with error bars indicating 95% confidence intervals.

4.3.4 Conclusions from Mendelian randomisation study investigating a potential causal link between mosaic loss of chromosome Y and idiopathic pulmonary fibrosis

This study aimed to investigate a potential causal relationship between mLOY and IPF. Results did not suggest that mLOY is likely to be on the causal pathway for IPF risk, or that IPF disease is likely to causally affect the intrinsic rate of mLOY. However, secondary analyses highlight the potential for a causal link between shortened telomeres and mLOY.

5.4 Discussion

This research aimed to use an MR approach to investigate a potential causal relationship between mLOY and IPF. The possibility that a causal link may exist between these traits is supported by the role of mLOY in driving cardiac fibrosis²⁵⁰ and primary biliary cholangitis²⁵¹, which has been classified as a fibrotic condition⁵². Further, recent unpublished research has highlighted a positive association between mLOY and pulmonary fibrosis at a single-cell level²⁵⁷. Several genetic risk variants, including variants mapped to *KNL1*, *SPDL1*, and *MAD1L1*^{90,109}, implicate disrupted spindle assembly and chromosomal regulation during mitosis as key potential pathways contributing to IPF. Importantly, chromosome mis-segregation during mitosis is also predicted to be a key process in the mLOY²⁴³. It is therefore plausible that mLOY may be on the causal pathway for IPF.

Overall, the results of this investigation suggest that mLOY and IPF are unlikely to be linked by direct causal effects acting from one trait on the other. Results of secondary analyses were not definitive but highlighted a potential relationship between shortened telomere length and mLOY. Shortened telomere length is predicted to be on the causal pathway for IPF and shortened telomere length is associated with worse survival in IPF cohorts^{7,167}. Telomere shortening has also been shown to be even more severe in AT2 cells associated

with fibrotic lesions, which supports the existence of a direct biological link²⁶⁴. Shortened telomeres have been proposed to lead to mLOY and a genetic variant mapping to the telomerase gene *TERT* has been identified in GWAS of mLOY. Shortened telomere length may therefore independently drive both IPF and mLOY²⁴³.

This methodology was strengthened by using complementary sensitivity analyses (weighed median, MR-Egger, MR-Lasso, and MR-PRESSO) to consider horizontal pleiotropy both generally across multiple variants and at a single variant level. In addition, the inclusion of positive and negative control analyses enabled assessment of whether the mLOY genetic instrument is an appropriate proxy for the effects of mLOY. It has already been established that mLOY is likely to be on the causal pathway for prostate cancer^{243,254}, therefore this exposure-outcome relationship was selected for a positive control analysis as recommended in current MR guidelines¹⁹⁷. As biologically female individuals do not generally have Y chromosomes, results from negative control analyses that suggest causal effects of mLOY on ovarian cancer, or on IPF in a biologically female-only IPF cohort, would highlight potential confounding. Results from control analyses suggesting a causal effect of mLOY on prostate cancer and no causal effect on ovarian cancer therefore give greater confidence that this study design is robust.

This two-sample MR study design used genetic association data from the Allen *et al.* (2022) GWAS of IPF susceptibility, which used only IPF cases diagnosed using European Respiratory Society (ERS) and ATS guidelines¹⁴. IPF cases were also taken from independent cohorts, rather than UK Biobank^{89,90}, which alleviates potential concern over sample overlap with the cohort used for detection of LOY in the respective mLOY GWAS²⁴³.

A key limitation of this study is that this MR approach may be limited in its capacity to represent specific or complex biological mechanisms. Firstly, the genetic associations used to proxy mLOY are based on LOY detected in blood. A more in-depth *in silico* investigation of the relationship between mLOY and IPF could hypothetically use genetic associations

with mLOY in lung tissue, should such data become available. The MR approach applied in this investigation also is not able to account for effects that are specific to a disease subtype, particularly as IPF genetics and biological mechanisms are complex. In addition, results from IWW-RE analyses highlighted high levels of heterogeneity in every exposure-outcome relationship investigated. The use of pleiotropy-robust MR methods such as MR-Lasso enabled assessment of causal estimates with a greatly reduced heterogeneity. However, the presence of such high heterogeneity across analyses raises concerns over the complexity of investigated traits and potentially undetected confounding.

In summary, this research aimed to investigate a possible causal relationship between mLOY and IPF. Overall, the results of this work do not suggest that mLOY is likely to cause IPF, or that IPF is likely to cause mLOY. If a correlation between mLOY and IPF exists, it may instead be driven by shared factors such as shortened telomere length, as indicated by MR analysis of telomere length on mLOY, or smoking^{231,265}. If shared factors underlie these traits, the development of mLOY and the development of fibrosis may occur in parallel. The presence of mLOY in IPF patients therefore remains a potentially important marker of advanced biological lung ageing that may inform clinical practice.

6: Discussion

IPF is a devastating disease of progressive lung scarring that has a poor prognosis and greatly affects patient quality of life. The therapeutic options available for IPF patients are limited, associated with considerable side effects, and do not halt or reverse disease progression. The biological and epidemiological factors contributing to IPF are incompletely understood, which limits the identification of treatment targets and drug repurposing opportunities. This thesis aimed to investigate whether IPF susceptibility variants are associated with other human traits to build understanding of variant biology and highlight mechanistic links with other phenotypes. This thesis also aimed to investigate shared genetic architecture across IPF and related traits. Finally, this thesis aimed to assess whether an epidemiological relationship (IPF and DM) is likely to be driven by direct causal links between disease phenotypes.

This chapter summarises the current understanding of IPF genetics and links to disease biology, the main results from this thesis, and how these results contribute to this research area. This chapter also describes the strengths and limitations of each methodological approach and outlines future research opportunities that stem from the findings of this thesis.

6.1 The current understanding of IPF genetics and links with other human phenotypes

In recent years, substantial progress has been made in identifying common genetic variants associated with IPF susceptibility^{86,87,89,90}, which are predicted to account for roughly 15% of disease liability⁹¹. The identification of variants such as rs35705950 (*MUC5B*) indicated

a role of altered mucin biology in IPF disease biology, which has since been supported by *in vitro* and *in vivo* models⁹⁷. Similarly, identification of variants such as rs7725218 (*TERT*) and rs9860874 (*TERC*) have implicated telomere biology, rs10808505 (*DEPTOR*) has implicated altered cellular metabolism, and rs12537430 (*MAD1L1*) and rs2292181 (*KIF15*) have implicated cell cycle control mechanisms in contributing to IPF disease. However, for many of the identified IPF susceptibility variants, such as rs2304645 (*IVD*), rs2609259 (*FAM13A*), or rs12610495 (*DPP9*), associated gene function doesn't clearly implicate a disease mechanism that can be linked to IPF disease.

It is unclear whether mechanisms implicated by IPF genetic risk variants affect other traits or phenotypes or if they are associated with fibrosis in non-lung tissues. It is also not known whether these mechanisms are linked to biomarkers that could improve IPF diagnosis, or if they are linked to biological traits with effective therapies that could be repurposed to treat IPF patients. Current genetic research in this area has identified genetic relationships between IPF and COVID-19 severity¹⁰¹ and systemic hypertension¹²⁷. Shared loci with opposite risk alleles for IPF and COPD have previously been identified¹³⁰, but the likelihood of shared causal variants has not been investigated through colocalisation analysis.

IPF is also known to be associated with a range of comorbidities including pulmonary hypertension, lung cancer, COPD, DM, and GORD^{39,155}. Research in this area has shown that GORD is likely to be on the causal pathway for IPF⁴². However, the origins of many of these epidemiological associations with IPF are not fully understood.

6.2 Summary of findings and biological insights

Chapter 2 describes a PheWAS study of IPF susceptibility variants. The first PheWAS stage identified a broad range of traits associated with 18 IPF susceptibility variants using a

publicly available GWAS resource. Prioritised association signals were subsequently evaluated for likelihood of shared causal variants with IPF risk signals through a colocalisation analysis. As an independent analysis, a secondary PheWAS was performed using the GE 100,000 GP dataset.

This PheWAS identified colocalised association signals with IPF risk and COVID-19 severity with rs12610495 (*DPP9*), as previously reported¹⁰¹. *DPP9* (Dipeptidyl peptidase 9) has been shown to be an inhibitory regulator of NLRP1 inflammasome activation and *DPP9* loss-of-function mutations are associated with various immune disorders and increased susceptibility to infections^{266,267}. This shared variant therefore indicates that disrupted immunity is still likely to play a role in IPF disease, despite anti-inflammatory treatments being detrimental to IPF patient outcomes²⁶⁸. The effect of the rs12610495 IPF risk variant is therefore likely to act through increased susceptibility to severe infection, possibly resulting in increased inflammatory damage to the alveolar epithelium. This relationship was additionally supported by the positive genetic correlation between IPF risk and COVID-19 severity identified in Chapter 3, suggesting that this shared risk effect is likely to extend beyond *DPP9*.

It has been previously indicated that the IPF risk variants rs2609259 (*FAM13A*) and rs2076295 (*DSP*) are also associated with COPD, but with opposite risk alleles for either disease¹³⁰. As part of this PheWAS, strong evidence of colocalisation was identified with IPF risk and FEV1/FVC < 0.7 association signals at both the rs2609259 and rs2076295 loci. This finding therefore supports this potential opposing genetic relationship between COPD and IPF by confirming a high likelihood of shared causal variants. A negative genetic correlation between IPF and FEV1/FVC < 0.7 (Chapter 3) is also suggestive of a broader relationship between these traits. However, a recent study has highlighted a dynamic AKT-mediated mechanism regulating *FAM13A* and downstream proliferation of lung epithelial cells in response to environmental stress^{269,270}. It is therefore possible that IPF and COPD

mechanisms only have opposing effects on disease risk in the absence of environmental stimuli, such as cigarette smoke.

A role of telomere biology in driving IPF disease has been indicated by associations with rs7725218 (*TERT*), rs9860874 (*TERC*), and rs41308092 (*RTEL1*), as well as biological and observational studies^{159,271,272}. This PheWAS and colocalisation analysis identified a strong likelihood that associations with IPF risk (rs9860874; *TERC*) and short telomere length share a causal variant, therefore supporting the hypothesised link between IPF genetics and telomere biology. This finding is additionally supported by a negative correlation between the genetics of telomere length and of IPF risk (Chapter 3). IPF risk association signals also colocalised with negative associations with benign tumours (uterine fibroids and neoplasms) at this locus, which may therefore be driven by opposing effects of telomere biology on fibrotic and neoplastic mechanisms.

The identification of colocalised signals associated with IPF risk and total testosterone levels at the *IVD* locus (rs2304645) represents an important novel finding. *IVD* encodes isovaleryl-CoA dehydrogenase, which has a predicted function in leucine catabolism²⁷³. These results are supported by negative associations with endometriosis and polycystic ovaries (PCOS) in the GE 100,000 GP PheWAS analysis, as both traits are associated with sex hormone variation. However, these findings are contradicted by opposing effects of testosterone on endometriosis and PCOS^{173,274,275}. Current research indicates a positive association between total testosterone and IPF, but a negative association with bioavailable testosterone^{131,132}. The inverse of these relationships is observed with telomere length, as total testosterone is strongly associated with shortened leukocyte telomere length¹³¹. As total testosterone and bioavailable testosterone have a great deal of shared genetic architecture¹⁷³, this opposing effect on IPF risk and telomere length may explain the null correlation observed between IPF risk and total testosterone in Chapter 3. Overall, this

finding highlights the rs2304645 risk allele may increase IPF risk through altered sex hormone levels leading to shortened telomere length.

An additional novel finding was the colocalisation of IPF risk with an *RPAP1* eQTL signal at the intronic locus rs12912339 (mapped to *KNL1*). *RPAP1* (encoding RNA Polymerase II-associated protein 1) is functionally linked to fibrosis-associated genes *TGFB3*, *VEGFA*, and *S100A4*^{138–141}. Trans-acting upregulation of *RPAP1* may therefore represent an alternative pathway by which rs12912339 is linked to IPF disease biology.

Genetic correlation analyses performed in Chapter 3 aimed to assess shared genetic architecture across IPE, traits with prior evidence of colocalisation with IPF risk signals, and a sample of fibrotic diseases. The only fibrotic disease suggested to share genetic architecture with IPF in this analysis was PBC (primary biliary cholangitis). PBC is an autoimmune condition characterised by chronic injury to the biliary epithelium, progressive fibrosis, and shares the risk factor of smoking with IPF¹⁷¹. This finding indicates that shared pathways or disease mechanisms may contribute to both IPF and PBC. As a possible example, $\alpha\text{v}\beta 6$ integrin has been implicated as a key regulator of profibrotic TGF β signaling in both diseases^{276,277}. This shared pathway is of particular interest, as $\alpha\text{v}\beta 6$ integrin inhibitors are under investigation as a potential antifibrotic treatment for IPF²⁷⁸ and could be effective in treating both diseases. Intriguingly, no strong evidence of shared genetic architecture was found between IPF and the remaining sample of fibrotic diseases. Similarly, PheWAS analyses conducted in Chapter 2 did not highlight any direct associations with other fibrotic disease phenotypes. This finding suggests that IPF may be genetically distinct from other fibrotic diseases. However, it must be considered that shared pathways across IPF and other fibrotic traits may only act in specific disease endotypes, under specific environmental or biological conditions, or may be driven by rare genetic variants.

Chapter 4 of this thesis details an MR study that aimed to assess whether the epidemiological relationship between IPF and DM is driven by causal effects of either disease on the other. Overall, no evidence of causal effects of DM on IPF, or IPF on DM was identified. Similarly, diabetes-associated factors HbA1c and fasting insulin level are unlikely to causally affect IPF risk. Univariable analyses indicated a positive causal relationship between BMI and IPF. However, it is unclear if the indicated effect of increased BMI on IPF susceptibility could be confounded by gene-lifestyle interactions associated with obesity GWAS SNPs, such as genetic interactions with smoking behaviour^{229,230}.

Chapter 5 of this thesis describes a further application of MR methods investigating if a potential association between mLOY and IPF is driven by causal effects. Overall, no direct causal relationship between these traits was suggested. However, results highlighted a potential effect of shortened telomere length on mLOY, which is already established as a strong risk factor for IPF^{7,159}. If mLOY and fibrosis share driving mechanisms, they may also occur in parallel. The presence of mLOY therefore has the potential to be an important marker of advanced biological lung ageing. Further understanding of the relationship between mLOY and IPF may also provide insights into IPF biology and contribute to clinical prediction of disease progression and survival.

6.3 Clinical implications

The findings from this thesis have the potential to lead to improved clinical care of IPF patients, particularly through greater understanding of IPF disease aetiology. Firstly, the finding of shared genetics associated with IPF susceptibility and total testosterone level indicates involvement of sex hormones in IPF disease. This finding suggests the increased incidence of IPF in biologically male patients⁶ could be driven by intrinsic biology, rather than environmental²⁷⁹ or diagnostic bias²⁸⁰.

Previous research has shown that diseases such as lung cancer, COPD, and DM are common comorbidities associated with IPF³⁹. Across this thesis, no strong evidence was found to suggest that shared genetic risk, or direct causal effects, may be driving these epidemiological relationships. While effects specific to certain disease endotypes or gene-environment interactions cannot be ruled out, it appears that these relationships are more likely to be driven by environmental factors.

Genetic correlation analyses did not identify strong evidence of shared genetic architecture across traits. Future research should focus on patients with multiple fibrotic conditions to identify genetics that are unique to fibrotic multimorbidity.

Overall, the biology associated with IPF risk genetics appears to be heterogeneous, which may lead to variable disease characteristics and treatment responses in IPF patients depending on their genetic profile. This thesis contributes to the biological understanding of IPF risk variants at multiple loci (*DPP9*, *FAM13A*, *TERC*, *IVD*, and *KNL1*), which will be vital in enabling personalised treatment approaches in response to genetic testing. While these findings represent progress in this area, further understanding of the biological implications of IPF genetics and potential disease endotypes are likely to be needed before personalised medicine can be successfully implemented in IPF patients.

6.4 Thesis strengths and limitations

The methods used throughout this thesis have several limitations. A main challenge in the PheWAS of public GWAS data, described in Chapter 2, was that IPF risk variants were found to be highly pleiotropic, resulting in hundreds of phenotypic associations per SNP in some cases. An optimal approach would have been to run colocalisation analyses with IPF risk and every associated phenotype, however this would have been a very intensive and time-consuming process. By prioritising a relatively small number of PheWAS associations for colocalisation analyses, many statistically significant and potentially important traits were

not included. Also, the approach used to prioritise PheWAS associations did not represent each IPF SNP equally, which may have limited an objective investigation of SNP biology.

The colocalisation method used (*coloc.abf*) is a powerful tool for most instances of overlapping genetic association signals, however this method assumes there is only one causal variant per trait. Particularly as the colocalisation approach used an arbitrary 500kb genomic region for efficiency purposes, it may be that traits with shared causal variants were incorrectly identified as having low evidence of colocalisation. Alternative methods, such as the ‘sum of single effects’ model (SuSiE)²⁸¹ relaxes the single causal variant assumption, which could in some instances results in more accurate assessment of colocalisation likelihood.

The LDSC genetic correlation method used in Chapter 3 is simple to use and provides powerful insights into shared genetic architecture, however follow-up with local correlation methods such as LAVA⁸⁰ could be advantageous. This would avoid issues such as heterogeneity resulting in a null correlation score despite strong correlation in localised regions. In addition, the identification of localised genomic regions associated with multiple traits would be more insightful for identifying shared biological mechanisms. The genetic correlation across fibrotic disease also only investigated a relatively small sample of traits, where mechanistic links may exist with other diseases that were not included.

While MR analyses described in Chapter 4 indicate that IPF and DM are unlikely to be linked by direct causal effects overall, this approach is limited by the assumption that both diseases are genetically and biologically homogeneous. This method therefore cannot rule causal effects that only act in a specific disease endotype or in combination with other biological factors.

A limitation across all methods used in this thesis is the use of data only from a GWAS of IPF susceptibility. As more GWAS are published that focus on disease progression rather

than simply disease risk, re-application of the methods used in this thesis may identify novel evidence of shared mechanisms. Identifying shared mechanisms relating to disease progression may also be more appropriate for highlighting drug repurposing opportunities, as treatments that slow or halt IPF disease progression are greatly needed.

A key strength of Chapter 2 is the use of GWAS summary data, which will typically be from studies with large sample sizes. A previous IPF PheWAS using the Vanderbilt repository BioVU¹¹³ and the Genomics England IPF PheWAS performed for this thesis were limited by statistical power. While the use of GWAS data may bias towards identifying associations with common traits, this is the first PheWAS to identify robust associations with a broad range of phenotypes for multiple IPF risk variants. The subsequent application of colocalisation analyses also provides greater confidence of shared causal variants.

The methodology of the MR study described chapter 4 aimed to produce a robust investigation of causal effects despite the biological and genetic complexity of DM and IPF. This study was strengthened by using continuous traits associated with DM and deliberate use of complementary sensitivity analyses to avoid common pitfalls of MR study designs.

The research conducted for this thesis investigates disease relationships at single SNP level through PheWAS (Chapter 2), on a genome-wide level through genetic correlation (Chapter 3), and investigates epidemiological relationships at a phenotypic level through MR (Chapter 4). This thesis was therefore particularly strengthened by use of a diverse range of methodological approaches to perform a multifaceted investigation of disease relationships with IPF. However, an important limitation to consider is that all methods applied in this thesis are based on genetic association data from one method (GWAS) applied to a single phenotype (IPF).

Many genetic variants identified through GWAS are common, often have a small effect size when considered individually, and do not account for a great deal of overall heritability

when considered collectively²⁸². As a relevant example, it has been demonstrated that fourteen of the established IPF genetic risk variants, including rs35705950 (*MUC5B*), only explain 12.4% of overall IPF disease liability in European populations⁹¹. In addition, the study by Leavy *et al.* (2021) highlighted that the individual contribution of most common genetic risk variants to IPF disease risk, except rs35705950 (*MUC5B*), is less than 1%. Insights from this thesis may therefore be limited because variability in variant contribution to tested phenotypes has not been considered. For example, while shared genetic associations identified at a single variant level (PheWAS, genetic colocalisation) are likely to be biologically informative, findings may be less meaningful if SNP contribution to overall heritability of either phenotype is low. Similarly, as previously discussed in Section 2.4, larger sample sizes may bias discovery towards traits that are more frequently tested, such as blood cell traits. Increased GWAS discovery through increasingly large sample sizes may increase the possibility of discovering genetic overlap, but any shared heritability identified may be based upon genetic variants with infinitesimally small effect sizes or those associated with generic, low-effect factors with non-specific effects on disease. This phenomenon may particularly affect post-GWAS methods that focus on broader shared heritability, such as genetic correlation and Mendelian randomisation. Going forward, it may be plausible to consider heritability thresholds in studies of genetic overlap to minimise these potential limitations. In addition, interactions between genetic variants, genes, and the environment are predicted to underlie a substantial proportion of the missing heritability of many phenotypic traits²⁸², therefore more complex and intricate statistical approaches may be required to understand epidemiological relationships between human diseases.

As the analyses performed throughout this thesis are primarily based on GWAS of a single phenotype (IPF), potential limitations associated with this phenotype should be discussed. For a patient to be diagnosed with IPE, a possible diagnosis of other ILDs must be excluded.

As IPF is a phenotype by exclusion, rather than a disease associated with a characteristic biological mechanism or single pathological feature, IPF cohorts are potentially heterogeneous and could be associated with a more diverse set of contributing environmental or genetic factors. This may explain the relative inconsistency of phenotypes found to be associated with IPF genetic risk variants in Chapter 2. Overall, this genetic and phenotypic heterogeneity may mean that some disease relationships with IPF are more complex than can be fully captured using the methods applied in this thesis.

However, despite these limitations the GWAS of IPF risk performed by Allen et al. (2022) represents the current highest published standard of research on common IPF genetic risk. Public GWAS data is a powerful and important resource, and the application of more targeted epidemiological approaches may have required a less comprehensive investigation. In addition, the methods applied in this thesis have produced novel findings that contribute to the overall understanding of IPF disease biology.

6.5 Future work

6.5.1 Expansion of colocalisation analyses

A core motivation of PheWAS is to understand the biology surrounding individual SNPs to provide insights into possible pathways of effect from genotype to phenotype. While the PheWAS described in Chapter 2 provided novel insights into the biology of multiple IPF risk variants, the prioritisation pipeline for follow up colocalisation analyses did not represent SNPs equally.

A first option to ensure equal representation of IPF SNPs would be to simply rank PheWAS results by p-value, then select an arbitrary number of independent associations per SNP for colocalisation analyses. Alternatively, a full colocalisation analysis of all associated

phenotypes may be a viable approach if the colocalisation method can be semi-automated. This possible approach is currently complicated by many publicly available GWAS having incomplete and inconsistently labelled data fields. However, databases such as the GWAS catalog are improving quality control standards of GWAS summary data by requiring mandatory data fields and using a centralised harmonisation pipeline. This possible approach is therefore becoming increasingly realistic⁶⁹.

6.5.2 Expanded investigation of shared genetic architecture across fibrotic disease traits

In Chapter 3, a sample of eight fibrotic diseases were selected for an exploratory genetic correlation analysis. However, a recent Delphi study classified 265 diseases as fibrotic⁵². While many fibrotic diseases may not have public GWAS summary data, a more systematic genetic correlation analysis of fibrotic disease traits may identify more traits with shared genetic architecture as identified between IPF and PBC.

6.5.3 Mendelian randomisation study assessing if sex hormones are causally associated with IPF risk

Work conducted in this thesis identified a shared causal variant driving a positive association between IPF risk and total testosterone levels, and unconfirmed overlap of signals associated with IPF risk and SHBG at *MAD1L1*, *TERC*, and *KNL1* loci. Total testosterone, bioavailable testosterone, and SHBG are genetically correlated¹⁷³ but may have independent effects on IPF risk. This therefore presents an opportunity to apply multivariable MR to discern if any of these factors are individually causally associated with IPF susceptibility. This study may provide insights into if the male predominance in IPF patient cohort²⁷⁹ is due to intrinsic causal effects. If no causal effects are suggested, this

would suggest that this effect is more likely to be driven by environmental factors such as occupational exposure or diagnostic bias^{279,280}.

6.6 Conclusion

In this thesis I have investigated disease relationships with IPF on a single variant, genome-wide, and phenotypic level. The analyses have performed have shown that IPF genetic risk variants are generally very pleiotropic, with the intriguing exception of rs35705950 (*MUC5B*) which appears to be almost exclusively associated with lung disease risk. This thesis contributes novel understanding of the biological effects of multiple IPF SNPs (*DPP9*, *FAM13A*, *TERC*, *IVD*, and *KNL1*) and supports the association between shortened telomeres and IPF. The findings of this thesis also highlight that a sample of fibrotic diseases are unlikely to share strong genetic overlap. Similarly, the epidemiological association between IPF and DM, and a potential biological association between mLOY and IPF, are shown here to be unlikely to be driven by direct causal effects. These findings contribute novel understanding of disease biology that may lead to improved treatments and better outcomes for IPF patients.

Reference list

1. Hutchinson, J., Fogarty, A., Hubbard, R. & McKeever, T. Global incidence and mortality of idiopathic pulmonary fibrosis: a systematic review. *European Respiratory Journal* **46**, 795–806 (2015).
2. Vancheri, C., Failla, M., Crimi, N. & Raghu, G. Idiopathic pulmonary fibrosis: a disease with similarities and links to cancer biology. doi:10.1183/09031936.00077309.
3. Plantier, L. *et al.* Physiology of the lung in idiopathic pulmonary fibrosis. *European Respiratory Review* **27**, (2018).
4. Nava, S. & Rubini, F. Lung and chest wall mechanics in ventilated patients with end stage idiopathic pulmonary fibrosis. *Thorax* **54**, 390–395 (1999).
5. Wijsenbeek, M., Suzuki, A. & Maher, T. M. Interstitial lung diseases. *The Lancet* **400**, 769–786 (2022).
6. Martinez, F. J. *et al.* Idiopathic pulmonary fibrosis. *Nature Reviews Disease Primers* **2017 3:1** **3**, 1–19 (2017).
7. Duckworth, A. *et al.* Telomere length and risk of idiopathic pulmonary fibrosis and chronic obstructive pulmonary disease: a mendelian randomisation study. *Lancet Respir Med* **9**, 285–294 (2021).
8. Müezziner, A., Zaineddin, A. K. & Brenner, H. A systematic review of leukocyte telomere length and age in adults. *Ageing Res Rev* **12**, 509–519 (2013).
9. Astuti, Y., Wardhana, A., Watkins, J. & Wulaningsih, W. Cigarette smoking and telomere length: A systematic review of 84 studies and meta-analysis. *Environ Res* **158**, 480–489 (2017).

10. Park, Y., Ahn, C. & Kim, T. H. Occupational and environmental risk factors of idiopathic pulmonary fibrosis: a systematic review and meta-analyses. *Scientific Reports* 2021 11:1 **11**, 1–10 (2021).
11. Molyneaux, P. L. & Maher, T. M. The role of infection in the pathogenesis of idiopathic pulmonary fibrosis. *European Respiratory Review* **22**, 376–381 (2013).
12. Wells, A. U. New insights into the treatment of CTD-ILD. *Nature Reviews Rheumatology* 2021 17:2 **17**, 79–80 (2021).
13. Schwaiblmair, M. *et al.* Drug Induced Interstitial Lung Disease. *Open Respir Med J* **6**, 63 (2012).
14. Raghu, G. *et al.* Diagnosis of Idiopathic Pulmonary Fibrosis. An Official ATS/ERS/JRS/ALAT Clinical Practice Guideline. *Am J Respir Crit Care Med* **198**, e44–e68 (2018).
15. Ley, B., Collard, H. R. & King, T. E. Clinical Course and Prediction of Survival in Idiopathic Pulmonary Fibrosis. <https://doi.org/10.1164/rccm.201006-0894CI> **183**, 431–440 (2012).
16. Ash, S. Y. & Washko, G. R. Interstitial lung abnormalities: risk and opportunity. *Lancet Respir Med* **5**, 95–96 (2017).
17. Patel, A. S. *et al.* Interstitial lung abnormalities in a large clinical lung cancer screening cohort: association with mortality and ILD diagnosis. *Respir Res* **24**, 1–8 (2023).
18. Hobbs, B. D. *et al.* Overlap of genetic risk between interstitial lung abnormalities and idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* **200**, 1402–1413 (2021).
19. Moll, M. *et al.* A Polygenic Risk Score for Idiopathic Pulmonary Fibrosis and Interstitial Lung Abnormalities. *Am J Respir Crit Care Med* **208**, 791–801 (2023).
20. Maher, T. M. PROFILEing idiopathic pulmonary fibrosis: rethinking biomarker discovery. *European Respiratory Review* **22**, 148–152 (2013).
21. Fainberg, H. P. *et al.* Forced vital capacity trajectories in patients with idiopathic pulmonary fibrosis: a secondary analysis of a multicentre, prospective, observational cohort. *Lancet Digit Health* **4**, e862–e872 (2022).

22. Organ, L. A. *et al.* Biomarkers of collagen synthesis predict progression in the PROFILE idiopathic pulmonary fibrosis cohort. *Respir Res* **20**, 1–10 (2019).
23. Maher, T. M. *et al.* An epithelial biomarker signature for idiopathic pulmonary fibrosis: an analysis from the multicentre PROFILE cohort study. *Lancet Respir Med* **5**, 946–955 (2017).
24. Richeldi, L. *et al.* Efficacy and safety of nintedanib in idiopathic pulmonary fibrosis. *N Engl J Med* **370**, 2071–2082 (2014).
25. King, T. E. *et al.* A phase 3 trial of pirfenidone in patients with idiopathic pulmonary fibrosis. *N Engl J Med* **370**, 2083–2092 (2014).
26. Nakazato, H., Oku, H., Yamane, S., Tsuruta, Y. & Suzuki, R. A novel anti-fibrotic agent pirfenidone suppresses tumor necrosis factor- α at the translational level. *Eur J Pharmacol* **446**, 177–185 (2002).
27. Spond, J. *et al.* Inhibition of experimental acute pulmonary inflammation by pirfenidone. *Pulm Pharmacol Ther* **16**, 207–214 (2003).
28. Fernandez, I. E. & Eickelberg, O. The impact of TGF- β on lung fibrosis: from targeting to biomarkers. *Proc Am Thorac Soc* **9**, 111–116 (2012).
29. Hisatomi, K. *et al.* Pirfenidone inhibits TGF- β 1-induced over-expression of collagen type I and heat shock protein 47 in A549 cells. *BMC Pulm Med* **12**, (2012).
30. Ruwanpura, S. M., Thomas, B. J. & Bardin, P. G. Pirfenidone: Molecular mechanisms and potential clinical applications in lung disease. *Am J Respir Cell Mol Biol* **62**, 413–422 (2020).
31. Wollin, L. *et al.* Mode of action of nintedanib in the treatment of idiopathic pulmonary fibrosis. *European Respiratory Journal* **45**, 1434–1445 (2015).
32. Corte, T. *et al.* Safety, tolerability and appropriate use of nintedanib in idiopathic pulmonary fibrosis. *Respir Res* **16**, 1–10 (2015).
33. Crestani, B. *et al.* Long-term safety and tolerability of nintedanib in patients with idiopathic pulmonary fibrosis: results from the open-label extension study, INPULSIS-ON. *Lancet Respir Med* **7**, 60–68 (2019).
34. Thabut, G. *et al.* Survival benefit of lung transplantation for patients with idiopathic pulmonary fibrosis. *Journal of Thoracic and Cardiovascular Surgery* **126**, 469–475 (2003).

35. Trachalaki, A., Sultana, N. & Wells, A. U. An update on current and emerging drug treatments for idiopathic pulmonary fibrosis. *Expert Opin Pharmacother* **24**, 1125–1142 (2023).
36. Selman, M. & Pardo, A. Revealing the pathogenic and aging-related mechanisms of the enigmatic idiopathic pulmonary fibrosis: An integral model. *Am J Respir Crit Care Med* **189**, 1161–1172 (2014).
37. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The Hallmarks of Aging. *Cell* **153**, 1194 (2013).
38. Collard, H. R. *et al.* Burden of illness in idiopathic pulmonary fibrosis. *J Med Econ* **15**, 829–835 (2012).
39. King, C. S. & Nathan, S. D. Idiopathic pulmonary fibrosis: effects and optimal management of comorbidities. *Lancet Respir Med* **5**, 72–84 (2017).
40. Mannino, D. M. & Buist, A. S. Global burden of COPD: risk factors, prevalence, and future trends. *Lancet* **370**, 765–773 (2007).
41. Leiter, A., Veluswamy, R. R. & Wisnivesky, J. P. The global burden of lung cancer: current status and future trends. *Nature Reviews Clinical Oncology* **2023 20:9** **20**, 624–639 (2023).
42. Reynolds, C. J. *et al.* The causal relationship between gastro-oesophageal reflux disease and idiopathic pulmonary fibrosis: a bidirectional two-sample Mendelian randomisation study. *European Respiratory Journal* **61**, (2023).
43. Ryerson, C. J. *et al.* Depression is a common and chronic comorbidity in patients with interstitial lung disease. *Respirology* **17**, 525–532 (2012).
44. Bai, L. *et al.* Idiopathic pulmonary fibrosis and diabetes mellitus: a meta-analysis and systematic review. *Respir Res* **22**, 1–10 (2021).
45. Kumar, V. & Nawroth, P. P. Is the association between diabetes mellitus and pulmonary fibrosis real? *Nature Reviews Endocrinology* **2021 17:12** **17**, 703–704 (2021).
46. Suissa, S., Kezouh, A. & Ernst, P. Inhaled Corticosteroids and the Risks of Diabetes Onset and Progression. *Am J Med* **123**, 1001–1006 (2010).
47. DeFronzo, R. A. *et al.* Type 2 diabetes mellitus. *Nature Reviews Disease Primers* **2015 1:1** **1**, 1–22 (2015).

48. Luppi, F., Kalluri, M., Faverio, P., Kreuter, M. & Ferrara, G. Idiopathic pulmonary fibrosis beyond the lung: understanding disease mechanisms to improve diagnosis and management. *Respiratory Research* 2021 22:1 **22**, 1–16 (2021).
49. Kreuter, M. *et al.* Impact of Comorbidities on Mortality in Patients with Idiopathic Pulmonary Fibrosis. *PLoS One* **11**, e0151425 (2016).
50. Henderson, N. C., Rieder, F. & Wynn, T. A. Fibrosis: from mechanisms to medicines. *Nature* 2020 587:7835 **587**, 555–566 (2020).
51. Distler, J. H. W. *et al.* Shared and distinct mechanisms of fibrosis. *Nature Reviews Rheumatology* 2019 15:12 **15**, 705–730 (2019).
52. Massen, G. M. *et al.* Classifying the unclassifiable—a Delphi study to reach consensus on the fibrotic nature of diseases. *QJM: An International Journal of Medicine* **116**, 429–435 (2023).
53. Seitz, H. K. *et al.* Alcoholic liver disease. *Nature Reviews Disease Primers* 2018 4:1 **4**, 1–22 (2018).
54. Wynn, T. A. & Ramalingam, T. R. Mechanisms of fibrosis: therapeutic translation for fibrotic disease. *Nature Medicine* 2012 18:7 **18**, 1028–1040 (2012).
55. Batlle, E. & Massagué, J. Immunity Review Transforming Growth Factor- β Signaling in Immunity and Cancer. *Immunity* **50**, 924–940 (2019).
56. Sano, S. *et al.* Hematopoietic loss of Y chromosome leads to cardiac fibrosis and heart failure mortality. *Science* (1979) **377**, 292–297 (2022).
57. Lleo, A. *et al.* Y chromosome loss in male patients with primary biliary cirrhosis. *J Autoimmun* **41**, 87–91 (2013).
58. Massen, G. M. *et al.* Using Routinely Collected Electronic Healthcare Record Data to Investigate Fibrotic Multimorbidity in England. *Clin Epidemiol* **16**, 433–443 (2024).
59. Piovesan, A. *et al.* Human protein-coding genes and gene feature statistics in 2019. *BMC Res Notes* **12**, (2019).
60. Omenn, G. S. Reflections on the HUPO Human Proteome Project, the Flagship Project of the Human Proteome Organization, at 10 Years. *Mol Cell Proteomics* **20**, 100062 (2021).

61. McCulloch, S. D. & Kunkel, T. A. The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. *Cell Res* **18**, 148 (2008).
62. Chatterjee, N. & Walker, G. C. Mechanisms of DNA damage, repair and mutagenesis. *Environ Mol Mutagen* **58**, 235 (2017).
63. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics* 2018 19:9 **19**, 581–590 (2018).
64. Uffelmann, E. *et al.* Genome-wide association studies. *Nature Reviews Methods Primers* 2021 1:1 **1**, 1–21 (2021).
65. Belmont, J. W. *et al.* A haplotype map of the human genome. *Nature* 2005 437:7063 **437**, 1299–1320 (2005).
66. Altshuler, D. L. *et al.* A map of human genome variation from population-scale sequencing. *Nature* 2010 467:7319 **467**, 1061–1073 (2010).
67. Dehghan, A. Genome-wide association studies. *Methods in Molecular Biology* **1793**, 37–49 (2018).
68. Mills, M. C. & Rahal, C. A scientometric review of genome-wide association studies. *Communications Biology* 2019 2:1 **2**, 1–11 (2019).
69. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res* **51**, D977–D985 (2023).
70. Sterne, J. A. C., Smith, G. D. & Cox, D. R. Sifting the evidence—what’s wrong with significance tests? *BMJ: British Medical Journal* **322**, 226 (2001).
71. Panagiotou, O. A. *et al.* What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *Int J Epidemiol* **41**, 273–286 (2012).
72. Fadista, J., Manning, A. K., Florez, J. C. & Groop, L. The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *European Journal of Human Genetics* 2016 24:8 **24**, 1202–1205 (2016).
73. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nature Genetics* 2015 47:8 **47**, 856–860 (2015).
74. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nature Genetics* 2019 51:9 **51**, 1339–1348 (2019).

75. Denny, J. C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics* **26**, 1205 (2010).
76. Wallace, C. Statistical Testing of Shared Genetic Control for Potentially Related Traits. *Genet Epidemiol* **37**, 802–813 (2013).
77. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* **10**, (2014).
78. Denny, J. C., Bastarache, L. & Roden, D. M. Phenome-Wide Association Studies as a Tool to Advance Precision Medicine. *Annu Rev Genomics Hum Genet* **17**, 353–373 (2016).
79. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nature Genetics* 2015 47:11 **47**, 1236–1241 (2015).
80. Werme, J., van der Sluis, S., Posthuma, D. & de Leeuw, C. A. An integrated framework for local genetic correlation analysis. *Nature Genetics* 2022 54:3 **54**, 274–282 (2022).
81. Gormley, M. *et al.* A multivariable Mendelian randomization analysis investigating smoking and alcohol consumption in oral and oropharyngeal cancer. *Nature Communications* 2020 11:1 **11**, 1–10 (2020).
82. Hartley, A. *et al.* Using multivariable Mendelian randomization to estimate the causal effect of bone mineral density on osteoarthritis risk, independently of body mass index. *Int J Epidemiol* **51**, 1254–1267 (2022).
83. Richardson, T. G. *et al.* Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable Mendelian randomisation analysis. *PLoS Med* **17**, e1003062 (2020).
84. Richardson, T. G., Sanderson, E., Elsworth, B., Tilling, K. & Smith, G. D. Use of genetic variation to separate the effects of early and later life adiposity on disease risk: mendelian randomisation study. *BMJ* **369**, (2020).
85. Sanderson, E. *et al.* Mendelian randomization. *Nature Reviews Methods Primers* 2022 2:1 **2**, 1–21 (2022).
86. Fingerlin, T. E. *et al.* Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. *Nat Genet* **45**, 613–620 (2013).

87. Allen, R. J. *et al.* Genetic variants associated with susceptibility to idiopathic pulmonary fibrosis in people of European ancestry: a genome-wide association study. *Lancet Respir Med* **5**, 869–880 (2017).
88. Noth, I. *et al.* Genetic variants associated with idiopathic pulmonary fibrosis susceptibility and mortality: A genome-wide association study. *Lancet Respir Med* **1**, 309–317 (2013).
89. Allen, R. J. *et al.* Genome-Wide Association Study of Susceptibility to Idiopathic Pulmonary Fibrosis. *Am J Respir Crit Care Med* **201**, 564–574 (2020).
90. Allen, R. J. *et al.* Genome-wide association study across five cohorts identifies five novel loci associated with idiopathic pulmonary fibrosis. *Thorax* **77**, 829–833 (2022).
91. Leavy, O. C. *et al.* Proportion of Idiopathic Pulmonary Fibrosis Risk Explained by Known Common Genetic Loci in European Populations. *Am J Respir Crit Care Med* **203**, 775–778 (2021).
92. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, (2016).
93. Seibold, M. A. *et al.* A Common MUC5B Promoter Polymorphism and Pulmonary Fibrosis. *N Engl J Med* **364**, 1503 (2011).
94. Furusawa, H. *et al.* Common idiopathic pulmonary fibrosis risk variants are associated with hypersensitivity pneumonitis. *Thorax* **77**, 508 (2022).
95. Platenburg, M. G. J. P. *et al.* The MUC5B promoter risk allele for idiopathic pulmonary fibrosis predisposes to asbestosis. *European Respiratory Journal* **55**, (2020).
96. Juge, P.-A. *et al.* MUC5B Promoter Variant and Rheumatoid Arthritis with Interstitial Lung Disease. *N Engl J Med* **379**, 2209 (2018).
97. Hancock, L. A. *et al.* Muc5b overexpression causes mucociliary dysfunction and enhances lung fibrosis in mice. *Nature Communications* **2018 9:1** **9**, 1–10 (2018).
98. Schwartz, D. A., Blumhagen, R. Z. & Fingerlin, T. E. Evolution of the Gain-of-Function MUC5B Promoter Variant. *Am J Respir Crit Care Med* **206**, 1189–1191 (2022).

99. Verma, A. *et al.* A MUC5B Gene Polymorphism, rs35705950-T, Confers Protective Effects Against COVID-19 Hospitalization but Not Severe Disease or Mortality. *Am J Respir Crit Care Med* **206**, 1220–1229 (2022).
100. van Moorsel, C. H. M. *et al.* The MUC5B Promoter Polymorphism Associates With Severe COVID-19 in the European Population. *Front Med (Lausanne)* **8**, 668024 (2021).
101. Allen, R. J. *et al.* Genetic overlap between idiopathic pulmonary fibrosis and COVID-19. *European Respiratory Journal* **60**, (2022).
102. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science (1979)* **347**, (2015).
103. The Human Protein Atlas. <https://www.proteinatlas.org/>.
104. D. L. Kellogg, 3rd, D. L. Kellogg, J., Musi, N. & Nambiar, A. M. Cellular Senescence in Idiopathic Pulmonary Fibrosis. *Curr Mol Biol Rep* **7**, 31 (2021).
105. Armanios, M. Y. *et al.* Telomerase Mutations in Families with Idiopathic Pulmonary Fibrosis. *New England Journal of Medicine* **356**, 1317–1326 (2007).
106. Barbaro, P. M., Ziegler, D. S. & Reddel, R. R. The wide-ranging clinical implications of the short telomere syndromes. *Intern Med J* **46**, 393–403 (2016).
107. Estrada, J. C. *et al.* Human mesenchymal stem cell-replicative senescence and oxidative stress are closely linked to aneuploidy. *Cell Death & Disease* **2013** 4:6 **4**, e691–e691 (2013).
108. Naylor, R. M. & Van Deursen, J. M. Aneuploidy in Cancer and Aging. <https://doi.org/10.1146/annurev-genet-120215-035303> **50**, 45–66 (2016).
109. Dhindsa, R. S. *et al.* Identification of a missense variant in SPDL1 associated with idiopathic pulmonary fibrosis. *Communications Biology* **2021** 4:1 **4**, 1–8 (2021).
110. Woodcock, H. V. *et al.* The mTORC1/4E-BP1 axis represents a critical signaling node during fibrogenesis. *Nature Communications* **2019** 10:1 **10**, 1–16 (2019).
111. Platé, M., Guillotin, D. & Chambers, R. C. The promise of mTOR as a therapeutic target pathway in idiopathic pulmonary fibrosis. *Eur Respir Rev* **29**, (2020).

112. Bastarache, L., Denny, J. C. & Roden, D. M. Phenome-Wide Association Studies. *JAMA* **327**, 75–76 (2022).
113. Claar, D. D. *et al.* A phenome-wide association study identifies a novel asthma risk locus near TERC. *Am J Respir Crit Care Med* **193**, 98–100 (2016).
114. Davis, L. Psychiatric Genomics, Phenomics, and Ethics Research In A 270,000-Person Biobank (BioVU). *European Neuropsychopharmacology* **29**, S739–S740 (2019).
115. Staley, J. R. *et al.* PhenoScanner: a database of human genotype-phenotype associations. *Bioinformatics* **32**, 3207–3209 (2016).
116. Kamat, M. A. *et al.* PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. *Bioinformatics* **35**, 4851–4853 (2019).
117. Elsworth, B. *et al.* The MRC IEU OpenGWAS data infrastructure. *bioRxiv* 2020.08.10.244293 (2020) doi:10.1101/2020.08.10.244293.
118. Torjesen, I. Genomes of 100 000 people will be sequenced to create an open access research resource. *BMJ* **347**, f6690 (2013).
119. Peplow, M. The 100 000 Genomes Project. *BMJ* **353**, (2016).
120. Turnbull, C. *et al.* The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ* **361**, (2018).
121. Rastegar-Mojarad, M., Ye, Z., Kolesar, J. M., Hebring, S. J. & Lin, S. M. Opportunities for drug repositioning from phenome-wide association studies. *Nature Biotechnology* 2015 33:4 **33**, 342–345 (2015).
122. Boettger, L. M., Handsaker, R. E., Zody, M. C. & Mccarroll, S. A. Structural haplotypes and recent evolution of the human 17q21.31 region. *Nat Genet* **44**, 881 (2012).
123. Mohammadi-Nejad, A. R. *et al.* Mapping brain endophenotypes associated with idiopathic pulmonary fibrosis genetic risk. *EBioMedicine* **86**, (2022).
124. Genomics England. National Genomic Research Library. Preprint at (2017).
125. Carroll, R. J., Bastarache, L. & Denny, J. C. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* **30**, 2375–2376 (2014).

126. Perkins, T. N. & Oury, T. D. The perplexing role of RAGE in pulmonary fibrosis: causality or casualty? *Ther Adv Respir Dis* **15**, (2021).
127. Parcesepe, G. *et al.* Shared genetic risk factors: implications for treatment of idiopathic pulmonary fibrosis and systemic hypertension. *ERJ Open Res* **9**, 00457–02023 (2023).
128. Tellechea, M. L. & Pirola, C. J. The impact of hypertension on leukocyte telomere length: a systematic review and meta-analysis of human studies. *Journal of Human Hypertension* 2017 31:2 **31**, 99–105 (2016).
129. Kuo, C. L., Pilling, L. C., Kuchel, G. A., Ferrucci, L. & Melzer, D. Telomere length and aging-related outcomes in humans: A Mendelian randomization study in 261,000 older participants. *Aging Cell* **18**, e13017 (2019).
130. Hobbs, B. D. *et al.* Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis. *Nature Genetics* 2017 49:3 **49**, 426–432 (2017).
131. Duckworth, A. *et al.* Study of the associations between short telomeres, sex hormones and pulmonary fibrosis. *medRxiv* 2022.09.29.22280270 (2022) doi:10.1101/2022.09.29.22280270.
132. Fang, C. *et al.* Relation between sex hormones and leucocyte telomere length in men with idiopathic pulmonary fibrosis. *Respirology* **25**, 1265–1273 (2020).
133. Kozlitina, J. & Garcia, C. K. Red Blood Cell Size Is Inversely Associated with Leukocyte Telomere Length in a Large Multi-Ethnic Population. *PLoS One* **7**, e51046 (2012).
134. Adams, C. D. & Boutwell, B. B. A Mendelian randomization study of telomere length and blood-cell traits. *Scientific Reports* 2020 10:1 **10**, 1–9 (2020).
135. Salvagno, G. L., Sanchis-Gomar, F., Picanza, A. & Lippi, G. Red blood cell distribution width: A simple parameter with multiple clinical applications. *Crit Rev Clin Lab Sci* **52**, 86–105 (2015).
136. Goldberg, I. *et al.* A Longitudinal Assessment of the Natural Change in Haemoglobin, Haematocrit, and Mean Corpuscular Volume with Age. *Acta Haematol* **146**, 206–213 (2023).

137. Patel, K. V. *et al.* Red Cell Distribution Width and Mortality in Older Adults: A Meta-analysis. *The Journals of Gerontology: Series A* **65A**, 258–265 (2010).
138. Lynch, C. J. *et al.* The RNA Polymerase II Factor RPAP1 Is Critical for Mediator-Driven Transcription and Cell Identity. *Cell Rep* **22**, 396–410 (2018).
139. Sun, T. *et al.* TGF β 2 and TGF β 3 isoforms drive fibrotic disease pathogenesis. *Sci Transl Med* **13**, 407 (2021).
140. Barratt, S. L. *et al.* Differential expression of VEGF-Axxx isoforms is critical for development of pulmonary fibrosis. *Am J Respir Crit Care Med* **196**, 479–493 (2017).
141. Xia, H. *et al.* Calcium-binding protein S100A4 confers mesenchymal progenitor cell fibrogenicity in idiopathic pulmonary fibrosis. *J Clin Invest* **127**, 2586–2597 (2017).
142. Wang, L., Babushkin, N., Liu, Z. & Liu, X. Trans-eQTL mapping in gene sets identifies network effects of genetic variants. *Cell genomics* **4**, (2024).
143. Vösa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat Genet* **53**, 1300–1310 (2021).
144. Salat, D. H. *et al.* Age-associated alterations in cortical gray and white matter signal intensity and gray to white matter contrast. *Neuroimage* **48**, 21–28 (2009).
145. Oldham, J. M. *et al.* Thyroid Disease Is Prevalent and Predicts Survival in Patients With Idiopathic Pulmonary Fibrosis. *Chest* **148**, 692–700 (2015).
146. Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D. J. & Richards, J. B. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nature Reviews Genetics* 2017 19:2 **19**, 110–124 (2017).
147. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five Years of GWAS Discovery. *The American Journal of Human Genetics* **90**, 7–24 (2012).
148. Bulik-Sullivan, B. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* 2015 47:3 **47**, 291–295 (2015).
149. Ni, G., Moser, G., Wray, N. R. & Hong Lee, S. Estimation of Genetic Correlation via Linkage Disequilibrium Score Regression and Genomic

Restricted Maximum Likelihood. *The American Journal of Human Genetics* **102**, 1185–1194 (2018).

150. Mosher, C. L. & Mentz, R. J. Cardiovascular implications of idiopathic pulmonary fibrosis: A way forward together? *Am Heart J* **226**, 69 (2020).
151. Fadista, J. *et al.* Shared genetic etiology between idiopathic pulmonary fibrosis and COVID-19 severity. *EBioMedicine* **65**, 103277 (2021).
152. Sheng, G. *et al.* Viral Infection Increases the Risk of Idiopathic Pulmonary Fibrosis: A Meta-Analysis. *Chest* **157**, 1175–1187 (2020).
153. Cilli, A. *et al.* Characteristics and outcomes of COVID-19 patients with IPF: A multi-center retrospective study. *Respir Med Res* **81**, 100900 (2022).
154. Mirza, S., Clay, R. D., Koslow, M. A. & Scanlon, P. D. COPD Guidelines: A Review of the 2018 GOLD Report. *Mayo Clin Proc* **93**, 1488–1502 (2018).
155. Raghu, G., Amatto, V. C., Behr, J. & Stowasser, S. Comorbidities in idiopathic pulmonary fibrosis patients: a systematic literature review. *Eur Respir J* **46**, 1113–1130 (2015).
156. Selman, M. & Pardo, A. Revealing the pathogenic and aging-related mechanisms of the enigmatic idiopathic pulmonary fibrosis: An integral model. *Am J Respir Crit Care Med* **189**, 1161–1172 (2014).
157. Selman, M., Martinez, F. J. & Pardo, A. Why does an aging smoker's lung develop idiopathic pulmonary fibrosis and not chronic obstructive pulmonary disease? *Am J Respir Crit Care Med* **199**, 279–285 (2019).
158. Nathan, S. D. *et al.* Using forced vital capacity (FVC) in the clinic to monitor patients with idiopathic pulmonary fibrosis (IPF): pros and cons. *Expert Rev Respir Med* **15**, 175–181 (2021).
159. Alder, J. K. *et al.* Short telomeres are a risk factor for idiopathic pulmonary fibrosis. *Proc Natl Acad Sci U S A* **105**, 13051–13056 (2008).
160. Adegunsoye, A. *et al.* Telomere length associates with chronological age and mortality across racially diverse pulmonary fibrosis cohorts. *Nature Communications* 2023 14:1 **14**, 1–13 (2023).
161. Northrop-Clewes, C. A. & Thurnham, D. I. Biomarkers for the differentiation of anemia and their clinical usefulness. *J Blood Med* **4**, 11 (2013).

162. George, G. *et al.* Short telomeres, telomeropathy, and subclinical extrapulmonary organ damage in patients with interstitial lung disease. *Chest* **147**, 1549–1557 (2015).
163. Kozlitina, J. & Garcia, C. K. Red Blood Cell Size Is Inversely Associated with Leukocyte Telomere Length in a Large Multi-Ethnic Population. *PLoS One* **7**, 51046 (2012).
164. Diaz De Leon, A. *et al.* Subclinical Lung Disease, Macrocytosis, and Premature Graying in Kindreds With Telomerase (TERT) Mutations. *Chest* **140**, 753 (2011).
165. Salvagno, G. L., Sanchis-Gomar, F., Picanza, A. & Lippi, G. Red blood cell distribution width: A simple parameter with multiple clinical applications. *Crit Rev Clin Lab Sci* **52**, 86–105 (2015).
166. Balasubramanian, A. *et al.* Haemoglobin as a biomarker for clinical outcomes in chronic obstructive pulmonary disease. doi:10.1183/23120541.00068-2021.
167. Stuart, B. D. *et al.* Effect of telomere length on survival in patients with idiopathic pulmonary fibrosis: An observational cohort study with independent validation. *Lancet Respir Med* **2**, 557–565 (2014).
168. Shen, M., Zhang, F. & Zhang, X. Primary biliary cirrhosis complicated with interstitial lung disease: A prospective Study in 178 patients. *J Clin Gastroenterol* **43**, 676–679 (2009).
169. Lee, H. E. *et al.* Histopathologic findings in lung biopsies from patients with primary biliary cholangitis. *Hum Pathol* **82**, 177–186 (2018).
170. Liu, B., Zhang, F. C., Zhang, Z. L., Zhang, W. & Gao, L. X. Interstitial lung disease and Sjögren's syndrome in primary biliary cirrhosis: A causal or casual association? *Clin Rheumatol* **27**, 1299–1306 (2008).
171. Gulamhusein, A. F. & Hirschfield, G. M. Primary biliary cholangitis: pathogenesis and therapeutic opportunities. *Nature Reviews Gastroenterology & Hepatology* 2019 17:2 **17**, 93–110 (2019).
172. Liu, B. *et al.* Aberrant TGF- β 1 signaling contributes to the development of primary biliary cirrhosis in murine model. *World Journal of Gastroenterology: WJG* **19**, 5828 (2013).

173. Ruth, K. S. *et al.* Using human genetics to understand the disease impacts of testosterone in men and women. *Nat Med* **26**, 252 (2020).
174. Petrovski, S. *et al.* An exome sequencing study to assess the role of rare genetic variation in pulmonary fibrosis. *Am J Respir Crit Care Med* **196**, 82–93 (2017).
175. Zhang, D. *et al.* Rare and Common Variants in KIF15 Contribute to Genetic Risk of Idiopathic Pulmonary Fibrosis. *Am J Respir Crit Care Med* **206**, 56–69 (2022).
176. Liu, J. *et al.* Trends in the incidence of diabetes mellitus: results from the Global Burden of Disease Study 2017 and implications for diabetes mellitus prevention. *BMC Public Health* **20**, 1–12 (2020).
177. Meier, J. J. *et al.* Functional Assessment of Pancreatic β -Cell Area in Humans. *Diabetes* **58**, 1595–1603 (2009).
178. Chatterjee, S., Khunti, K. & Davies, M. J. Type 2 diabetes. *The Lancet* **389**, 2239–2251 (2017).
179. Kahn, S. E., Cooper, M. E. & Del Prato, S. Pathophysiology and treatment of type 2 diabetes: perspectives on the past, present, and future. *The Lancet* **383**, 1068–1083 (2014).
180. Lindström, J. & Tuomilehto, J. The Diabetes Risk Score A practical tool to predict type 2 diabetes risk. *Diabetes Care* **26**, 725–731 (2003).
181. Li, C. *et al.* Associations Between Diabetes and Idiopathic Pulmonary Fibrosis: a Study-level Pooled Analysis of 26 Million People. *J Clin Endocrinol Metab* **106**, 3367–3380 (2021).
182. Hyldgaard, C., Hilberg, O. & Bendstrup, E. How does comorbidity influence survival in idiopathic pulmonary fibrosis? *Respir Med* **108**, 647–653 (2014).
183. Tuleta, I. & Frangogiannis, N. G. Fibrosis of the diabetic heart: Clinical significance, molecular mechanisms, and therapeutic opportunities. *Adv Drug Deliv Rev* **176**, 113904 (2021).
184. Russo, I. & Frangogiannis, N. G. Diabetes-associated cardiac fibrosis: Cellular effectors, molecular mechanisms and therapeutic opportunities. *J Mol Cell Cardiol* **90**, 84–93 (2016).

185. Park, J. *et al.* Risk of liver fibrosis in patients with prediabetes and diabetes mellitus. *PLoS One* **17**, e0269070 (2022).
186. Ampuero, J. *et al.* Significant fibrosis predicts new-onset diabetes mellitus and arterial hypertension in patients with NASH. *J Hepatol* **73**, 17–25 (2020).
187. Loomba, R. *et al.* Association between diabetes, family history of diabetes, and risk of nonalcoholic steatohepatitis and fibrosis. *Hepatology* **56**, 943–951 (2012).
188. Karihaloo, A. Anti-fibrosis therapy and diabetic nephropathy. *Curr Diab Rep* **12**, 414–422 (2012).
189. Tuleta, I. & Frangogiannis, N. G. Diabetic fibrosis. *Biochim Biophys Acta Mol Basis Dis* **1867**, (2021).
190. Talakatta, G. *et al.* Diabetes induces fibrotic changes in the lung through the activation of TGF- β signaling pathways. *Scientific Reports* 2018 8:1 **8**, 1–15 (2018).
191. Vasbinder, A. *et al.* Inflammation, Hyperglycemia, and Adverse Outcomes in Individuals With Diabetes Mellitus Hospitalized for COVID-19. *Diabetes Care* **45**, 692–700 (2022).
192. Erener, S. Diabetes, infection risk and COVID-19. *Mol Metab* **39**, 101044 (2020).
193. Lombardi, A., Agarwal, S., Schechter, C. & Tomer, Y. In-Hospital Hyperglycemia Is Associated With Worse Outcomes in Patients Admitted With COVID-19. *Diabetes Care* **45**, 2683–2688 (2022).
194. Smith, G. D. & Ebrahim, S. ‘Mendelian randomization’: can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* **32**, 1–22 (2003).
195. Hartwig, F. P., Davies, N. M., Hemani, G. & Smith, G. D. Two-sample Mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique. *Int J Epidemiol* **45**, 1717 (2016).
196. Smith, G. D. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet* **23**, R89 (2014).
197. Burgess, S. *et al.* Guidelines for performing Mendelian randomization investigations: update for summer 2023. *Wellcome Open Res* **4**, 186 (2019).

198. Burgess, S. & Cronjé, H. T. Incorporating biological and clinical insights into variant choice for Mendelian randomisation: examples and principles. *eGastroenterology* **2**, e100042 (2024).
199. Elsayed, N. A. *et al.* 2. Classification and Diagnosis of Diabetes: Standards of Care in Diabetes—2023. *Diabetes Care* **46**, S19–S40 (2023).
200. Marshall, J. Coarsening Bias: How Coarse Treatment Measurement Upwardly Biases Instrumental Variable Estimates. *Political Analysis* **24**, 157–171 (2016).
201. Tudball, M. J. *et al.* Mendelian randomisation with coarsened exposures. *Genet Epidemiol* **45**, 338 (2021).
202. Burgess, S. & Labrecque, J. A. Mendelian randomization with a binary exposure variable: interpretation and presentation of causal estimates. *Eur J Epidemiol* **33**, 947–952 (2018).
203. Garfield, V., Salzmänn, A., Burgess, S. & Chaturvedi, N. A Guide for Selection of Genetic Instruments in Mendelian Randomization Studies of Type 2 Diabetes and HbA1c: Toward an Integrated Approach. *Diabetes* **72**, 175–183 (2023).
204. Sanderson, E. Multivariable Mendelian Randomization and Mediation. *Cold Spring Harb Perspect Med* **11**, 1–12 (2021).
205. Moss, S. *et al.* Assessing Causal Relationships Between Diabetes Mellitus and Idiopathic Pulmonary Fibrosis. *medRxiv* 2024.01.04.24300827 (2024) doi:10.1101/2024.01.04.24300827.
206. Onengut-Gumuscu, S. *et al.* Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nature GeNetics VOLUME* **47**, 14 (2015).
207. Xue, A. *et al.* Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nature Communications* **2018 9:1** **9**, 1–14 (2018).
208. Bonàs-Guarch, S. *et al.* Re-analysis of public genetic data reveals a rare X-chromosomal variant associated with type 2 diabetes. *Nat Commun* **9**, (2018).
209. Wheeler, E. *et al.* Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide meta-analysis. *PLoS Med* **14**, e1002383 (2017).

210. Chen, J. *et al.* The Trans-Ancestral Genomic Architecture of Glycemic Traits. *Nat Genet* **53**, 840 (2021).
211. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Hum Mol Genet* **27**, 3641–3649 (2018).
212. Machiela, M. J. & Chanock, S. J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31**, 3555–3557 (2015).
213. Yavorska, O. O. & Burgess, S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *Int J Epidemiol* **46**, 1734–1739 (2017).
214. Bowden, J., Smith, G. D. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol* **44**, 512–525 (2015).
215. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet Epidemiol* **40**, 304–314 (2016).
216. Verbanck, M., Chen, C. Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature Genetics* **50**, 693–698 (2018).
217. Rees, J. M. B., Wood, A. M., Dudbridge, F. & Burgess, S. Robust methods in Mendelian randomization via penalization of heterogeneous causal estimates. *PLoS One* **14**, (2019).
218. Slob, E. A. W. & Burgess, S. A comparison of robust Mendelian randomization methods using summary data. *Genet Epidemiol* **44**, 313 (2020).
219. Burgess, S. & Thompson, S. G. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *Am J Epidemiol* **181**, 251–260 (2015).
220. Grant, A. J. & Burgess, S. Pleiotropy robust methods for multivariable Mendelian randomization. *Stat Med* **40**, 5813–5830 (2021).

221. Rees, J. M. B., Wood, A. M. & Burgess, S. Extending the MR-Egger method for multivariable Mendelian randomization to correct for both measured and unmeasured pleiotropy. *Stat Med* **36**, 4705–4718 (2017).
222. Kang, Q., Ren, J., Cong, J. & Yu, W. Diabetes mellitus and idiopathic pulmonary fibrosis: a Mendelian randomization study. *BMC Pulm Med* **24**, 1–11 (2024).
223. Ma, X., Zhang, Y., Wu, F., Liu, X. & Zhang, W. Mendelian randomisation highlights type 1 diabetes as a causal determinant of idiopathic pulmonary fibrosis. *Diabetology & Metabolic Syndrome* 2024 16:1 **16**, 1–7 (2024).
224. Martinez, F. J. *et al.* Idiopathic pulmonary fibrosis. *Nature Reviews Disease Primers* 2017 3:1 **3**, 1–19 (2017).
225. Zhu, J. *et al.* A Causal Atlas on Comorbidities in Idiopathic Pulmonary Fibrosis: A Bidirectional Mendelian Randomization Study. *Chest* **164**, 429–440 (2023).
226. Faverio, P. *et al.* Nutritional assessment in idiopathic pulmonary fibrosis: a prospective multicentre study. *ERJ Open Res* **8**, (2022).
227. Alakhras, M., Decker, P. A., Nadrous, H. F., Collazo-Clavell, M. & Ryu, J. H. Body Mass Index and Mortality in Patients With Idiopathic Pulmonary Fibrosis. *Chest* **131**, 1448–1453 (2007).
228. Blüher, M. Obesity: global epidemiology and pathogenesis. *Nature Reviews Endocrinology* 2019 15:5 **15**, 288–298 (2019).
229. Albuquerque, D., Nóbrega, C., Manco, L. & Padez, C. The contribution of genetics and environment to obesity. *Br Med Bull* **123**, 159–173 (2017).
230. Justice, A. E. *et al.* Genome-wide meta-analysis of 241,258 adults accounting for smoking behaviour identifies novel loci for obesity traits. *Nature Communications* 2017 8:1 **8**, 1–19 (2017).
231. Bellou, V., Belbasis, L. & Evangelou, E. Tobacco Smoking and Risk for Pulmonary Fibrosis: A Prospective Cohort Study From the UK Biobank. *Chest* **160**, 983–993 (2021).
232. Cheng, F. *et al.* Shortened Leukocyte Telomere Length Is Associated With Glycemic Progression in Type 2 Diabetes: A Prospective and Mendelian Randomization Analysis. *Diabetes Care* **45**, 701–709 (2022).

233. Grunnet, L. G. *et al.* Leukocyte telomere length is associated with elevated plasma glucose and HbA1c in young healthy men independent of birth weight. *Scientific Reports* 2019 9:1 **9**, 1–8 (2019).
234. Verhulst, S. *et al.* A short leucocyte telomere length is associated with development of insulin resistance. *Diabetologia* **59**, 1258–1265 (2016).
235. Njajou, O. T. *et al.* Shorter telomeres are associated with obesity and weight gain in the elderly. *International Journal of Obesity* 2012 36:9 **36**, 1176–1179 (2011).
236. Tamura, Y. *et al.* β -Cell Telomere Attrition in Diabetes: Inverse Correlation Between HbA1c and Telomere Length. *J Clin Endocrinol Metab* **99**, 2771–2777 (2014).
237. Ardestani, A., Lupse, B., Kido, Y., Leibowitz, G. & Maedler, K. Cell Metabolism mTORC1 Signaling: A Double-Edged Sword in Diabetic β Cells. *Cell Metab* **27**, 314–331 (2018).
238. Yuan, T. *et al.* Reciprocal regulation of mTOR complexes in pancreatic islets from humans with type 2 diabetes. *Diabetologia* **60**, 668–678 (2017).
239. Raghu, G. *et al.* Idiopathic Pulmonary Fibrosis (an Update) and Progressive Pulmonary Fibrosis in Adults: An Official ATS/ERS/JRS/ALAT Clinical Practice Guideline. <https://doi.org/10.1164/rccm.202202-0399ST> **205**, E18–E47 (2022).
240. Li, J. X. & Cummins, C. L. Fresh insights into glucocorticoid-induced diabetes mellitus and new therapeutic directions. *Nature Reviews Endocrinology* 2022 18:9 **18**, 540–557 (2022).
241. Machiela, M. J. & Chanock, S. J. The Ageing Genome, Clonal Mosaicism and Chronic Disease. *Curr Opin Genet Dev* **42**, 8 (2017).
242. Guo, X. *et al.* Mosaic loss of human Y chromosome: what, how and why. *Human Genetics* 2020 139:4 **139**, 421–446 (2020).
243. Thompson, D. J. *et al.* Genetic predisposition to mosaic Y chromosome loss in blood. *Nature* **575**, 652–657 (2019).
244. Fukami, M. & Miyado, M. Mosaic loss of the Y chromosome and men's health. *Reprod Med Biol* **21**, e12445 (2022).

245. Forsberg, L. A. *et al.* Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. *Nature Genetics* 2014 46:6 **46**, 624–628 (2014).
246. Hallast, P. *et al.* Assembly of 43 human Y chromosomes reveals extensive complexity and variation. *Nature* 2023 621:7978 **621**, 355–364 (2023).
247. Wright, D. J. *et al.* Genetic variants associated with mosaic Y chromosome loss highlight cell cycle genes and overlap with cancer susceptibility. *Nat Genet* **49**, 674–679 (2017).
248. Zhou, W. *et al.* Mosaic loss of chromosome Y is associated with common variation near TCL1A. *Nat Genet* **48**, 563–568 (2016).
249. Jakubek, Y. A. *et al.* Genomic and phenotypic correlates of mosaic loss of chromosome Y in blood. *The American Journal of Human Genetics* (2025) doi:10.1016/J.AJHG.2024.12.014.
250. Sano, S. *et al.* Hematopoietic loss of Y chromosome leads to cardiac fibrosis and heart failure mortality. *Science* **377**, 292–297 (2022).
251. Lleo, A. *et al.* Y chromosome loss in male patients with primary biliary cirrhosis. *J Autoimmun* **41**, 87–91 (2013).
252. Tatler, A. L. & Jenkins, G. TGF- β Activation and Lung Fibrosis. <https://doi.org/10.1513/pats.201201-003AW> **9**, 130–136 (2012).
253. Calver, J. F. *et al.* Defining the mechanism of galectin-3–mediated TGF- β 1 activation and its role in lung fibrosis. *Journal of Biological Chemistry* **300**, (2024).
254. Noveski, P. *et al.* Loss of Y Chromosome in Peripheral Blood of Colorectal and Prostate Cancer Patients. *PLoS One* **11**, e0146264 (2016).
255. Guo, X. *et al.* Mosaic loss of human Y chromosome: what, how and why. *Human Genetics* 2020 139:4 **139**, 421–446 (2020).
256. Brown, D. W. *et al.* Genetically predicted telomere length is associated with clonal somatic copy number alterations in peripheral leukocytes. *PLoS Genet* **16**, e1009078 (2020).
257. Wang, D. *et al.* Association between mosaic loss of chromosome Y and pulmonary fibrosis susceptibility and severity. *bioRxiv* 2024.05.25.595885 (2024) doi:10.1101/2024.05.25.595885.

258. Leavy, O. C. *et al.* Genome-wide SNP-sex interaction analysis of susceptibility to idiopathic pulmonary fibrosis. *medRxiv* **11**, 2024.01.12.24301204 (2024).
259. Liu, A. *et al.* Genetic drivers and cellular selection of female mosaic X chromosome loss. *Nature* **631**, 134–141 (2024).
260. de Araújo Lima, L. & Wang, K. PennCNV in whole-genome sequencing data. *BMC Bioinformatics* **18**, 49–56 (2017).
261. Schumacher, F. R. *et al.* Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat Genet* **50**, 928–936 (2018).
262. Phelan, C. M. *et al.* Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. *Nat Genet* **49**, 680–691 (2017).
263. Codd, V. *et al.* Polygenic basis and biomedical consequences of telomere length variation. *Nat Genet* **53**, 1425–1433 (2021).
264. Snetselaar, R. *et al.* Short telomere length in IPF lung associates with fibrotic lesions and predicts survival. *PLoS One* **12**, e0189467 (2017).
265. Dumanski, J. P. *et al.* Smoking is associated with mosaic loss of chromosome Y. *Science* (1979) **347**, 81–83 (2015).
266. Hollingsworth, L. R. *et al.* DPP9 sequesters the C terminus of NLRP1 to repress inflammasome activation. *Nature* 2021 592:7856 **592**, 778–783 (2021).
267. Harapas, C. R. *et al.* DPP9 deficiency: An inflammasomopathy that can be rescued by lowering NLRP1/IL-1 signaling. *Sci Immunol* **7**, (2022).
268. Prednisone, Azathioprine, and N -Acetylcysteine for Pulmonary Fibrosis . *New England Journal of Medicine* **366**, 1968–1977 (2012).
269. Gong, L. *et al.* AKT Phosphorylates FAM13A and Promotes Its Degradation via CUL4A/DDB1/DCAF1 E3 Complex. *Am J Respir Cell Mol Biol* **68**, 577–590 (2023).
270. Polverino, F. Genetic Susceptibility to Chronic Obstructive Pulmonary Disease and Idiopathic Pulmonary Fibrosis The Missing Link? *Am J Respir Cell Mol Biol* **68**, 482–484 (2023).
271. McDonough, J. E. *et al.* A role for telomere length and chromosomal damage in idiopathic pulmonary fibrosis. *Respir Res* **19**, 1–8 (2018).

272. Adegunsoye, A. *et al.* Telomere length associates with chronological age and mortality across racially diverse pulmonary fibrosis cohorts. *Nature Communications* 2023 14:1 **14**, 1–13 (2023).
273. Brown, E. A. *et al.* Three linked variants have opposing regulatory effects on isovaleryl-CoA dehydrogenase gene expression. *Hum Mol Genet* **33**, 270–283 (2024).
274. Dinsdale, N., Nepomnaschy, P. & Crespi, B. The evolutionary biology of endometriosis. *Evol Med Public Health* **9**, 174–191 (2021).
275. McGrath, I. M., Montgomery, G. W. & Mortlock, S. Polygenic risk score phenome-wide association study reveals an association between endometriosis and testosterone. *BMC Med* **21**, 1–14 (2023).
276. Patsenker, E. *et al.* Inhibition of Integrin α v β 6 on Cholangiocytes Blocks Tgfbeta Activation and Retards Biliary Fibrosis Progression. *Gastroenterology* **135**, 660 (2008).
277. Decaris, M. L. *et al.* Dual inhibition of α v β 6 and α v β 1 reduces fibrogenesis in lung tissue explants from patients with IPF. *Respir Res* **22**, 1–14 (2021).
278. John, A. E. *et al.* Translational pharmacology of an inhaled small molecule α v β 6 integrin inhibitor for idiopathic pulmonary fibrosis. *Nature Communications* 2020 11:1 **11**, 1–14 (2020).
279. Kawano-Dourado, L., Glassberg, M. K., Assayag, D., Borie, R. & Johannson, K. A. Sex and gender in interstitial lung diseases. *Eur Respir Rev* **30**, (2021).
280. Assayag, D., Morisset, J., Johannson, K. A., Wells, A. U. & Walsh, S. L. F. Patient gender bias on the diagnosis of idiopathic pulmonary fibrosis. *Thorax* **75**, 407–412 (2020).
281. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J R Stat Soc Series B Stat Methodol* **82**, 1273–1300 (2020).
282. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics* 2019 20:8 **20**, 467–484 (2019).