Understanding the genetic basis of disease endotypes in idiopathic pulmonary fibrosis

Thesis submitted for the degree of Doctor of Philosophy at the University of Leicester

by

Luke Michael Kraven BSc MSc

Department of Health Sciences

University of Leicester

May 2022

Abstract

Understanding the genetic basis of disease endotypes in idiopathic pulmonary fibrosis

Luke Michael Kraven

Idiopathic pulmonary fibrosis (IPF) is a rare, incurable disease of unknown cause characterised by progressive scarring of the lungs. The prognosis of IPF is poor with a median survival time of approximately 4 years and current treatment options are limited. The aim of the analyses in this thesis was to utilise genomic and transcriptomic data to improve the understanding of the pathogenesis of IPF, which could aid drug development and lead to improvements in treatments.

This thesis describes the first genetic analyses of the age at which IPF is first developed. First, genome-wide association studies were performed to identify common genetic variants that are associated with the age-of-onset of IPF. Following this, gene-based collapsing analyses were performed to investigate the role of rare genetic variation in the age-of-onset of IPF. These analyses highlighted some suggestively significant genes of potential interest as well as some important factors to consider when studying this phenotype.

A series of transcriptomic analyses were conducted to identify groups of IPF patients that could represent endotypes of the disease. New bioinformatics methods were utilised in these analyses to combine and cluster multiple datasets. This approach allowed for the largest transcriptomic cluster analysis in IPF to-date to be performed, which revealed three distinct groups of patients with IPF. These findings were consistent with the theory of multiple endotypes of IPF; significant differences in lung function and survival were found between clusters and gene enrichment analysis implicated metabolic changes, apoptosis, cell cycle and the immune system in the development of these potential IPF endotypes. Supervised machine learning was used to develop a gene expression-based classifier with the ability to assign patients with IPF to one of the three clusters. With further development, this classifier could be a useful clinical tool for outcome prediction and patient stratification in IPF.

Acknowledgements

There are a number of people that I would like to thank for their help and support during my PhD. First and foremost, I would like to thank Professor Louise Wain for her invaluable guidance, advice and reassurance. If Louise weren't my primary supervisor, I have no doubt that my PhD experience would have been half as enjoyable and rewarding as it was.

I would also like to thank my other PhD supervisors, Professor Gisli Jenkins, Dr Astrid Yeo, Dr Billy Fahy and Adam Taylor, for their enthusiasm and advice with the scientific analyses in this thesis. I would particularly like to thank Adam for taking me under his wing during my placement at GSK, where I produced some of the material in this thesis for which I am most proud.

In addition, I would like to express my gratitude to everyone in the Wain-Tobin team at the University of Leicester. I am fortunate to have worked with such a talented and wonderful group of individuals and it is a pity that we did not get more time to work together in-person during my PhD. In particular, I would like to thank my GWAS guru Dr Richard Allen. Without Richard's help, I think I would probably still be trying to get my first analysis to run.

Last but not least, I would like to thank my family for their love and support during my PhD, including my parents (who are always there for me when I need them), my Grandpa Ivan (who started me on a mathematical career path when I was five by cutting up my sausages to teach me fractions) and my wonderful partner Renie (who brightened every single day).

COVID impact statement

I consider myself fortunate that as a data scientist I was able to conduct my research from home and therefore the research conducted for this PhD thesis did not change drastically from the original plan as a result of the COVID-19 pandemic. However, I must note that the coronavirus did likely impact my work in some ways, both directly and indirectly. For example, my placement at GlaxoSmithKline was cut short in early 2020 due to the lockdown, which meant that I was never given the opportunity to work directly with Dr Astrid Yeo in the Human Genetics department. In addition, completing my thesis in isolation came with some challenges, such as the mental impact of working away from my team members and living apart from my family and friends.

Table of Contents

List of Figures

List of Tables

Chapter 1 – Introduction

Idiopathic pulmonary fibrosis (IPF) is a chronic, progressive lung disease with a poor prognosis and limited treatment options. This PhD thesis details genetic and transcriptomic analyses that were performed to improve the understanding of the pathogenesis of IPF, which could inform drug development and improve treatment options for patients. This initial chapter introduces IPF, genetics and transcriptomics. Following this, the specific aims of the thesis are outlined.

1.1 Idiopathic Pulmonary Fibrosis

1.1.1 Introduction

Idiopathic pulmonary fibrosis (IPF) is a type of interstitial lung disease (ILD). It is a specific form of chronic, progressive interstitial pneumonia of unknown cause, characterised by scarring in the lungs. An IPF diagnosis requires the exclusion of other forms of interstitial pneumonia, including other idiopathic interstitial pneumonias (IIPs) and ILDs associated with environmental exposure, medication, or systemic disease[1]. Typically, symptoms begin to present in individuals in their sixties and seventies, and the incidence of the disease increases with older age[2]. The most common symptoms are shortness of breath and a dry, persistent hacking cough. Over time, other signs and symptoms may develop, including rapid, shallow breathing, unintended weight loss, fatigue, malaise (a general feeling of being unwell), aching muscles/joints and clubbing of the tips of the fingers or toes. Complications can include pulmonary hypertension, heart failure, pneumonia, or pulmonary embolism.

The life expectancy of the disease is poor, with studies from the late 1990s and early 2000s often reporting median post-diagnosis survival times of 2.5-3.5 years[3,4,4,5,5]. However, in recent years, two anti-fibrotic therapies (pirfenidone and nintedanib) have been approved for the treatment for IPF as they have been shown to slow disease progression and improve survival[6,7,7]. As such, recent studies have estimated slightly greater median post-diagnosis survival times, typically between 3-5 years[8,9,9].

In 2008, it was estimated that there were approximately 5 million people worldwide who suffer from IPF and that incidence appears to be increasing[10]. The incidence and prevalence of IPF varies significantly across countries[11] and are both widely reported to be higher in males. For example, for every 100,000 individuals in the UK in 2012, 16 males and 9 females were newly diagnosed with IPF and 62 males and 40 females had been previously diagnosed with IPF at some time in their life[12].

The clinical course of IPF can be highly variable[13]. The rate of decline and progression to death in patients may take several clinical forms: slow deterioration with worsening severity of dyspnoea (difficult or laboured breathing), rapid deterioration and progression to death, or periods of relative stability interposed with periods of sudden respiratory decline (called acute exacerbations and sometimes exhibited by hospitalizations for respiratory failure).

Whilst the cause of IPF is currently unknown, several risk factors associated with the disease have been identified. Firstly, cigarette smoking has shown a significant association with the disease. In a meta-analysis of six studies (totalling 784 IPF cases and 1,397 controls), which investigated the link between smoking and IPF, it was found that patients with IPF were significantly more likely than controls to have previously smoked (OR = 1.58; 95% CI = [1.27, 1.97]) and estimated that 49% of IPF cases would be prevented if smoking was eliminated as an exposure[14]. Secondly, other environmental and occupation exposures have also been shown to increase the risk of developing IPF. These include exposure to metal dust; wood dust; coal dust; silica; stone dust; biological dusts coming from hay, mould spores or other agricultural products; and occupations related to farming/livestock[15]. This could partly explain the discrepancies in prevalence and incidence between males and females, as men are traditionally more likely to work in jobs in which such exposures occur and therefore more likely to develop the disease. In addition, a recent study[16] has shown that the differences in IPF prevalence and incidence between males and females may partly be due to diagnostic bias, with males being more likely to be diagnosed with IPF than females.

Additionally, many studies have investigated the possible role of chronic viral infection in the aetiology of IPF, focusing on viruses such as hepatitis C and herpesviruses[17]. However, evaluating the associations between IPF and microbes such as viruses was hindered by many confounding factors, for example when patients were receiving immunosuppressive therapy[18]. Consequently, both positive associations and negative associations between these viruses and IPF have been reported, and until recently, strong conclusions about the role of infection in IPF could not be drawn. However, a recent meta-analysis of 20 case-control studies from 10 countries (totalling 1,287 IPF cases) reported that a viral infection was associated with a significant increase in the odds of developing IPF (OR = 3.48; 95% CI = [1.61, 7.52], P = 0.001), therefore supporting the idea that viral infection is a risk factor for IPF[19].

A further risk factor for IPF is genetic predisposition. IPF is considered a ‘polygenic’ disease, where several genetic signals have been shown to contribute to disease susceptibility. As a polygenic disease, the disease is likely the result of complex interactions between the genetic and environmental factors[20].

1.1.2 Disease mechanism

The current model for the pathogenesis of IPF[21] (Figure 1.1), is as follows: first, a host becomes susceptible due to their genetic composition and age. Then when the alveolar epithelium is injured (by cigarette smoke, industrial dusts, viral infection etc.), an abnormal wound healing response is triggered, which results in the activation of inflammatory cells, increased vascular permeability and the release of profibrotic cytokines. This creates an environment that is supportive of exaggerated fibroblast and myofibroblast activity and leads to an increase in the deposit of extracellular matrix within the lung parenchyma, which impairs gas exchange.

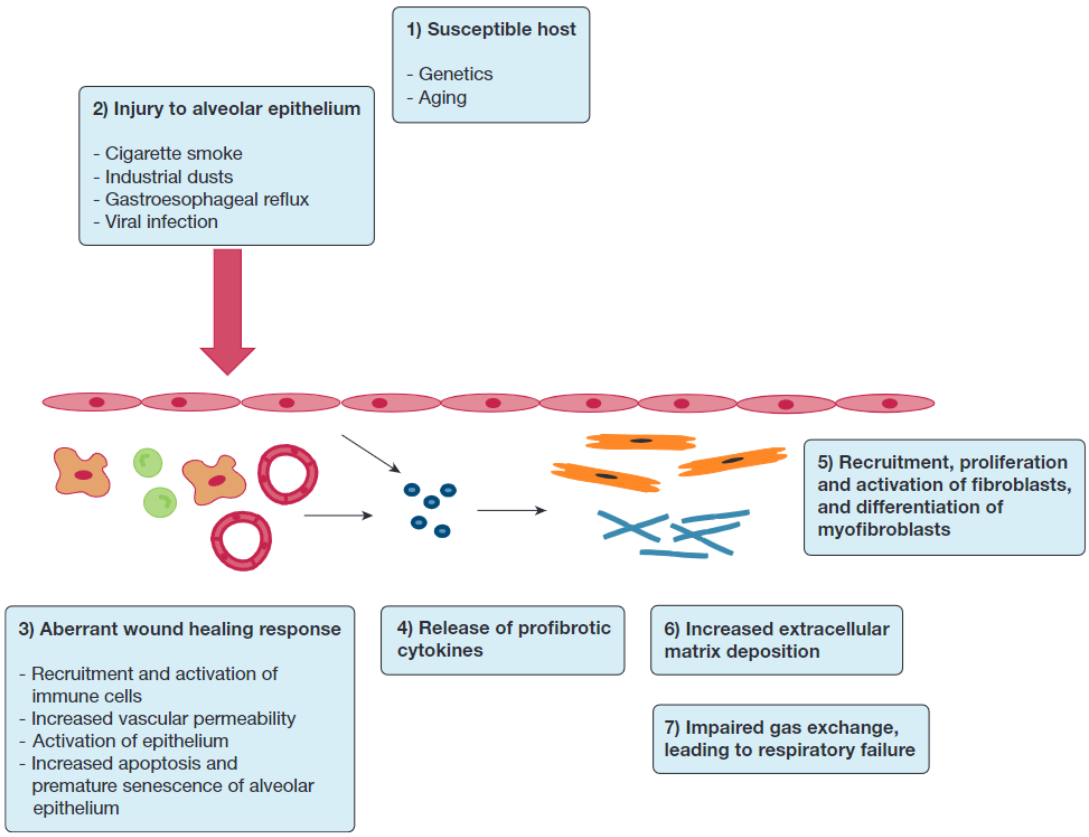


FIGURE 1.1: *A diagram showing the current model for the disease mechanism of IPF. Figure taken from Goodwin et al.*[21].

Several cell types and signalling pathways have been implicated in the pathogenesis of the disease, including dysregulated epithelial repair, host defence, the immune response, fibroproliferative responses linked to aberrant kinase activation, transforming growth factor-β and its downstream pro-fibrogenic pathways, and developmental pathway reactivation[22]. Furthermore, dysfunction of alveolar type II (AT2) cells appears to play a role in the development of IPF, with telomere shortening and increased endoplasmic reticulum stress being proposed as factors that may contribute to this[23].

Another important pathogenic mechanism in IPF is cellular senescence, which significantly contributes to chronic matrix remodelling and fibrosis[23]. A major driver of cellular senescence is DNA damage, which is consistent with the finding that telomere shortening is linked to IPF susceptibility, as shortened telomeres predispose cells to DNA damage. Importantly, clearing of senescent cells in mice protected from lung fibrosis, indicating that senotherapeutics which target these cells by eliminating them or putting them back on the right repair track could be a novel therapeutic approach in IPF[24,25,25].

The highly heterogeneous clinical course of IPF may suggest that the disease consists of different subtypes. Subtypes of a disease that are defined by a particular pathophysiological mechanism are known as endotypes. These are in contrast to disease phenotypes, which are particular observable characteristics of a disease but are not directly suggestive of a particular underlying mechanism. It has

been speculated previously that IPF may consist of multiple endotypes[26,27,27], though these are not yet well understood. Several different biomarkers have been developed to define clinically significant subtypes of IPF, which could support the theory that multiple IPF endotypes exist. For example, serum biomarkers have been found that are able to predict prognosis in IPF, such as serum matrix metalloproteinase (MMP)7[28,29,29], surfactant protein D, CA19-9, CA-125[30] and serum levels of the collagen synthesis neoepitopes PRO-C3 and PRO-C6[31].

1.2 Genetics

1.2.1 Background

The human genome is 99.9% identical between any two unrelated individuals[32]. However, as it is incredibly large, consisting of approximately 3.3 billion base pairs[33], many genetic differences can still be found between two such individuals. As genetics play a part, to a greater or lesser extent, in all diseases[34], this genetic variation can be leveraged to help us gain a better understanding of how diseases work and can even provide insight into how they may be treated. Rare 'Mendelian' disorders such as cystic fibrosis and haemophilia are defined by variants (or 'mutations') within a single gene, whilst more common, complex 'polygenic' diseases, such as heart disease and type 2 diabetes, are typically the result of interactions between environmental factors and genetic variants from multiple genes. Whilst disease status for a Mendelian disease is usually directly attributable to the presence of a particular mutation, genetic variants that are associated with a polygenic disease each typically confer a relatively small change in the risk of developing that disease. Still, the small individual effects from such variants can accumulate and can give an individual a considerably increased odds of developing a polygenic disease, compared to individuals who do not possess those variants.

Genetic epidemiology is a relatively new scientific field that aims to study the genetic and environmental risk factors of complex diseases through the application of epidemiological methodology to genetic data. Genetic data is measured in two main ways: deoxyribonucleic acid (DNA) sequencing and targeted probe-based assays (Section 1.2.4). The first type of assays to be widely used were panels that were designed to detect microsatellites, which are small sequences of DNA that are repeated many times and are among the most variable types of sequence in the genome[35]. A common type of study in this early period was linkage analysis, which considered and compared variant signatures (for example, microsatellites) in multiplex families (a family in which more than one family member was afflicted with a particular disease) to identify genes that are influencing the disease. These studies were largely successful at determining the genes linked to a disease of interest when the disease was Mendelian, but were far less successful when the disease was polygenic, due to the greater genetic heterogeneity[36].

Towards the end of the 20th century, the field started to shift away from studying multiplex families and towards cohorts of unrelated individuals. In addition, microarray technology started to focus on single-nucleotide polymorphisms (SNPs, Section 1.2.3) instead of microsatellites. As genetic variants that lay

close together on the same chromosome are often correlated (linkage disequilibrium, Section 1.2.3), this enabled the development of strategically designed SNP microarrays, which together with imputation reference panels (Section 1.2.4), allowed for genetic variation across the entire genome to be studied without an individual's whole genome needing to be directly measured. This paved the way for genome-wide association studies (GWAS, Section 2.2), which utilise genome-wide genetic data to identify genetic variants that are associated with a trait for a polygenic disease.

The other main approach to measuring DNA is with high-throughput sequencing. This involves measuring an individual's entire genetic code, including non-variant positions. However, sequencing the whole human genome was originally an expensive and time-consuming process: it took the Human Genome Project, an international consortium of multiple research labs, over 10 years to sequence the whole genomes of just a few individuals. However, the time and cost of DNA sequencing have dropped dramatically in recent years, and it is becoming increasingly feasible to sequence the exomes (the parts of the genome comprised of exons, approximately 1-2% of the genome) or even the entire genomes for all individuals in a study. Studying sequencing data has some advantages over studying genotyping data, including the ability to discover new variants as well as investigate the impact of the rarest variants, which are not usually considered in GWAS.

Since the publication of the first GWAS in 2005[37], at least 5,600 GWAS have been performed and over 71,000 unique variant-trait associations have been reported[38]. In addition, sample sizes in genetic studies are increasing, with some recent GWAS meta-analyses reaching over 1 million participants[39,40,40]. Findings from genetic studies can help us gain a better understanding of the genetic architecture of a disease trait, such as the number of independent genetic loci that are associated with the disease, the frequency and effect sizes of the suspected causal variants (those that are exerting a genuine biological effect on the phenotype) in the study population and the disease's level of heritability[41]. Importantly, combining genetic results with data from external resources (such as gene expression data) can implicate genes and biological mechanisms in the pathogenesis of the disease. This can inspire testable hypotheses and molecular experiments, such as the effect on the disease trait in a mouse model when a particular suspected causal gene is 'knocked out', which could ultimately lead to improved treatment options in humans.

Genetic findings can also inform drug development. Clinical drug development is an expensive and time-consuming process that many drugs fail due to a lack of efficacy, but drugs that are developed with genetic support are twice as likely to pass clinical development as those without it[42,43,43]. Furthermore, chemical compounds that target the protein products of genes that were identified in a GWAS can be promising candidates for drug repurposing. One example of this is CDK4/CDK6 inhibitors for the treatment of rheumatoid arthritis[44]. In addition, some genetic variants have been linked

to how well a person tolerates certain drugs, with some that have been found to render certain drugs ineffective, or even toxic[45].

Findings from genetic studies also have applications in preventative healthcare. Once multiple disease susceptibility-associated genetic variants have been discovered for a particular disease, these can be used to develop a 'polygenic risk score', which can provide an individual with an estimate of their relative genetic risk for that disease. This information can be used to identify at-risk populations to inform screening strategies or to target preventative interventions or lifestyle modifications[46,47,47].

1.2.2 Other 'omics'

The scientific field which studies the structure, function, evolution, mapping and editing of genomes is referred to as genomics. But the genome is not the only informative biological system that can be leveraged to help us gain a better understanding of the pathogenesis of human diseases. It is also possible to study an individual's entire collection of ribonucleic acid (RNA) transcripts, proteins, metabolites or epigenetic modifications. These systems are often tissue-specific and more dynamic than the genome as they can change in response to certain stimuli, such as a therapeutic intervention, as well as being influenced by genetic variation. Each named with the suffix *–omics* to mirror the term genomics, the study of these systems are respectively referred to as transcriptomics, proteomics, metabolomics and epigenomics.

Measurements of these 'omic' systems provide a snapshot of the system at a particular time point, which can be used to infer the biological processes that are being activated and may be important in the development of a disease. In addition, omic data can be used to develop predictive or prognostic models of a particular disease that are more accurate than those obtained using standard clinical approaches[48]. However, due to the dynamic nature of these biological systems, they are subject to some issues that are not present when studying the genome, such as reverse causation, confounding (by factors other than ancestry, see Section 2.1.4) and batch effects (Section 5.13).

1.2.3 Genetic variation and linkage disequilibrium

The term genotype is used to refer to the genetic constitution of an individual at a particular locus. The sequences of DNA that differ between individuals at a particular locus are known as alleles. SNPs are the most common type of genetic variation found in humans[49] and each SNP represents a difference in a single nucleotide. For example, where one individual may have an adenine on the 5' strand of one or both chromosomes (paired with a thymine on the 3' strand), another may have a guanine on the 5' strand (paired with a cytosine on the 3' strand). These occur almost once in every 1,000 nucleotides on average, which means that there are roughly 3 to 4 million SNPs in a person's genome[50]. These variations may be unique or occur in many individuals. Some SNPs have functional consequences, such as changing amino acids and mRNA, or disrupting known regulatory regions. As humans have two copies of each chromosome, for a SNP with two possible alleles (which is usually the case), a person

can have one of three genotypes, two that are homozygous and one that is heterozygous (Figure 1.2). The prevalence of a SNP is given in terms of the minor allele frequency (MAF), which is the frequency of the less common allele in a population. As each person possesses two alleles at a genetic locus (one inherited from each parent in meiosis), the MAF of a population for a particular SNP can be calculated as follows:

$$\text{MAF} = \frac{Number\ of\ least\ common\ allele\ in\ population}{2*(Number\ of\ people\ in\ population)} \quad (1.1)$$

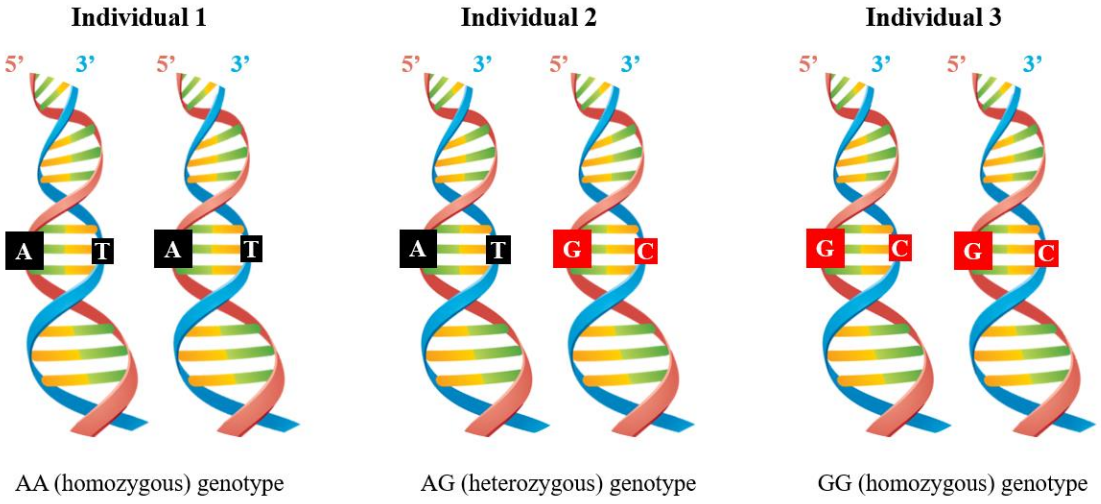


FIGURE 1.2: *Diagram showing the three possible genotypes at a locus with two alleles, A and G.*

Linkage disequilibrium (LD) is a key concept in the field of population genetics. SNPs that lay close together on a chromosome are less likely to be separated by a recombination event during meiosis than SNPs that lay farther apart. As such, they are more likely to be inherited together. Two genetic loci are in LD if, across the whole population, they are found together on the same haplotype (the allelic configuration along a single chromosome) more often than expected. This is useful because this means that tests of association may be informative even when the true causal variant has not been measured[51].

An important point to note is that variant frequency and LD differ between ancestral populations. Evidence suggests that this is because all modern human populations have a common origin (believed to be in Africa) and each ancestral population is the result of groups that migrated out of Africa at different times[52]. More ancient ancestral populations have greater genetic diversity and finer LD structure between genetic markers than ancestral populations that migrated out of Africa more recently. If these differences in ancestry are unaccounted for, confounding by ancestry can occur, which is also known as population stratification (Section 2.1.4). As a result, studies considering LD between SNPs are often performed in one particular ancestral population.

LD between two SNPs is most commonly measured in terms of D' or r^2, both of which measure the deviation from random association. Both of these measures are derived from the coefficient of linkage disequilibrium, D. The level of LD between two alleles, say allele A and allele B, is defined as follows:

$$D_{AB} = p_{AB} - p_A p_B \tag{1.2}$$

Where p_{AB} is the frequency with which both allele A and allele B occur on the same haplotype, p_A is the allele frequency of A and p_B is allele frequency of B. The two alleles are in LD when $D_{AB} \neq 0$. D is normalised to produce D', which allows for the comparison of the level of LD between different pairs of alleles. Alternatively, the correlation coefficient *r* can be calculated as follows:

$$r = \frac{D}{\sqrt{p_A(1-p_A)p_B(1-p_B)}} \tag{1.3}$$

An r^2 value of 1 means that one SNP is always observed with - or perfectly 'tags' - the other. Therefore, SNPs that are in high LD with, and thus representative of, other SNPs in a particular haplotype are called tag SNPs. Due to the presence of tag SNPs, it is possible to identify genetic variation and an association with a phenotype without genotyping every SNP in a chromosomal region. This, together with imputation (next section), reduces the need to directly measure every single SNP across the genome.

1.2.4 Collecting genetic data

Genotyping is the process of determining which genetic variants an individual possesses through the analysis of a sample of their DNA once it has been extracted from a tissue sample (e.g. whole blood, saliva or buccal swab). This can be done via a number of methods, such as a genotyping array, polymerase chain reaction (PCR) or next-generation sequencing (NGS). Genotyping arrays report the genotypes of the tested individuals for a large number of SNPs (for example, the UK BiLEVE Axiom Array by Affymetrix probes 807,411 variants and the UK Biobank Axiom Array probes 825,927[53]) by returning two allele probe intensities (i.e. an intensity for each allele) for each variant. These intensities are plotted and genotype calling software assigns each individual to a genotype group (Figure 1.3). Genotyping chips are an effective method of measuring many different variants at once, especially common variants. However, they require prior knowledge of the location and alleles of the variants of interest.

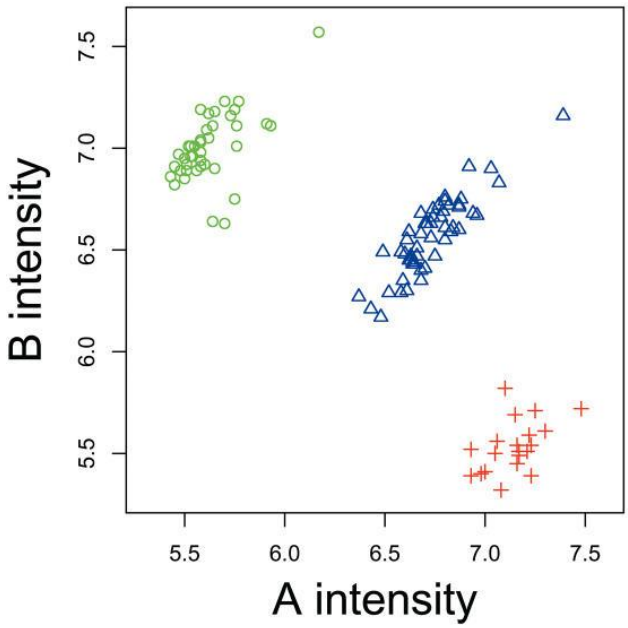


FIGURE 1.3: *Example of a genotype cluster plot, taken from Phillipe et al[54]. The green circles are samples that have been called as BB genotypes, blue triangles are samples that have been called as AB genotypes and the red plus signs are samples that have been called as AA genotypes.*

By utilising the fact that variants close to each other are inherited together in haplotype blocks, the genotypes of variants not included on a genotyping array can be inferred through imputation. This is done by comparing the alleles of the genotyped individuals to a large reference panel of sequenced individuals, estimating which haplotype an individual is likely to have and therefore which alleles for the non-genotyped variants they are likely to possess. Imputation allows a GWAS to test millions of variants when relatively few were directly genotyped. This is hugely beneficial as it improves the coverage of the genome, can identify genotyping errors and can allow for studies with different genotyping arrays to be combined, which increases the statistical power.

Examples of imputation reference panels include the 1000 Genomes Project, which has collected genetic information from over 2,500 individuals from 26 different populations at 84.7 million SNPs[55], and UK10K which contains information on 3,781 individuals at over 42 million SNPs[56]. More recently, the Haplotype Reference Consortium (HRC) has combined multiple reference panels (including 1000 Genomes Project and UK10K) to construct a reference panel of approximately 65,000 human haplotypes at around 39 million SNPs[57].

Although most GWAS utilise imputation to vastly increase the number of variants tested, some variants may not be imputed well. Imputation quality for a variant is usually denoted R^2 and is scored from 0 to 1, with 0 indicating the lowest possible imputation quality and 1 being the highest possible quality[58].

Whilst genotyping is used to determine genotypes for pre-specified genetic variants, sequencing is a method that is used to determine the whole DNA sequence, including non-variant positions. This is done by generating many small overlapping reads of the DNA and then reconstructing the DNA from

these using a reference sequence for assembly. Each part of the DNA sequence can be read numerous times and the more times it is read, the more accurate the sequencing becomes. However, this also increases costs.

1.2.5 Genetic research into IPF

There are a few different lines of evidence which suggest that the development of pulmonary fibrosis is partly determined by genetic factors. The primary implicating factor is that clustering of cases of pulmonary fibrosis (an uncommon disease) have been widely reported between members of genetically related individuals, including: monozygotic twins raised in different environments, in consecutive generations in the same families, and in family members separated at an early age[59]. Familial pulmonary fibrosis (FPF), also referred to as familial interstitial pneumonia and familial idiopathic pulmonary fibrosis, is identified when two or more members of a family have an idiopathic interstitial pneumonia. Up until early in the 21st century it was believed that FPF represented a rare subset of IIP, comprising 3-5% of cases[60,61,61,62,62]. However, more recent estimates put this figure as high as 20%[63,64,64].

Early genetic studies found that rare mutations in the *TERT* and *TERC* genes were associated with familial pulmonary fibrosis[65,66,66]. *TERT* is a gene that encodes telomerase reverse transcriptase, which together with the RNA component of telomerase (*TERC*), is required to maintain telomere integrity. Since these FPF studies, common and low-frequency polymorphisms in or near telomere genes (including *TERT*) have been found to be associated with susceptibility to IPF[67,68,68]. It is believed that mutations in *TERT* or *TERC* that result in telomere shortening over time confer a dramatic increase in susceptibility to IPF.

Many genetic studies have since been performed to identify genetic loci associated with IPF susceptibility. Early studies showed that IPF has a strong association with a common variant on chromosome 11 called rs35705950, which lies in the promoter region of a gene called *MUC5B*[69,70,70]. This has been widely replicated in many GWAS[71,72,72,73,73]. In a meta-analysis that combined nine studies of IPF (with a total of 2,733 IPF patients and 5,044 controls)[74], it was found that each copy of the minor (T) allele of rs35705950 was associated with a 5-fold increased risk of IPF compared with the major (G) allele (OR 4.85, 95% CI [3.79–6.21], $P=5.88\times10^{-36}$).

In addition to its association with susceptibility to IPF, the effect of the *MUC5B* promoter polymorphism on an individual's disease progression has been the subject of investigation. Peljto et al.[75] (n=586) reported that the number of copies of the T allele at this locus was significantly associated with an increased survival time in patients with IPF. However, Dudbridge et al.[76] have suggested that the paradoxical association between the strong susceptibility variant rs35705950 and increased survival time may be the result of index event bias. Index event bias can occur when the phenotype of interest is an event subsequent to disease onset and the individuals in a study are selected based on their disease status, but common causes of incidence and prognosis are not accounted for. After adjusting for index

event bias using an effect size correction that utilises the residuals from the regression of genetic effects on prognosis on genetic effects on incidence, the authors found that rs35705950 was instead associated with decreased survival (n=565).

After the discovery of the association between *MUC5B* and IPF risk, a subsequent GWAS[72] (with 542 IPF cases and 542 controls in the discovery stage and 544 IPF cases and 687 controls in replication) found genome-wide significant associations between IPF susceptibility and three SNPs within the Toll-interacting protein (*TOLLIP*) locus, which lies in the same region on chromosome 11 as *MUC5B* (11p15). The SNPs in the *TOLLIP* region were found to be in low LD with rs35705950 and therefore indicated that these were independent associations. However, they were later found to be non-independent in a larger GWAS meta-analysis[77]. In addition, this study reported genome-wide associations between IPF susceptibility and variants near the *MDGA2* and *SPPL2C* regions on chromosomes 14 and 17 respectively.

Fingerlin et al.[71] (2013) conducted a case-control GWAS (with 1,616 IPF cases and 4,683 controls in discovery, 876 cases and 1,890 controls in replication) to identify additional genetic risk factors linked with IIPs (of which IPF is the most common). In their study they confirmed the associations between IIP and the previously reported variants found in the *MUC5B*, *TERC* and *TERT* gene regions. In addition, the authors identified seven novel genetic risk loci associated with risk of IIP, including one located in the desmoplakin (*DSP*) gene.

In 2017, Allen et al. published a two-stage GWAS of IPF susceptibility[73] (stage 1: 602 IPF cases and 3,366 controls, stage 2: 2,158 IPF cases and 5,195 controls) and identified a novel variant near A-kinase anchoring protein 13 (*AKAP13*), as well as confirming the associations between IPF susceptibility and the *DSP* and *MUC5B* variants. More recently, Allen et al.[77] used polygenic risk scores to show that there is still a significant amount of genetic variation in IPF risk which has not been explained by the previously reported genetic variants. The authors then performed the largest GWAS meta-analysis of IPF susceptibility to-date (with a total of 2,668 IPF cases and 8,591 controls), in which they identified and successfully replicated three new genome-wide significant signals of association with IPF susceptibility (near *KIF15*, *MAD1L1* and *DEPTOR*) and confirmed associations at 11 previously reported loci.

In a recent case-control study by Lorenzo-Salazar et al.[78], genetic data from individuals with IPF were sequenced at three genetic loci, 11p15.5 (the locus containing *MUC5B*), 14q21.3 (*MDGA2*) and 17q21.31 (*SPPL2C*). The discovery stage of this study (181 IPF cases and 501 controls) identified 36 correlated variants that reached genome-wide significance. Three of these had MAF <5%, suggesting a minor impact of low-frequency variants in IPF susceptibility in these loci. The most strongly associated variant in the study was rs35705950, the previously reported variant in the *MUC5B* promoter region. Additionally, their results suggested that the *MUC5AC* gene could also be contributing to IPF risk, as

two variants in this region were found to be associated with IPF, were nominally significant when adjusting for rs35705950 and were successfully replicated in an additional independent cohort.

Moore et al.[79] also performed a sequencing study, though theirs was larger with 3,624 cases and 4,442 controls targeting 10 genetic loci that were chosen based on results from previous GWAS. In addition to identifying for the first time that rare variation in *FAM13A* is associated with disease, the authors confirmed the role of rare variation in the *TERT* and *RTEL1* gene regions in the risk of IPF and found that the *FAM13A* and *TERT* regions have independent common and rare variant signals.

Most recently, an exome-wide association study of 752 sporadic IPF cases and 119,055 UK Biobank controls[80] identified a novel IPF susceptibility signal in the form of a single rare missense variant in the *SPDL1* region, at which each copy of the minor allele was estimated to increase the odds of developing IPF by 2.9 times. Table 1.1 shows a summary of the genetic loci that have displayed genome-wide significant associations with IPF susceptibility as of November 2021.

TABLE 1.1: *List of genetic variants that have been reported to be associated with IPF at genome-wide significance. EAF = effect allele frequency, OR = odds ratio, CI = confidence interval. Summary statistics are from the largest GWAS meta-analysis of IPF susceptibility performed to date by Allen et al[77]. All co-ordinates are according to human genome build 19.*

SNP/Allele	Chr.	Position	Gene	EAF	OR (95% CI)	References
rs78238620	3	44902386	Intergenic (8kB from *KIF15*, 1kB from *TMEM42*)	5.5%	1.51 [1.30, 1.75]	77
rs12696304	3	169481271	Intergenic (100kB from *MECOM*, 1kB from *TERC*)	28.1%	1.30 [1.21, 1.41]	65†
rs2609255	4	89811195	*FAM13A* (intronic)	22.5%	0.79 [0.73, 0.85]	71†
rs7725218	5	1282414	*TERT* (intronic)	32.5%	0.72 [0.67, 0.77]	67†
rs116483731	5	169015479	*SPDL1* (missense)	0.9%	2.18 [1.53, 3.09]	80
rs2076295	6	7563232	*DSP* (intronic)	46.9%	1.46 [1.37, 1.56]	71,73,73†
*HLA-DQB1**06:02	6	06:02	*HLA-DQB1*			81
rs12699415	7	1909479	*MAD1L1* (intronic)	42.0%	0.78 [0.73, 0.83]	77
rs28513081	8	120934126	*DEPTOR* (intronic)	52.7%	1.20 [1.14, 1.27]	77
rs11191865	10	105672842	*OBFC1* (intronic)	49.1%	1.15 [1.08, 1.23]	71
rs7934606	11	1093945	*MUC2* (intronic)	44.9%	0.95 [0.88, 1.03]*	71
rs35705950	11	1241221	*MUC5B* (promoter)	14.9%	4.84 [4.37, 5.36]	70,71,71,72,72,73,73†
rs111521887	11	1312706	*TOLLIP* (intronic)	19.8%	1.00 [0.91, 1.10]*	72
rs9577395	13	113534984	*ATP11A* (intronic)	20.7%	0.77 [0.71, 0.83]	71†
rs7144383	14	48040375	*MDGA2* (intronic)	11.0%	0.90 [0.81, 1.00]	72
rs2034650	15	40717302	Intergenic (7kb from *IVD*, 33kB from *BAHD1*)	47.0%	1.30 [1.21, 1.39]	82†
rs62025270	15	86300198	Intergenic (12kB from *AKAP13*, 11kB from *KLHL25*)	5.3%	1.54 [1.38, 1.73]	73,77,77†
rs17690703	17	43925297	*MAPT-AS1* (intronic)	24.5%	0.78 [0.72, 0.85]	72
rs1981997	17	44056767	*MAPT* (intronic)	20.9%	0.74 [0.68, 0.80]	71
rs12610495	19	4717672	*DPP9* (intronic)	30.5%	1.31 [1.22, 1.42]	71†

†: A previously reported signal that was confirmed in Allen et al[77].

*: Previously reported variants that were not independently associated with IPF susceptibility in Allen et al.[77] after conditioning on rs35705950 genotype.

In summary, several studies have identified independent genetic signals for IPF susceptibility. Over time, as technology and methodology advance, data collection costs decrease and study sample sizes grow larger, these studies will likely continue to identify novel signals that are associated with IPF risk. However, there has been relatively little genetic research into other IPF phenotypes, such as the age-of-onset, survival or disease progression, as there is less data available for these traits as they are not as commonly recorded for research use. Studying these underutilised phenotypes could reveal novel genetic signals that are important to the aetiology of IPF but which would be missed in studies of IPF susceptibility. As such, the identification of genes and biological processes involved in the development of these phenotypes could be informative for elucidating the pathogenesis of IPF and discovering novel treatments.

1.3 Transcriptomics

1.3.1 The transcriptome

During the initial stage of gene expression, a sequence of DNA is read by an enzyme called RNA polymerase, which binds to the DNA (together with one or more transcription factors) and separates the two strands of the DNA double helix. RNA polymerase then produces a complementary strand of RNA which is identical to one of the strands of DNA but with each thymine nucleotide base replaced with a base of uracil. This RNA molecule is known as a primary transcript, which can then be further processed into messenger RNA (mRNA) that could ultimately be translated into a protein. The entire collection of transcripts that are present within a tissue or a cell are known as the transcriptome. The transcriptome can be measured to learn about the level of gene expression within that tissue/cell at a certain time point.

Part of the processing of mRNAs involves a process known as splicing, in which the introns (the sections that do not directly code for a protein) of a gene are removed and the exons (the protein coding regions) are combined in different ways to create different proteins. Importantly, this means that a single gene can code for multiple protein ‘isoforms’. mRNA transcripts from the same gene that are alternatively spliced to code for different protein isoforms are known as splice variants.

The measurement of gene expression within a sample was first made possible in 1977 through the development of the Northern blot, in which the RNA from a gene is isolated and detected using a hybridisation probe. This was followed by the development of reverse transcription-PCR, which is faster than northern blotting and provides a better quantitative estimate of the level of gene expression[83]. However, both methods only allowed one gene (or very few) to be analysed at a time. This issue was addressed through the development of microarrays (Section 1.3.2), which revolutionised the field as they allowed for thousands of genes to be analysed simultaneously. Microarrays were the primary method of choice for the measurement of transcriptomic data until the recent rise in popularity of RNA-

sequencing (RNA-seq, Section 1.3.2), which uses high-throughput NGS to measure the entire transcriptome.

The transcriptome can be affected by genetic variation in the genome in several important ways. For instance, missense variants result in missense mRNAs, which bear one or more mutated codons that code for a different amino acid sequence than the naturally occurring sequence and will lead to the translation of a different protein. Similarly, stop-gain mutations cause the transcript (and resulting protein) to be abnormally shortened by the presence of a premature stop codon. Frameshift mutations also commonly alter the first stop codon, which can lead to the amino acid chain being abnormally short (or long). Fortunately, there is a mechanism known as nonsense-mediated RNA decay, which safeguards the quality of the transcriptome by eliminating transcripts which contain premature stop codons[84].

When genetic mutations occur in non-coding sequences, it is often less clear how they affect gene expression. Although, some mechanisms through which non-coding variants can affect gene expression are known. For example, a mutation that lays in the promoter sequence of a gene can influence gene expression by disturbing the recruitment of transcription factors at the promoter. Alternatively, a mutation in the splicing site of an intron can interfere with correct splicing of the transcribed mRNA, even when the mutation is synonymous[85]. Plus, a mutation in a transcription factor binding site (the region of the gene where transcriptional machinery binds to the protein) can change the rate of efficiency of transcription, which in turn can alter the levels of mRNA and the resulting protein[86].

In addition to being influenced by variation in the genome, the transcriptome can also change in response to external stimuli, such as a drug. As such, it can be difficult to infer the direction of cause-and-effect when investigating the effect of a particular disease on the transcriptome. In addition, the measurements in a transcriptomic analysis can be affected by non-biological factors, which gives rise to the issue of batch effects (Section 5.1.4).

1.3.2 Data collection

Broadly, the transcriptome is usually measured in similar ways to the genome, namely with microarray assays that are comparable to genotyping arrays or with RNA-seq that is comparable to whole-exome/genome sequencing. Microarrays assays contain many probes, each designed to detect and measure the prevalence of a particular transcript within a tissue/cell sample. Each probe contains a DNA fragment for a particular sequence of interest. The RNA within the sample is extracted, converted into complementary DNA (cDNA) and then labelled with a fluorescent tag. If there are any cDNA molecules present in the sample that are complementary to the sequence of the probe, they will bind together during hybridisation. The cDNA is then amplified through PCR and the fluorescence intensity of the amplification reaction is monitored to quantify the prevalence of that particular transcript within the

sample. As microarrays require prior knowledge of the RNA sequences, they cannot be used to discover new structural variations in the transcripts.

Alternatively, RNA-seq can be used to measure all of the transcripts present in the transcriptome and so provides a more comprehensive examination of the system than microarrays. RNA-seq is performed by isolating the RNA, breaking it into fragments, converting it into cDNA and then applying next-generation sequencing to the fragments. These sequences are then aligned and mapped to a reference genome. As RNA-seq requires no a-priori knowledge of RNA sequences, it is equipped to detect variations in the transcripts, such as splice variants. Despite being more expensive than microarrays, RNA-seq provides greater sensitivity and has become the most popular choice of transcriptomic technology in recent years[87].

1.3.3 Normalisation

An important step in a transcriptomic analysis is normalisation, in which the raw data are adjusted to account for factors that could lead to bias when comparing expression measures across different probes or samples[88]. One of the most common methods to remove between-probe variation is quantile normalisation[89]. Quantile normalisation is applied to a set of arrays and transforms the distribution of probe intensities for each array so that all arrays follow the same distribution, whilst maintaining the order of probe intensities within each array.

In addition, when using RNA-seq data the number of reads that map to a gene are affected by the length of the gene and the sequencing depth. Therefore, normalised metrics that account for these factors, such as reads per kilobase million (RPKM), fragments per kilobase million (FPKM) and transcripts per million (TPM), are usually preferred to standard read counts. Also, as read counts can vary greatly, with some that can be very large, they are often commonly transformed. The $\log_2$ scale is the most common transformation as this allows for the simple calculation of fold change between measurements.

1.3.4 Transcriptomic analyses

Studying the transcriptome can provide information on how genes are regulated and can also help to infer the functions of previously unannotated genes[87]. Comparison of the transcriptome between groups (e.g. a group of individuals with a disease and a group of healthy individuals, or the same individuals but at different time points) allows for the identification of differentially expressed genes (DEGs), which can highlight important genes, biological pathways and mechanisms related to a disease. In addition, DEGs can be used to develop diagnostic or prognostic biomarkers capable of making predictions about disease status or outcome based on the levels of gene expression for those DEGs[90,91,91].

Clustering of transcriptomic data can allow for the identification of genes that are often expressed together (‘co-expressed’), which can be useful as this can highlight biological pathways and processes that are being activated at particular times or under certain stimuli[92,93,93]. Another use for clustering in

transcriptomics is that this can identify groups of patients with similar patterns in gene expression, which can represent patients with the same disease or a particular disease subtype[92,94,94]. It can then be helpful to study the DEGs between these groups, for the reasons described above.

1.3.5 Transcriptomic research into IPF

Studies of gene expression have also been important in increasing the understanding of IPF (see Chapter 5). For instance, transcriptomic data is well suited to the investigation of disease endotypes as individuals with the same disease and similar transcriptomic profiles have genes that are co-expressed, which suggests that the same biological processes are being activated. Whilst this does not confer information about causality, this could suggest that those individuals are experiencing the same form of the disease (i.e. the same endotype). As such, several studies have used gene expression data to identify subtypes of IPF patients and some have used gene expression-derived subtypes to develop diagnostic and prognostic biomarkers for IPF (see Section 5.1.2 for a description of these studies). Whilst most transcriptomic studies in IPF have historically used microarrays or bulk RNA sequencing, which do not allow for cell-specific expression to be studied, in recent years the field has seen an increase in the number of studies that use single-cell RNA sequencing (scRNAseq). This has resulted in the development of the IPF Cell Atlas, which provides public access to four of the largest single-cell IPF data sets produced to-date, along with several visualization tools for differential gene expression analysis[95].

1.4 Aims of this thesis

The pathogenesis of IPF remains unclear and treatment options remain limited for patients. Therefore, the overall objective of this thesis is to utilise genomic and transcriptomic data to improve our understanding of the pathogenesis of IPF, which could aid drug development and lead to improvements in treatments. More specifically, there are two aims that will be addressed in this thesis: i) to define the genetic determinants of age-of-onset of IPF and ii) to use transcriptomic data to define potential endotypes of IPF.

1.4.1 Aim 1: To define the genetic determinants of age-of-onset of IPF

There have been several GWAS to-date that have been performed to identify genes that are associated with IPF susceptibility. However, there has been relatively little genetic research into other phenotypes, such as the age-of-onset of IPF. Studying this novel phenotype could reveal genes and biological processes that are involved in the pathogenesis of IPF, which could be potential drug targets. As such, the first aim in this thesis is to identify genetic variants and genes that are associated with the age-of-onset of IPF. The objectives of this aim are:

- To identify common genetic variants that are significantly associated with the age-of-onset of IPF.

- To identify genes in which a burden of rare genetic variants is significantly associated with the age-of-onset of IPF.

1.4.2 Aim 2: To use transcriptomic data to define potential endotypes of IPF

The considerable clinical heterogeneity in IPF may suggest that the disease consists of multiple clinically distinct endotypes. If such endotypes do exist, the identification of these could allow for the development of prognostic biomarkers and for more tailored approaches to treatment for patients. In addition, this could implicate particular biological mechanisms and pathways in the development of the different endotypes of the disease, which could inform drug development. As gene expression data can be used to identify transcriptomic profiles that are associated with particular disease characteristics and trajectories, the second aim in this thesis is to identify clinically distinct endotypes of IPF through the application of clustering to transcriptomic data from IPF patients. There were three specific objectives to this aim:

- To identify groups of IPF patients that could be representative of distinct and clinically relevant endotypes of IPF
- To use any putative endotypes to develop prognostic biomarkers for IPF
- To investigate the genetic basis of any putative IPF endotypes

Chapter 2 – General methods

This chapter describes methods which pertain to more than one of the subsequent chapters in the thesis, including genetic association studies, genome-wide association studies (GWAS) and time-to-event analysis.

2.1 Genetic association studies

Genetic association studies are performed to identify genes or genome regions that contribute to the risk of a specific disease by testing for a correlation between genetic variation and a phenotype of interest[96]. A phenotype is a particular trait or measurable characteristic of an individual, aside from their genetic information. Phenotypes may be binary traits, such as disease status, or quantitative characteristics (also commonly referred to as continuous traits), such as systolic blood pressure or lung function measurements. For example, in a study where the phenotype of interest is disease status, a particular allele may be found to appear significantly more frequently in a diseased population than in a healthy population. This can be interpreted as meaning that possessing this particular allele increases an individual's risk of developing the disease. Association studies are a major tool for investigating genetic components of complex diseases that are influenced by many genes, with each contributing a modest effect to overall disease pathogenesis.

2.1.1 Quality control

Without thorough quality control (QC), genetic association studies will not generate reliable results because raw genotype data can contain errors[97]. These errors can arise for numerous reasons, such as poor quality of DNA samples, poor DNA hybridization to the array, poorly performing genotype probes, sample mix-ups and sample contamination. QC consists of filtering out SNPs (this is especially important in a genome-wide analysis, see Section 1.3) and individuals based on several qualities, for example, an individual's genotype call rate.

Genotypes that are missing at random will not bias a test but non-random missingness, with one specific genotype having a lower call rate, may bias tests of association. In addition, a poor call rate for a subject may indicate wider issues with the data and that all the data for that individual may be unreliable. Additionally, if the cases and controls in a case-control study are drawn from separate studies or genotyped separately, different call rates between these groups may also lead to spurious results. Therefore, SNPs that are missing in a large proportion of subjects and individuals who have high rates of genotype missingness should be removed during the QC stage of a genetic association study.

Sample mix-ups can be identified by checking that the reported sex-at-birth of an individual is consistent with the sex that is derived from their genetic data[98], which can be easily predicted as males should have zero true heterozygote calls on the X chromosome and females should have many. In addition, evaluating the proportion of genotype calls that are heterozygous for all individuals can indicate samples

that have been contaminated, as contaminated samples are likely to have a far greater proportion of heterozygous calls than uncontaminated samples[99].

Genotypes within a population are expected to appear in certain ratios. The Hardy-Weinberg equilibrium (HWE) is the principle that in the absence of evolutionary influences, such as: mate choice, inbreeding, mutation, selection and genetic drift, the proportions of alleles and genotypes in an infinitely large diploid population will be the same from one generation to the next, given that there is no migration and that the allele frequencies are equal in the sexes.

Say that for a particular SNP, there are two possible alleles, A and B. The following equation can be derived from the HWE principle:

$$p^2 + 2pq + q^2 = 1 \tag{2.1}$$

Where p is the allele frequency of A and q is the allele frequency of B. This equation states that if the population is in HWE, we would expect to see a proportion of p^2 AA homozygotes, q^2 BB homozygotes and $2pq$ AB heterozygotes.

The HWE principle is useful in genetic association studies as it can be used to test whether the observed genotype frequencies in a population differ from the frequencies predicted by Equation 1.4. Significant deviations from HWE are often the consequence of missingness or genotyping error, and HWE tests are an efficient way of detecting genotyping error[100]. Therefore, variants that depart from HWE should potentially be excluded from the analysis, especially if the variant departs from HWE in healthy controls (who should represent the general population).

Depending on the study design and sample size of the study, it can also be important to filter SNPs based on MAF because rare variants are often less well measured/imputed than more common variants and statistical power is low for rare SNPs unless the sample size is very large[98]. As a result, SNPs with low MAF (e.g. less than 1%) may be excluded from genome-wide analyses (Section 2.2). In addition to removing spurious results, this lightens the computational burden and reduces the number of tests to correct for. Related individuals may also need to be removed as some GWAS approaches assume that all subjects are unrelated and so the inclusion of relatives could lead to biased estimations of standard errors of SNP effect sizes[98]. Finally, when using imputed data, poorly imputed variants are usually excluded based on imputation quality to ensure that spurious associations are not the result of poor imputation.

2.1.2 Genetic models

If we again consider a particular locus that contains a SNP with alleles A and B where the B allele confers an increase in risk, under a dominant genetic model, those with an AB genotype and those with a BB genotype would both have an n-fold risk of disease. Whereas under a recessive model, only those

with a BB genotype would have an n-fold risk of disease. Under an additive model, it is assumed that the risk of disease is increased n-fold for those of genotype AB and 2n-fold for those with genotype BB. In genetic association studies an additive genetic model is often assumed for all genetic loci. This is because the true underlying genetic models are often unknown and the additive model has reasonable power to detect both additive and dominant effects[101].

2.1.3 Testing for genetic association

The analysis of genetic data varies depending on the choice of study design, with the two most common designs being case-control studies and quantitative trait studies. The case-control approach compares two groups of individuals, usually one healthy control group and one disease-afflicted case group. Often, the control group is matched to the diseased group on covariates such as sex, age and smoking history to reduce any biases which may be introduced through differences between the groups. At each SNP it is tested whether the allele frequency is significantly different between the case and the control group using standard statistical approaches such as contingency table methods or logistic regression to produce odds ratios (ORs).

Alternatively, a genetic association study may be designed to investigate a quantitative trait. This type of study often has more power than a case-control study consisting of the same number of disease cases and are typically analysed using linear regression models. When using a linear model, it is assumed that there exists a linear relationship between the independent variables and the mean of the dependent (outcome) variable, that the individuals in the study are independent and that the model residuals are normally distributed with constant variance.

One benefit of using regression methods is that they allow for the adjustment of important covariates such as age and sex. In addition, genetic principal components (Section 2.1.4) are also commonly included as covariates in the regression model to adjust for fine-scale population structure within an ancestral population, as this can confound the results.

Returning to the example in the previous sections, say that at a genetic locus an individual may have either an AA, AB or BB genotype and that B is the risk/effect allele. To investigate a quantitative trait, the following linear regression model could be used:

$$Y_i = \beta_0 + \beta_1 G_i + \beta_2 X_{1i} + \beta_3 X_{2i} + \cdots + \varepsilon_i \qquad (2.2)$$

Where Y_i is the value of the quantitative phenotype for individual i, β_j is the change in the phenotype for a one unit increase in the j^{th} variable, X_{1i} are the values for covariates for individual i that are being adjusted for, ε_i is an error term which we assume is normally distributed around a mean of zero and β_0 is the intercept term. G_i corresponds to the coding for the genotype of the i^{th} individual, which is the value given under the genotypic model used in the study. If the variant was directly genotyped, G_i can be either 0, 1 or 2. For example, under a dominant model individuals with an AA genotype will be

coded 0 and those with AB and BB genotypes will all be coded 1, whereas under an additive model an AA genotype will be coded 0, an AB genotype will be coded 1 and a BB genotype will be coded 2.

If the variant was not directly genotyped but was instead imputed, the value of G_i is often estimated in two main ways. Say that for an individual the probability of an AA genotype at this locus is estimated to be 0.1, the probability of an AB genotype is estimated to be 0.4 and the probability of a BB genotype is estimated to be 0.5 (note that they must sum to 1). If we assume 'hard calls', we assume that the true genotype corresponds to the genotype with the greatest probability (which in this case is BB, corresponding to G_i being coded as 2). If we assume genotype 'dosages', we perform the following simple linear transformation of the genotype probabilities to calculate G_i:

$$G_i = \big(0 \times P(AA\ genotype)\big) + \big(1 \times P(AB\ genotype)\big) + \big(2 \times P(BB\ genotype)\big) \qquad (2.3)$$

Which would result in G_i being coded as 1.4 in our example.

Similarly to the linear regression model (Equation 1.5), we may use logistic regression to analyse a binary trait (e.g. case-control status) with the following statistical model:

$$logit(p_i) = log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 G_i + \beta_2 X_{1i} + \beta_3 X_{2i} + \cdots + \varepsilon_i \qquad (2.4)$$

Where the outcome is coded as 0 or 1 and p_i is the probability of individual i having an outcome of 1, i.e. the probability that they are in the case group. In this case, the logit function provides the log-odds of p_i. The values of β_j correspond to the log odds ratio for every unit increase in the j_{th} variable. Of particular interest is the value of $e^{\beta 1}$ which corresponds to the odds ratio for the genetic effect. When testing for evidence of a genetic association (using either Equation 1.5 or 1.7), we test the null hypothesis that the genetic variant is not associated with the phenotype, i.e. β_1=0. This is most commonly done using either the score test, the Wald test or the likelihood ratio test.

2.1.4 Population stratification

In a genetic association study, there is one true confounder that must be controlled for: confounding by ancestry, also known as population stratification. Population stratification is particularly important to account for in case-control studies and occurs when the two disease status groups have poorly matched ancestry, leading to differences in variant frequencies between the groups that are not caused by the disease status and therefore produces spurious associations. To help combat this issue, studies often only include subjects from one particular ancestral group, such as those of European ancestry. However, there will still be an underlying population structure in a group of individuals from the same ancestral group, which could impact the results of an analysis.

Principal components analysis (PCA) is a dimensionality-reduction technique that can be applied to genetic data to generate genetic principal components (also known as principal components of

ancestry), which are a relatively small number of uncorrelated variables that explain as much of the total variance within the data as possible. Novembre et al.[102] showed that a plot of the first two genetic principal components for a group of European individuals resembled a map of Europe, with those from the same countries often grouped closely together (Figure 1.4). This study suggested that population structure correction may be important even in seemingly closely related populations, such as Europeans. As such, in a genetic association study, genetic principal components are usually included as covariates in the regression model to help explain some of the variation in the data caused by the differences in ancestry and reduce the risk of spurious associations[103].

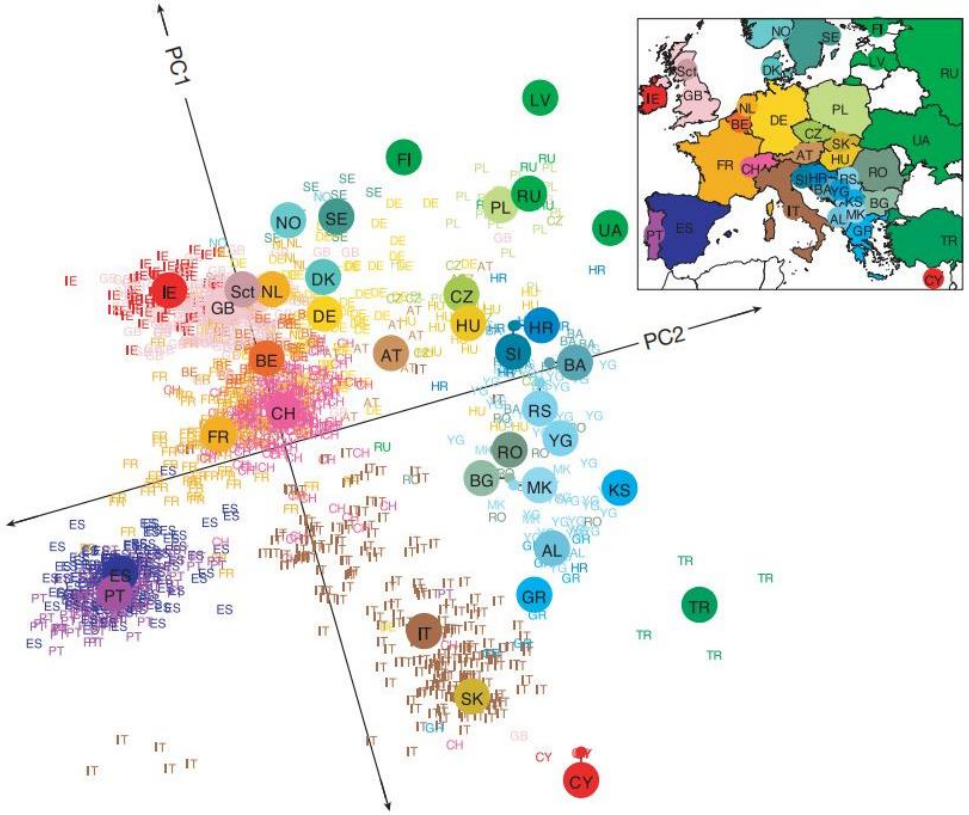


FIGURE 2.1: *Population structure within individuals of European ancestry. Figure taken from Novembre et al.*[102]*. It shows a statistical summary of genetic data from 1,387 Europeans based on the first genetic principal component (PC1) and the second genetic principal component (PC2). Small coloured labels represent individuals and large coloured points represent median PC1 and PC2 values for each country. The inset map provides a key to the labels. The axes are rotated to emphasize the similarity to the geographic map of Europe.*

2.2 Genome-wide Association Studies

A genome-wide association study (GWAS) is a large-scale genetic association study, in which the entire human genome, rather than particular candidate gene regions, is scanned to identify variants that are associated with a certain phenotype. As such, the methods for testing the genetic association of a single variant with a phenotype of interest are largely the same as those described in Section 2.1. However, as

they do not require candidate regions, GWAS have been characterized as 'hypothesis-free' and are thus less likely to miss important variants that may have been overlooked in a candidate gene/variant study.

2.2.1 Testing genome-wide genetic variation

As genotyping and imputation technology advance, the number of SNPs that are able to be included in a GWAS is growing larger. However, false positive results occur more frequently in a study as the number of tests increases. As such, adjustment for multiple testing is vital to reduce the false positive rate in a GWAS. The Bonferroni correction, the most commonly used adjustment in GWAS, works by testing each individual hypothesis at a significance level of α/m, where α is the desired overall significance level (usually 0.05) and m is the number of independent tests. The Bonferroni correction is easy to implement but assumes that the tests are independent, and SNPs in LD are not independent. Therefore, rather than correcting for the total number of SNPs tested, the standard threshold for genome-wide statistical significance in a GWAS is $P<5\times10^{-8}$, which is equivalent to a Bonferroni correction for an assumed 1 million independent variants across the human genome[104].

2.2.2 GWAS results

Even after QC, the results of a GWAS should still be evaluated for evidence of systematic biases, such as population stratification. This is often performed by plotting the p-values of the GWAS in a quantile-quantile (Q-Q) plot (Figure 1.5) and by calculating the genomic inflation factor (λ), which is a measure that quantifies the level of excess false positives within the results of a GWAS. λ is calculated by taking the p-values from the association tests for all genetic variants and then using each of these to calculate a corresponding test statistic from a chi-square test with one degree of freedom. Then, the median of all the test statistics is divided by the expected median of the chi-squared distribution with one degree of freedom and the resulting value is the genomic inflation factor.

A λ value of 1 is desirable as it indicates that there is no inflation of the test statistic due to systematic biases and thus the results require no adjustment. In practice, λ is deemed to be acceptable if it is between 0.9 and 1.1 in a GWAS. If there is evidence of systematic bias, the results can be adjusted using a method called genomic control[105].

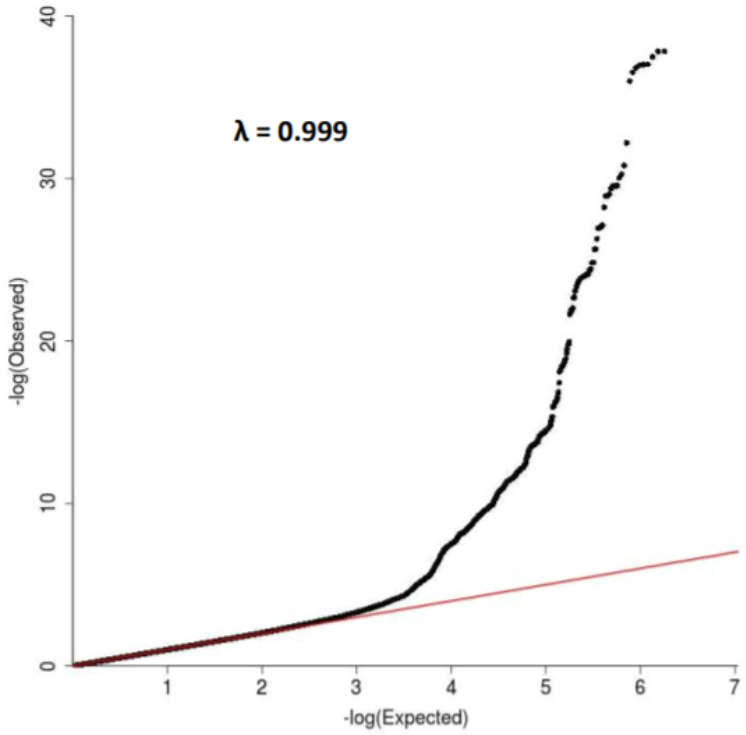


FIGURE 2.2: *Example of a Q-Q plot. Figure taken from Allen et al.*[77]*. Log-transformed expected P-values are on the x-axis and log-transformed observed P-values are on the y-axis. The red line shows where the expected distribution equals the observed distribution. There are approximately 8 million SNPs in this plot, with most laying in the bottom left-hand corner. Despite the number of highly statistically significant SNPs shown on the right-hand side of the plot, the genomic inflation factor (λ) is low because the vast majority of SNPs closely follow the red line.*

Genomic control is performed as follows: first, the standard error of each SNP is multiplied by the square root of the genomic inflation factor. Then, a new z-statistic is calculated for each SNP by dividing the beta coefficient for that SNP by the adjusted value of the standard error. Finally, a corrected p-value for each SNP is calculated by performing a Z-test using the new z-statistic for that SNP.

The results of GWAS are usually visualised using a Manhattan plot (Figure 1.6), which is a type of scatter plot where each point represents one SNP, the x-axis represents the position of the SNP along its chromosome and the value on the y-axis shows the negative base-10 logarithm of the p-value from the association test for that SNP. Named after its resemblance to the Manhattan skyline, this type of plot will ideally show numerous peaks - indicating genetic loci that have displayed strong associations with the phenotype of interest - which tower over the less significantly associated genetic variants. The horizontal red line in Figure 1.6 represents the threshold for genome-wide statistical significance ($P<5\times10^{-8}$) and each point that lies above this line represents a SNP that is genome-wide significantly associated with the disease. The peaks are observed because there are many SNPs close together along the chromosome and they are in LD, therefore being found together on the same haplotype more often than expected. Because of this, the 'sentinel' SNP at the top of a peak (the most statistically significant

SNP of that genetic locus) is not necessarily the variant that is influencing the disease and the challenge lies in determining the true underlying causal variant (or variants).

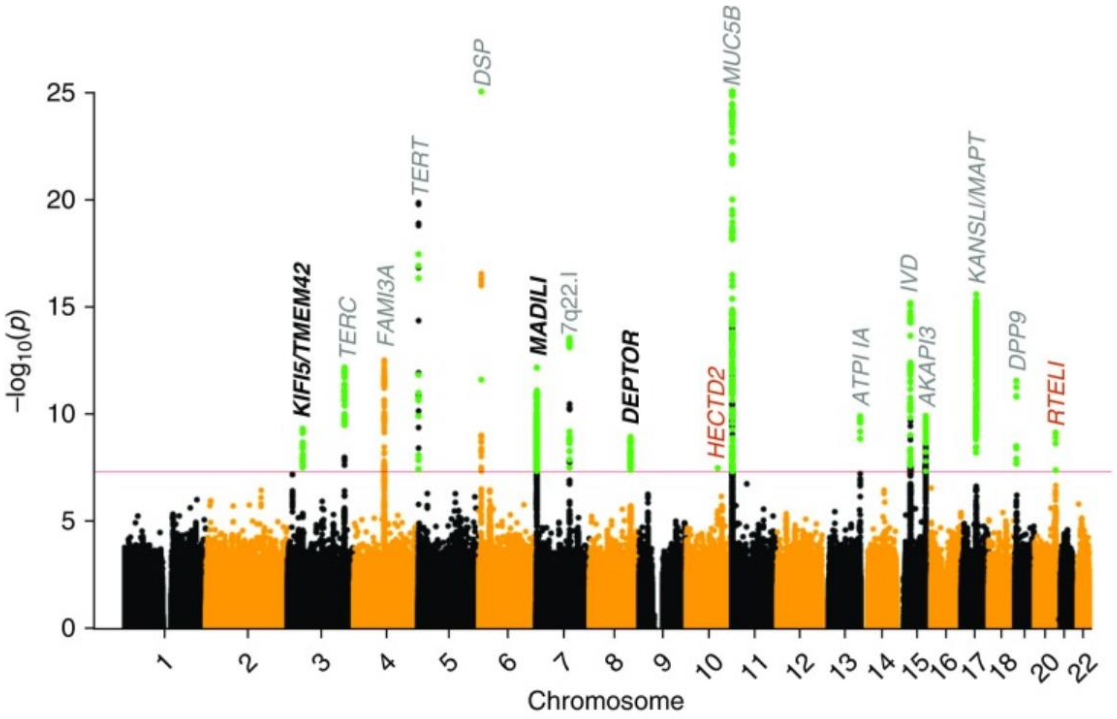


FIGURE 2.3: *Example of a Manhattan plot, a commonly used visualisation tool in genome-wide association studies. This figure was taken from Allen et al[77]. The horizontal red line represents a genome-wide threshold for statistical significance, set at $P<5\times10^{-8}$ here.*

Regional association plots (Figure 1.7) provide a closer look at an area of a Manhattan plot and can be used to examine a genetic locus of interest in greater detail. Regional association plots also show the LD relationship between a particular variant (usually the sentinel SNP) and the other variants in the region. For example, in Figure 1.7 the most strongly associated variant in the region is at position 130,184,065 and several SNPs in the region are in LD with this variant, as indicated by the various colours. Additionally, regional association plots can show recombination frequency information for the region as well as the genes within or close to the association peak.

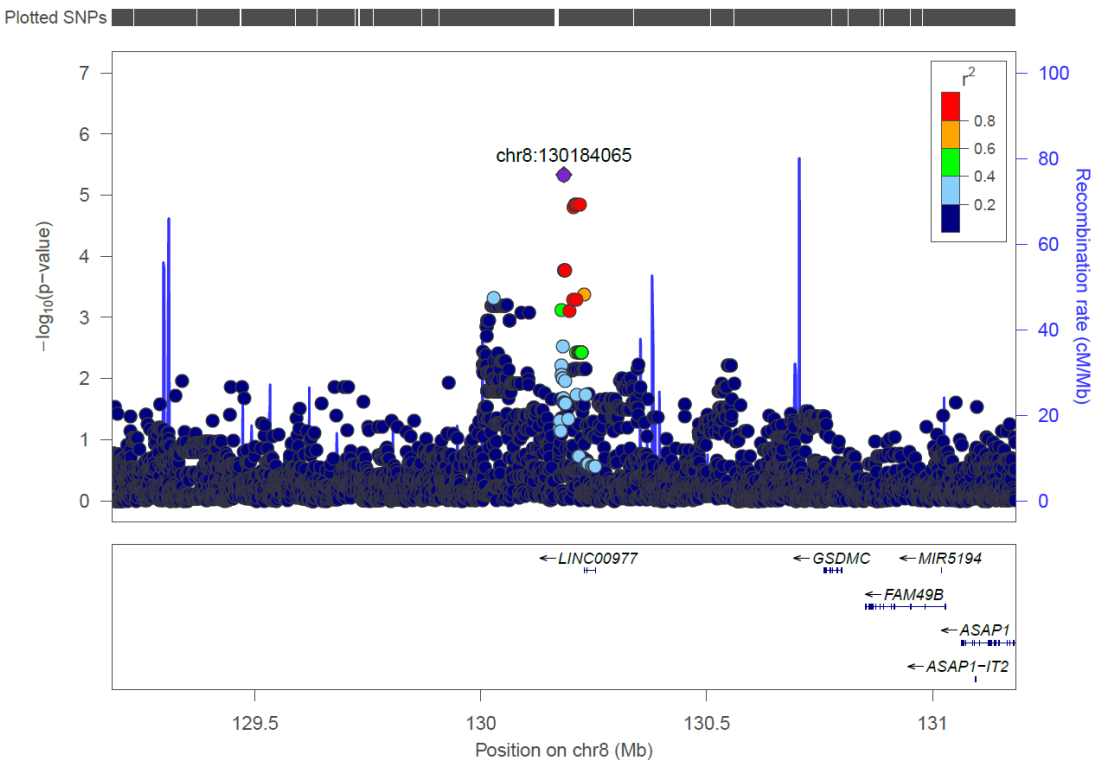


FIGURE 2.4: *Example of a regional association plot, a commonly used plot in genome-wide association studies.*

Due to the high frequency of false positive results found in a GWAS, a study should aim to verify any genetic signals that have displayed strong statistical significance in the genome-wide scan to rule out false positive associations. A traditional discovery and replication study design is commonly used, in which the genome-wide scan is performed in one cohort to identify genome-wide significant associations (the discovery stage), and then any proposed associations are tested in an additional independent cohort (the replication stage). Any genetic signals that reach a Bonferroni-corrected threshold in the replication stage are considered to have successfully replicated (i.e. to represent true and verified associations).

However, genetic variants often suffer from 'winner's curse', which is a phenomenon where the effect size of an association is overestimated in the first study to report it and is found to be lower when investigated in subsequent analyses. Winner's curse can lead to underpowered replication analyses and so a replication sample should ideally be larger than the sample that the association was discovered in. There must also be homogeneity between the initial and replication cohorts; the subjects should be drawn from the same ancestral population and they should have identical phenotype criteria.

Alternative variations of this study design are often used. For example, a study may choose to use a 2-stage design. In this type of study, stage 1 is a genome-wide scan in which SNPs that reach statistical significance at a certain predefined threshold (i.e. $P<5\times10^{-6}$) qualify for follow-up in stage 2. Stage 2 subsequently tests these variants in an additional independent sample and then the stage 1 and stage 2

results are meta-analysed. Variants that reach genome-wide significance in the meta-analysis are considered to be verified. This type of study design can be particularly useful in low powered studies with relatively small sample sizes in the discovery stage.

2.2.3 Investigating genetic association signals

If a significant association is detected between a genetic variant and a disease, this does not always mean that this variant is directly influencing the disease, as the true causal variant may be in LD with the detected SNP. Bayesian fine-mapping (Section 3.5.1) can be used to produce a set of variants (referred to as the 95% credible set) for each putative signal that are 95% likely to contain the underlying causal variant, assuming that the causal variant has been analysed[106]. However, fine-mapping has three important requirements: all variants in the region must be either genotyped or have high imputation quality, the variant quality control must have been stringent and the sample size must be large enough to provide sufficient power to differentiate between SNPs in high LD.

One approach that is used to explain the functional link between a SNP and the risk of a disease is expression quantitative trait loci (eQTL) analysis (Section 3.5.1). Genotype and gene expression data from particular tissues of interest (e.g. blood or lung tissue) are queried to identify whether a genetic variant that has been found to be associated with a disease trait is also significantly associated with changes in expression for any genes. This eQTL evidence gives an indication of which genes might be functionally relevant to a signal, although conclusive evidence for a causal relationship between SNP genotype and gene expression can only be obtained through further bespoke molecular experiments.

2.3 Time-to-event analysis

In time-to-event analysis, there are several useful functions that can be estimated. Of note are the survival function (Equation 2.5), which is the probability of an event of interest occurring after time *t*, and the hazard function (Equation 2.6), which is the instantaneous probability of the event of interest occurring at time *t*, given that the event has not already occurred. If *T* is the time that the event of interest occurs, these can be written as:

$$S(t) = \mathrm{P}(\mathrm{T} \geq \mathrm{t}) \tag{2.5}$$

$$h(t) = \lim_{\delta \to 0} \frac{P(t \leq T < t + \delta)}{\delta} \tag{2.6}$$

The survival function can be estimated from the data using the Kaplan-Meier estimator[107]. At the k^{th} time point in the data, the Kaplan-Meier estimator can be written as:

$$\hat{S}(t_k) = \prod_{i=1}^{k} \frac{n_i - d_i}{n_i} \tag{2.7}$$

where d_i is the number of events at time point i and n_i is the number of individuals still at risk of experiencing the event at time *i*. Plotting $\hat{S}(t_k)$ against time gives a Kaplan-Meier curve. On a Kaplan-Meier plot, the median survival time of a group of individuals is the point in time at which their Kaplan-Meier curve crosses 0.5 on the y-axis (i.e. when $\hat{S}(t_k) = 0.5$).

One of the most common approaches in time-to-event analysis is the Cox PH model[108], which is as follows:

$$h(t) = h_0(t) \times exp\ (\beta_1 X_1 + \beta_2 X_2 + \cdots) \quad (2.8)$$

where $h_0(t)$ denotes the baseline hazard function, which describes how the hazard function changes over time for an individual with baseline levels of covariates.

The Cox PH model assumes that the covariates are multiplicatively related to the hazard, which means that the hazard function in one group over time is always equal to the hazard function in another group multiplied by a constant value. This assumption is known as the PH assumption. Two groups with different covariate values can be compared using the hazard ratio (HR), which is the hazard function in one group divided by the hazard function in the other. As the time-dependent effects in Equation 2.8 will cancel each other out in this calculation, the HR will be a constant and the Cox PH model can be fit without making any assumptions about the underlying distribution of the baseline hazard function.

Evidence of breaches to the PH assumption can be checked using Schoenfeld residuals[109]. For each subject and for each covariate, Schoenfeld residuals are calculated as the difference between the observed covariate values minus the expected covariate values at each failure time. Under the PH assumption, these residuals should not show a trend when plotted against time.

Chapter 3 – Genome-wide analyses to identify genetic determinants of the age-of-onset of IPF

Previous GWAS have identified several genetic variants that are associated with IPF susceptibility, but none have investigated the age-of-onset. This chapter describes the first analyses to investigate the genetic basis of the age-of-onset phenotype in non-familial IPF. The objective of these analyses was to identify common (MAF >5%) and low-frequency (MAF 1-5%) genetic variants that were significantly associated with the age-of-onset of IPF, and to verify findings via follow-up in additional independent samples or through internal validation methods.

3.1 Introduction

In 2019, Krauss et al.[110] showed that individuals with familial pulmonary fibrosis (FPF) had a significantly lower average age-of-onset than those with sporadic cases of IPF. In addition, the authors found that in cases of FPF the younger generation tended to manifest disease at a younger age compared to their counterparts in the older generation. These findings suggested that there is a genetic component to the age-of-onset of FPF and it stands to reason that the age-of-onset of Idiopathic PF may also possess a genetic component.

Three cohorts of subjects with IPF were included in these analyses. These were the PROFILE, Trent Lung Fibrosis (TLF) and UK Biobank (UKB) cohorts (Section 3.2.1). The age-of-onset of IPF is difficult to determine exactly as it takes place prior to the development of symptoms and to the IPF diagnosis. It was assumed that a subject's diagnosis would have been the earliest accurately recorded time after the development of the disease and onset of symptoms. As such, each subject's age-at-diagnosis was considered to be a reasonable proxy for their age-of-onset in these analyses.

In the first analysis, the age-at-diagnosis of IPF was modelled as a continuous outcome using linear regression (Section 3.2). This GWAS was performed using a two-stage study design[111]. In the first stage, a genome-wide analysis was performed to identify variants that were associated with the age-at-diagnosis of IPF in a discovery cohort (PROFILE). In the second stage, the strongest genetic signals of association were tested further in the two remaining independent cohorts (TLF and UKB). Following this, the results from both stages were meta-analysed to identify whether any variants were genome-wide significantly associated with the age-of-diagnosis of IPF when all three cohorts of IPF subjects were considered together.

The analysis was then repeated with two key improvements regarding the study design and the choice of statistical model (Section 3.3). Firstly, the age-at-diagnosis of IPF was modelled using time-to-event analysis methods, which are better suited to model the age-of-onset than linear regression, as the age-of-onset could be considered time-to-event data. In addition, time-to-event methods are not affected by

deviations from normality within the distributions for the age-of-onset proxies, which were encountered during the first analysis. Secondly, a 3-way GWAS meta-analysis study design was utilised, which allowed all available data to be incorporated in the discovery analysis. In addition, criteria specifying support for signals from multiple contributing studies were applied to improve the credibility of any findings by ensuring that any novel genetic signals must have showed consistent association in each independent cohort.

3.2 Two-stage GWAS analysis

3.2.1 Datasets and study design

Three cohorts of individuals with IPF from separate studies were included in this analysis. These studies were selected as they had either reported the age-at-diagnosis of their subjects or they had recruited all subjects within six months of their IPF diagnosis and reported the age of their subjects at the time of enrolment. All three studies originated from the United Kingdom. The first was the Prospective Observation of Fibrosis in the Lung Clinical Endpoints (PROFILE) study[112]. The primary aim of the PROFILE study was to develop and validate novel prognostic biomarkers for IPF patients. Across two centres (the University of Nottingham and the Royal Brompton Hospital), a total of 560 newly diagnosed individuals were recruited to the study.

The second cohort originated from the Trent Lung Function (TLF) study[113], a prospective study which aimed to investigate hypercoagulability in IPF patients. Patients were eligible for inclusion in the study if they were newly diagnosed with IPF in the six months before the start of the study or throughout the recruitment period (from January 2010 to February 2012). 211 incident cases of IPF were recruited to the study during this time.

The third cohort of IPF subjects were part of the UK Biobank (UKB) study[114]. Approximately 500,000 British individuals aged 40-69 years were recruited into the study from 2006 to 2010. Of these, 121,271 individuals later responded to a questionnaire that asked whether they had been diagnosed with IPF by a doctor and their age-at-diagnosis (if applicable). 108 of the participants said they had been diagnosed by a doctor to have IPF (data field 22135) and provided their age-at-diagnosis (data field 22155).

This analysis consisted of three parts. Stage 1 was the discovery analysis, in which association testing was conducted genome-wide in individuals with IPF from the PROFILE study. In stage 2, any variants that met $P<10^{-5}$ in stage 1 were tested in the additional independent IPF cases from the TLF and UKB studies. Following this, the results from stage 1 and stage 2 were meta-analysed.

3.2.2 Methods

In this study, the age at which each subject was diagnosed with IPF was used as a proxy for their age-of-onset. The IPF cases in UK Biobank self-reported their age-at-diagnosis via a questionnaire. For the

individuals in the PROFILE and TLF studies, their age-at-enrolment was used as a proxy for their age-at-diagnosis.

To address the issue of population stratification, individuals were excluded from this analysis if they had self-reported non-European ancestry. In addition, PCA was used to identify and exclude individuals who were not of European Ancestry based on their genotype data; individuals were considered to be of non-European ancestry and excluded if they were visually determined to be an outlier on a plot of the first two genetic principal components (i.e. if they laid far apart from the main group of overlapping individuals). Additionally, individuals in UKB with outlier values for age-at-diagnosis (<35 years) were removed as these were self-reported and thus subject to error.

KING relationship inference software[115] was used to check for duplicate subjects and related individuals across the three cohorts. Subjects present in more than one cohort and those found to be second degree related (or higher) to another individual in the study were excluded. In cases where a pair of identical or related individuals were identified across cohorts and one of the individuals was in the PROFILE cohort, the identical/related individual was removed from the other study (TLF/UKB). This was done to maintain the largest possible sample size in the PROFILE cohort, therefore maximising the statistical power of the discovery stage of the analysis. In all other cases where related individuals were identified, one member of the related pair was selected at random and excluded.

Sex information was available in all cohorts, with all subjects being self-reported as either male or female. In the PROFILE cohort, individuals were recruited from either Nottingham or Brompton. All subjects in PROFILE were genotyped using the UK Biobank Axiom array, all subjects in TLF were genotyped using the closely related UK BiLEVE Axiom array and UKB used both of these arrays. Genotype imputation for the PROFILE and TLF cohorts was performed for a previously published GWAS of IPF susceptibility[116]. Genotype imputation for the UKB cohort was performed as described in Bycroft et al.[117]. Smoking history information was reported as ‘never smoker’, ‘former smoker’ or ‘current smoker’ in the PROFILE study and the TLF study. In the UK Biobank study, each subject’s smoking status was reported as either ‘smokes on most or all days’, ‘never smoker’, ‘ex-smoker’ or ‘smokes occasionally’. In both cases, these were included as categorical variables in the statistical model.

In each study cohort, a linear regression model was used to test for genetic associations, with the age-at-diagnosis of IPF as the response variable. Sex, smoking status, recruitment centre (only applicable to those in PROFILE) and genotyping array (only applicable to those in UKB) were included as covariates in the linear regression model. Additionally, the first 10 principal components of ancestry were included in the model to account for fine-scale population structure. For each genetic locus, the number of copies of the risk allele present for each individual was input into the model as an imputed

dosage and the primary variable of interest was the genetic effect at that locus. An additive genetic model was assumed. The equation for the linear model is shown in Equation 3.1.

$$Age\ variable = \beta_0 + \beta_1 Number\ of\ copies\ of\ risk\ allele + \alpha_1 Sex + \alpha_2 Smoking\ Status + \alpha_3 Recruitment\ Centre + \alpha_4 Array + \beta_2 PC1 + \cdots + \beta_{11} PC10 + \varepsilon \quad (2.1)$$

Quality control was performed in the discovery stage by excluding any genetic variants that could have led to spurious results due to being poorly imputed, by being too rare in the discovery population or by being out of HWE in that population. Only the variants that were well imputed ($R^2 > 0.5$), that were common/low frequency (MAF ≥ 0.01) and that were in Hardy-Weinberg equilibrium ($P > 1\times10^{-6}$) were included in the analysis.

The genome-wide analysis of stage 1 (PROFILE cohort only) was performed using SNPTEST v2.5.2. The genome-wide results were visualised in a Manhattan plot using R (v4.0.0 and the 'qqman' package). A Q-Q plot and the genomic inflation factor (λ, Section 2.2.2) were used to determine whether there was unadjusted population structure present within the stage 1 results. If $\lambda \geq 1.1$, the results were considered to require adjustment, which was undertaken using genomic control (Section 2.2.2).

Suggestive statistical significance for the purpose of defining independent signals within the data was set as $P < 5\times10^{-5}$ based on visualisation of the stage 1 results. Sentinel SNPs were defined as those that met the suggestive threshold and were at least 1 megabase (Mb) away from any other variants that showed a more significant association with the age-at-diagnosis (i.e. those with a lower p-value). LocusZoom[118] software was used to create a regional association plot for each sentinel SNP that visualised the genetic region 1Mb to either side of the sentinel SNP, as well as the LD structure between that SNP and the other genetic variants in that region. These plots were used to check each region for evidence of additional independent signals, as well as to identify any spurious signals (defined as those with abnormal LD structure or LD structure that was not consistent with the MAF of the sentinel variant). Signals that were found to be spurious according to these criteria were disqualified from follow-up in stage 2 of the analysis.

Conditional analyses were then conducted using SNPTEST v2.5.2 to test for independent signals within regions ±1Mb around the sentinel SNPs. The association test at each region was repeated with the same covariates as described previously, whilst additionally including the number of copies (as a dosage) of that region's sentinel SNP coded allele in the linear model. The other signals in that region were then evaluated to see if they were significantly associated with the age-at-diagnosis of IPF after the adjustment for the first signal. If at least one of these variants reached suggestive statistical significance, the number of copies of the coded allele of the most significantly associated SNP in the region was then added to the linear model. This process was repeated until no variants reached the suggestive threshold.

Any independent suggestively significant ($P<5\times10^{-5}$) signals were investigated further using LocusZoom regional association plots.

Of all independent signals that met $P<5\times10^{-5}$ in stage 1 following conditional analyses, those that met a slightly stricter threshold of $P<10^{-5}$ were prioritised for follow up in stage 2. Association tests for these variants were then performed in the individuals in the TLF and UKB cohorts. This was done in each cohort separately and using the same statistical approach and software as described for stage 1. These results were then meta-analysed using a fixed-effect inverse variance model. Genome-wide statistical significance was defined as $P_{meta}<5\times10^{-8}$.

To assess whether there was any overlap between the genetic determinants of the age-at-diagnosis of IPF and those of IPF susceptibility, a lookup was performed within the stage 1 results. The variants in this lookup consisted of the most significantly associated SNPs from the 14 genome-wide significant signals in the largest GWAS meta-analysis of IPF susceptibility to-date[77], as well as the rare variant in the *SPDL1* region that was recently identified as being genome-wide significantly associated with IPF susceptibility[119] (Additional Table B.3.1 in Appendix B). A variant was considered to be associated with both IPF susceptibility and the age-at-diagnosis of IPF if the corresponding P-value for that variant in the age-at-diagnosis lookup was lower than the Bonferroni-corrected threshold for 15 tests at a significance level of 0.05 (i.e. $P<0.003$).

3.2.3 Results

After exclusions, there were 465 individuals with IPF in the PROFILE cohort, 210 in the TLF cohort and 98 in the UKB cohort (Table 3.1). There was only one individual in the UKB cohort who reported that they were a current smoker, so this category was removed and this individual was moved into the former smoker group. In all three cohorts, there were a higher proportion of males than females and on average males were slightly older when they were diagnosed/enrolled into their study. Individuals in the TLF cohort had the greatest mean age-at-diagnosis/enrolment at 73 years for females and 74 years for males. Those in the UKB cohort had the lowest mean age-at-diagnosis/enrolment at 63 years for females and 65 years for males. In the PROFILE cohort, those who were recruited from the Royal Brompton Hospital were on average 4 years younger than those who were recruited from the University of Nottingham at the time of recruitment (Additional Figure A.3.1 in Appendix A).

The age-at-enrolment of PROFILE and TLF both appeared approximately normally distributed around a mean of 70-75 years, whilst the distribution of the UKB cohort appeared to be truncated (Figure 3.1). This was likely because only individuals who were between 40 and 69 years of age were recruited into the study and so there were no subjects whose age-at-diagnosis could have been greater than 75 years at the time the information was gathered.

TABLE 3.1: *Demographics for the individuals with IPF from the three cohorts that were included in the analysis.*

		PROFILE (n=465)		**Trent Lung Fibrosis (n=210)**		**UK Biobank (n=98)**	
Phenotype		**Count (%)**	**Mean age-at-enrolment (years) (sd)**	**Count (%)**	**Mean age-at-enrolment (years) (sd)**	**Count (%)**	**Mean age-at-diagnosis (years) (sd)**
Sex	Female	108 (23.2%)	70.1 (7.6)	65 (31.0%)	72.8 (8.7)	43 (43.9%)	63.3 (7.7)
	Male	357 (76.8%)	71.0 (8.3)	145 (69.0%)	73.8 (9.1)	55 (56.1%)	64.5 (7.6)
Smoking status	Never	144 (31.0%)	71.5 (7.4)	56 (26.7%)	74.7 (9.5)	33 (33.7%)	64.4 (7.6)
	Former	296 (63.7%)	70.9 (8.1)	136 (64.8%)	73.7 (8.6)	65 (66.3%)	63.1 (7.7)
	Current	25 (5.4%)	64.6 (9.4)	18 (8.6%)	68.1 (8.4)	-	-
Recruitment centre	Brompton	205 (44.1%)	68.5 (8.1)	-	-	-	-
	Nottingham	260 (55.9%)	72.5 (7.7)	-	-	-	-
Genotyping array	UK BiLEVE	-	-	210 (100%)	73.5 (9.0)	12 (12.2%)	61.8 (9.1)
	Axiom	465 (100%)	70.8 (8.1)	-	-	86 (87.8%)	64.3 (7.4)

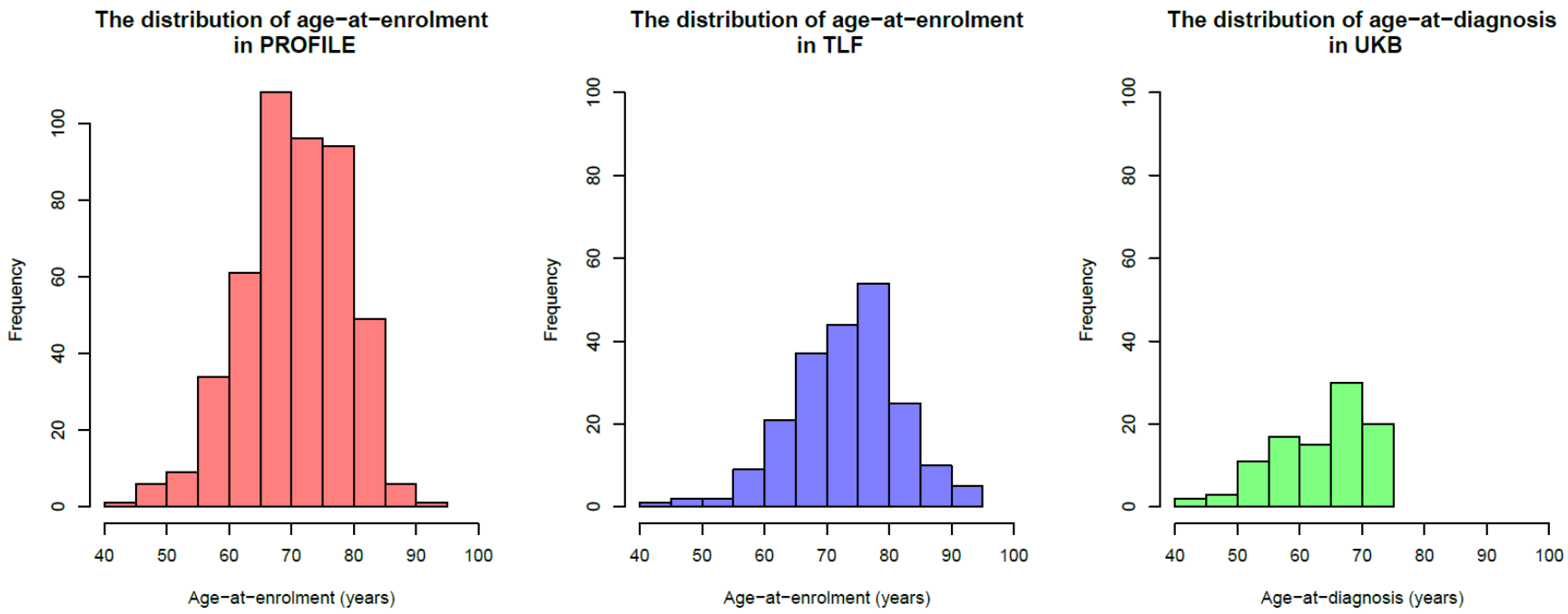


FIGURE 3.1: *Histograms showing the distributions of the age-at-enrolment into study of the PROFILE and Trent Lung Fibrosis (TLF) cohorts and the distribution of the self-reported age-at-diagnosis of the UK Biobank (UKB) cohort.*

Prior to variant QC, there were 39,131,578 variants available for the discovery analysis. Of these, 10,858,143 variants were polymorphic in the PROFILE cohort. Of those, 9,958,982 had good imputation quality (R^2>0.5), of which 7,647,898 had MAF ≥ 0.01. 2,674 of those SNPs were found to be out of HWE ($P<1\times10^{-6}$) and were excluded, leaving 7,645,224 SNPs that were included in the discovery analysis.

Association testing was performed genome-wide on the 7,645,224 SNPs in the PROFILE cohort. The p-values of the association tests did not suggest inflation due to unadjusted population structure (Figure 3.2). This was supported by a λ of 1.008, and no further adjustment was made to account for population stratification.

The genetic variant with the most statistically significant association with the age-at-diagnosis was found on chromosome 8 with a p-value of 2.87×10^{-7} (Figure 3.3). In total there were 82 sentinel SNPs (represented by the green points in Figure 3.3).

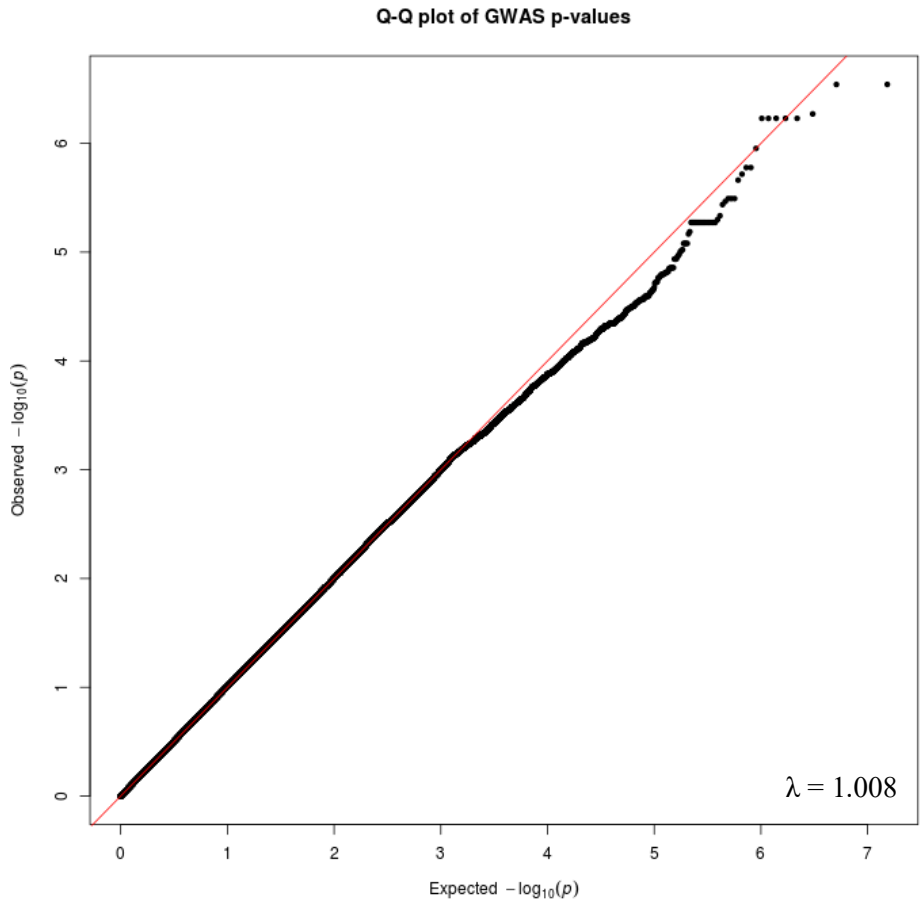


FIGURE 3.2: *A quantile-quantile plot of the p-values in the discovery analysis (performed in 465 subjects from the PROFILE study). λ = the genomic inflation factor.*

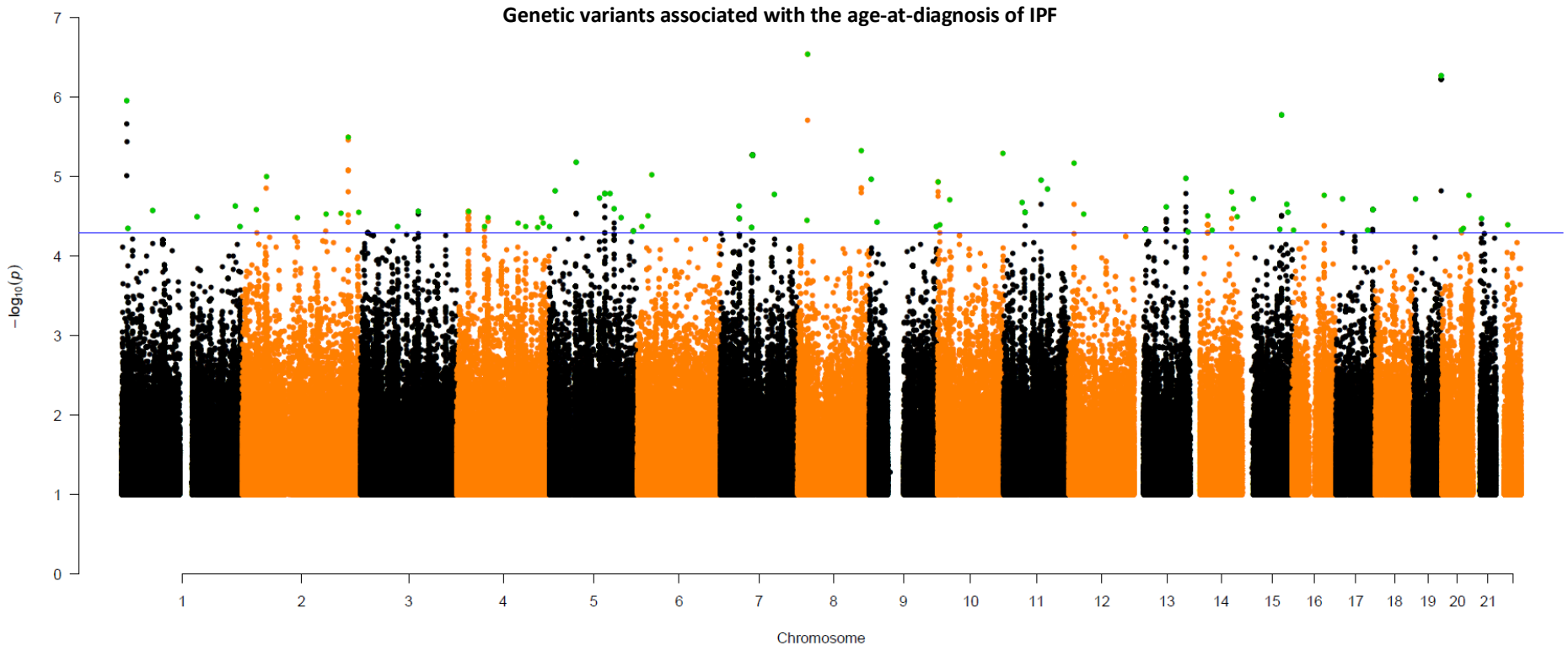


FIGURE 3.3: *A Manhattan plot showing the results of the genome-wide analysis, in which each genetic variant that passed quality control was tested for an association with the age-at-diagnosis of IPF in the PROFILE cohort. The 82 sentinel SNPs are highlighted green and the threshold for suggestive significance ($P<5\times10^{-5}$) is represented by the blue horizontal line. All variants with a p-value greater than 0.1 were removed to reduce the computational burden of the plot.*

Conditional analyses for the 82 sentinel SNPs identified six additional independent genetic signals, producing a total of 88 independent genetic signals of interest (Additional Table B.3.2). None of the regional association plots for the 88 signals indicated that any variants should be disqualified from further study due to abnormalities in the LD structure or inconsistencies between variant frequency and LD structure.

Of the 88 independent genetic signals, 14 variants had $P<10^{-5}$ and were prioritised for follow-up in stage 2 (Table 3.2). For all 14 variants, each copy of the minor allele corresponded to a younger expected age-at-diagnosis of IPF. The variant with the greatest effect size was rs114791520, where each copy of the effect allele corresponded to a younger age-at-diagnosis of 1.3 years.

TABLE 3.2: *Summary statistics from the stage 1 results for the 14 genetic variants that were eligible for follow-up in stage 2 of this study. EAF= effect allele frequency, SE= standard error, R^2 refers to the imputation quality of the variant.*

rsid	Chr.	Position	Locus	Reference/ effect allele	EAF	P-value	Beta	SE	R^2
rs78672887	1	11473150	Intergenic	C / T	3.4%	1.11×10^{-6}	-0.83	0.170	0.862
rs12471179	2	49568428	*LOC105374595* (intron)	A / C	32.4%	9.89×10^{-6}	-0.29	0.065	0.972
rs7599256†	2	171495421	*MYO3B* (intron)/ *LOC100130256* (intron)	G / T	54.4%	6.80×10^{-6}	0.28	0.062	0.998
rs72958256	2	217676681	*LOC101928278* (intron)	C / T	21.2%	3.22×10^{-6}	-0.35	0.075	0.967
rs72749864	5	54507085	Intergenic	T / A	3.3%	6.49×10^{-6}	-0.73	0.162	0.967
rs114791520	6	29487974	Intergenic	T / C	1.1%	9.48×10^{-6}	-1.32	0.297	0.906
rs76259754	7	65599188	*CRCP* (intron)	C / T	1.4%	5.34×10^{-6}	-1.19	0.262	0.933
rs75681116	8	20248930	Intergenic	C / T	2.9%	2.87×10^{-7}	-0.95	0.184	0.998
rs12155839	8	130184065	Intergenic	C / T	17.3%	4.63×10^{-6}	-0.37	0.081	0.987
rs182317201	10	133919390	*JAKMIP3* (intron)	C / T	1.7%	5.02×10^{-6}	-1.08	0.237	0.974
rs74715174	12	10065926	*CLEC2A* (intron)	A / T	3.9%	6.81×10^{-6}	-0.71	0.157	0.950
rs113262525‡	12	10291177	Intergenic	T / G	8.0%	4.63×10^{-6}	-0.48	0.104	0.977
rs61459715	15	80777127	*ARNT2* (intron)	G / A	1.6%	1.67×10^{-6}	-1.17	0.244	0.883
rs117388035	19	57260406	*LOC105372472* (intron)	G / A	2.6%	5.36×10^{-7}	-0.98	0.195	0.981

†: Conditional on the number of copies of the rs192643964 coded allele.
‡: Conditional on the number of copies of the rs74715174 coded allele.

In stage 2, the 14 variants listed in Table 3.2 were tested for an association with the age-at-diagnosis of IPF in the subjects from the TLF cohort (n=210) and the UKB cohort (n=98). One variant, rs76259754, could not be tested in UKB as there was only one instance of the alternate allele, which meant that the allele count for this SNP was below the minimum allele count required by SNPTEST. The results from stages 1 and 2 were then meta-analysed (Table 3.3).

There were no genetic variants that reached genome-wide statistical significance ($P<5\times10^{-8}$) in the meta-analysis. However, there were three SNPs (rs114791520, rs75681116 and rs182317201, indicated by the highlighted rows in Table 3.3) that had a consistent direction of effects across all three cohorts. All three of these variants were low frequency, with MAF in the 1-5% range in all cohorts. rs75681116 and rs182317201 maintained suggestive statistical significance in the meta-analysis at a threshold of $P_{meta}<10^{-5}$. However, none of these three variants had greater statistical significance in the meta-analysis than in the stage 1 discovery analysis. Of the three, the variant with the greatest estimated effect size was rs182317201 (β=-0.83), where each copy of the effect allele corresponded to an estimated younger age-of-onset of IPF by approximately 10 months. Only one variant, rs72958256, maintained suggestive significance in the meta-analysis despite not having a consistent direction of effects in each cohort. This SNP, which is found on chromosome 2, had a lower p-value in the meta-analysis ($P_{meta}=3.0\times10^{-6}$) than in the discovery analysis ($P=3.2\times10^{-6}$) but did not have a consistent direction of effects across the three cohorts as the estimated effect size was negative in PROFILE and TLF but positive in UKB, though very close to zero (β= 0.001). The estimated effect size of this SNP was -0.28, which meant that each copy of the T allele at this locus corresponded to a younger estimated age-of-onset of IPF by approximately 3 months. Forest plots for these four variants (rs114791520, rs75681116, rs182317201 and rs72958256) are shown in Figure 3.4.

TABLE 3.3: *Stage 2 and meta-analysis results for the 14 sentinel variants that had $P<10^{-5}$ in stage 1. EAF = effect allele frequency, SE = standard error. The symbols in the direction of effects column show whether the effect size of that variant was positive or negative in the PROFILE cohort, the Trent Lung Fibrosis cohort and the UK Biobank cohort respectively. Variants with a consistent direction of effect across all three studies are highlighted.*

			Stage 2: Trent Lung (n=210)				**Stage 2: UK Biobank (n=98)**				**Meta-analysis of stages 1 & 2 (n=773*)**			
rsid	**Chr.**	**Reference / effect allele**	**EAF**	**Beta**	**SE**	**P-value**	**EAF**	**Beta**	**SE**	**P-value**	**Direction of effects**	**Beta**	**SE**	**P-value**
rs78672887	1	C / T	2.4%	0.058	0.283	0.838	1.1%	-0.708	0.673	0.293	- + -	-0.599	0.143	2.70×10^{-5}
rs12471179	2	A / C	31.1%	0.129	0.104	0.216	32.3%	0.006	0.154	0.692	- + +	-0.144	0.052	0.005
rs7599256†	2	G / T	51.4%	-0.073	0.089	0.414	58.2%	0.234	0.144	0.104	+ - +	0.172	0.048	3.22×10^{-4}
rs72958256	2	C / T	18.9%	-0.236	0.121	0.051	18.2%	0.001	0.171	0.996	- - +	-0.279	0.060	3.01×10^{-6}
rs72749864	5	T / A	4.3%	0.199	0.224	0.373	2.1%	-0.192	0.487	0.694	- + -	-0.396	0.127	0.002
rs114791520	6	T / C	1.2%	-0.006	0.435	0.988	1.0%	-0.126	0.682	0.853	- - -	-0.811	0.231	4.47×10^{-4}
rs76259754	7	C / T	1.2%	0.275	0.435	0.527	<1%	-	-	-	- +	-0.801	0.224	3.52×10^{-4}
rs75681116	8	C / T	2.6%	-0.240	0.297	0.420	4.1%	-0.453	0.350	0.196	- - -	-0.701	0.143	9.71×10^{-7}
rs12155839	8	C / T	16.5%	-0.205	0.128	0.110	19.1%	0.011	0.171	0.949	- - +	-0.277	0.063	1.25×10^{-5}
rs182317201	10	C / T	2.2%	-0.291	0.327	0.374	1.0%	-1.137	0.676	0.094	- - -	-0.833	0.185	6.30×10^{-6}
rs74715174	12	A / T	2.6%	-0.118	0.272	0.664	3.1%	0.065	0.400	0.871	- - +	-0.495	0.129	1.20×10^{-4}
rs113262525‡	12	T / G	6.9%	0.160	0.192	0.403	6.1%	0.114	0.293	0.696	- + +	-0.292	0.087	0.001
rs61459715	15	G / A	1.4%	0.979	0.398	0.014	1.0%	0.469	0.675	0.487	- + +	-0.491	0.199	0.014
rs117388035	19	G / A	4.8%	0.095	0.226	0.673	1.5%	-0.355	0.557	0.524	- + -	-0.508	0.143	3.64×10^{-4}

†: Conditional on the number of copies of the rs192643964 coded allele.
‡: Conditional on the number of copies of the rs74715174 coded allele.
*: n=675 for variant rs76259754 as this had a MAF of less than 1% in the UKB cohort, n=773 for all other variants.

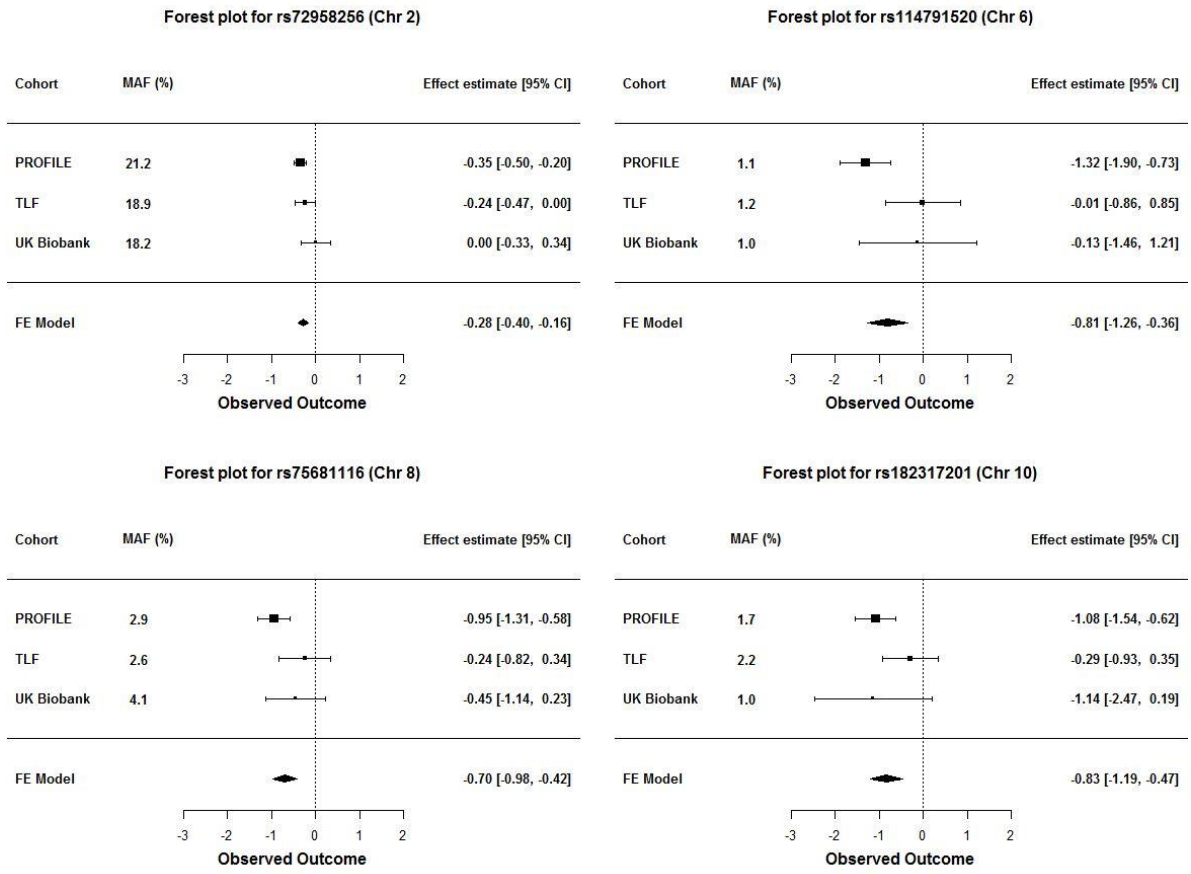


FIGURE 3.4: *Forest plots for the four genetic variants that showed consistent directions of effects across all three cohorts or maintained suggestive significance in the meta-analysis ($P_{meta}<10^{-5}$). These plots show the effect sizes and confidence intervals of each variant in each cohort and in the meta-analysis using a fixed-effects model.*

Regional association plots for the same four variants (Figure 3.5) showed that the common variant rs72958256 (MAF=21.2% in PROFILE) was in LD with several other variants within the same genetic region. The other three variants were less common with minor allele frequencies ranging from 1-5% and were each in LD with very few variants at the same genetic loci.

The results of the lookup for the 15 SNPs that were previously identified as being genome-wide significantly associated with IPF susceptibility are shown in Table 3.4. None of these variants were significantly associated with the age-at-diagnosis of IPF in the PROFILE cohort at a Bonferroni-corrected significance threshold (all P>0.003).

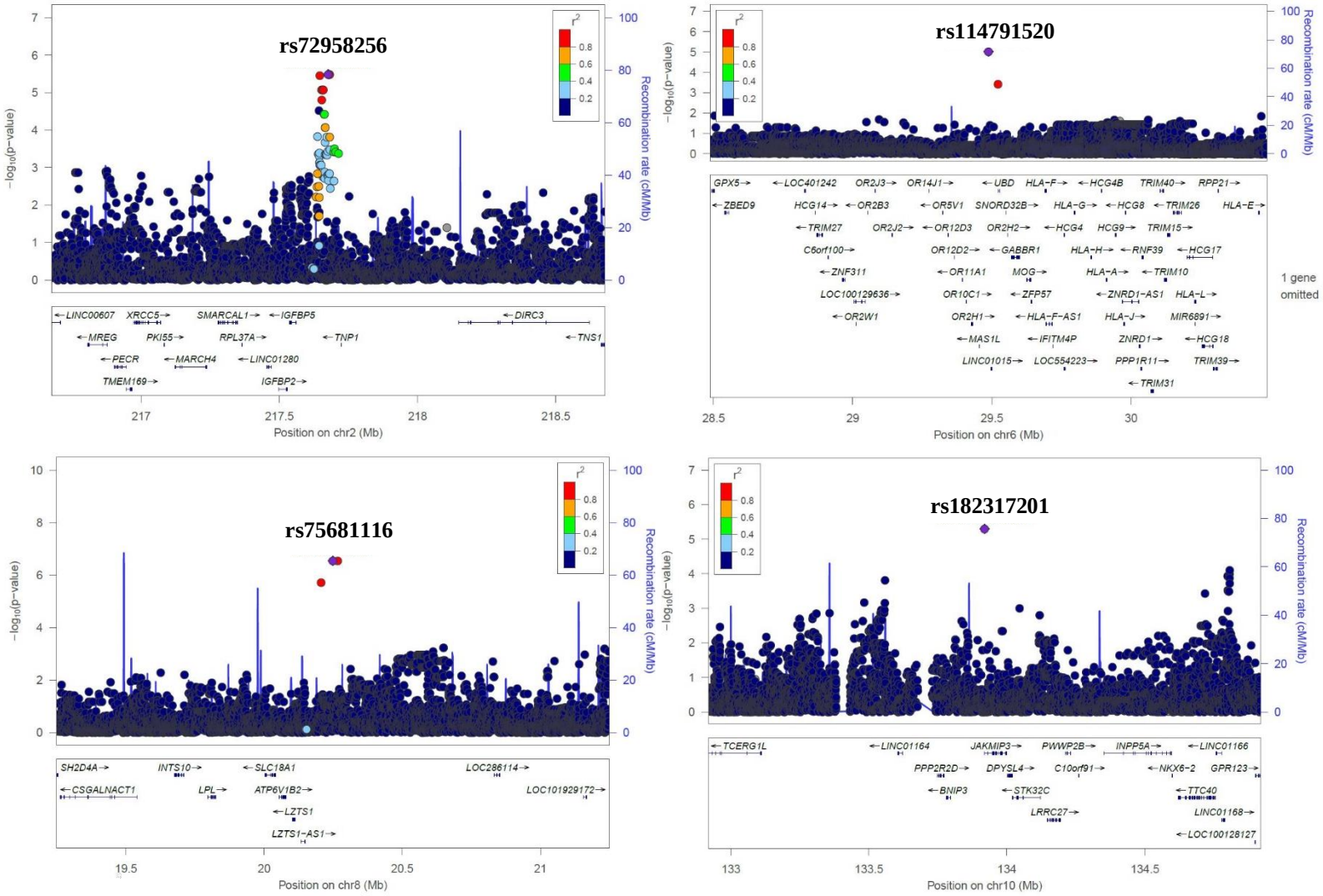


FIGURE 3.5: *Regional association plots for the four genetic variants that showed consistent directions of effects across all three cohorts or maintained suggestive significance in the meta-analysis ($P_{meta}<10^{-5}$). Plots were produced using stage 1 results.*

TABLE 3.4: *Summary statistics from the lookup for the 15 SNPs that have been previously identified as being genome-wide associated with IPF susceptibility. EAF = effect allele frequency, SE = standard error.*

Chr.	Position	EAF	rsid	Locus	Beta	SE	P-value
3	44902386	6.2%	rs78238620	*KIF15*	0.032	0.132	0.811
3	169481271	32.9%	rs12696304	*LRRC34/TERC*	-0.009	0.067	0.888
4	89885086	44.7%	rs2013701	*FAM13A*	0.052	0.065	0.427
5	1282414	28.1%	rs7725218	*TERT*	0.131	0.071	0.067
5	169015479	2.5%	rs116483731	*SPDL1*	0.090	0.197	0.647
6	7563232	56.6%	rs2076295	*DSP*	-0.055	0.061	0.366
7	1909479	56.6%	rs12699415	*MAD1L1*	0.130	0.063	0.040
7	99630342	43.4%	rs2897075	*7q22.1*	0.013	0.064	0.840
8	120934126	40.2%	rs28513081	*DEPTOR*	-0.043	0.064	0.495
11	1241221	34.7%	rs35705950	*MUC5B*	0.073	0.072	0.313
13	113534984	17.5%	rs9577395	*ATP11A*	0.092	0.083	0.270
15	40720542	57.6%	rs59424629	*IVD*	-0.045	0.064	0.478
15	86097216	31.2%	rs62023891	*AKAP13*	0.100	0.067	0.138
17	44214888	16.0%	rs2077551	*MAPT*	0.299	0.192	0.121
19	4717672	34.7%	rs12610495	*DPP9*	-0.043	0.063	0.499

3.3 Meta-analysis of three GWAS of age-at-diagnosis of IPF using time-to-event methods

Three significant issues were encountered during the 2-stage GWAS. Firstly, the analysis had limited power due to the relatively small sample size of the discovery cohort (n=465 IPF cases). Secondly, the findings of the study had limited credibility due to a lack of support in the stage 2 cohorts. Thirdly, the age-at-diagnosis of those in the UKB cohort was not normally distributed due to being truncated at approximately 75 years and so linear regression may not have been the most effective approach to model this data.

In this next analysis, the first issue was addressed through the implementation of a '3-way' GWAS meta-analysis study design, in which a separate GWAS was performed in each cohort of IPF subjects (PROFILE, TLF and UKB), with age-at-diagnosis of IPF as the phenotype of interest. Following this, the association summary statistics for each SNP were meta-analysed. The rationale for adopting this approach was that this would maximise the available statistical power for the discovery of genetic associations. To address the second issue, study-level thresholds were applied to each signal after the meta-analysis to exclude signals that were only present in one of the three studies, thus ensuring that any novel genetic signals were not being driven by a strong false positive association in only one cohort. Finally, in this subsequent analysis, time-to-event analysis methods were applied in place of linear regression. The method that was used, a Cox proportional-hazards (PH) model, does not assume a particular underlying distribution for the outcome of interest or that the residuals in the model are

normally distributed, and was therefore robust to the non-normality of the UKB cohort's age-at-diagnosis distribution.

3.3.1 Methods

This analysis was conducted using the same data from the same individuals as described in Section 3.2.1, i.e. the individuals with IPF from the PROFILE, TLF and UKB cohorts. Subject exclusion criteria were identical regarding ancestry, relatedness and outliers. The same proxies for the age-of-onset of IPF were used (age-at-enrolment as an approximation of age-at-diagnosis diagnosis in PROFILE and TLF and self-reported age-at-diagnosis in UKB). In addition, the same variant-level QC thresholds were applied as described in Section 3.2.2.

Time-to-event methods (Section 2.3) were utilised in this analysis to model the proxy for the age-of-onset of IPF. Genome-wide association testing was performed in each cohort separately to assess the association between the age-at-diagnosis of IPF and each available genetic locus. All genetic association testing was performed using a Cox PH model with the diagnosis of IPF as the event of interest. This was performed in R v4.0.0 using the 'survival' package. Again, genotype dosages were used and an additive model was assumed for the genetic effect of each variant. The same covariates as in the previous analysis were included in the model. The equation for the Cox PH model is shown in Equation 2.2.

$$h(t) = h_0(t) \times exp\,(\beta_1 Number\ of\ copies\ of\ risk\ allele + \alpha_1 Sex + \alpha_2 Smoking\ Status + \alpha_3 Recruitment\ Centre + \alpha_4 Array + \beta_2 PC1 + \cdots + \beta_{11} PC10) \quad (2.2)$$

For each covariate, a Chi-square test based on the correlation between time and the scaled Schoenfeld residuals for that variant was used to assess whether there was evidence that the PH assumption was being violated.

The genomic inflation factor (λ) and Q-Q plots were used to evaluate whether there was evidence of inflation within the results for each cohort. In the case where there was evidence of inflation ($\lambda>1.1$) for a particular cohort, genomic control (Section 2.2.2) was used to correct the results of that cohort. Q-Q plots were then used to assess the efficacy of the genomic control corrections.

All variants that were measured and passed QC in all three cohorts were then meta-analysed using a fixed-effect inverse variance model, as in the previous analysis. Variants that were at least nominally significantly associated with the age-at-diagnosis of IPF in each cohort ($P<0.05$, post-genomic control) and had a consistent direction of effects across cohorts were considered to meet our internal validation criteria. All variants that did not meet these criteria were excluded. Genome-wide significance was defined as $P_{meta}<5\times10^{-8}$ and suggestive significance was defined as $P_{meta}<5\times10^{-6}$. A sensitivity analysis

was performed to assess whether any of the suggestively significant signals would have reached genome-wide significance if genomic control had not been applied.

Finally, a lookup of previously reported IPF susceptibility signals was conducted within the results of the meta-analysis, as described in Section 3.2.2.

3.3.2 Results

In each cohort, more than 7.5 million variants passed QC and were tested for an association with the age-at-diagnosis of IPF. However, the results from all three cohorts showed evidence of genomic inflation (Figure 3.6) and the genomic inflation factor for each cohort was greater than 1.1 ($\lambda_{PROFILE}$ = 1.155, λ_{TLF} = 1.132 and λ_{UKB} = 1.194).

As no inflation was observed within the results when the linear regression model was applied to the PROFILE cohort in the previous analysis (Section 3.2.3) and the same covariates were adjusted for (including the first 10 genetic principal components), the inflation observed in this analysis was unlikely to be due the presence of unadjusted population structure within the data. Instead, it appeared more likely that the inflation was a result of the application of time-to-event analysis methods in place of the linear regression model.

The removal of all variants in which the genetic effect was found to break the PH assumption at a nominal level ($P<0.05$) did not significantly lower the genomic inflation factor of any of the cohorts (Table 3.5). Likewise, more stringent thresholds for imputation quality and MAF did not reduce the inflation to acceptable levels ($\lambda<1.1$) for any cohort. As no cause of the inflation was identified, genomic control was applied to the results of each cohort (Figure 3.7).

After genomic control (Figure 3.8), there were two genome-wide significant signals ($P<5\times10^{-8}$) in the PROFILE cohort and five in the TLF cohort. All seven sentinel variants for these signals were uncommon, with MAF<3%.

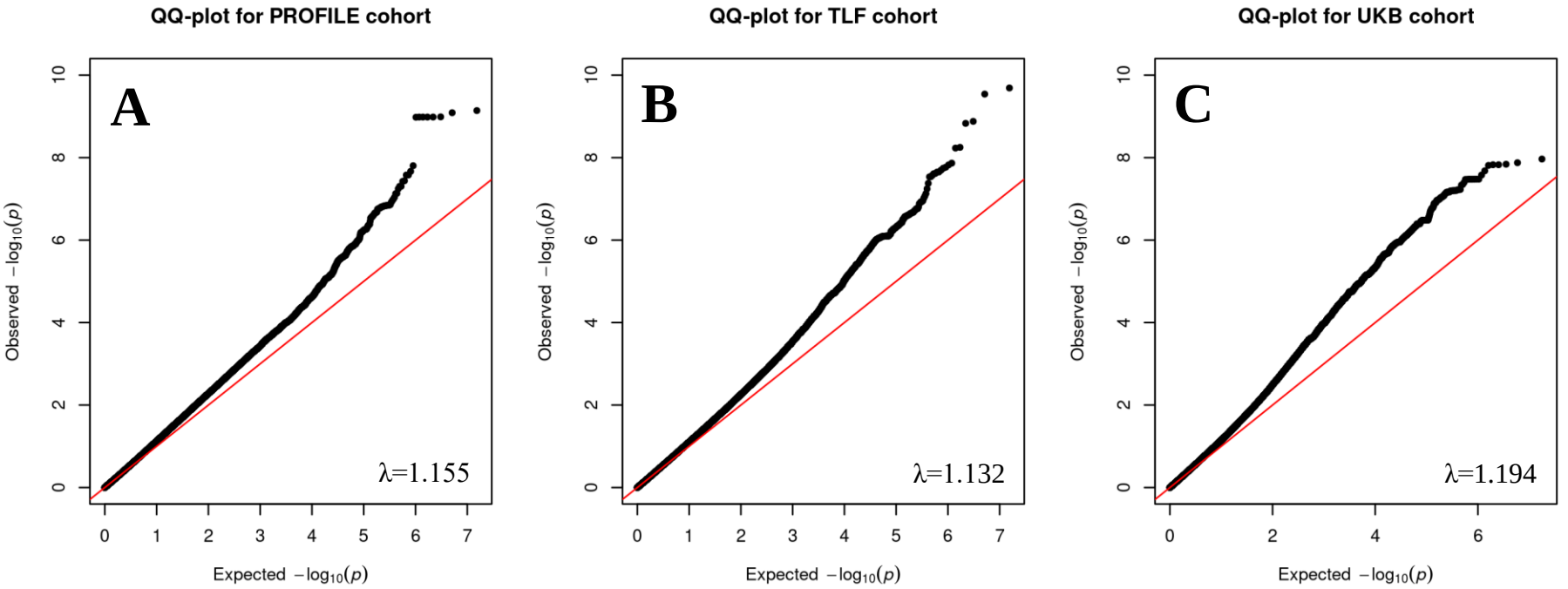


FIGURE 3.6: *Quantile-quantile plots showing the presence of genomic inflation in the results of the PROFILE cohort (A), the Trent Lung Fibrosis cohort (B) and the UK Biobank cohort (C). λ = genomic inflation factor. There were 7,645,226 variants tested in the PROFILE cohort, 7,714,782 variants tested in the Trent Lung Fibrosis cohort and 8,857,166 variants tested in the UK Biobank cohort.*

TABLE 3.5: *The results of the investigation into the cause of the inflation within the results of each genome-wide analysis. Various filters were applied to the results of each cohort to study the effect of excluding certain genetic variants. The remaining number of variants and corresponding genomic inflation factor is shown for each cohort. The proportional hazards p-value is from a Chi-square test based on the correlation between time and the scaled Schoenfeld residuals for that variant. R^2 refers to the imputation quality of the genetic variants.*

Criteria for inclusion	**Number of remaining variants (genomic inflation factor)**		
	PROFILE	**TLF**	**UKB**
None (all variants post-quality control)	7,645,226 (1.155)	7,714,782 (1.132)	8,857,165 (1.194)
Proportional-hazards p-value > 0.05	7,193,642 (1.155)	6,942,873 (1.130)	8,502,005 (1.194)
$R^2 > 0.9$	7,058,965 (1.156)	6,884,951 (1.126)	8,155,498 (1.189)
$R^2 > 0.99$	3,208,805 (1.155)	3,182,168 (1.122)	5,007,509 (1.181)
Minor allele frequency > 0.05	5,410,313 (1.159)	5,455,668 (1.115)	6,116,205 (1.169)
Minor allele frequency > 0.1	4,285,762 (1.168)	4,329,547 (1.110)	4,868,863 (1.164)

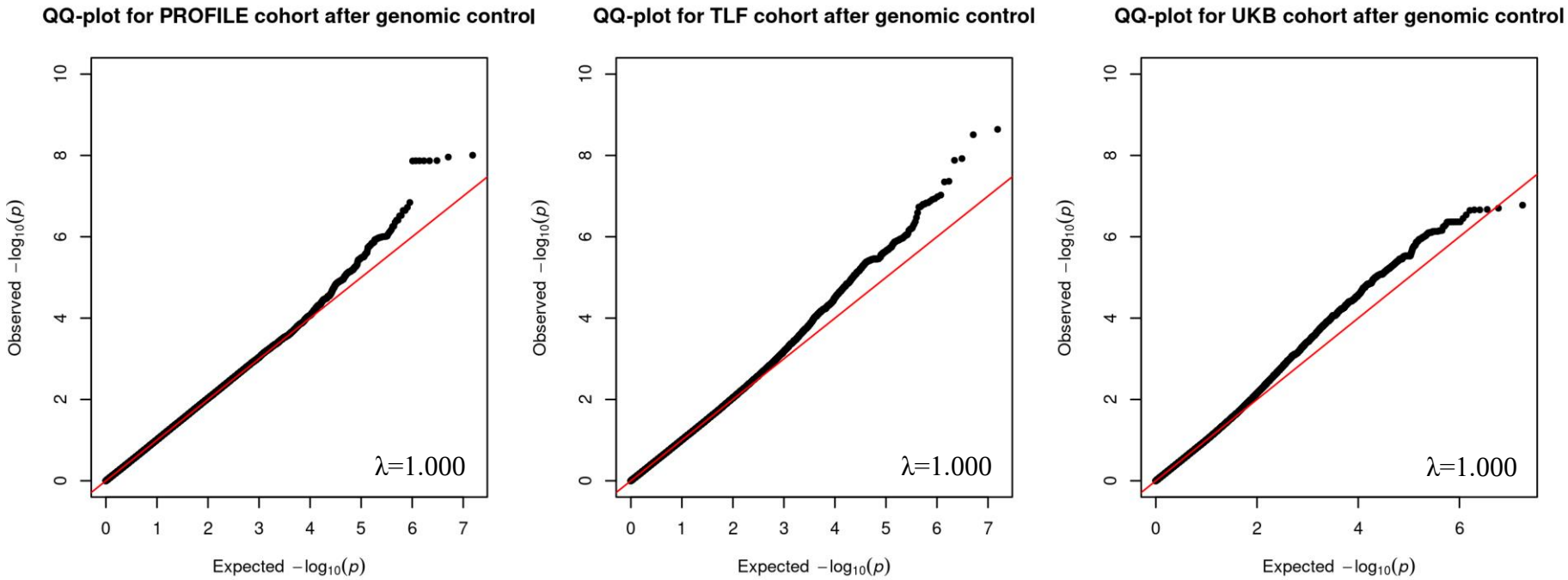


FIGURE 3.7: *Quantile-quantile plots displaying the p-values of the PROFILE cohort (A), the Trent Lung Fibrosis cohort (B) and the UK Biobank cohort (C) after genomic control. λ = genomic inflation factor.*

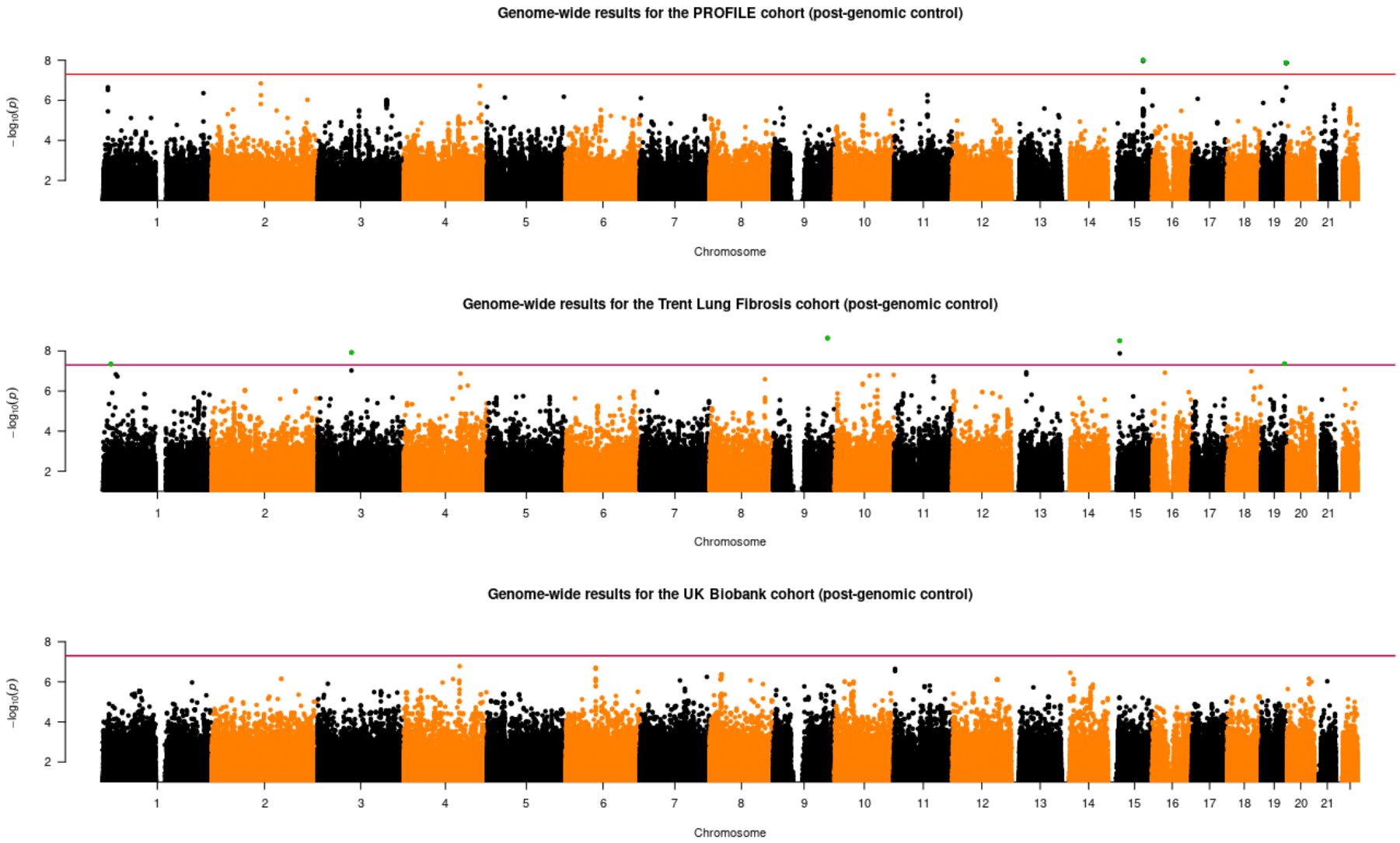


FIGURE 3.8: *Manhattan plots showing the association between each genetic variant and the age-at-diagnosis of IPF in each cohort. Sentinel SNPs for genetic signals that displayed genome-wide statistical significance ($P<5\times10^{-8}$, threshold indicated by the red line on each plot) are highlighted green. All variants with a p-value greater than 0.1 were removed to reduce the computational burden of the plots.*

7,179,468 variants were measured in all three study cohorts and passed QC in each cohort. These were meta-analysed (Additional Figure A.3.2). 248 variants had a consistent direction of effects across the three studies and showed nominal significance (P<0.05) in each cohort. All variants that did not meet these criteria were excluded.

None of the 248 variants reached genome-wide significance ($P_{meta}<5\times10^{-8}$) in the meta-analysis (Figure 3.9), though there were five genetic signals that reached suggestive significance ($P_{meta}<5\times10^{-6}$, Table 3.6). Of the five genetic signals that reached suggestive significance ($P_{meta}<5\times10^{-6}$), the variant with the strongest association with the age-of-onset of IPF in the meta-analysis was rs183759512 ($P_{meta}=1.05\times10^{-6}$). This variant is found on chromosome 7 within an intron of *DOCK4* and had a MAF of approximately 4% in all three study cohorts. The HR for this variant was 2.2, which means that at any follow-up time, individuals with one additional copy of the risk allele at this locus were estimated to be 2.2 times as likely to develop IPF compared to those with one fewer copy. Due to the additive model assumed, individuals who possess two copies of the risk allele at this locus were estimated to be 4.8 times as likely to develop IPF as those with no copies, at any follow-up time.

The other four sentinel variants included two uncommon SNPs that are found within introns of the genes *RBM17* (MAF=1.9%) and *RNF121* (MAF=1.2%), a common SNP found within an exon of the gene *FARP1* (MAF=28.5%) and one uncommon intergenic SNP (MAF=3.3%). The HRs for these variants ranged from 1.38-4.00. None of the five suggestively significant signals reached genome-wide significance in the sensitivity analysis (Additional Table B.3.3).

The results of the lookup for the 15 SNPs that were previously identified as being genome-wide significantly associated with IPF susceptibility are shown in Table 3.7. None of these variants were significantly associated with the age-at-diagnosis of IPF in the meta-analysis at a Bonferroni-corrected significance level (all P>0.003).

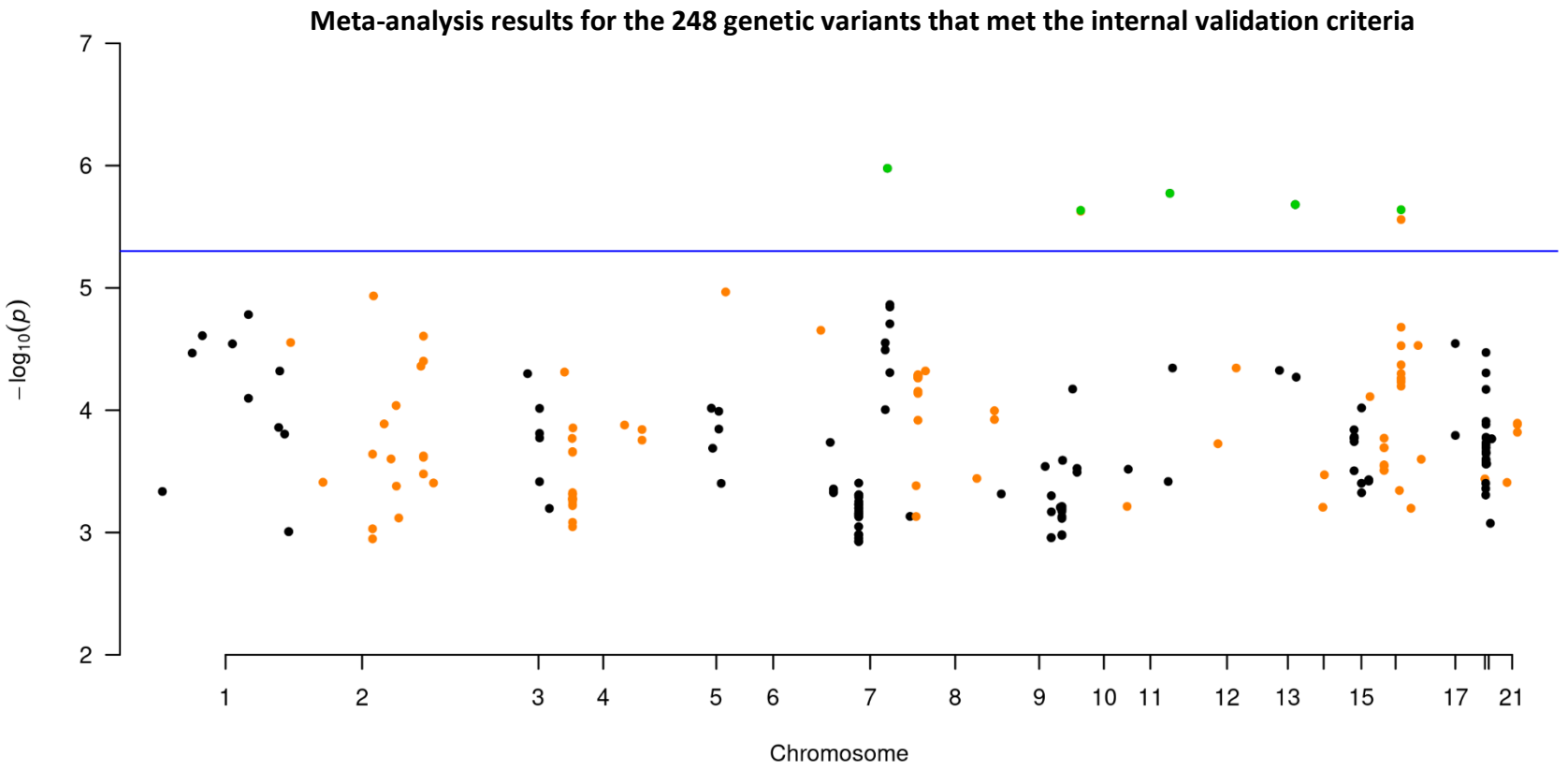


FIGURE 3.9: *A sparse Manhattan plot showing the statistical significance of the associations between the age-at-diagnosis of IPF and the 248 genetic variants that passed the internal validation criteria after the results for the PROFILE, TLF and UKB cohorts were meta-analysed. Sentinel SNPs for genetic signals that reached suggestive statistical significance ($P_{meta}<5\times10^{-6}$, threshold indicated by the blue line on the plot) are highlighted green.*

TABLE 3.6: *Summary statistics for the sentinel SNPs of the five independent genetic signals that were suggestively significant in the meta-analysis. EAF = effect allele frequency, HR = hazard ratio.*

					PROFILE (n=465)				**Trent Lung Fibrosis (n=210)**				**UK Biobank (n=98)**				**Meta-analysis (N=773)**		
rsid	**Chr.**	**Position**	**Gene**	**Ref/effect allele**	**EAF**	**HR**	**95% CI**	**P-value**	**EAF**	**HR**	**95% CI**	**P-value**	**EAF**	**HR**	**95% CI**	**P-value**	**HR**	**95% CI**	**P-value**
rs183759512	7	111708942	*DOCK4* (intron)	C / T	4.1%	2.11	(1.38, 3.22)	0.001	4.0%	2.09	(1.24, 3.90)	0.007	4.1%	2.64	(1.13, 6.18)	0.025	2.20	(1.60, 3.02)	1.05×10^{-6}
rs41295127	10	6134617	*RBM17* (intron)	A / T	1.6%	2.00	(1.13, 3.54)	0.016	2.7%	4.05	(2.09, 7.82)	9.3×10^{-5}	1.5%	4.42	(1.04, 18.8)	0.045	2.77	(1.81, 4.22)	2.32×10^{-6}
rs3915628	11	71682613	*RNF121* (intron)	C / T	1.3%	2.15	(1.08, 4.31)	0.030	1.1%	12.29	(6.09, 24.8)	4.5×10^{-5}	1.1%	19.69	(3.42, 113.1)	8.4×10^{-4}	4.00	(2.27, 7.07)	1.68×10^{-6}
rs9513422	13	99083935	*FARP1* (exon)	C / T	27.8%	1.27	(1.08, 1.49)	0.004	30.0%	1.65	(1.47, 1.87)	2.9×10^{-4}	28.1%	1.60	(1.03, 2.48)	0.035	1.38	(1.21, 1.58)	2.08×10^{-6}
rs118122250	16	54209057	Intergenic	G / A	3.4%	2.09	(1.40, 3.12)	3.1×10^{-4}	3.4%	2.08	(1.09, 3.97)	0.026	2.6%	3.84	(1.24, 11.9)	0.020	2.20	(1.58, 3.04)	2.30×10^{-6}

TABLE 3.7: *Summary statistics from the meta-analysis for the 15 SNPs that have been previously identified as being genome-wide associated with IPF susceptibility. EAF = effect allele frequency, SE = standard error.*

Chr.	Position	EAF	rsid	Locus	$Beta_{meta}$	SE_{meta}	P_{meta}
3	44902386	6.9%	rs78238620	*KIF15*	0.077	0.112	0.492
3	169481271	32.3%	rs12696304	*LRRC34/TERC*	-0.033	0.062	0.601
4	89885086	44.9%	rs2013701	*FAM13A*	-0.029	0.061	0.639
5	1282414	27.5%	rs7725218	*TERT*	-0.047	0.062	0.447
5	169015479	2.2%	rs116483731	*SPDL1*	-0.183	0.201	0.363
6	7563232	55.2%	rs2076295	*DSP*	0.054	0.054	0.321
7	1909479	53.9%	rs12699415	*MAD1L1*	-0.049	0.055	0.370
7	99630342	41.5%	rs2897075	*7q22.1*	-0.025	0.060	0.679
8	120934126	40.0%	rs28513081	*DEPTOR*	-0.072	0.059	0.217
11	1241221	33.4%	rs35705950	*MUC5B*	-0.109	0.065	0.093
13	113534984	17.7%	rs9577395	*ATP11A*	-0.040	0.077	0.603
15	40720542	58.2%	rs59424629	*IVD*	0.033	0.054	0.541
15	86097216	32.6%	rs62023891	*AKAP13*	-0.003	0.061	0.962
17	44214888	16.4%	rs2077551	*MAPT*	-0.094	0.117	0.424
19	4717672	35.2%	rs12610495	*DPP9*	0.003	0.060	0.965

3.4 Comparison of genome-wide results between linear regression and Cox proportional-hazards models in the PROFILE cohort

Two different approaches have been used in this thesis chapter to model the age-at-diagnosis of IPF: linear regression and Cox proportional-hazards. When using the Cox PH model, a time-to-event method, genomic inflation was observed within the results of the genome-wide analyses. As both types of model have been applied genome-wide to the PROFILE cohort, a direct comparison between the results from each approach could be informative as to whether this inflation was spurious and therefore whether it was necessary to correct the time-to-event results using genomic control.

There were 7,657,086 genetic variants that were tested for an association with the age-at-diagnosis of IPF in PROFILE using both methods. The beta coefficients from the linear regression model were strongly negatively correlated (Pearson's r = -0.837) with the log-hazard ratios from the Cox PH model (Figure 3.10). This negative correlation is due to the fact that an estimated increase in the age-at-diagnosis is denoted by a positive beta under the linear regression model but a negative hazard ratio under the linear regression model and vice versa for an estimated decrease in the age-of-onset. Despite the appearance of a slight skew on the plot, the line of best fit (blue) closely resembled the line $y = -x$ (red), indicating that the effect sizes from the Cox model were unbiased in comparison to the effect sizes from the linear regression.

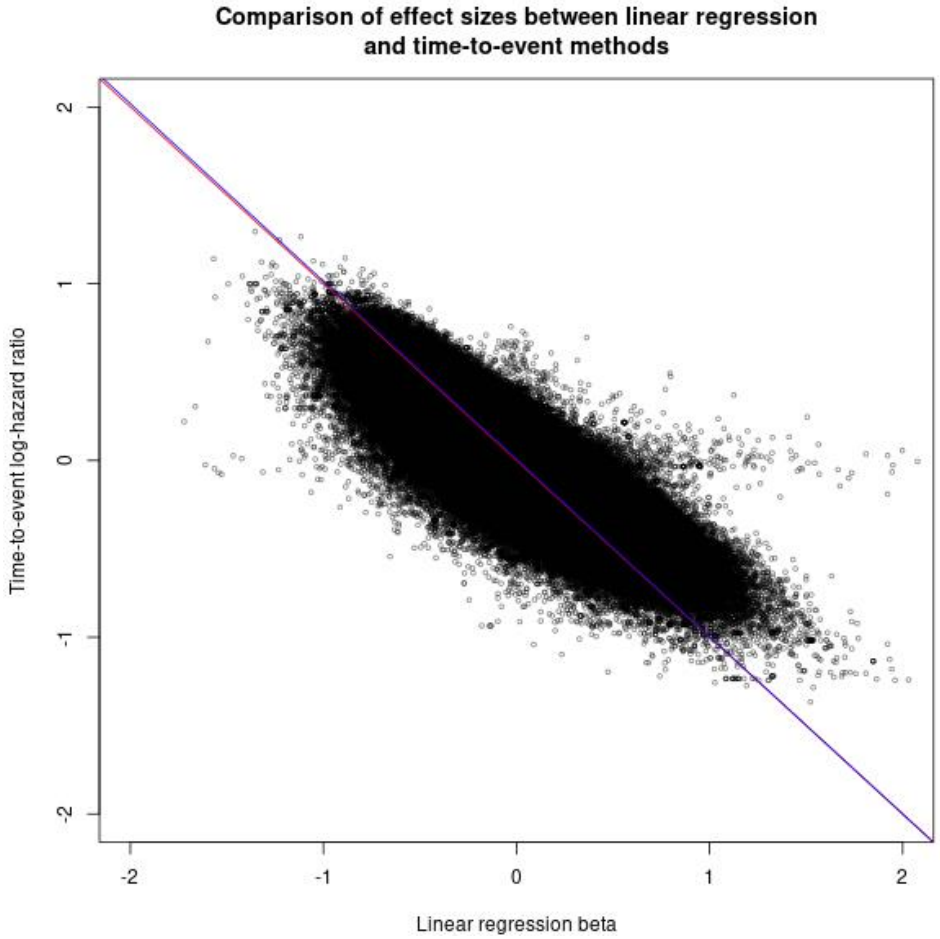


FIGURE 3.10: *A scatterplot comparing effect sizes between linear regression and time-to-event methods in the PROFILE cohort. There are two lines on the plot which are nearly completely overlapping; the red line corresponds to y=-x and the blue line indicates the linear line of best fit between the two sets of results.*

Figure 3.11A shows the comparison of p-values between the two methods, which were positively correlated (Pearson's r = 0.570). This plot indicated that the variants that had the most statistically significant p-values in the linear regression tended to have even stronger statistical significance using the time-to-event model. However, the line of best fit on the plot (blue) suggests that most genetic variants had a lower p-value under the linear regression model than the time-to-event model. The histograms in Figure 3.11B show that the p-values from the linear regression model appeared to be uniformly distributed (as expected), but the distribution of p-values under the Cox model was right-skewed, which resulted in an overabundance of p-values that were close to zero. This suggests that the Cox model overestimated the statistical significance for a considerable number of genetic variants and supports the basis for correcting the results of the time-to-event GWAS using genomic control.

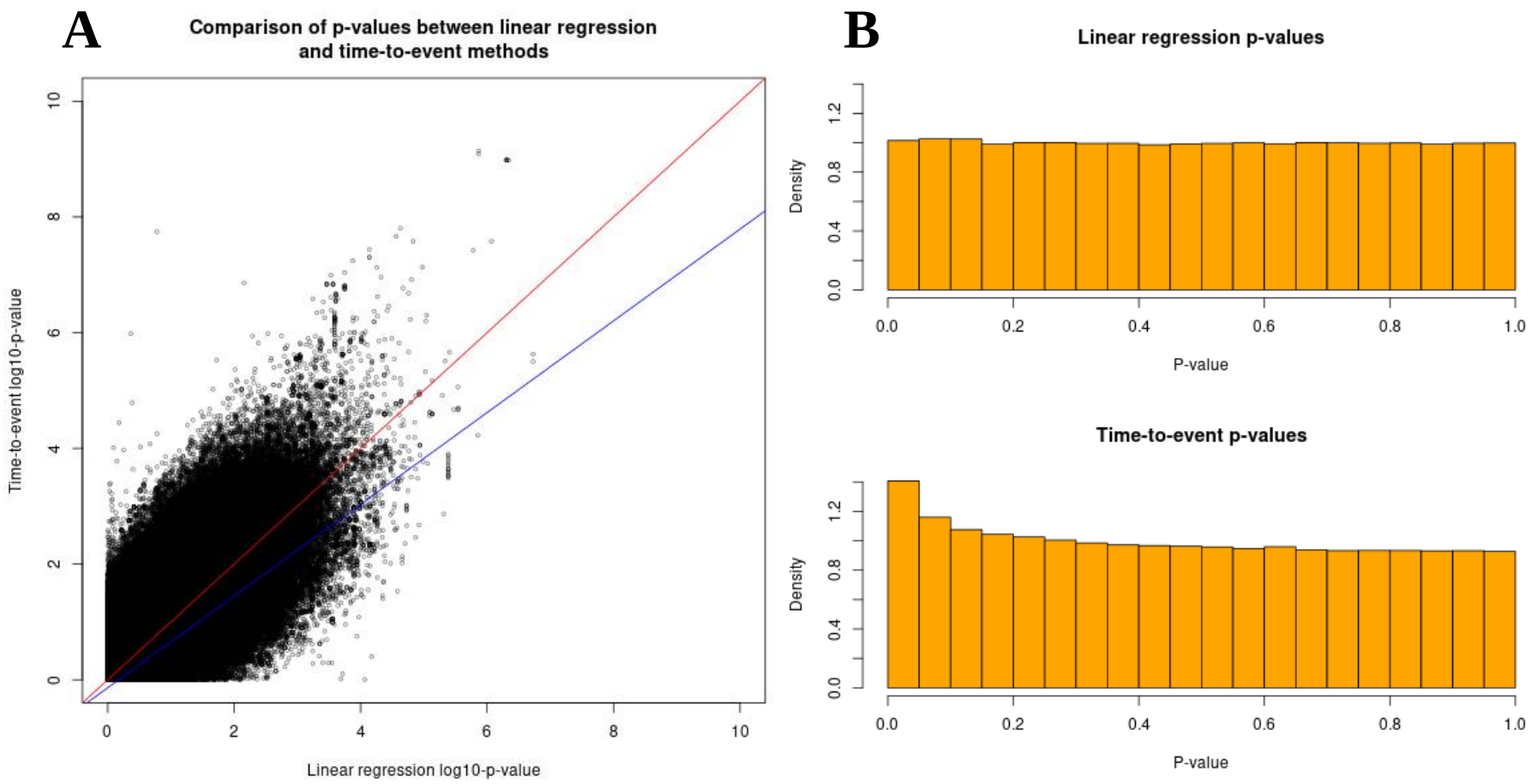


FIGURE 3.11: *A scatterplot (A) and histograms (B) comparing p-values between linear regression and time-to-event methods in the PROFILE cohort. In Figure A the red line corresponds to y=x and the blue line indicates the linear line of best fit between the two sets of results.*

3.5 Signal refinement and functional follow-up of suggestively significant signals of association

Five signals of genetic association were found to be suggestively significant in the meta-analysis of three GWAS for the age-at-diagnosis of IPF (Section 3.3). Functional follow-up of these signals could enable the development of testable hypotheses about how these variants may be exerting an effect on the age-of-onset of IPF as well as identifying potential drug targets. This section describes the signal refinement and functional follow-up that was performed for each suggestively significant signal.

3.5.1 Methods

Bayesian fine mapping was performed to generate a set of variants that was 95% likely to contain the causal variant for each signal, assuming that the causal variant had been analysed. The standard Bayesian approach is to assume that there is a single causal variant per genetic region and to calculate a 95% credible set of variants such that this set is at least 95% likely to contain the causal variant[124]. This approach involves computing a Bayes factor for each SNP, which is a measure of the strength of the association between that SNP and the trait of interest. The Bayes factor is defined as:

$$Bayes\ factor = \frac{P(Data \mid H_0)}{P(Data \mid H_1)} \tag{3.3}$$

Where H_0 is the null hypothesis (that the phenotype of interest is independent of genotype at this SNP) and H_1 is the alternative hypothesis (that the phenotype of interest is associated with this SNP). As such, a Bayes factor >1 suggests that there is more evidence for the null hypothesis and a Bayes factor <1 suggests that there is more evidence for the alternative hypothesis. The posterior odds of H_0 equals the Bayes factor multiplied by the prior odds of H_0.

However, the Bayes factor cannot always be calculated or it may be computationally intensive to do so. Therefore, some methods, such as that proposed by Wakefield[125], use an approximation of the Bayes factor. The approximate Bayes factor (ABF) used in the Wakefield method is calculated using the following formula:

$$ABF = \sqrt{\frac{V+W}{V}} \exp\left(-\frac{\hat{\theta}^2}{2}\frac{W}{V(V+W)}\right) \tag{3.4}$$

Where θ is a parameter of interest (e.g. the effect size for the genetic effect), which has variance V. It is assumed that $\theta \sim N(0, W)$ and so W is a measure of the strength of the genetic association, conditional upon the existence of one.

The approximate posterior probability of a variant being causal is calculated as the ABF for that variant divided by the sum of all ABFs in that signal, as the sum of the probabilities must equal one. A 95% credible set for each signal is then produced by taking the variants with the greatest approximate posterior probabilities until the sum of the approximate posterior probabilities exceeds 0.95.

Bayesian fine-mapping was performed for each of the five suggestively significant signals using the Wakefield method. This was performed using all variants within 1Mb of the sentinel SNP with $P_{meta}<0.001$, regardless of whether they had passed the internal validation criteria. A prior value of 0.04 was chosen for the Wakefield prior (W in Equation 3.4), which is equivalent to a 95% belief that the hazard ratio lays between $\frac{2}{3}$ and $\frac{3}{2}$[125]. Following this, functional annotation of each variant in the credible sets was performed using SnpEff v.5.0[126].

The Open Targets Genetics Portal[127] was used to investigate whether the SNP with the greatest posterior probability in each 95% credible set had been associated with any other disease trait in a previous phenome-wide association study (PheWAS). The Open Targets Genetics Portal recommends a Bonferroni-corrected threshold of approximately 1×10^{-5} for their PheWAS data and so any variant-trait association that reached this threshold was reported.

An expression quantitative trait locus (eQTL) is a genetic locus that explains a proportion of the variance of a gene expression phenotype (i.e. a genetic variant that is associated with the level of expression for a gene)[128]. Variants that act on nearby genes (conventionally defined as within 1Mb of the gene) are referred to as cis-eQTLs whereas those that act on genes farther away or on a different chromosome are referred to as trans-eQTLs.

Data from the eQTLGen consortium[129], a large resource containing genetic and blood-derived transcriptomic information from 31,684 individuals, were used to assess whether each variant in the 95% credible sets were significant cis-eQTLs in blood (defined as having a false discovery rate [FDR] <0.05). If so, a colocalisation analysis was performed to investigate whether the age-of-onset GWAS signal was likely to share a causal variant with the gene expression signal, which may suggest that the SNP is exerting an effect on the age-of-onset of IPF via expression of that gene.

The colocalisation analysis was performed using the method introduced by Giambartolomei et al.[130], in which a Bayesian test is applied to the summary statistics from the analyses of the two traits. This method jointly tests five different hypotheses and calculates the probability that each is true. These hypotheses are:

- H_0: There are no causal variants for either trait in the region of interest.
- H_1: There is a causal variant for the first trait in the region of interest but not the second trait.
- H_2: There is a causal variant for the second trait in the region of interest but not the first trait.
- H_3: There is a causal variant for the first trait in the region of interest and a causal variant for the second trait in the region, but these are different variants.
- H_4: There is a variant in the region of interest that is causal for both traits.

The signals were considered to have successfully colocalised if the probability of H_4 was calculated to be ≥ 80%. This test was implemented in R v.4.0.0 using the 'coloc' package. For each signal, the region

of interest was defined as the region ±1Mb around the gene under consideration. The age-at-diagnosis of IPF was regarded to be the first trait and the level of expression for the gene under consideration was regarded as the second trait. The summary statistics for these two traits were then jointly visualised in a mirror plot (also known as a Miami plot) using the 'mirrorplot' R package.

3.5.2 Results

Bayesian fine-mapping was used to generate a 95% credible set for each of the five suggestively significant signals (Additional Table B.3.4). In brief, there were 13 variants within the credible set for the signal on chromosome 7, 20 within the credible set for the chromosome 10 signal, 18 variants within the credible set for the chromosome 16 signal and the credible sets for the chromosome 11 and 13 signals each contained a single variant. Of the variants with the greatest posterior probability in each credible set, none had been significantly associated with a disease trait in a previous PheWAS at the Bonferroni-corrected threshold of $P=1\times10^{-5}$ according to the Open Targets Genetics Portal. Of these five variants, two were found to be significant cis-eQTLs (FDR <0.05): rs9513422 (chromosome 13) was a significant eQTL for *FARP1* and rs118122250 (chromosome 16) was an eQTL for *IRX3*. The five credible sets also contained eQTLs for *IFRD1, PRKCQ-AS1, RP11-554I8.1* and *RP11-5N23.3*.

The variant with the greatest posterior probability for the signal on chromosome 7 (rs183759512) lays within an intron of *DOCK4*. *DOCK4* encodes a guanine nucleotide exchange factor protein which is involved in regulation of adherens junctions between cells. Interestingly, adherens junctions have been implicated in the pathogenesis of IPF, with the epithelia of patients with IPF having increased expression of tight junction proteins compared with healthy controls[131,132,132]. rs183759512 flanks the promoter region for *DOCK4* and so this variant may affect gene expression via interactions at the promoter. However, it was not found to be a significant eQTL for *DOCK4* or any other genes (in blood) in this study.

The variant with the greatest posterior probability for the signal on chromosome 10 (rs41295127) is an intronic variant for *RBM17*. *RBM17* encodes an RNA binding protein which is part of the spliceosome complex and functions in the second catalytic step of mRNA splicing[133]. However, this variant was not a significant cis-eQTL for *RBM17* (or any other genes).

For the chromosome 11 signal, the sole variant in the credible set was rs3915628, an intronic variant for *RNF121* that is predicted to have a role in nonsense-mediated decay. *RNF121* encodes a ring finger protein, which are involved in protein-protein and protein-DNA interactions[134]. However, rs3915628 was not found to be a significant eQTL for *RNF121* (or any other genes).

The only variant in the credible set for the chromosome 13 signal was rs9513422, a protein coding variant in *FARP1* that is predicted to lie within a binding site for CTCF (a transcriptional repressor). The FARP1 protein plays a role in the formation of dendritic filopodia and dendritic spines, regulation

of dendrite length and ultimately the formation of synapses. rs9513422 was found to be a significant cis-eQTL for *FARP1* in blood, which could implicate *FARP1* (and thus the development of dendritic cells) in the age-of-onset of IPF. This would be an interesting finding as circulating dendritic cells are markedly depleted in IPF patients at the time of diagnosis[135].

Finally, for the chromosome 16 signal, the variant with the greatest posterior probability was rs118122250. Despite being an intergenic variant, rs118122250 was found to be a significant cis-eQTL for *IRX3* (which plays a role in an early step of neural development[136]) and so may have a role in gene regulation. Although, this gene does not appear to have immediate biological relevance to IPF pathology.

The results of the colocalisation analyses were not suggestive of a shared causal mechanism between any of the age-of-onset signals and the level of expression of the aforementioned genes (Table 3.8). The probability that a single causal variant for both signals was within the region of interest ranged from 1.9-12.6%. Instead, the Bayesian model predicted that it was much more likely that in each region of interest there was a causal variant for the eQTL signal but not the age-at-diagnosis of IPF, or that there was a causal variant for the age-at-diagnosis of IPF and a causal variant for the eQTL signal but that these were different variants. The lack of successful colocalisation was supported by the mirror plots (Additional Figures A.3.3-8), which showed that the signals from the age-at-diagnosis of IPF GWAS did not fully overlap with the signals from the eQTLGen cis-eQTL analyses.

TABLE 3.8: *The results from the colocalisation analyses between the age-of-onset of IPF 3-way GWAS signals and each gene for which a variant in the 95% credible set for that signal was an eQTL. H_0 = in the region of interest there are no causal variants for either trait, H_1 = in the region there is a causal variant for the age-at-diagnosis of IPF but not a causal variant for the eQTL signal, H_2 = in the region there is a causal variant for the eQTL signal but not the age-at-diagnosis of IPF, H_3 = in the region there is a causal variant for the age-at-diagnosis of IPF and a causal variant for the eQTL signal but these are different variants, H_4 = in the region there is a variant that is causal for both traits. As all H_4 < 80%, none of the signals were considered to have successfully colocalised.*

GWAS Signal	Gene	Probability of hypothesis being true (%)				
		H_0	H_1	H_2	H_3	H_4
Chr. 7	*IFRD1*	0.0%	0.0%	57.6%	39.2%	3.2%
Chr. 10	*PRKCQ-AS1*	0.0%	0.0%	58.1%	37.8%	4.1%
Chr. 10	*RP11-554I8.1*	0.0%	0.0%	54.5%	35.1%	10.4%
Chr. 10	*RP11-5N23.3*	0.0%	0.0%	53.0%	34.4%	12.6%
Chr. 13	*FARP1*	0.0%	0.0%	46.2%	51.9%	1.9%
Chr. 16	*IRX3*	0.0%	0.0%	56.2%	35.2%	8.6%

3.6 Discussion

The two analyses described in this chapter were conducted with the objective of identifying genetic determinants of the age-of-onset of IPF. As these were the first GWAS in IPF research to investigate

the age-of-onset phenotype, they have provided some important insight into the genetic aetiology of the disease. Firstly, the findings in this chapter suggest that the genetic variants associated with the age-of-onset of IPF appear to be different from those reported to confer susceptibility to the disease. Secondly, five signals of association were suggestively significant ($P<5\times10^{-6}$) in the 3-way GWAS meta-analysis, despite the low sample size (n=773 IPF cases total) and the use of genomic control, and could be true positive signals for which there was insufficient power to achieve genome-wide significance.

However, there were no genome-wide significant genetic associations with the age-at-diagnosis of IPF identified in these studies. One possible reason for this could be due to the relatively small sample size (773 IPF cases in total across the three cohorts) which meant that these GWAS were greatly underpowered. In a sample of 773 unrelated individuals, the additive genetic effects from a single variant would need to explain 5.1% of the phenotypic variance for the study to have an 80% probability of detecting that variant at a genome-wide significance level ($P<5\times10^{-8}$). 5.1% of the phenotypic variance is greater than an individual variant typically explains, although the existence of a single SNP that exerts a much larger effect on a trait than other variants is not unheard of in IPF, as the *MUC5B* promoter variant rs35705950 (the single largest genetic risk factor for IPF susceptibility) explains 5.9-9.4% of disease liability in the general population whilst all 13 non-*MUC5B* SNPs that have been discovered through GWAS collectively explain 1.8–2.9%[120]. Therefore, if the age-of-onset of IPF had a similar genetic aetiology to IPF risk, these studies may have had sufficient power to detect genome-wide significant associations.

Additionally, the two-stage study presented in Section 3.2 had some important limitations. Firstly, the choice of study design meant that the discovery stage did not utilise all available data and so this part of the analysis was greatly underpowered, with a sample size of 465 individuals. Secondly, studies with low sample size (i.e. <3,000 individuals) that apply linear regression to non-normally distributed data are subject to a loss of accuracy and an increase in the uncertainty of estimates[121]. This means that the non-normality shown in the distribution of the age-at-diagnosis for the subjects in UK biobank could have restricted the potential for the linear regression model in the two-stage GWAS (Section 3.2) to detect a genuine genetic effect in those individuals. Thirdly, the statistical significance of the associations for the 14 variants in the meta-analysis were clearly being driven by the strong associations found in stage 1. For example, the variant with the strongest association with the age-at-diagnosis of IPF in the meta-analysis was rs75681116 with a p-value of 9.7×10^{-7}. The effect size of this variant was -0.95 in the PROFILE cohort yet was considerably smaller in both stage 2 cohorts (β_{TLF}=-0.24 and β_{UKB}=-0.45). Therefore, it is possible that the novel suggestive associations between the age-of-onset of IPF and each of the variants reported in this study could be false positive results. Although, the effect sizes for these variants being lower in the replication cohorts than reported in the discovery cohort (PROFILE) could be partly explained by the 'winner's curse' phenomenon.

A second GWAS was performed to address these limitations (Section 3.3). Firstly, the adoption of a 3-way GWAS meta-analysis study design allowed for more individuals to be tested genome-wide, increasing the statistical power of the study for discovery. However, as noted previously, even with all three cohorts combined (N=773) the study was underpowered to detect variants at a genome-wide significance level unless the additive effects of a variant could explain a large proportion of the phenotypic variance (approximately 5%).

Secondly, the Cox PH model used in this study did not assume an underlying distribution within the data and so was less likely to be affected by the non-normality within the distribution for the age-at-diagnosis in the UKB cohort than linear regression. Lastly, this study had the important strength that there were internal validation criteria which reduced the likelihood that a novel signal was being driven purely by a strong association within one cohort, thereby reducing the likelihood of false positives. Through these improvements to methodology and study design, this study was able to identify five genetic variants that were suggestively significantly associated with the age-at-diagnosis of IPF ($P_{meta}<5\times10^{-6}$). Still, none reached genome-wide significance in the meta-analysis ($P_{meta}<5\times10^{-8}$) and therefore the inclusion of additional independent data is needed.

An unexpected issue that arose during the 3-stage GWAS was the unusually high genomic inflation factor (greater than 1.1 for all three cohorts). It was thought to be unlikely that the inflation was a result of unadjusted population structure within the data, as the 2-stage analysis (Section 3.2) also tested the PROFILE cohort genome wide, adjusted for the same covariates, yet had a genomic inflation factor within the acceptable range. Therefore, it was suspected that the inflation was an artefact of the time-to-event methods being applied genome-wide.

However, the exact cause of the inflation remained unknown as the removal of variants that broke the PH assumption did not reduce the genomic inflation factor to acceptable levels, nor did the removal of poorly imputed or uncommon variants. As a result, the findings from each genome-wide analysis required adjustment prior to the meta-analysis and internal validation. The method used to correct the results from each cohort was genomic control, which can often be over-conservative[122] and therefore could have prevented some variants that are genuinely associated with the age-at-diagnosis of IPF from meeting the internal validation criteria.

The cause of the inflation could be investigated further through a simulation study. In fact, a recent simulation study[123] compared the performance of the Cox PH model to a logistic regression model when both are used in a genome-wide analysis, finding that the Cox model had greater power with approximately the same level of type I error. As such, the genomic inflation factors for the Cox model tended to be greater than those from the logistic regression model. It stands to reason that this would also be true when the Cox model is used in place of linear regression, as in this study. If this were the case, this may suggest that the application of genomic control in this study was an overcorrection.

However, as evidenced in the sensitivity analysis, there were no genetic signals that would have reached genome-wide significance had genomic control not been applied.

The exact age-of-onset of a disease such as IPF is difficult to determine precisely as the onset of the disease occurs at an unrecorded time prior to the diagnosis and takes place even before the development of symptoms. Chest computed tomography (CT) scans can be a useful tool for the detection of early IPF as individuals with interstitial lung abnormalities (ILA) often go on to develop IPF[137]. However, most ILA cases do not develop to IPF and the prevalence of ILA has been estimated to be 50-200 times greater than the prevalence of IPF[138].

As such, the exact age-of-onset of IPF was not known for any of the individuals that were included in these analyses and proxies for the age-of-onset were relied upon instead. These proxies, the self-reported age-at-diagnosis for the subjects in the UK Biobank cohort and the age-at-enrolment for the individuals in the PROFILE and Trent Lung Function cohorts, were deemed to be appropriate substitutes for the age-of-onset but the use of proxies in place of the actual phenotype would have been detrimental to these analyses in a few important ways. Firstly, the time between the onset of the disease and the diagnosis/enrolment into a study could vary greatly from person to person and could be subject to some important biasing factors such as sex[139] and smoking status[140]. Whilst sex and smoking status were adjusted for in these analyses, other unmeasured variables could have biased the results. For example, a considerable difference in the average age-at-diagnosis was observed between individuals in the PROFILE study who were enrolled from two different recruitment centres, one in London and the other in Nottingham. This difference could be the result of unmeasured factors, such as differences in the levels of pollution between the two cities or the socioeconomic status of the individuals from the two areas, which has been found to affect waiting times for hospital procedures in England[141], as well as being linked to the levels of occupational exposure to IPF risk factors of the individuals from each area[142]. However, differences in referral processes between the two regions likely played a large part in the disparity in the average age-at-enrolment between recruitment centres.

Secondly, for individuals in the UK Biobank cohort, the age-at-diagnosis of IPF was self-reported and as a result these values could have been impacted by recall bias. Third, the proxies for the age-of-onset in all cohorts were likely affected by survival bias, as only those who survived long enough to enrol into each study were available for inclusion in the analysis. Fourth, some individuals in the PROFILE and TLF studies could have enrolled into their study soon after their diagnosis, whilst others could have been recruited six months after their diagnosis. This could have affected the accuracy of the age-at-enrolment proxy and impacted the statistical power of these parts of the analysis, as well as significantly increasing the survival bias in these cohorts due to the short median survival time of IPF. Fifth, GWAS that use events that are subsequent to the development of disease as the phenotype of interest are vulnerable to index event bias[143], where the associations detected in the study can be biased by factors

relating to the selection of subjects into the study. However, as the genetic determinants for the age-at-diagnosis of IPF appear to be different to those for IPF susceptibility and the sentinel variants for the suggestive signals had not been strongly associated with any traits in a previous PheWAS, it appears unlikely that these signals are strongly associated with factors related to study selection, and thus are unlikely to be the result of index event bias.

Two statistical approaches were used in this chapter to model the proxies for the age-of-onset of IPF: linear regression in the first study and a Cox PH model in the second. Linear regression had the advantage that it was simple to implement with widely available and well supported software, as well as producing results that were arguably easier to interpret under the additive genetic model that was assumed. On the other hand, the Cox PH model was robust to the non-normality observed in the distribution for the age-at-diagnosis of the UK Biobank cohort and appeared to have greater statistical power, as some variants in the PROFILE cohort reached genome-wide significance in the second study but did not in the first. However, it appeared that the Cox model had overestimated the statistical significance for many of the most significant variants, which meant that the results required adjustment using genomic control.

As discussed previously, there was a considerable difference in the average age-at-diagnosis between individuals in the PROFILE study who were enrolled from the two different recruitment centres. This was adjusted for in this study by including recruitment centre as an interaction term in the statistical model. However, it may have been more appropriate to use a linear mixed model with recruitment centre included as a random effect, as this can prevent false-positive associations that may arise due to population structure and can increase statistical power by applying a correction that is specific to the sample structure[144].

Due to the meta-analysis approach that was taken in both studies, additional, independent cohorts of IPF cases could be easily added to these analyses to increase statistical power. For this reason, it may be of interest for future work to focus on updating these studies if data for additional IPF cohorts can be obtained. This could potentially boost the power of these analyses to the point that novel variants with genuine associations with the age-of-onset of IPF could be identified at a genome-wide significance level. However, this would likely require thousands of additional IPF cases. The next age-of-onset GWAS should aim to have a sample size of at least 4,000 unrelated individuals as this would provide 80% power for the detection of a genetic effect that accounts for 1% of the phenotypic variance.

In these analyses, a single variable from the UK Biobank was used to identify individuals who had been previously diagnosed with IPF by a doctor (data field 22135), which originated from a questionnaire. However, this questionnaire was completed by less than a quarter of the total number of UKB participants and so there are likely many additional individuals with IPF who have contributed their genetic data to the UKB project but were not included in these analyses. Many of these individuals

could be identified by using Hospital Episode Statistics (HES) data and then could be incorporated into future analyses. However, for a patient with IPF to be in the HES database they must have been admitted to hospital, but the first instance of this may have occurred a long time after their initial IPF diagnosis and therefore their age at this time may not be a good proxy for their age-at-diagnosis. A further difficulty with this approach is that UK Biobank do not use a HES code that specifies only cases of IPF. Therefore, a more general code (such as J84.9: unspecified interstitial pulmonary disease) would need to be used instead, resulting in the need to disentangle the IPF cases from the non-IPF interstitial lung disease cases, which may prove challenging.

Alternatively, primary care data could potentially be used to determine an IPF subject's age when they first visited the doctor after developing symptoms for IPF. This information could provide an even closer estimate for the age-of-onset of IPF than the age-at-diagnosis or age-at-enrolment, and would arguably make for the best possible proxy for the age-of-onset. Plus, UK Biobank have made primary care data available for around half of their participants. However, this is dependent on the use of primary care symptom codes and deciding which symptom codes to use and how best to use these to define the onset of IPF would require careful consideration. A recent study of 462 individuals with IPF[145] found that the most common primary care symptom patterns observed within a year of the diagnosis of IPF were dyspnoea (in 48.7% of patients), cough (40.9%) and cough with dyspnoea (23.4%). 50% of the IPF cases in this study were diagnosed within 5 years from their first recorded cough and within 3 years from their first recorded dyspnoea. However, 31% of the patients were not recorded as having any symptoms in the 1 year prior to the IPF diagnosis, and 15% did not have any symptom codes at all prior to their IPF diagnosis. Additionally, this study did not consider other conditions or co-morbidities which may have been causing the recorded symptoms. Therefore, using primary care data to estimate the age-of-onset of IPF could improve the accuracy of the estimate for some individuals, but would not be possible for everybody.

In addition, future larger studies may wish to estimate the level of heritability of the age-of-onset of IPF and compare this to the level of heritability for IPF susceptibility. Furthermore, it would be of interest to assess whether the heritability of the age-of-onset of IPF is greater in individuals with younger disease onset, which would be consistent with the finding of Krauss et al.[110] that on average, individuals with FPF were developing the disease younger than sporadic IPF cases. If this is the case, therapeutics that are developed to target mechanisms involved in the pathogenesis of the age-of-onset of IPF may prove to be more effective in individuals who developed the disease at an early age. This could therefore help inform the design of clinical trials for such therapeutics.

To conclude, the first GWAS to investigate the age-of-onset of IPF have highlighted important factors to consider when analysing this phenotype, such as the use of time-to-event analysis methods and the choice of suitable proxies for the age-of-onset. These studies suggest that there could be a genetic basis

to the age-of-onset as five genetic signals of suggestive association were discovered in the meta-analysis of the second study, but larger studies must be conducted before these can be confirmed at a genome-wide significance level.

Chapter 4 – Rare variant analyses to identify genes associated with the age-of-onset of IPF

Rare genetic variants (MAF<1%) are known to play an important role in the development of IPF and it stands to reason that rare variants may similarly influence the age at which IPF is developed. However, statistical power is low when studying individual rare variants. As such, rare variants within the same gene are often grouped at the gene level to increase statistical power. The aim of the analyses in this chapter was to use this approach to identify genes in which an excess of rare genetic variants was associated with the age-of-onset of IPF.

4.1 Introduction

Chapter 3 of this thesis described two GWAS of the age-of-onset of IPF. The GWAS approach, which utilizes imputed genotype data, is useful for detecting common (MAF >5%) and low-frequency (MAF 1-5%) genetic variants associated with a phenotype of interest. However, the loci detected in GWAS rarely explain more than a small fraction of the genetic variance of the trait of interest, which has led to speculation of a 'missing heritability problem'[146,147,147]. One possible explanation for this missing heritability is that much of the unexplained genetic variance could be down to rare variants (MAF<1%) that exert large effect sizes on the trait[148], but which are usually not well-covered in GWAS as they are more prone to genotyping and imputation errors[99]. However, rare variants have been found to play an important role in the genetic architecture of complex diseases[149,150,150], including type 1 diabetes[151] and heart failure[152].

Due to recent advances in technology and falling costs, it is becoming increasingly feasible to use whole-genome sequencing (WGS), which has better measurement accuracy than genotyping[153] and can therefore allow for rare variants to be tested. However, tests for single rare variants have low statistical power unless the sample size or effect size is very large[154]. Additionally, the number of independent rare variants is far greater than the number of independent common variants and so if all single rare variants are tested genome-wide, a more stringent multiple-testing correction may be needed than the typical threshold for genome-wide significance used in a GWAS ($P<5\times10^{-8}$), which would further reduce statistical power[154]. As such, methods have been proposed that involve grouping rare variants together in order to raise statistical power. Most commonly, rare variants are grouped together ('collapsed') at the gene-level to create a burden variable that can be regressed against the phenotype of interest to test for the cumulative effects of rare variants within that gene.

There are several statistical approaches that can be used to collapse and test the aggregation of rare variants within a gene or other genomic region of interest[155]. Early approaches, labelled burden tests, assumed that all rare variants within the region of interest are causal with the same direction of effect. Burden tests usually require a frequency threshold to be selected in order to define rare variants (e.g.

rare variants could be defined as those with MAF lower than 1% or 5%). Alternatively, there are methods known as non-burden tests that do not assume that all of the rare variants in the region of interest are causal or that they all have the same direction of effect. Non-burden tests do not require the selection of thresholds and can be applied to both rare and common genetic variants, but with greater weights given to rare variants if these are expected to have larger effect sizes than common variants or are more likely to be causal.

As discussed in Section 1.2.5, rare variants are known to be important in the development of IPF. One sequencing study of 3,624 IPF cases and 4,442 control subjects[156] found that the *FAM13A* and *TERT* gene regions, which had been previously identified in GWAS of IPF susceptibility, actually contain a combination of independent common and rare genetic signals that each contribute to IPF susceptibility. More recently, an exome-wide association study of 752 sporadic IPF cases and 119,055 UK Biobank controls[80] identified a novel IPF susceptibility signal in the form of a single rare missense variant in the *SPDL1* region, for which each copy of the minor allele was estimated to increase the odds of developing IPF by 2.9 times. Additionally, a gene-based collapsing analysis was successful in identifying three genes (*TERT*, *RTEL1* and *PARN*) in which an excess of rare variants were contributing to the pathogenesis of sporadic (non-familial) IPF[157].

To date, there have been no rare variant studies in IPF that have investigated the age-of-onset phenotype. In Chapter 3 of this thesis, many of the sentinel SNPs from the suggestively significant signals were low-frequency and appeared to exert large effect sizes on the age-at-onset of IPF. It stands to reason that there may be rare variants with similarly large effect sizes (or perhaps greater) that are associated with the age-of-onset of IPF, and that pooling these variants into genes could provide sufficient statistical power to detect them. As such, this chapter describes the first gene-based collapsing analyses in IPF to study the age-of-onset phenotype. The objective of these studies was to identify genes in which an aggregated excess of rare genetic variants was associated with the age-of-onset of IPF at a study-wide significance level. Two different statistical methods were used to collapse genetic variants at the gene level and test the statistical significance of the collapsed variable. A burden test was used in the first study (Section 4.2) and a non-burden test was utilised in the second study (Section 4.3).

4.2 Gene-based collapsing analysis using a burden test

4.2.1 Methods

As in Chapter 23, the age-at-diagnosis of IPF was considered a suitable proxy for the age-of-onset. This study was performed in individuals with IPF from the PROFILE study who were enrolled into PROFILE within six months of their IPF diagnosis (Section 3.2.1). Therefore, age-at-enrolment was considered a suitable proxy for the age-at-diagnosis for these individuals.

WGS, alignment (to human genome assembly GRCh38) and variant calling for the PROFILE cohort was performed externally, as previously described by Dhindsa et al.[80]. In brief, the DNA underwent

paired-end 150bp WGS and the average coverage was a 42-fold read depth. More than 98% of the reference bases had at least 10x coverage.

Many quality metrics that are used to filter out spurious variant calls utilise Phred quality scores, which are measures that assess the quality of the identification of the nucleotide bases. A Phred quality score (Q) is linked to the probability of an erroneous call (P) as follows:

$$Q = -10 \log_{10} P \tag{4.1}$$

So, for example, if there was a 10% probability of an erroneous call, the Phred-scaled quality score would equal 10, and if there was a 1% probability that the call was erroneous, the Phred-scaled quality score would equal 20.

The following metrics were used to quality control the data:

- QUAL: The Phred-scaled probability that a polymorphism exists at a particular site.
- DP: The combined read depth across all samples.
- GQ: The Phred-scaled probability that the genotype assignment for a sample is correct. Specifically, GQ is the difference between the Phred-scaled likelihood of the most likely genotype and the Phred-scaled likelihood of the second most likely genotype.
- FS: Phred-scaled p-value for strand bias, estimated using Fisher's exact test.
- MQ: Root mean square of the mapping quality of reads supporting the variant call, across all samples.
- MQRankSum: An approximation of the Z-score from a rank-sum test comparing the mapping qualities of the reads supporting the reference allele and the reads supporting the alternate allele.
- ReadPosRankSum: An approximation of the Z-score from a rank-sum test comparing whether the positions of the reference and alternate alleles are different within the reads.

Variant calls with QUAL $\leq$ 30, DP $\leq$ 10, GQ $\leq$ 30, FS $\geq$ 200, MQ $\leq$ 40, MQRankSum $\leq$ -8 or ReadPosRankSum $\leq$ -2 were removed from the analysis.

The individuals in the PROFILE cohort were then quality controlled to reduce the risk of bias within the results due to population stratification or relatedness. This was re-performed for this study (as opposed to simply using the same individuals as in the age-of-onset GWAS) as the samples had been re-analysed using WGS, which re-introduced the possibility of contamination and mix-ups during sample handling. In addition, the re-analysis of the PROFILE samples meant that some individuals who had been excluded from the GWAS because their array-based genotype data had failed QC may be re-included in this WGS analysis, which would increase the sample size and statistical power of the study. This sample QC was performed using peddy[158].

First, the proportion of variant calls for each sample that were heterozygous was calculated and a histogram was used to visualise the distribution of these proportions. Any samples that were visually determined to be an outlier because their proportion of heterozygosity was much greater than the other individuals in the cohort were considered likely to be contaminated and were excluded from the analysis.

Second, genotype information from 2,504 individuals in the 1000 Genomes Project[159] was used to build a classifier that can predict the most likely ancestry of additional samples. This was done for all remaining individuals in the PROFILE cohort and any individuals that were predicted to be of non-European ancestry were excluded from the study. Additionally, any individuals with a predicted probability of European ancestry ≤ 0.9 were also excluded. The efficacy of this filtering was visualised using plots of the first two genetic principal components for the individuals in PROFILE.

Third, the coefficient of relatedness between all individuals in the cohort was calculated. If a pair of individuals had a coefficient of relatedness ≥ 0.125 (indicating at least a third-degree relationship), one individual was selected at random and excluded from the analysis.

Finally, by utilising the fact that males should have zero true heterozygote calls on the X chromosome and females should have many, the sex of each sample was predicted from the genetic data. Any samples whose self-reported sex did not match the sex predicted using the genetic data were excluded.

Of the genetic variants that passed quality control, only rare variants that met strict criteria relating to population frequency and functionality were included in the analysis. These variants were termed qualifying variants (QVs). Two different models were used for the selection of QVs (Table 4.1). The primary model considered only likely deleterious variants, defined as those that were annotated as frameshift, missense, start loss, stop gain or stop loss variants. The negative control model considered only synonymous variants. Whilst some synonymous variants can exert effects on a trait through disruption of regulatory elements, it was considered unlikely that several synonymous variants within the same gene would act in this way and therefore any genes which were to show a strong association with the age-of-onset of IPF could indicate the presence of bias within the results. This approach was previously implemented by Dhindsa et al. in their gene-based collapsing analysis for IPF susceptibility[80]. Variant annotation was performed using SnpEff v5.0[126].

Two types of variant MAF were considered when selecting QVs: an ‘external’ MAF (which represents that variant’s MAF in the general population) and an ‘internal’ MAF (which is the frequency of the minor allele in the study population). In this study, the external MAF of a variant was defined as the MAF in the Genome Aggregation Database (gnomAD)[160] v3.1, a resource that contains sequencing data from the genomes of 76,156 unrelated individuals. A strict external MAF threshold of 0.05% was selected to increase the specificity of the model by minimizing the background variation within a gene and allowing genuine genetic-risk alleles to become prominent in the test. In addition, the use of a strict

external MAF threshold reduced the risk that multiple QVs within the same gene were in LD and not independent. The internal MAF of each variant was calculated as the MAF in the PROFILE cohort after sample QC and a maximum threshold of 1.0% was implemented.

TABLE 4.1: *The criteria for the two different collapsing models used in this study. MAF = minor allele frequency.*

Collapsing model	External MAF	Internal MAF	Variant type
Primary model	<0.05% in gnomAD	<1% in PROFILE	Frameshift, missense, start loss, stop gain, stop loss
Negative control	<0.05% in gnomAD	<1% in PROFILE	Synonymous

In this analysis, the age-at-diagnosis of IPF was modelled in a linear regression framework as a function of the proportion of QVs at each gene for which an individual carries a minor allele, as first described by Morris and Zeggini[161]. This was performed using RVTESTS software[162]. The Morris-Zeggini approach assumes that all individuals in the study cohort are unrelated and that the phenotype is a normally distributed quantitative trait. As a burden test, this approach assumes that all QVs are causal and are all acting on the phenotype of interest with the same direction of effect. A burden test was selected for this analysis as it was expected that rare mutations would usually be associated with poorer patient outcomes, corresponding to a decrease in the age-at-diagnosis of IPF. This rationale was supported by the results of the age-of-onset GWAS in Chapter 3, as for each of the most statistically significant variants in the two studies in that chapter, the minor allele was associated with a younger age-of-onset. In addition, the results from a burden test approach (with simpler assumptions) are arguably easier to interpret than the results produced by a non-burden test.

For each gene, let n denote the number of QVs that have been found to lay within that gene across the study population and let r_i denote the number of those QVs for which individual i carries at least one copy of the minor allele.

The phenotype of the i^{th} individual, y_i, can be modelled as:

$$y_i = \alpha + \lambda \frac{r_i}{n} + \beta x_i + \varepsilon_i \qquad (4.2)$$

Where x_i denotes a vector of covariate measurements for the i^{th} individual, with corresponding regression coefficients β. The parameter λ is the covariate of interest and represents the expected increase in the phenotype for an individual carrying at least one minor allele for all possible QVs compared to an individual carrying none. The statistical significance of the association between the accumulation of QVs within the gene and the phenotype was assessed using the likelihood ratio test, in

which a model representing the null hypothesis (where $\lambda=0$) was compared to a model representing the alternate hypothesis (where $\lambda\neq0$).

Both models were adjusted for the following covariates: sex (male or female), recruitment centre (Nottingham or Brompton), smoking status (never smoker, former smoker or current smoker) and the first 10 genetic principal components. Any individuals with missing information for any of these covariates were removed from the study.

Genes for which there were fewer than two individuals carrying minor alleles for QVs within that gene were removed. Q-Q plots were used to visualise the results for the remaining genes. The genomic inflation factor, λ, was calculated as described in Section 2.2.2 and was used to assess whether there was evidence of genomic inflation within the results. If λ was calculated to be above 1.1, the results were corrected using genomic control (Section 2.2.2). The threshold for study-wide significance was defined as $P<2.6\times10^{-6}$, as this corrects the standard significance level of 0.05 for approximately 19,000 protein coding genes, as recommended in Povysil et al.[150].

Three sensitivity analyses were performed to assess the robustness of the results. In sensitivity analysis 1, the gene-based burden testing was repeated but any individuals who had extreme values for the age-at-enrolment (determined visually using a histogram) were excluded from the analysis. This was done for the both the primary model and negative control model.

In sensitivity analysis 2, two additional gene collapsing models (Table 4.2) were tested to investigate the effect that changes to the internal MAF and definition of deleterious variants may have on the results. First, a 'strict' model was implemented by reducing the internal MAF threshold to 0.5%, whilst all other variables remained the same. Second, a 'lenient' model was implemented by keeping the original internal MAF unchanged at 1% but instead considering additional types of variants to be deleterious and thus qualifying for inclusion in the gene collapsing model. Individuals that were identified as age outliers in the first sensitivity analysis were not included in these analyses. All other methods were the same as previously described.

Lastly, in sensitivity analysis 3, a 'leave-one-out' analysis was performed for all genes that remained study-wide significant ($P<2.6\times10^{-6}$) in either of the previous sensitivity analyses. This was performed to assess whether those signals were being driven by a single QV. In this analysis, the statistical test for each study-wide significant gene was repeated but each of the QVs within that gene were excluded from the model in turn. A signal was considered to be robust if that gene maintained study-wide significance despite the removal of any single QV within that gene (i.e. all $P<2.6\times10^{-6}$).

TABLE 4.2: *The qualifying variant criteria under the primary collapsing analysis model and the two models used in sensitivity analysis 2. MAF = minor allele frequency.*

Collapsing model	External MAF	Internal MAF	Variant type
Primary model	<0.05% in gnomAD	<1% in PROFILE	Frameshift, missense, start lost, stop gained, stop lost
Strict model	<0.05% in gnomAD	<0.5% in PROFILE	Frameshift, missense, start lost, stop gained, stop lost
Lenient model	<0.05% in gnomAD	<1% in PROFILE	Frameshift, missense, start lost, stop gained/lost, 3'/5' untranslated region, untranslated region premature start gain, gene fusion, inframe insertion/deletion, splice donor/acceptor, splice region

4.2.2 Results

WGS data were available for 541 individuals with IPF in the PROFILE cohort. One sample had a proportion of heterozygous calls that was much higher than the other individuals in the PROFILE cohort (Figure 4.1), which indicated that this sample may have been contaminated. This sample was excluded from the study.

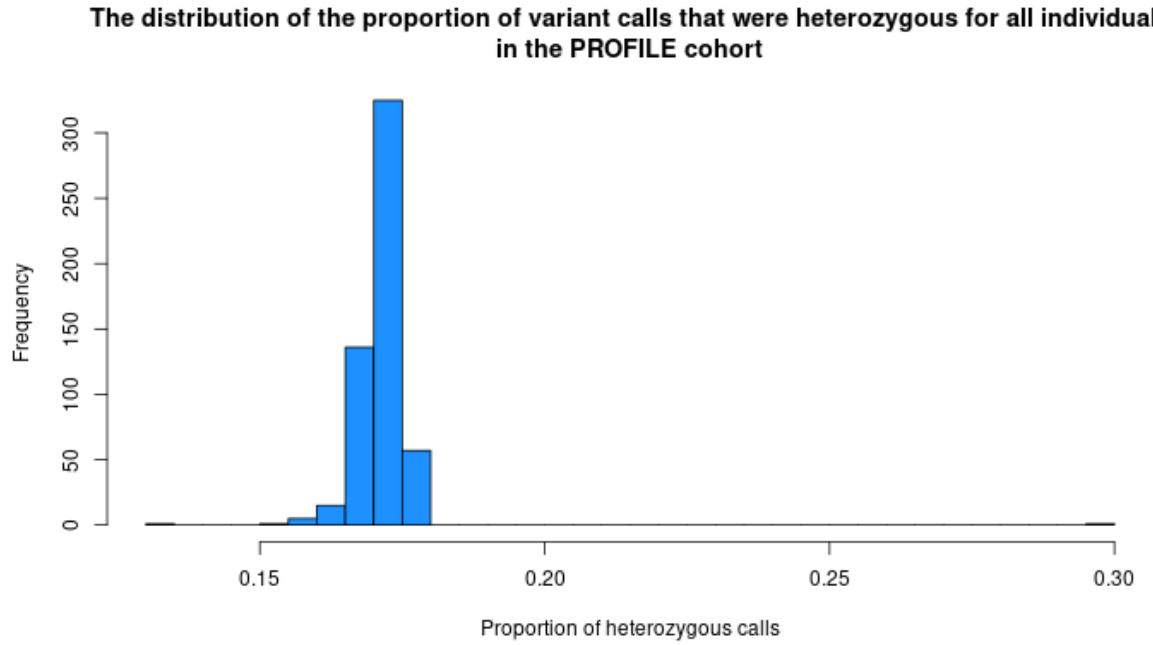


FIGURE 4.1: *A histogram showing the proportion of variant calls that were heterozygous for each individual in the PROFILE cohort.*

While most individuals in the PROFILE cohort were predicted to be of European ancestry, there were also some that appeared to be of South Asian, American or African descent, and one individual that was predicted to be of European ancestry but laid far apart from the other European subjects (Figure 4.2A).

Of all individuals in PROFILE, 519 were estimated to be of European ancestry with a probability greater than 90% (Figure 4.2B) and all other individuals were excluded from the study. The European outlier was removed by this filtering, along with all individuals of non-European ancestry.

Of the remaining subjects, two pairs of individuals had a coefficient of relatedness above 0.125, suggesting at least a third-degree relationship. One individual from each pair was selected at random and excluded from the study, which left 517 IPF cases remaining. Of these 517 individuals, there were 507 whose study-reported sex matched their genetically predicted sex and the 10 sex mismatches were excluded.

493 subjects had complete data for all covariates and were included in the gene-based collapsing analysis (Table 4.3). As observed in Section 3.2.3, on average, males were slightly older than females when enrolled into the PROFILE study and those who were recruited in Nottingham were older than those who were recruited in Brompton. Additionally, current smokers had a far lower mean age-at-enrolment than former smokers and never smokers.

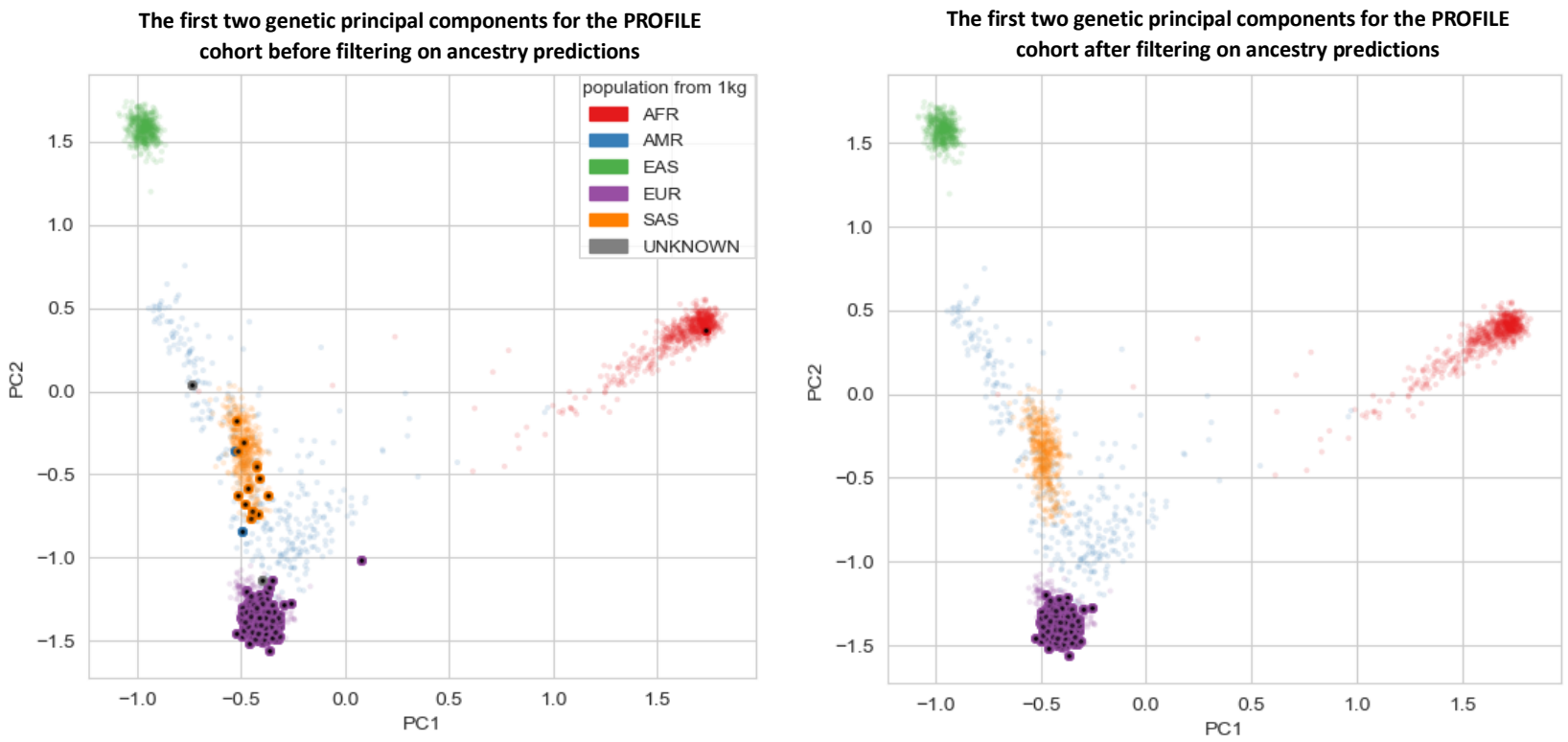


FIGURE 4.2: *Scatter plots of the first two genetic principal components for the individuals in PROFILE before (A) and after (B) filtering based on genetic data-derived ancestry predictions. Solid points represent individuals in the PROFILE cohort and transparent points represent individuals of known ancestry in the 1000 Genomes Project. The individuals in the PROFILE cohort in Figure 4.2B were predicted to be of European ancestry with a probability greater than 90%. AFR = African ancestry, AMR = American ancestry, EAS = East Asian ancestry, EUR = European ancestry, SAS = South Asian ancestry.*

TABLE 4.3: *Demographics for the individuals with IPF from the PROFILE cohort that were included in the analysis.*

Phenotype		PROFILE (n=493)	
		Count (%)	**Mean age-at-enrolment (years) (sd)**
Sex	Female	116 (23.5%)	69.8 (8.5)
	Male	377 (76.5%)	70.7 (8.4)
Smoking status	Never	158 (32.0%)	71.3 (8.4)
	Former	306 (62.1%)	70.7 (8.1)
	Current	29 (5.9%)	64.0 (8.9)
Recruitment centre	Brompton	205 (41.6%)	68.6 (8.0)
	Nottingham	288 (58.4%)	71.8 (8.5)

As mentioned previously, the Morris-Zeggini collapsing analysis method expects the phenotype of interest to be normally distributed. The age-at-enrolment into PROFILE was approximately normally distributed (Figure 4.3), though there was one individual whose age-at-enrolment was moderately lower than that of the other subjects. This individual was not removed from the main analysis but a sensitivity analysis was performed by repeating the study with that subject excluded.

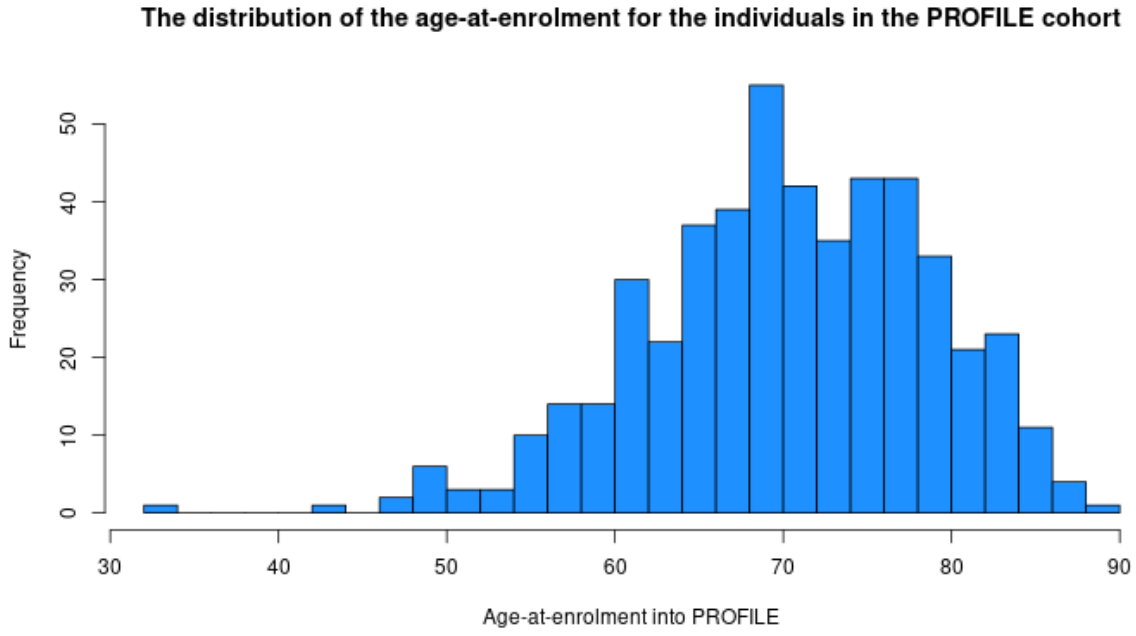


FIGURE 4.3: *histogram showing the distribution of the age-at-enrolment for the subjects in the PROFILE cohort.*

Primary model (deleterious variants)

Across the genome, there were a total of 41,025 rare, likely deleterious variants that met the QV criteria under the primary model (Table 4.4). There were 8,603 genes that contained at least 2 QVs, with a median of 3 QVs per gene. The genomic inflation factor (λ) was 1.061, indicating that the results did

not require adjustment due to genomic inflation (Figure 4.4). Two genes, *IGF2BP2* and *RELT*, reached the threshold for study-wide statistical significance ($P<2.6\times10^{-6}$).

TABLE 4.4: *The count and percentage for each type of variant that were considered in the collapsing analysis under the primary model.*

Variant type	Count	Percentage
Frameshift	934	2.3%
Missense	39,201	95.6%
Start loss	63	0.2%
Stop gain	794	1.9%
Stop loss	33	0.1%
Total	41,025	

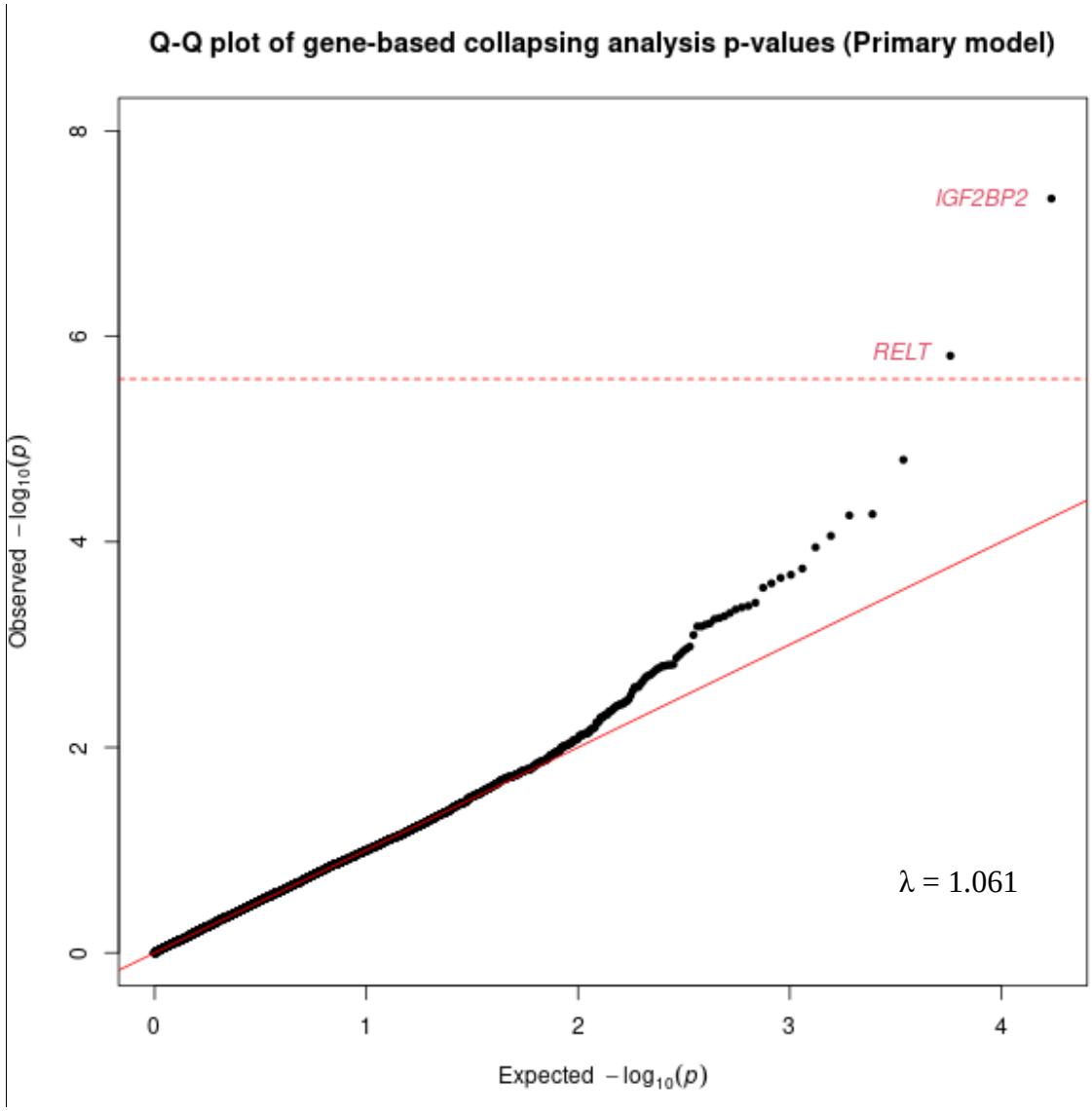


FIGURE 4.4: *A quantile-quantile plot displaying the p-values from the gene-based collapsing analysis under the primary model. The dashed line represents the threshold for study-wide statistical significance ($P=2.6\times10^{-6}$). λ = the genomic inflation factor.*

The gene with the greatest statistical significance under the primary model was *IGF2BP2* (Insulin-Like Growth Factor 2 MRNA-Binding Protein 2, $P=4.5\times10^{-8}$). There were three individuals in PROFILE that carried a minor allele for a QV within *IGF2BP2*. The mean age-at-diagnosis for these three individuals was 48.7 years, which was lower than the mean age-at-diagnosis for those who carried no QVs within *IGF2BP2* (70.6 years).

The other gene that reached study-wide statistical significance under the primary model was *RELT* (Tumor necrosis factor receptor superfamily member 19L, $P=1.5\times10^{-6}$). Five individuals in PROFILE carried minor alleles for QVs within *RELT* and these five individuals had a mean age-at-diagnosis of 54.8 years (compared to 70.7 years for those with no QVs within *RELT*).

Negative control model (synonymous variants)

There was a total of 22,848 synonymous variants that were considered QVs in the analysis under the negative control model. There were 5,719 genes that contained at least 2 QVs and a median of 2 QVs per gene. There was no evidence of genomic inflation ($\lambda = 0.981$) and none of the genes reached study-wide statistical significance ($P<2.6\times10^{-6}$) (Figure 3.5). However, several of the observed p-values on the right-hand side of the plot deviated from the expected line. As this was under the negative control model, which in theory should not detect any genes that are strongly associated with the phenotype of interest, this was considered irregular and motivated sensitivity analysis 1.

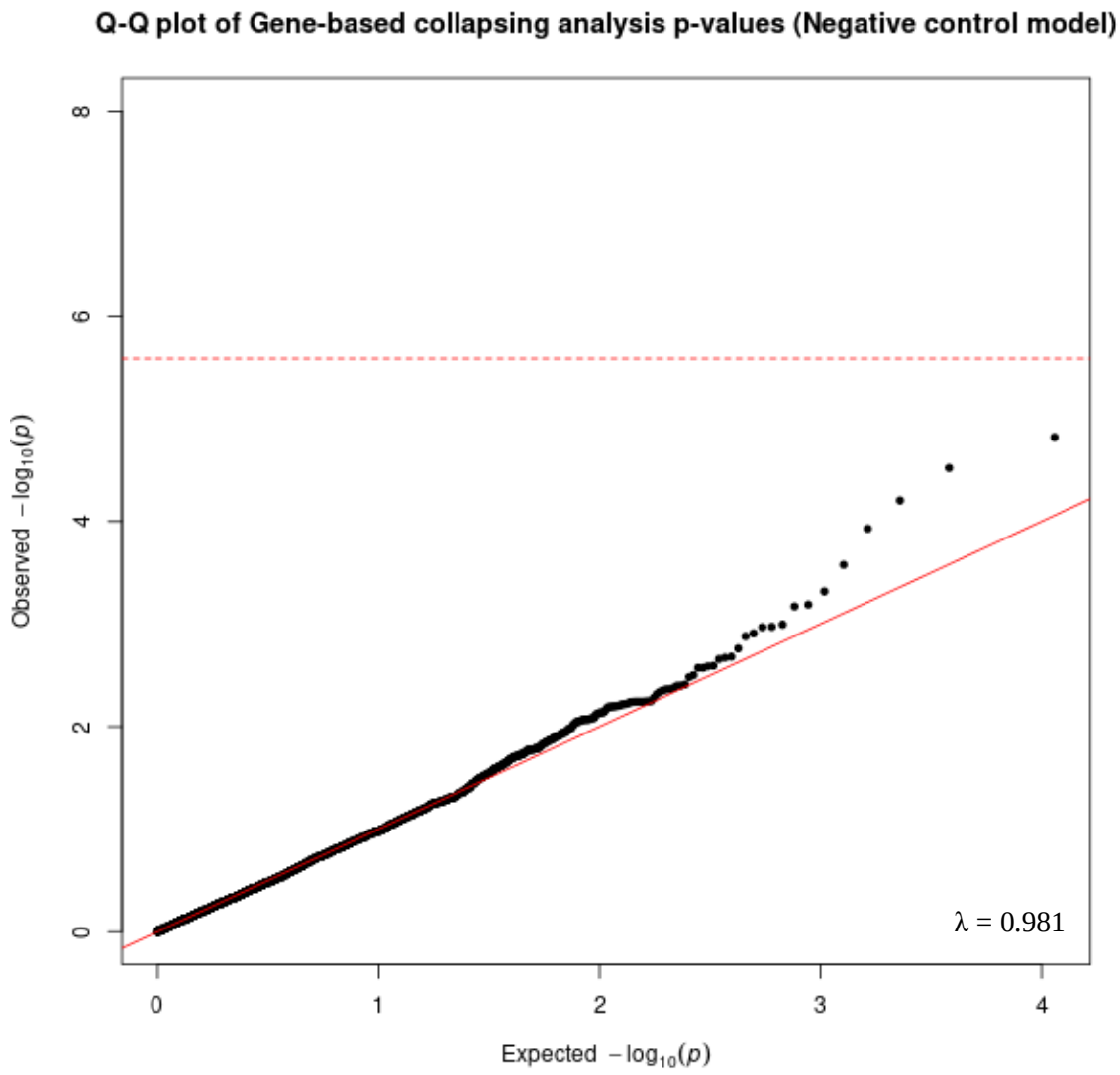


FIGURE 4.5: *A quantile-quantile plot displaying the p-values from the gene-based collapsing analysis under the negative control model. The dashed line represents the threshold for study-wide statistical significance ($P=2.6\times10^{-6}$). λ = the genomic inflation factor.*

Sensitivity analyses

492 individuals remained in the PROFILE cohort for sensitivity analysis 1 after the exclusion of the sole age outlier (Figure 4.3). Under the primary model, there were now 8,592 genes that contained at least 2 QVs, with a median of 3 QVs per gene. The genomic inflation factor was 1.082 (Figure 4.6). No genes reached the threshold for study-wide statistical significance, which suggested that the significant results found previously were largely being driven by the presence of the age outlier. The two genes that were previously study-wide significant, *IGF2BP2* and *RELT*, had p-values of 0.001 and 0.003 respectively in this sensitivity analysis.

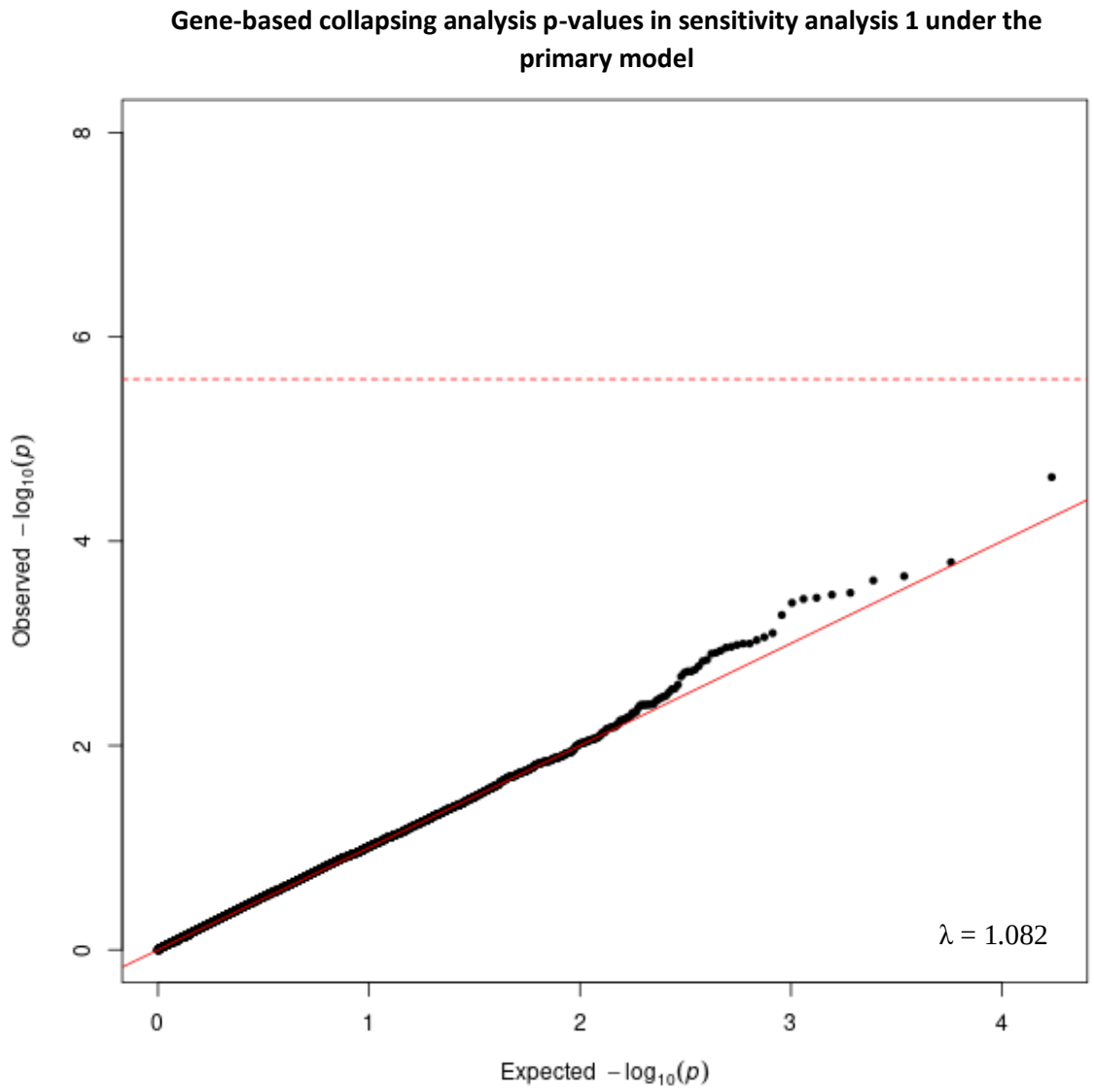


FIGURE 4.6: *A quantile-quantile plot showing the p-values from sensitivity analysis 1 under the primary model. The dashed line represents the threshold for study-wide statistical significance ($P=2.6\times10^{-6}$). λ = the genomic inflation factor.*

Figure 4.7 shows the Q-Q plot for the results of sensitivity analysis 1 under the negative control model. The observed p-values in Figure 4.7 laid closer to the expected line than in Figure 4.5, therefore suggesting that the inclusion of the age outlier was causing spurious results.

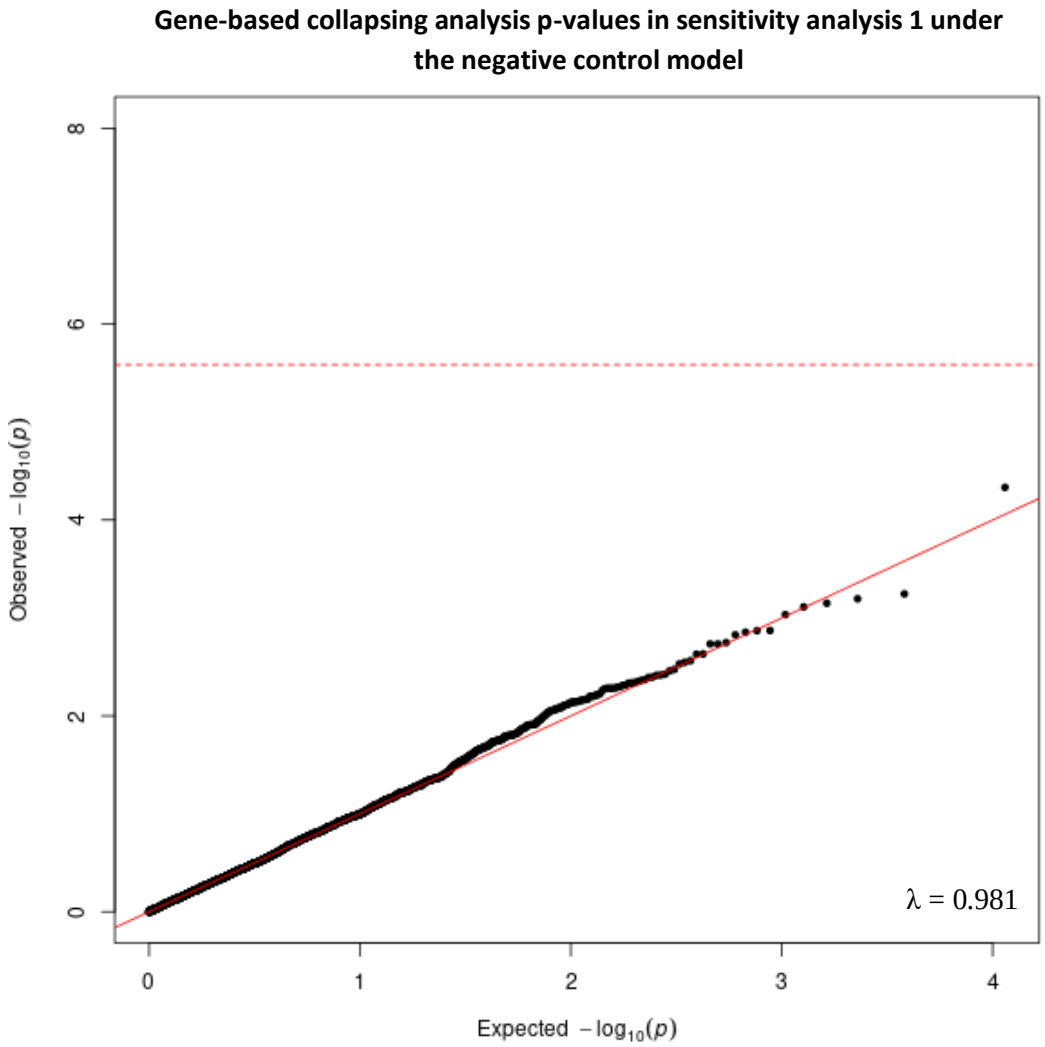


FIGURE 4.7: *A quantile-quantile plot showing the p-values from sensitivity analysis 1 under the negative control model. The dashed line represents the threshold for study-wide statistical significance ($P=2.6\times10^{-6}$). λ = the genomic inflation factor.*

In sensitivity analysis 2, there were 40,954 QVs under the strict model and 123,413 under the lenient model (Table 4.5). The distribution of variant types for the strict model closely resembled the distribution under the primary model (Table 4.3), whereas the majority of the QVs under the lenient model were variants that lay in an untranslated region. Under the strict model, there were 8,582 genes that contained at least 2 QVs and a median of 3 QVs per gene. Under the lenient model, there were 11,219 genes that contained at least 2 QVs, with a median of 3 QVs per gene. Neither set of results required adjustment for genomic inflation (λ_{strict} = 1.082 and $\lambda_{lenient}$ = 1.062), however none of the genes reached the study-wide significance threshold under either model (Figure 4.8). As there were no genes in sensitivity analyses 1 and 2 that reached the study-wide significance threshold, no leave-one-out analyses were performed for sensitivity analysis 3.

TABLE 4.5: *The count and percentage for each type of variant that were considered in the collapsing analysis under the primary model.*

Variant type	Strict model	Lenient model
Frameshift	931 (2.3%)	934 (0.8%)
Missense	39,135 (95.6%)	39,201 (31.8%)
Start loss	63 (0.2%)	63 (0.1%)
Stop gain	793 (1.9%)	794 (0.6%)
Stop loss	32 (0.1%)	33 (0.0%)
3’ untranslated region	0 (0%)	64,253 (52.1%)
5’ untranslated region	0 (0%)	9,311 (7.5%)
Untranslated region premature start gain	0 (0%)	1,513 (1.2%)
Gene fusion	0 (0%)	2 (0.0%)
Inframe insertion	0 (0%)	149 (0.1%)
Inframe deletion	0 (0%)	564 (0.5%)
Splice donor	0 (0%)	322 (0.3%)
Splice acceptor	0 (0%)	228 (0.2%)
Splice region	0 (0%)	6,046 (4.9%)
Total number of variants	40,954	123,413

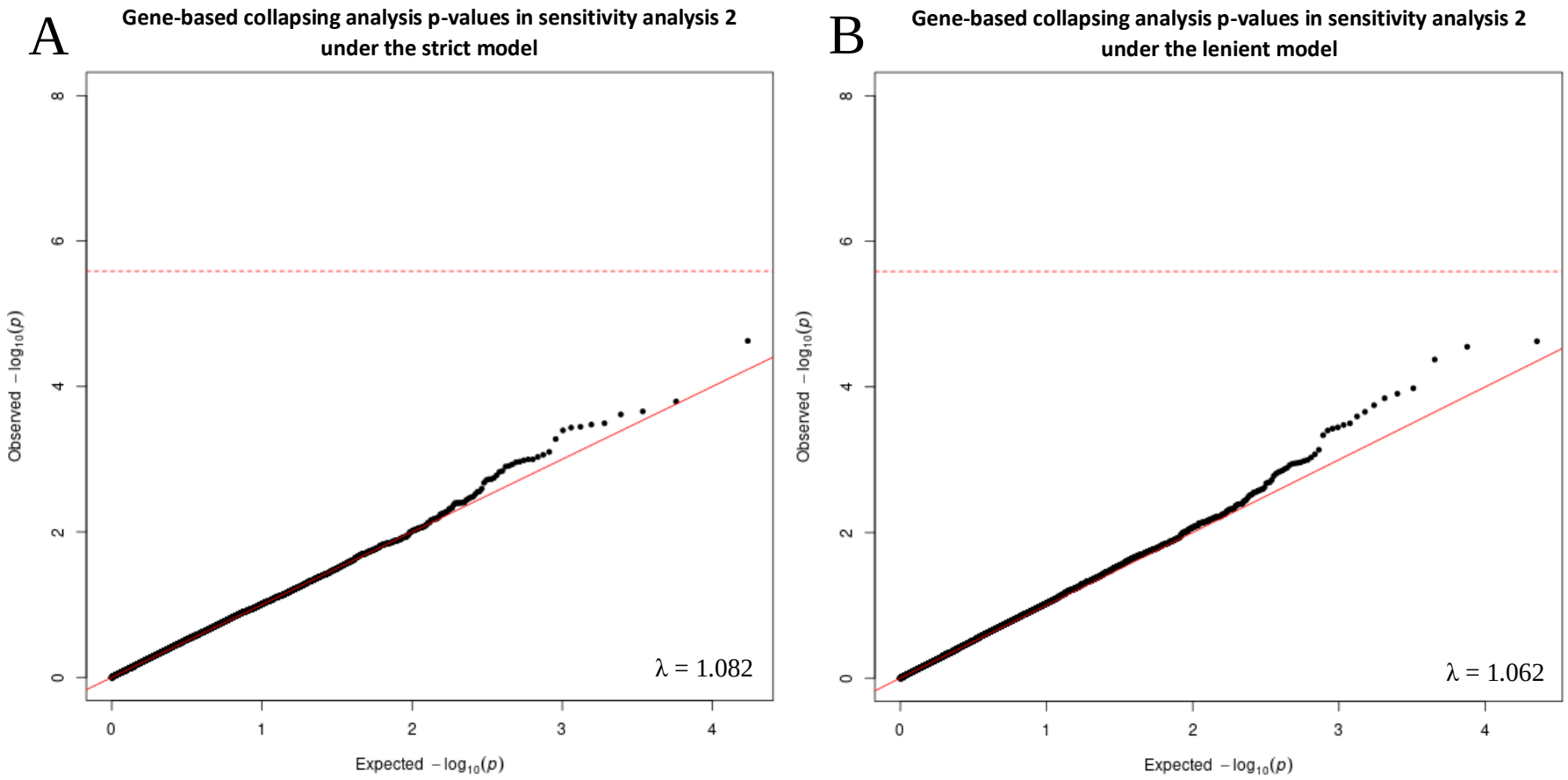


FIGURE 4.8: *Quantile-quantile plots showing the results sensitivity analysis 2, under the strict model (A) and under the lenient model (B). The dashed lines represent the threshold for study-wide statistical significance ($P=2.6\times10^{-6}$). λ = the genomic inflation factor.*

4.3 Gene-based collapsing analysis using a non-burden test

As discussed in the previous section, using a non-burden test to perform the gene-based collapsing analysis may be preferable to a burden test in certain scenarios and could lead to an increase in statistical power. As such, the analysis was repeated using the Sequencing Kernel Association Test (SKAT)[164], a non-burden test that performs well when there are non-causal, deleterious and protective variants present within the gene of interest.

4.3.1 Methods

Variant calling and QC were performed as described in Section 4.2.1. The methods for sample QC were the same as described in Section 4.2.1 with one exception: the sole age outlier was not included in this analysis.

The Sequencing Kernel Association Test (SKAT) was used to test for associations between the joint effects of genetic variants within a gene and the age-at-diagnosis of IPF. This was performed using RVTESTS software[162]. As SKAT does not assume that all variants are causal, genetic variants did not need to meet any criteria based on frequency or function to be included in the analysis. A weighting function was used to increase the importance of rarer variants and to decrease the importance of more common variants in the analysis (see below for details).

Like the Morris-Zeggini method, SKAT uses a regression framework and allows for the adjustment of covariate factors. Assume that sequencing data are available for *n* individuals and there are *p* variant sites observed within a gene. For subject *i*, y_i denotes the phenotype of that individual, $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \dots, X_{im})$ denotes a vector of *m* covariate values for that individual and $\boldsymbol{G}_i = (G_{i1}, G_{i2}, \dots, G_{ip})$ denotes that individual's genotype for the *p* variants.

The following linear model was used to model the age-at-diagnosis of IPF:

$$y_i = \alpha_0 + \boldsymbol{\alpha}^T \boldsymbol{X}_i + \boldsymbol{\beta}' \boldsymbol{G}_i + \varepsilon_i \qquad (4.3)$$

Where α_0 is an intercept term, $\boldsymbol{\alpha}^T = (\alpha_1, \alpha_2, \dots, \alpha_m)^T$ is a vector of regression coefficients for the *m* covariates, $\boldsymbol{\beta}^T = (\beta_1, \beta_2, \dots, \beta_p)^T$ is a vector of regression coefficients for the *p* variants within the gene of interest and ε_i is an error term that has a mean of zero and a variance of σ^2. To evaluate whether the genetic variants within each gene of interest are influencing the age-at-diagnosis of IPF, the null hypothesis $H_0: \boldsymbol{\beta} = 0$ (i.e. $\beta_1, = \beta_2 = \cdots = \beta_p = 0$) was tested. This was done by assuming that each β_j follows an arbitrary distribution with a mean of zero and a variance of $w_j\tau$, where τ is a variance component and w_j is a pre-specified weight for variant j. Testing the null hypothesis $H_0: \boldsymbol{\beta} = 0$ is equivalent to testing whether the variance component τ is equal to 0 (i.e. $H_0: \tau = 0$), which can be tested using a score test. The variance component score test statistic for each gene was calculated as:

$$Q = (\boldsymbol{y} - \widehat{\boldsymbol{\mu}})^T \boldsymbol{K} (\boldsymbol{y} - \widehat{\boldsymbol{\mu}}), \quad (4.4)$$

Where $\widehat{\boldsymbol{\mu}}$ is the predicted mean of $\boldsymbol{y}$ under H_0 (equal to $\widehat{\alpha_0} + \boldsymbol{X}\widehat{\boldsymbol{\alpha}}$, where $\widehat{\alpha_0}$ and $\widehat{\boldsymbol{\alpha}}$ are estimated under the null model by regressing $\boldsymbol{y}$ on only the covariates $\boldsymbol{X}$). Here $\boldsymbol{K} = \boldsymbol{G}\boldsymbol{W}\boldsymbol{G}^{\boldsymbol{T}}$, where $\boldsymbol{G}$ is an $n \times p$ matrix with the (i, j)-th element being the genotype of variant j of subject i, and $\boldsymbol{W} = diag(\omega_1, \omega_2, \ldots, \omega_p)$ is a diagonal matrix that contains the weights of the p variants. $\boldsymbol{K}$ is an $n \times n$ matrix with the (i, i')-th element equal to $K(G_i, G_{i'}) = \sum_{j=1}^{p} \omega_j G_{ij} G_{i'j}$. $K(.,.)$ is known as the weighted linear kernel function and $K(G_i, G_{i'})$ measures the genetic similarity between subjects i and i' for the p variants in the gene of interest.

The variant weights ω_j were pre-specified as a function of the MAF of each variant. In particular, $\sqrt{\omega_j}$ was set as $Beta(MAF_j; 1, 25)$ (i.e. the Beta distribution density function with shape parameters 1 and 25 evaluated at the sample MAF for variant j). The shape parameters 1 and 25 were selected as this increases the weight of rare variants but maintains moderate non-zero weights for uncommon variants (1% < MAF < 5%) (Additional Figure A.4.1).

The same covariates were included in the model as in the previous analysis (i.e. sex, recruitment centre, smoking status and the first 10 genetic principal components) and an additive genetic model was assumed.

Any genes for which fewer than two individuals in the PROFILE cohort carried genetic variants were excluded from the analysis. Q-Q plots were used to visualise the results for the remaining genes and the genomic inflation factor (λ) was calculated as described in Section 2.2.2. Applying genomic control to the output of SKAT and generating adjusted p-values is non-trivial and so in the case where λ was calculated to be above 1.1, genomic control was not applied. The threshold for study-wide significance was again defined as $P<2.6\times10^{-6}$.

4.3.2 Results

In total there were 16,505,366 genic genetic variants that were present in the 492 individuals in the PROFILE cohort. There were 24,818 genes for which at least two individuals in PROFILE carried variants within those genes, with a median of 234 variants per gene. These genes were analysed using SKAT (Figure 4.9). Despite evidence for the presence of genomic inflation within the results (λ=1.144), none of the genes reached the study-wide significance threshold of $P<2.6\times10^{-6}$. The gene with the greatest statistical significance in the analysis was *LOC101929550*, with a p-value of 1.0×10^{-4}.

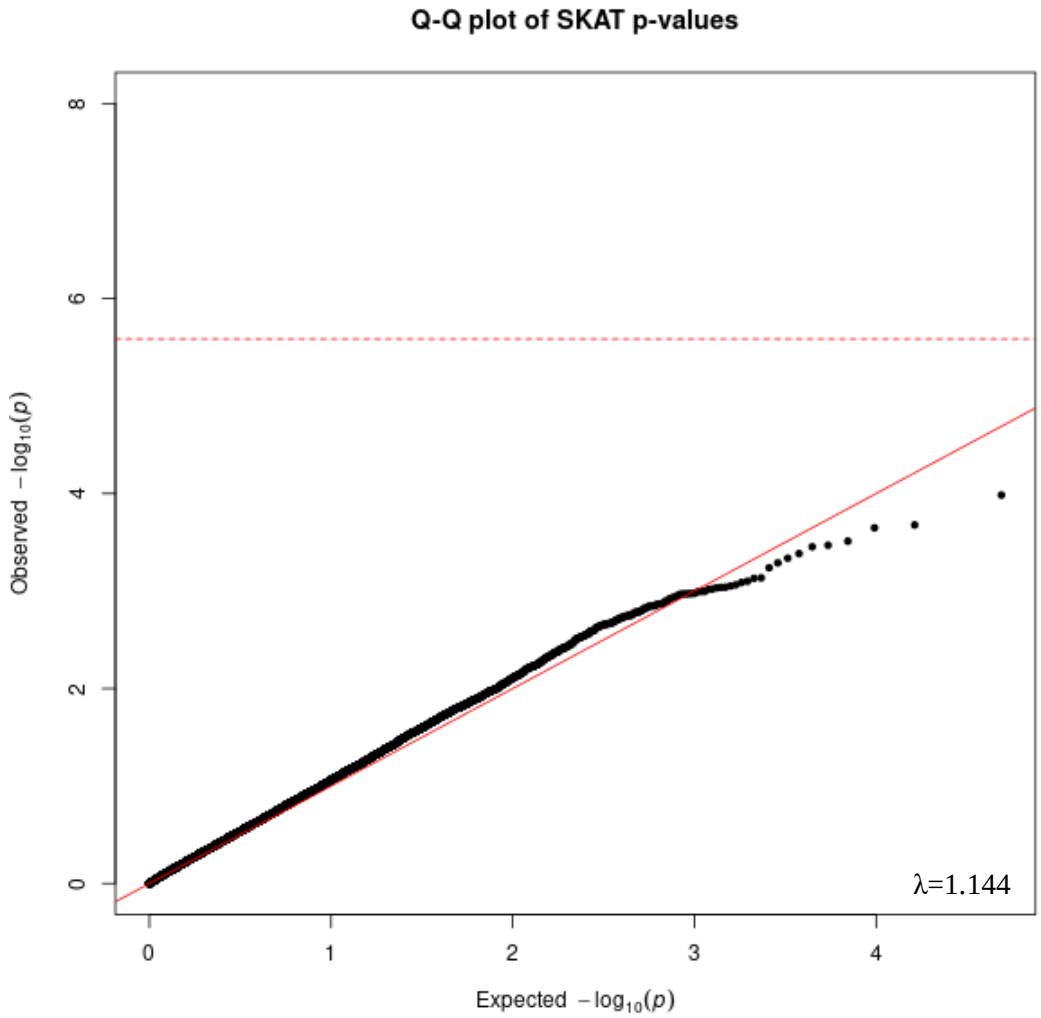


FIGURE 4.9: *Quantile-quantile plot showing the results of the analysis using SKAT. The dashed lines represent the threshold for study-wide statistical significance ($P=2.6\times10^{-6}$). λ = the genomic inflation factor.*

4.4 Discussion

This chapter described the first gene-based collapsing analyses in IPF to investigate the age-of-onset phenotype. By utilizing WGS data, this study was able to consider rare variants (MAF<1%) that were not included in the age-of-onset of IPF GWAS meta-analysis (Chapter 3).

The first approach was to use the Morris-Zeggini method, a burden test that assumes that all variants within a gene are causal and are acting with the same direction of effect. Therefore, it was crucial to select genetic variants such that only those that were most likely to be causal were included in the analysis. Rare, likely deleterious variants were selected for inclusion in the primary model and the Morris-Zeggini method was used to test each gene for an association between the aggregated burden of these variants and the age-of-diagnosis of IPF. There were initially two genes that reached the threshold for study-wide statistical significance ($P<2.6\times10^{-6}$) but it was discovered in a sensitivity analysis that these findings were being driven by a single individual who had been enrolled into the PROFILE study at a much younger age than the other individuals in the cohort. Two additional collapsing analysis

models were then tested in an additional sensitivity analysis by altering the definition of a qualifying variant. Despite these changes, neither model detected any study-wide significant associations.

Although the two genes that were initially study-wide significant were not robust to the removal of the age outlier, it may be worth speculating on the plausibility of an association between these genes and the age-of-onset of IPF. Before the removal of the age outlier, the gene with the greatest statistical significance under the primary model was *IGF2BP2* (Insulin-Like Growth Factor 2 MRNA-Binding Protein 2, $P=4.5\times10^{-8}$). This gene encodes IGF2BP2, which binds to insulin-like growth factor 2 (IGF2) and regulates its translation. IGF2 plays a key role in regulating fetoplacental development and it is involved in glucose metabolism in adipose tissue, skeletal muscle and liver in adults. Interestingly, both IGFBP2 and IGFBP1 (another insulin-like growth factor binding protein) have been proposed as biomarkers for IPF[166–168] and recent findings have shown that epigenetic repression of *IGFBP2* promotes pulmonary fibrosis in mice, suggesting that restoring IGFBP2 in fibrotic lungs could be effective for the treatment of IPF[169]. Our results before the removal of the age outlier would support this hypothesis. Therefore, an association between *IGFBP2* and the age-of-onset of IPF is plausible and this gene could be a good candidate for future studies of this phenotype, preferably once larger study sizes are available.

The second study-wide significant gene was *RELT*, a gene which is believed to play a role in apoptosis[170]. This gene has not been previously linked to the pathogenesis of IPF, although it is plausible that apoptosis may influence the age-of-onset of IPF; it has been previously reported that apoptosis is increased in alveolar epithelial cells of IPF patients but decreased in myofibroblasts[171], with this imbalance contributing to the development of IPF[172]. Furthermore, the use of therapies that can selectively manipulate apoptosis have been proposed[173]. These results could support the hypothesis that apoptosis is an important factor in the development of IPF, though again additional data is needed to confirm this finding.

There could be several reasons to explain why no genes reached the threshold for study-wide significance ($P<2.6\times10^{-6}$) when using the Morris-Zeggini burden test after the removal of the single age outlier. Firstly, the relatively low sample size of the study (n= 492 IPF cases in the sensitivity analyses) meant that it may have been underpowered to detect associations between genes and the age-of-onset of IPF if the effect sizes were small. In fact, a simulation study that compared different types of collapsing-analysis methods, which had a sample size of 697 unrelated individuals, found that the Morris-Zeggini approach (labelled RVT1 in the study) had an average power of only 17% when considering gene-based variants for a quantitative phenotype and a MAF threshold of 1%[163].

Secondly, as the Morris-Zeggini method assumes that all variants are causal, it was important to functionally annotate the variants and include only those that are most likely to be deleterious. However, these functional annotations are merely predictions and can vary from one tool to another. As the

Morris-Zeggini approach is reliant on high quality annotations, any errors in the annotations of SnpEff could have led to a loss in statistical power.

Thirdly, as a burden test, the Morris-Zeggini method assumes that all QVs within a gene are causal and are acting with the same direction of effect. Therefore, in a scenario where these assumptions are not met, this method will have reduced statistical power. If so, a non-burden test may have been more appropriate. Non-burden approaches also have the advantage that they can include uncommon and common variants in the calculations (often with lower weights than the rare variants), which may be more appropriate in this instance, given the low sample size of the study and low counts of rare variants.

Therefore, for completeness, the analysis was then repeated using SKAT, a type of non-burden test that can perform well when the assumptions of a burden test are not met. As SKAT can allow for non-causal variants within the gene of interest, there was no need to exclude variants based on frequency or function prior to the analysis. Instead, a weighting function was used to increase the weight of rare variants and to decrease the weight of common variants in the calculations. SKAT was used to test each gene for an association between the joint effects of all genetic variants within the gene and the age-of-diagnosis of IPF. However, this did not reveal any statistically significant genes that were associated with the age-at-diagnosis of IPF. This suggests that the reason that no statistically significant genes were identified when using the Morris-Zeggini method was not due to the presence of non-causal genetic variants or the presence of harmful and protective variants within the same gene. It therefore appears more likely that these analyses did not detect any statistically significant results due to a lack of statistical power as a result of the relatively low sample size.

It is possible that the statistical power of the SKAT analysis could have been increased by changing the variant weighting function. For example, a variant weighting function that gives even greater weight to rare variants and less weight to uncommon variants may have been a more appropriate choice. However, the pre-specified parameters for the weighting function are fairly arbitrary and performing follow-up analyses with different choices of parameters could be considered data dredging. A better approach to boost the statistical power of the SKAT analysis would be to incorporate functional information for the variants in an effort to give greater weights to functional variants that are more likely to be causal. However, it is not yet possible to implement this using RVTESTS software.

Genomic inflation was observed in the results when using SKAT. This has been previously reported to occur in analyses with an insufficient number of samples and has led to the development of adaptive procedures, such as AP-SKAT[165], that offer more accurate p-values in this scenario. Regardless, none of the genes in this analysis reached the study-wide significance threshold and so correcting the p-values for the genomic inflation would not change the conclusions of this analysis.

A threshold for study-wide significance ($P<2.6\times10^{-6}$) was used to correct the standard significance level of P=0.05 for the presence of approximately 19,000 protein coding genes. However, this may have been

over-conservative as only genes for which multiple individuals in PROFILE carried at least one QV in that gene were considered in each analysis. For example, there were only 8,604 genes that met this criterion under the primary model and a Bonferroni correction for this number of genes would give a less strict study-wide significance threshold of $P=5.8\times10^{-6}$. On the other hand, the study-wide significance threshold was not corrected for the fact that after the inclusion of the two sensitivity analysis models, a total of four different models were used in this study (excluding the negative control models), and so the study-wide significance threshold of $P=2.6\times10^{-6}$ may have actually been slightly under-conservative. Regardless, no genes came close to reaching this threshold after the exclusion of the age outlier, under any model.

The lack of statistical power due to the relatively low sample size was therefore the primary limitation of these analyses. However, the role of rare variants in the genetic architecture of the age-of-onset of IPF had not previously been investigated and it was possible that multiple causal variants within the same genes could have been exerting such large effects on the age-of-onset that the collapsing models could have detected them. As no such genes were detected in this study, future studies on this topic will likely need to aim to obtain a larger sample size to increase the statistical power of the collapsing models. If this is not possible, it may be necessary to consider candidate genes to reduce the search space and lower the multiple testing burden.

Individuals with IPF were excluded from the study if they had any missing values for any of the covariates that were included in the model. One method to increase the statistical power of the study would be to impute the missing covariate data and include these individuals. However, that would only increase the sample size by 14 individuals, which would be unlikely to substantially increase the statistical power.

The optimal way to significantly increase statistical power would be to obtain and test additional independent IPF WGS datasets and meta-analyse these with the results of the PROFILE cohort. However, this type of data remains costly to measure and so obtaining a suitably large dataset may prove challenging. Alternatively, collapsing the genes into known pathways and regressing the rare variant burden of those pathways against the age-of-onset of IPF could increase the statistical power of the study by further reducing the multiple testing burden. Given the current available sample size, this may be a more viable approach to assessing whether rare variants are influencing the age-of-onset of IPF.

Another weakness of this study is that genetic variants located outside of gene regions were not considered in these analyses, despite them being measured and present within the WGS dataset. Using SKAT, it was possible to have searched areas across the entire genome for regions associated with the age-at-diagnosis of IPF by using moving windows as the regions of interest rather than genes. However, this would have greatly increased the multiple testing burden and the low power of the study meant that

increasing the testing space any further would have likely been counter-productive towards identifying statistically significant results.

However, this study did have some important strengths. Firstly, the use of both a burden test and a non-burden test meant that the effects of rare variants on the age-of-onset of IPF could be investigated under two opposing sets of assumptions. In the scenario where many rare variants are causal and acting in the same direction on the age-of-onset of IPF, the Morris-Zeggini method would have been relatively well powered compared to SKAT. Conversely, in the scenario where not all variants are causal and those that are causal are acting in opposing directions, SKAT would likely have outperformed the Morris-Zeggini method. Whilst this thoroughness was an asset to our study, it is worth noting that an optimised method that combines burden and SKAT statistics, referred to as SKAT-O, has been developed and has been shown to maintain statistical power in both scenarios[174]. As such, it may have been preferable to use SKAT-O rather than perform the burden and non-burden tests separately.

Another strength is that both the Morris-Zeggini method and SKAT allowed for the adjustment of important covariate factors. Adjusting for the non-confounding covariates (sex, recruitment centre and smoking status) should have reduced bias within the results and could have explained more of the phenotypic variance, thereby increasing the statistical power of the linear model[175]. Additionally, adjusting for genetic principal components should have reduced the risk of confounding within the results due to population stratification. However, including 10 genetic principal components in the models (as is common practice in a GWAS) may have been an over-adjustment, given the small sample size and the fact that less population structure underlies rare variants than common variants[150]. Therefore, it may have been more appropriate to include fewer genetic principal components in the models, which could have increased their statistical power.

As discussed previously, a non-burden test such as SKAT has a few advantages compared to a burden test such as the Morris-Zeggini method. The use of a variant weighting function meant that MAF thresholds were not needed to define rare variants, which allowed information from uncommon and common variants to be included in the calculations. In addition, SKAT was not dependent on high quality functional effect predictions, which was a source of potential error and power loss when using the Morris-Zeggini method. Still, after the exclusion of the age outlier the results of the two methods were similar and no genes reached the study-wide significance threshold in either analysis. Therefore, it is difficult to conclude that one method was a more appropriate choice than the other in this instance.

In conclusion, a greater sample size is required, or alternative methods must be applied in order to increase the statistical power of the analysis. Until then, the extent to which rare variants are playing a role in the age-of-onset of IPF is uncertain.

Chapter 5 – Cluster analysis in multiple transcriptomic datasets to identify endotypes of IPF

Considerable heterogeneity in disease progression, survival and response to therapy in IPF suggests that a range of subtypes of the disease may exist. The aim of this chapter is to identify these endotypes by utilizing gene expression data and two new statistical methods. The first is a method of data co-normalization that enables multiple transcriptomic datasets to be combined, whilst the second is a method of clustering that considers data from various clustering algorithms in order to identify the most robust subgroups of subjects within the pooled data.

5.1 Introduction

5.1.1 Identification of endotypes through cluster analysis of transcriptomic data

Omic analyses allow researchers to perform a close examination of a biological system to gain a deeper understanding of its underlying mechanisms, which is achieved through taking a huge number of molecular measurements within a tissue or cell. As omics technologies continue to emerge and develop, the number of molecular measurements in these analyses continue to increase, often producing complex, high-dimensional datasets. These high-dimensional datasets can potentially contain tens of thousands of measurements for each subject and as a result, they are often so large that they can become troublesome to analyse or visualise. Additionally, as there are so many variables to consider, recognising patterns in the data and identifying subgroups of subjects with similar biological measurements often cannot be done without the use of a computational method.

One approach to defining subgroups in large, multi-dimensional datasets is cluster analysis. Cluster analysis is a method of unsupervised machine learning, in which objects (such as individuals, or genes) that are similar are put into groups called clusters, whilst dissimilar objects are put into separate clusters. Cluster analysis can be applied to transcriptomic data to define disease endotypes. Individuals with the same disease and similar transcriptomic profiles have genes that are co-expressed, which indicates that the same biological processes are being activated. This does not confer information about causality, but it does suggest that the individuals within the same clusters may be experiencing forms of the disease with similar underlying biological mechanisms.

Previous transcriptomic analyses have been particularly successful in defining clinically significant subgroups of cancer patients that have led to improvements in treatment. For example, one study[176] used microarrays to extract gene expression data from tumours in 117 breast cancer patients who did not have tumour cells at their local lymph nodes at diagnosis. Hierarchical clustering was applied to the dataset (consisting of approximately 5,000 genes), which revealed two groups of tumours, one that was associated with a favourable prognosis and one that was associated with a poor prognosis. From this, the authors created a 70-gene signature that was able to predict a poor prognosis in breast cancer patients

and has been successfully validated in multiple subsequent studies[177,178,178,179,179]. This led to the development of the MammaPrint assay[180], which is a prognostic tool that uses the same 70-gene signature to help physicians decide whether a patient would benefit from chemotherapy. The assay is currently used in the U.S. to spare patients that are at a low risk of developing distant metastases from needlessly experiencing the severe side effects of chemotherapy. Other examples include lung cancer[181] and bowel cancer[182].

5.1.2 Previous work on this topic in IPF

There have been some previous studies in IPF that have aimed to identify subgroups of subjects by utilizing gene expression data. Another common aim for these studies is to then take their subgroups and use them to try to develop prognostic biomarkers in the form of gene signatures. For example, Boon et al.[183] used gene expression collected from the lung parenchyma to show that the transcriptomic profiles of 12 IPF samples (6 clinically defined as stable IPF and 6 progressive IPF) were distinguishable from healthy control samples. However, the authors were not able to successfully validate a gene signature (of 134 genes) that they had built to distinguish between the progressive and stable IPF samples, perhaps due to the low sample size of the validation cohort (n=8 IPF subjects, 4 clinically defined to have slow progression and 4 clinically defined to have accelerated progression). Additionally, not all the genes included in their progression signature were measured in their validation dataset, which would have greatly hindered the signature's efficacy. More recently, Wang et al.[184] used gene expression from whole lung tissue to build a gene signature (of 392 genes) with the ability to differentiate between subjects that had mild IPF (with lung tissue from biopsy) and those who had severe IPF (with lung tissue from explant). However, when looking at an additional cohort of IPF cases, they were not able to build a signature that could successfully differentiate between patients who had stable IPF and IPF with acute exacerbations.

In 2013, Herazo-Maya et al.[185] studied the peripheral blood transcriptome of 45 individuals with IPF and found 52 genes that were significantly associated with transplant-free survival. The authors used these 52 genes to build a gene signature that was applied to a validation cohort of 75 IPF cases. The gene expression data was then clustered, showing two main clusters of IPF subjects. There were significant differences in survival found between clusters (median transplant-free survival time was 3.4 years in cluster 1 and 1.6 years in cluster 2), but the clusters did not differ significantly in terms of other measured clinical traits. Pathway analysis was used to show that the biological pathway that was most strongly associated with transplant-free survival was the 'costimulatory signal during T cell activation' pathway.

This 52-gene signature was later validated further in a large international multi-centre study[186] where it was applied to 425 IPF cases, who were classed as either high risk (those more likely to die or require a lung transplant) or low risk (those less likely to die or require a lung transplant) using a method called

The Scoring Algorithm for Molecular Subphenotypes (SAMS). There were significant differences in survival time between the two risk groups in all six cohorts and the authors demonstrated that the gene signature could be added to the Gender, Age and Lung Physiology (GAP) index[187] to substantially improve prediction accuracy when incorporated with currently used clinical tools. However, one weakness of their 2013 study is that the two groups were not well characterized in terms of the biological mechanisms that may be driving the differences in survival between groups. Perhaps this aspect of the study was held back by the fact that once all the genes that were not associated with transplant-free survival in the initial study had been discarded, only a relatively small number of genes remained for pathway analysis. Pathway analysis uses a list of genes to identify biological pathways that are enriched (when the genes in the list lay along that pathway more than would be expected by chance), and so a small list of genes may limit the potential for the pathway analysis to detect any significant results. Another important weakness of their 52-gene signature is that its clinical use is limited by the requirement that gene expression from a whole cohort of IPF patients would be needed in order to predict whether a single new patient is high risk or low risk, due to the way that the gene expression data must be normalised within-cohort before calculations can be made.

5.1.3 Improvements to transcriptomic analysis methodology

A key issue with transcriptomic analyses is that non-biological experimental variation or 'batch effects' are commonly observed across multiple batches of microarray experiments[188]. These technical differences between batches can be caused by many factors, including: the batch of amplification reagent used, the time of day when the gene expression is measured or even the atmospheric ozone level[189]. This creates an issue when comparing gene expression data that were collected at different times within the same study, and an even greater issue when comparing data from separate studies. Furthermore, differences in microarray technology and data collection procedures between studies also contribute to the non-biological variation, making transcriptomic data from different studies even more difficult to compare or combine. This, compounded with the fact that it can be extremely expensive/difficult to gain access to the relevant tissue/cell types, means that many transcriptomic studies end up with a low sample size.

In addition to limiting the statistical power of the analysis, a low sample size presents a problem for cluster analysis, as this often leads to the disease heterogeneity not being fully captured and results in findings that cannot not be successfully validated. Another reason that the results from many cluster analyses fail in the validation stage is that small changes in the methodology can lead to very different results. However, recent advances in clustering and data pooling methods can help to solve these problems.

An example of these improvements in methodology can be found in a study by Sweeney et al.[94], where the authors demonstrated new methods that allowed them to increase the sample size of their study by

pooling publicly available gene expression data from 14 studies (providing a total of 700 sepsis cases) and then performing a cluster analysis to identify three distinct clusters of sepsis patients. These clusters were clinically distinct, with the first cluster comprising of less severe patients, while the second and third clusters separated the more severe patients into a younger and an older group. The authors then used gene ontology analysis to further characterise the three clusters. The authors stated that since they had not used clinical data whatsoever in their clustering, discovering differences in mortality between the clustered patients suggests that the clusters may represent distinct pathophysiologic states of clinical relevance. They then validated their findings using nine independent validation datasets (n=600), in which they assigned each of the validation sepsis cases into one of the three clusters and found that the same clinical and molecular phenotypes were observed. When comparing their findings to previously reported endotypes of sepsis, Sweeney et al. found that other groups had reported subtypes similar to two of their clusters, but the substantially larger size of their study had allowed them to discover the third cluster.

This chapter describes a transcriptomic cluster analysis of IPF cases that was performed to identify clinically relevant and distinct endotypes of IPF. Sweeney et al.'s new method of data co-normalisation[190], which aims to reduce the technical differences between datasets, was adopted in this analysis. This allowed for multiple publicly available datasets to be combined, thus providing a relatively large sample size of IPF cases for the analysis. Additionally, a new method of unsupervised clustering[191] was utilised. This approach combined multiple clustering algorithms over a range of genes and validation measures to identify the most robust number of clusters in the pooled dataset of IPF cases. Following this, a gene expression-based classifier with the ability to assign new individuals into one of these clusters was developed, which allowed for validation of findings in additional independent datasets.

5.2 Study design

There were three main parts to this analysis (Figure 5.1). First, a systematic search of publicly available IPF gene expression datasets was performed to collect the data to include in the analysis (Section 5.3). After these data were collected, the selected datasets were partitioned into those that would be analysed during the discovery stage and those that would be used for validation, based on the criteria described in the following section. In the discovery stage, the transcriptomic data from the discovery studies were co-normalised, pooled and clustered. Comparisons of phenotypic traits between clustered subjects and enrichment analysis on clustered genes were performed to characterize each of the clusters. Following this came the development of a gene-expression based cluster classifier with the ability to accurately assign additional independent samples in to one of the discovery clusters, whilst using far fewer genes than the original clustering. In the validation stage, this cluster classifier was applied to the validation datasets and again phenotypic traits were compared between clustered subjects to evaluate whether the

clinical/demographic differences between clusters were consistent with those observed in the discovery stage.

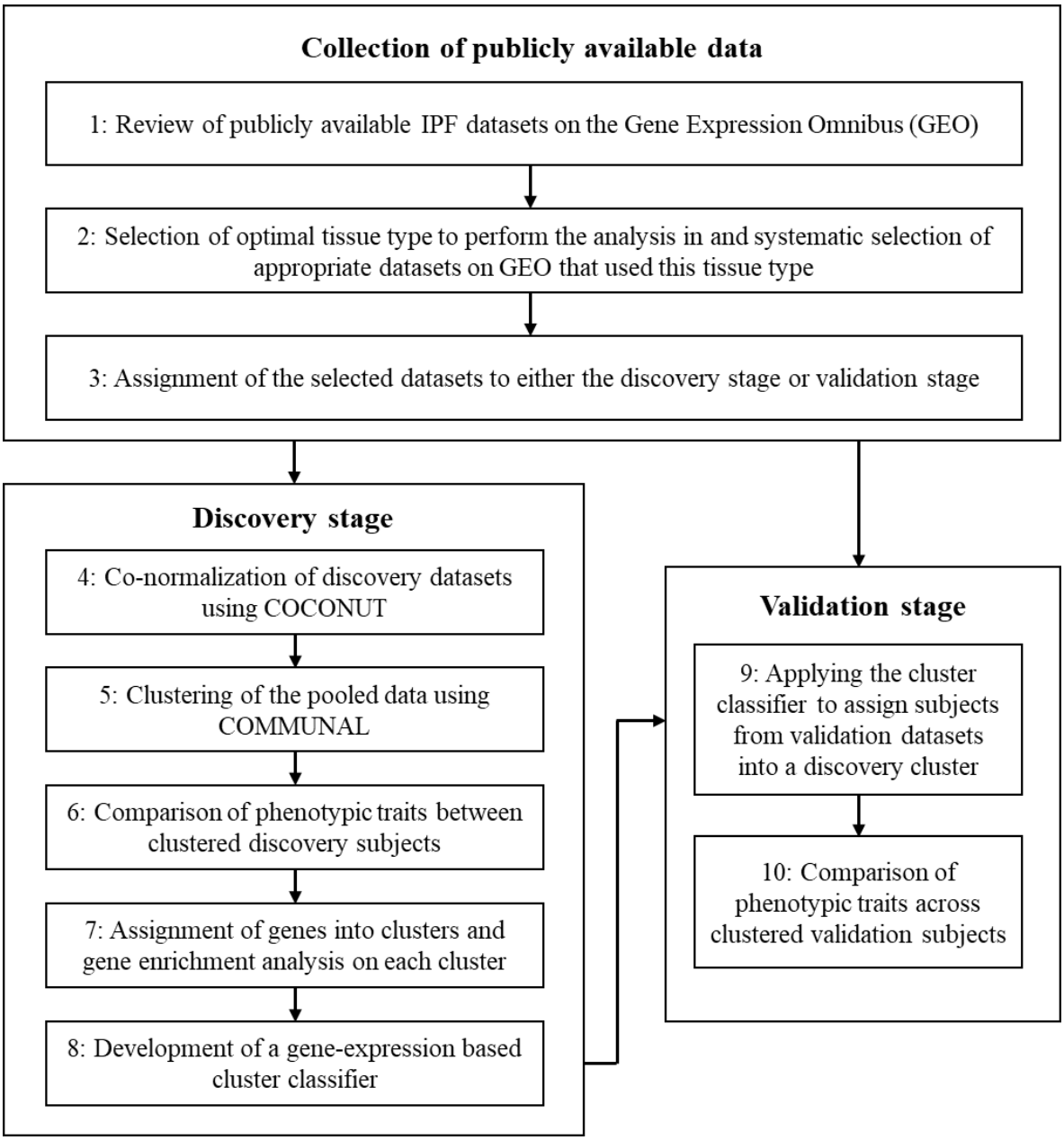


FIGURE 5.1: *Flowchart showing the study design for this analysis.*

5.3 Systematic selection of publicly available transcriptomic datasets

The Gene Expression Omnibus (GEO)[192] is an international public repository that archives and freely distributes transcriptomic data submitted by the research community. This includes data obtained via microarray, next-generation sequencing or other forms of high-throughput functional genomics. Authors may also upload 'meta' data to accompany the gene expression data, which contains additional information for the samples, such as demographic information, clinical data or details about the

collection process of the samples. A systematic search of GEO was performed in March 2020 to select the datasets that were suitable for inclusion in the cluster analysis.

5.3.1 Methods

Multiple sets of transcriptomic data from independent cohorts were required for the analysis. GEO was searched for all collections that contained the term 'IPF' and all collections that did not contain data from human tissue were excluded. The GEO search was then restricted to collections with at least 30 samples, which allowed for the inclusion of the largest datasets with the most IPF cases and healthy control subjects. These datasets were the most likely to successfully co-normalise due to the higher counts of healthy control subjects (Section 5.5). The systematic search was not restricted by platform.

Each of the remaining collections were then reviewed to assess whether they contained data for IPF cases, as it was possible that a collection could have contained the term 'IPF' but did not contain data originating from individuals with IPF. All collections that did not contain data for IPF subjects were excluded.

The method for data co-normalisation, COmbat CO-Normalization Using conTrols (COCONUT), requires data from healthy controls to perform the first step of the co-normalisation, as well as disease cases from the same study (Section 5.5). COCONUT assumes that the gene expression profiles of the healthy subjects in all cohorts come from the same statistical distribution. As gene expression varies by tissue/cell-type, the co-normalisation would have been most successful if the gene expression data from each dataset had been measured from the same tissue/cell type. As such, the GEO search was then restricted to collections that had measured gene expression data from the single most appropriate tissue/cell type to use for the analysis. The choice of the optimal tissue/cell type was based on the following criteria:

- More than one of the remaining collections must have studied this tissue/cell type.
- Of these, multiple collections must have contained data for healthy control subjects in addition to the IPF subjects, so that the data for the IPF cases could be co-normalised with COCONUT.
- The optimal tissue/cell type should have provided a suitable sample size to fully capture the heterogeneity of IPF. As such, if there were fewer than 100 IPF cases across all collections for a tissue/cell type, this tissue/cell type was excluded.
- The collections that contained transcriptomic data measured from the optimal tissue/cell type must have additionally reported meta data for the subjects, as meta data was required for the discovery analysis to characterise the clusters and to identify any clinical or demographical differences between the subjects in each cluster. Meta data was also important for use in the validation stage (Section 5.10). For a phenotypic trait to be used in both stages, it must have been reported for the subjects in multiple independent cohorts, at least one of which must have

included healthy controls. For a tissue/cell type to be selected as the optimal choice, several traits should have met this criterion.

The relevance of the tissue/cell type to IPF was also considered in selecting the optimal tissue/cell type, with the most relevant being considered a preferable choice compared to those that are less relevant. However, the four factors listed above took precedence over this and none of the tissue/cell types were excluded based on this factor.

After an optimal tissue/cell type was selected, all collections that did not include data from this tissue/cell type were excluded. In addition, any cohorts that were found to have fewer than 10 IPF cases were excluded. Collections that contained gene expression data from the optimal tissue/cell type that did not include data for healthy controls were not discarded as these were used in the validation stage, which did not require the data to be co-normalised.

As multiple transcriptomic datasets were to be combined, it was important to check across collections for samples that originated from the same individual, as one person may have participated in multiple studies, or separate studies could have analysed the same tissue sample to test different hypotheses. The subjects in each collection were checked for unique study identification codes. If samples from common subjects were identified, all but one of these were removed as they were not independent and therefore could have biased the results of the study.

5.3.2 Results

The results of the systematic selection of suitable datasets is summarised in a flow diagram (Figure 5.2). There were 143 collections on GEO with gene expression data available that featured the term ‘IPF’, 125 of which contained data from human cohorts. 38 collections remained after these were filtered by sample count. The distribution of collection sizes is shown in Appendix B (Table B.5.1).

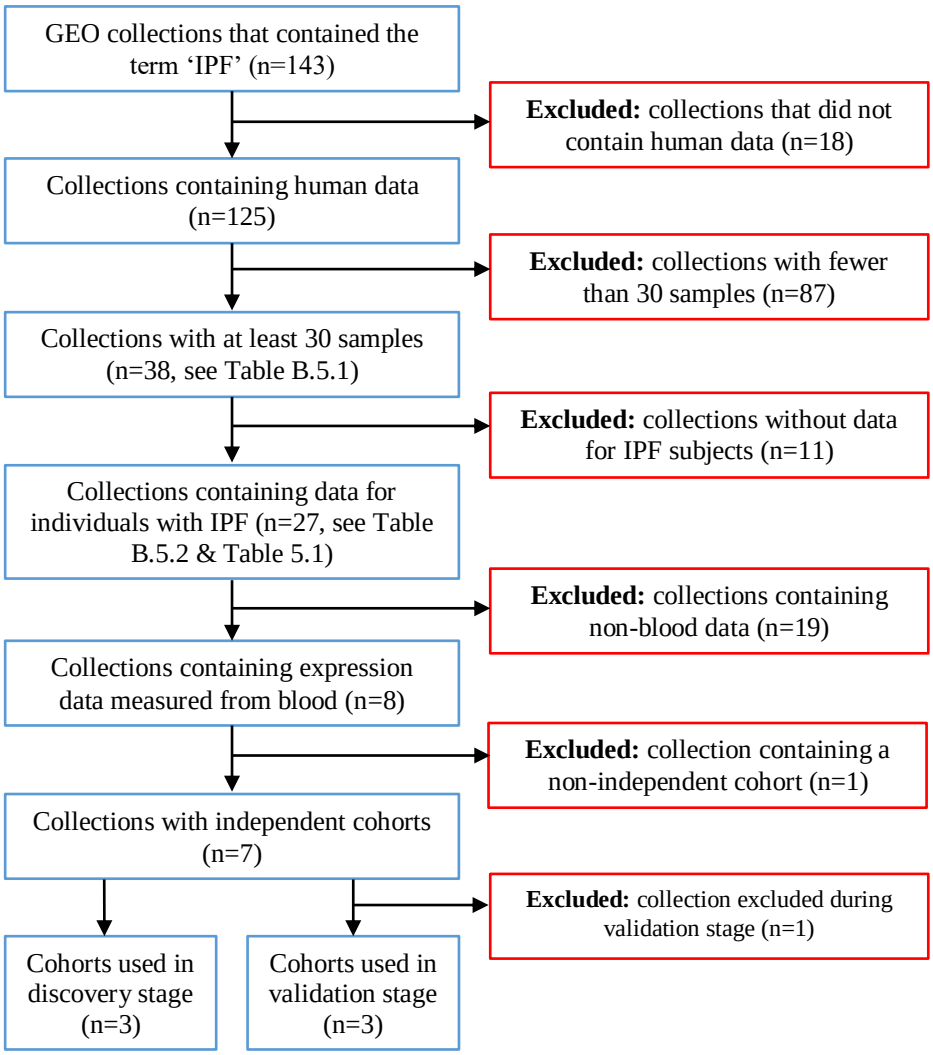


FIGURE 5.2: *Flow diagram showing the process used for the systematic selection of publicly available IPF gene expression datasets from the Gene Expression Omnibus for use in this analysis.*

The transcriptomic data from these collections were measured from many different tissue types (including blood, lung explants and lung fibroblasts) and using a wide range of platforms (including both microarray and high-throughput sequencing platforms). The oldest collection originated from a study published in 2009 and the newest collections were not yet published but were uploaded to GEO in 2020. The collections that were uploaded to GEO before December 2016 were all microarray-based. After this date, an increasing number of collections had measured gene expression data using RNA-sequencing instead of microarrays, as the cost of this method fell, making it increasingly feasible to sequence the entire human transcriptome.

27 collections contained data for IPF cases and so were suitable for further consideration (Table 5.1 and Additional Table B.5.2). Some of these collections included gene expression data collected from several different types of lung tissue, including whole lung, lung fibroblasts, epithelial cells and lung squamous cell carcinoma. However, for a tissue/cell type to be chosen as the optimal choice, multiple collections

of data from that tissue/cell type must have included healthy controls. This left only blood, whole lung and lung fibroblasts in consideration. The total number of IPF cases in the collections with lung fibroblasts (n=35) was deemed too low to fully capture the disease heterogeneity and this cell type was excluded. At this stage, lung tissue was considered a preferable choice of tissue type to blood, as it is more relevant to IPF, a disease of the lungs. Additionally, the cell-types that are most relevant to IPF, such as epithelia and fibroblasts, should have been represented within whole-lung samples. Two of the 11 lung collections had fewer than 10 IPF cases so these studies were excluded.

TABLE 5.1: *A summary by tissue type for the 27 collections on the Gene Expression Omnibus that contained gene expression data for individuals with IPF with at least 30 samples. The table is stratified by whether each collection included healthy control data.*

Human tissue type	Number of collections with controls (number of IPF cases)	Number of collections without controls (number of IPF cases)	Total IPF cases
Alveolar macrophages	1 (15)	0 (0)	15
Blood	4 (295)	4 (245)	540
Bronchoalveolar lavage	1 (176)	0 (0)	176
Epithelial cells	1 (325)	0 (0)	325
Whole lung	9 (318)	2 (19)	337
Lung fibroblasts	3 (22)	1 (13)	35
Lung squamous cell carcinoma	1 (10)	0 (0)	10

Clinical and demographic traits were not widely reported in the collections of lung tissue data that had passed the inclusion criteria (Table 5.2), with many collections not having any meta data to accompany the gene expression data on GEO. Sex was the most commonly reported trait and yet was only available for three of the nine studies. Despite lung tissue being the preferred tissue type, the lack of commonly reported traits across cohorts meant that it would not have been feasible to perform the analysis using the lung collections without additional data being made available by the study authors. The study authors of the lung collections were contacted in March 2020 to request additional clinical and demographic data that they would be willing to share.

TABLE 5.2: *Clinical and demographic traits that were reported in at least one of the lung tissue data collections, and their availability across collections. The ✔ symbol indicates that the trait was reported in that collection and the ✗ symbol indicates that the trait was not reported in that collection. St George's score = total score on the St George's Respiratory Questionnaire for IPF*[246]*, FVC=Forced vital capacity, D_{LCO} = Diffusing capacity for carbon monoxide.*

	Collections with control data							**Collections without control data**	
Trait	GSE10667	GSE124685	GSE134692	GSE110147	GSE92592	GSE53845	GSE32537	GSE48149	GSE24988
Age	✗	✗	✔	✗	✗	✗	✔	✗	✗
Ancestry	✗	✗	✔	✗	✗	✗	✗	✗	✗
Sex	✗	✗	✔	✗	✗	✔	✔	✗	✗
Smoking status	✗	✗	✔	✗	✗	✗	✔	✗	✗
Height	✗	✗	✔	✗	✗	✗	✗	✗	✗
Weight	✗	✗	✔	✗	✗	✗	✗	✗	✗
Smoking pack years	✗	✗	✗	✗	✗	✗	✔	✗	✗
St George's score	✗	✗	✗	✗	✗	✗	✔	✗	✗
Predicted FVC	✗	✗	✗	✗	✗	✗	✔	✗	✗
Predicted D_{LCO}	✗	✗	✗	✗	✗	✗	✔	✗	✗

Clinical and demographic traits were more widely reported in the collections that contained transcriptomic data from blood (Table 5.3). Age and sex were available for the individuals in all eight of the blood cohorts and there were also several other traits that were sufficiently commonly reported across studies to be used for validation, including ancestry; forced vital capacity (FVC); diffusing capacity for carbon monoxide (D_{LCO}); survival time; forced expiratory volume in one second (FEV_1) and the Gender, Age and Physiology (GAP) index for IPF mortality[193]. Genotype information for the *MUC5B* promoter variant rs35705950 (the strongest genetic risk factor for IPF) was also available for some individuals in collections GSE33566, GSE93606 and GSE132607. As a result, blood was selected as the optimal tissue/cell type for the analysis. The rationale was that the analysis could be repeated in lung tissue once sufficient clinical data had been collected, or that the lung tissue datasets could possibly be used to validate any findings that were discovered from performing the analysis in blood.

It was discovered that two of the blood collections, GSE132607 (n=74) and GSE85268 (n=68), both contained subjects from the Correlating Outcomes With Biochemical Markers to Estimate Time-progression in Idiopathic Pulmonary Fibrosis (COMET) study (ClinicalTrials.gov identifier: NCT01071707). Both collections had reported the COMET identification numbers of their subjects as meta data on GEO. Using these, it was found that there were 59 IPF subjects in common between the two cohorts. Both studies were performed using the Affymetrix Human Gene Expression Array by the same team of researchers. However, as they were not independent, GSE85268 was excluded as it was the collection with fewer IPF subjects and fewer clinical traits in common with the other blood collections. This left seven collections, four that included healthy control samples and three that did not.

The seven remaining collections of data were uploaded by research groups from across the USA (including the University of Virginia, Yale University, the University of Nevada and the University of Colorado) and the UK (Imperial College London). GSE27957 and GSE28042 were uploaded by the Kaminski Lab in Yale. These two collections were both used in the same study[185], where GSE27957 was used as discovery data and GSE28042 was used as independent replication data. Similarly, the data found in GSE133298 and GSE132607 were uploaded by researchers at the University of Virginia and were used as independent cohorts in the same study (unpublished as of April 2022, both collections uploaded to GEO in September 2019). All remaining collections were uploaded by separate research groups and no additional evidence of common individuals across cohorts was found so the seven cohorts of IPF subjects were deemed independent. However, the possibility that subjects could be common in two or more studies cannot be ruled out. The collection GSE133298 was removed during the validation stage (Section 5.10).

TABLE 5.3: *Clinical and demographic traits that were reported in at least one of the blood data collections, and their availability across collections. The ✔ symbol indicates that the trait was reported in that collection and the ✗ symbol indicates that the trait was not reported for that collection. FVC=Forced vital capacity, D_{LCO} = Diffusing capacity for carbon monoxide, GAP index = the Gender, Age and Physiolic index for IPF mortality*[187].

	Collections with healthy control data				**Collections without healthy control data**			
Trait	GSE38958	GSE33566	GSE93606	GSE28042	GSE133298	GSE132607	GSE85268	GSE27957
Age	✔	✔	✔	✔	✔	✔	✔	✔
Ancestry	✔	✗	✗	✔	✔	✔	✗	✔
Sex	✔	✔	✔	✔	✔	✔	✔	✔
Smoking status	✗	✗	✔	✗	✔	✔	✗	✗
FVC	✔	✔	✔	✔	✔	✔	✗	✔
D_{LCO}	✔	✔	✔	✔	✔	✔	✗	✔
Height	✗	✗	✗	✗	✗	✗	✔	✗
Weight	✗	✗	✗	✗	✗	✗	✔	✗
Survival time	✗	✗	✔	✔	✗	✗	✗	✔
GAP index	✔	✔	✔	✔	✔	✔	✗	✔
FEV_1	✗	✗	✔	✔	✗	✗	✗	✗

5.4 Discovery stage studies

The datasets were then divided into those that would be used in the discovery stage and those that would be used in the validation stage. Cohorts used in the discovery analysis were required to have included healthy controls in order to enable the gene expression co-normalisation. Of the four cohorts that included healthy control subjects, the three with the greatest number of controls were selected to use in the discovery stage, as theoretically the co-normalisation would have worked more effectively with studies with a greater number of healthy controls. The fourth study was saved for use in the validation stage to boost the statistical power of that part of the analysis.

The first cohort used in the discovery analysis, with GEO accession code GSE38958, came from an American observational study[194] that was investigating the relationship between sphingosine-1-phosphate lyase and pulmonary fibrosis. To this end, the authors studied gene expression data from peripheral blood mononuclear cells of IPF subjects (n=70) and compared this to gene expression from healthy controls (n=45). IPF cases were recruited from the University of Chicago.

The second gene expression dataset, GSE33566, contained data for 123 IPF subjects and 30 healthy controls. The IPF cases were recruited through the ILD or the FPF Programs conducted at National Jewish Health and Duke University. A subset of this data was used in an American observational study[195], where the authors hypothesised that a peripheral blood biomarker for IPF would be able to identify the disease in its early stages and allow for disease progression to be monitored. In the study, 40 IPF cases were split into groups based on their predicted FVC and D_{LCO}, then the authors looked for differentially expressed genes between groups. In addition, these groups were compared to a group of 27 non-diseased family members who acted as age and sex matched healthy controls. A larger subset of the GSE33566 dataset (89 IPF cases and 26 healthy controls) was later used in a study that aimed to develop a gene signature with the ability to diagnose IPF[196].

The third collection was GSE93606, which contained data from a subset of participants in the PROFILE study[112] (n=57 IPF subjects and n=20 healthy age, sex and smoking history matched controls). The transcriptomic data in GSE93606 is from a study that had the objective of examining host-microbial interactions in IPF subjects over time[197]. In this study, gene expression data from peripheral blood and lung function measurements were collected at multiple time points. However, in the cluster analysis described within this thesis, only the gene expression/lung function data that had been collected at baseline was used. IPF patient survival was also recorded up to a maximum follow-up time of 34 months (Additional Figure A.5.1). More than 50% of subjects were still alive at the end of the study, so a median survival time could not be calculated.

In each study, the average ages of the IPF cases and control subjects were both between 60 and 70 years (Table 5.4). The proportion of males in each disease status group varied more across cohorts, being most different in GSE38958, where the IPF group was comprised of 83% males and the control group

was only 60% males. In each cohort, the lung function measurements percent predicted FVC and D_{LCO} were reported for the IPF cases but not the control subjects. These also varied greatly between cohorts. Curiously, the IPF subjects in GSE93606 had the highest mean percent predicted FVC (suggesting healthier individuals) but the lowest percent predicted D_{LCO} (suggesting unhealthier individuals). One limitation of combining data from separate studies is that different studies, particularly those originating from different countries, may have inconsistent procedures for data collection and different study recruitment criteria, which could partly explain the conflicting lung function statistics.

All three studies were microarray-based (Table 5.5) and each dataset came with a corresponding gene mapping file. In each dataset, probes that did not map to a gene were removed. In the instance where multiple probes mapped to the same gene, only the probe with the greatest mean expression was included in the analysis. Each dataset was then quantile normalised to reduce any technical differences between the gene probes within a study. Following this, each dataset was $\log_2$-scaled so that all expression data was in a consistent form prior to co-normalisation. Genes were matched across studies based on their gene symbols, giving 9,371 that were commonly measured across all three studies.

TABLE 5.4: *Summary statistics for the IPF and control subjects in each of the discovery stage studies.*

	GSE38958		**GSE33566**		**GSE93606**	
Study reference	[194]		[195]		[197]	
Country	USA		USA		UK	
Disease status	**IPF**	**Control**	**IPF**	**Control**	**IPF**	**Control**
Sample size	70	45	93	30	57	20
Age (years, sd)	68.2 (7.2)	69.3 (9.3)	67.2 (11.4)	62.4 (14.3)	67.4 (8.0)	66.0 (10.6)
Sex (% male)	82.6%	60.0%	65.6%	46.7%	66.7%	60.0%
Ancestry (% European)	82.8%	71.1%	Unknown	Unknown	Unknown	Unknown
FVC % predicted (sd)	62.4 (15.0)	Unknown	62.0 (28.8)	Unknown	72.2 (20.3)	Unknown
D_{LCO} % predicted (sd)	43.3 (18.7)	Unknown	52.1 (27.9)	Unknown	39.2 (14.1)	Unknown
Mortality (%)	Unknown	Unknown	Unknown	Unknown	40.4%	Unknown

TABLE 5.5: *Information about the transcriptomic data in the discovery datasets and the platform used in each study.*

	GSE38958	**GSE33566**	**GSE93606**
Microarray platform	Affymetrix Human Exon 1.0 ST Array	Agilent-014850 Whole Human Genome Microarray	Affymetrix Human Gene 1.1 ST Array
Number of gene probes	44,280	32,850	33,297
Number of unique genes	17,256	12,171	20,254

5.5 Data co-normalisation

5.5.1 Methods

COmbat CO-Normalization Using conTrols (COCONUT) was used to reduce the technical differences between the three discovery transcriptomic datasets, therefore enabling a cluster analysis to be performed on the pooled, co-normalized data. COCONUT is an unbiased co-normalisation method which assumes that all healthy controls across studies come from the same statistical distribution. It uses the healthy controls in each study to calculate correction factors that remove the technical differences in the data for the disease cases, without introducing bias based on the number of disease cases present. The method is adapted from the ComBat empiric Bayes normalization method[188], which is often used to adjust for batch effects within a study.

Data for each study was input into COCONUT by providing a gene expression matrix of common genes against subjects. These were accompanied by an indicator variable that specified which individuals were cases and which were controls. This was performed in R v.4.0.0 with the 'COCONUT' package v.1.0.2. The procedure used by COCONUT is described below in three main steps.

Step 1: Standardize the healthy control data

First, the data for the disease cases is separated from the healthy control subject data across all studies. Steps 1 and 2 use only the data for the healthy control subjects, and step 3 uses only the data from the disease cases. The gene expression data for each study should be normalised and on the same scale (e.g. the $\log_2$ scale). The expression value for gene g for sample j from study i is modelled as:

$$Y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\varepsilon_{ijg}. \quad (5.1)$$

Where α_g represents the overall gene expression for gene g, X is a design matrix for sample conditions, β_g is the vector of regression coefficients corresponding to X and ε_{ijg} are the error terms, which are assumed to follow $\varepsilon_{ijg} \sim N(0, \sigma_g^2)$. γ_{ig} and δ_{ig} are the additive and multiplicative technical effects of study i for gene g, respectively.

The data is then standardised gene-wise so that all genes have similar overall mean and variance. This is done by first estimating the variance of gene g, as follows:

$$\hat{\sigma}_g^2 = \frac{1}{N}\sum_{ij}\left(Y_{ijg} - \hat{\alpha}_g - X\hat{\beta}_g - \hat{\gamma}_{ig}\right)^2. \quad (5.2)$$

Where N is the total number of samples. The standardised data, Z_{ijg}, are then calculated as:

$$Z_{ijg} = \frac{Y_{ijg} - \hat{\alpha}_g - X\hat{\beta}_g}{\hat{\sigma}_g}. \quad (5.3)$$

Step 2: Obtain technical effect parameter estimates using parametric empirical priors

It can be assumed that $Z_{ijg} \sim N(\gamma_{ig}, {\delta_{ig}}^2)$, but note that the γ_{ig} parameters are not the same as those in Equation 4.1. Additionally, the parametric forms for prior distributions on the technical effect parameters are assumed to be: $\gamma_{ig} \sim N(\gamma_i, \tau_i^2)$ and ${\delta_{ig}}^2 \sim$ Inverse Gamma (λ_i, θ_i). The hyperparameters $\gamma_i, \tau_i^2, \lambda_i$ and θ_i are then estimated empirically from the standardized data using the method of moments.

The technical effect parameters γ_{ig} and ${\delta_{ig}}^2$ are then estimated using the following equations:

$$ {\gamma_{ig}}^* = \frac{n_i\, \bar{\tau}_i^2 \hat{\gamma}_{ig} + {\delta_{ig}}^{2*} \bar{\gamma}_i}{n_i\, \bar{\tau}_i^2 + {\delta_{ig}}^{2*}} \qquad \text{and} \qquad {\delta_{ig}}^{2*} = \frac{\bar{\theta}_i + \frac{1}{2}\sum_j (Z_{ijg} - {\gamma_{ig}}^*)^2}{\frac{n_j}{2} + \bar{\lambda}_i - 1}. \tag{5.4} $$

Where n_i is the number of studies and n_j is the number of healthy control samples.

Step 3: Adjust the diseased samples data for the technical differences between studies

With the estimated parameters, ${\gamma_{ig}}^*$ and ${\delta_{ig}}^{2*}$, now obtained for the control samples, the adjustment is then applied to the data for the corresponding disease cases, D_{ijg}. This forces the disease components of all cohorts to be from the same background distribution, while retaining their relative distance from the control component. The adjusted disease cases data, ${D_{ijg}}^*$, is obtained as follows:

$$ {D_{ijg}}^* = \frac{\hat{\sigma}_g}{\hat{\delta}_{ig}^*} (D_{ijg} - {\hat{\gamma}_{ig}}^*) + \hat{\alpha}_g + X\hat{\beta}_g \tag{5.5} $$

After the application of COCONUT, the efficacy of the co-normalisation was visualised using plots of the first two principal components of the gene expression data. PCA was performed in R v4.0.0.

5.5.2 Results

COCONUT was applied to the total discovery dataset consisting of 220 IPF cases and 90 healthy control subjects from the three studies. After this, all control subjects were removed from further analysis. Prior to COCONUT co-normalisation (Figure 5.3A), the data from the three cohorts were entirely separated in high-dimensional space due to technical differences between the studies. Therefore, clustering this data would simply recapitulate the study clusters. Whereas after COCONUT (Figure 5.3B), the data for the IPF cases from the three studies were overlapping in high-dimensional space, indicating that the technical differences between datasets had been reduced and that the pooled dataset was now suitable for clustering. However, there remained a degree of separation between the blue and yellow points in Figure 5.3B, suggesting that there may still be technical differences present between the co-normalised datasets GSE38958 and GSE93606.

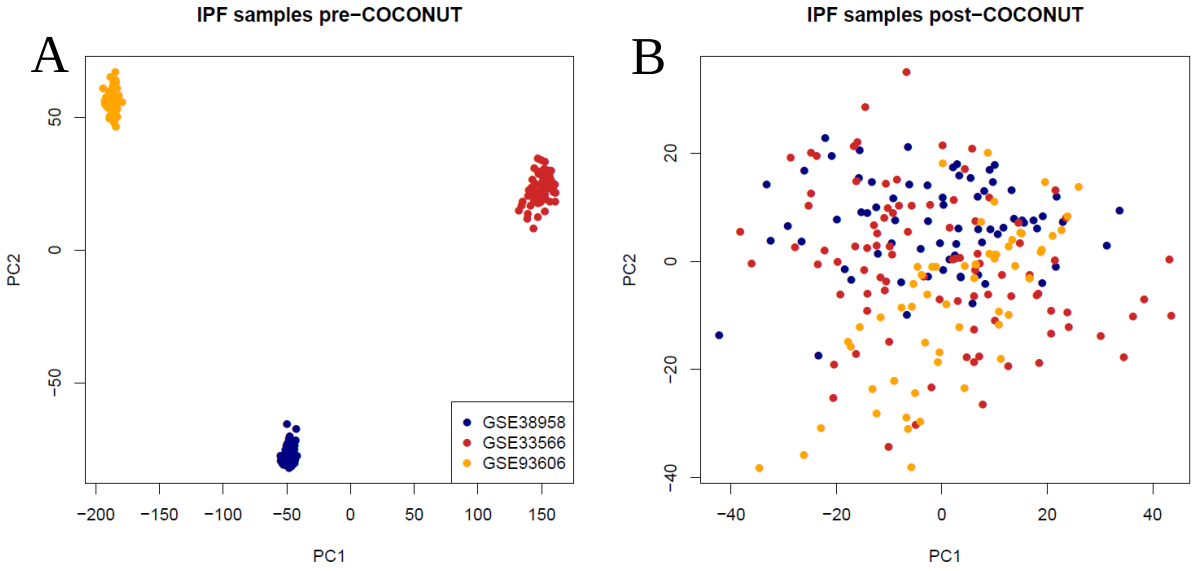


FIGURE 5.3: *Plots of the first two principal components of the gene expression data for the IPF samples from the three studies, before (A) and after (B) COCONUT co-normalisation. PC1 = the first principal component of the data, PC2 = the second principal component of the data.*

5.6 Clustering

5.6.1 Methods

Applying a single clustering algorithm (e.g. hierarchical clustering) for a chosen validity measure (e.g. connectivity) can often yield unstable, non-reproducible results[198]. Therefore, methods that attempt to find more robust solutions are becoming more popular, such as consensus clustering. Consensus clustering is an approach where multiple iterations of the chosen clustering method are performed on sub-samples of the dataset.

To identify the optimal number of clusters in the pooled transcriptomic data, the Combined Mapping of Multiple clUsteriNg ALgorithms (COMMUNAL)[191] approach was used. COMMUNAL is a method of unsupervised clustering that integrates data from multiple clustering algorithms, across a range of input variables (in this case, genes) and evaluates the validity of each number of clusters using multiple validity measures. COMMUNAL then outputs a 3-dimensional (3D) map that can be used to select the optimal number of clusters in the data, as well as the optimal number of genes to use in the clustering.

In this study, COMMUNAL was run using consensus clustering versions of two algorithms, K-means clustering and Partitioning Around Medoids (PAM). K-means clustering was selected for its efficiency in large datasets[199] and PAM, also known as k-medoids clustering, was selected as it is robust to noisy datasets and outliers[200]. The two methods are similar; both are iterative and attempt to split the data into groups by minimising the distance between the points labelled to be in a cluster and the centre point of that cluster. In k-means clustering this centre point is the average between the points in that cluster, whilst in PAM a data point (labelled a medoid) is chosen as the centre point. Both methods are unsupervised, although the number of clusters must be specified beforehand.

Five different metrics were used to assess the validity of the clustering for different numbers of clusters and genes. These were chosen automatically by COMMUNAL. All validity measures were standardized so that they were on a comparable scale, then the mean of all standardised validity measures was used to decide the optimal clustering. The first validity metric was the gap statistic[201], which is the most widely used method for determining the optimal number of clusters in applied statistics and has been shown to perform well with gene expression data[202]. The gap statistic compares the total within-cluster variation for different values of K (the number of clusters) with their expected values under an appropriate null reference distribution. The estimate of the optimal number of clusters will be the value that maximizes the gap statistic, which indicates that the clustering structure is the most different from a random uniform distribution of points.

The second validity measure used in this study was connectivity[203]. This metric indicates the degree of connectedness of the clusters and has a value between zero and infinity. In a system where the clustering has been effective, the clusters will be clearly separated and thus the connectivity will be minimised. The third validity metric was the average silhouette width[204]. The silhouette width coefficient ranges from -1 to 1 and indicates how well each data point lies within its cluster. A value close to 1 for a data point implies that it is a part of the correct cluster, whereas a value close to -1 means that the data point is assigned to the wrong cluster. Therefore, a high average silhouette width across all of the data points indicates that the data has been well-clustered.

The fourth validity measure was the G3 metric[205], which compares dissimilarity across clusters and will be minimised in a good clustering assignment. The final measure used to assess the validity of the clustering was Pearson's gamma coefficient, which is a normalised version of Hubert's gamma statistic[206], where values lie between -1 and 1 and values close to 1 indicate a strong clustering.

Figure 5.4 illustrates how these five validity measures vary for clustering assignments of different quality. The three plots in Figure 5.4 each show an example of a clustering assignment for omics data from the same 13 individuals, where clustering assignment 1 is very poor, clustering assignment 2 is acceptable and clustering assignment 3 is excellent. For clustering assignment 1, the worst assignment, we see that the connectivity and G3 metric are at their greatest whilst the average silhouette width and Pearson's gamma coefficient are at their lowest. As the clustering assignments improve, the connectivity and G3 metric decrease towards 0 whilst the average silhouette width and Pearson's gamma coefficient increase towards 1. The gap statistic only varies depending on the number of clusters and not the quality of a particular assignment, so it is the same for clustering assignment 1 and 2 (both of which have 4 clusters) but is greater in cluster assignment 3 (2 clusters).

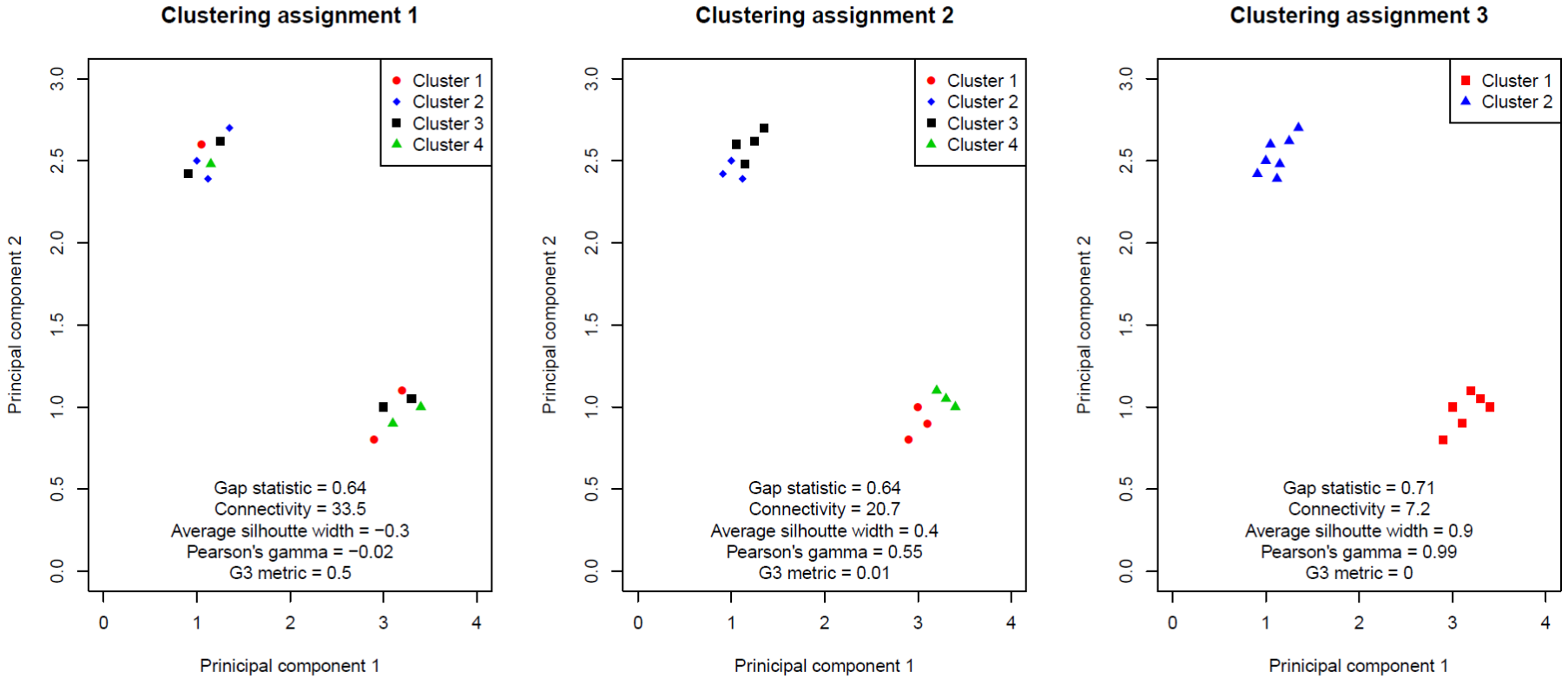


FIGURE 5.4: *Plots illustrating how the five validity measures used in the COMMUNAL clustering vary for different quality clustering assignments. Each plot shows the first two principal components of some example omics data for 13 subjects. Clustering assignment 1 is an example of a very poor clustering assignment, whilst clustering assignment 2 is acceptable and clustering assignment 3 is excellent.*

The COMMUNAL R package outputs a 3D map, which allows the user to choose the best clustering assignment for the data. The map shows the mean of standardized values of each validity measure across the entire tested space. On the 3D map, blue squares indicate a potentially optimal clustering at a certain number of genes by finding the assignment where the mean combined validation metric is greatest. The absolute maximum K for any consensus subset is marked with a red square. The points where the blue and red squares overlap indicate stable optima. If stable optima at K clusters are seen over most of the tested space, this indicates the presence of a strong, consistent biological signal at this number of clusters. Once the optimal number of clusters had been selected, the optimal number of genes was then chosen as the lowest number of genes for which there were stable optima at this number of clusters, in order to minimise the amount of noise or redundant signal.

Before applying COMMUNAL to the pooled data, the genes were ranked in order of variance, with the 'top' 100 genes referring to the 100 genes with the greatest variance. The COMMUNAL algorithm was then applied (in R v.3.4.0 with the 'COMMUNAL' package v.1.1.0), using a range of input genes, from the top 100 to the top 5,000. The genes with the greatest variance were used as those were the most likely to be informative, so as to minimise the number of non-informative genes and increase the signal-to-noise ratio.

PCA and heatmaps (using R v.3.4.0 and the 'gplots' v.3.1.3 package) were used to visualise the separation of the clusters from the optimal cluster assignment in high-dimensional space. Heatmaps are a data visualisation technique where numerical values (which in this case are the levels of gene expression for each gene for each clustered individual) are represented by different colours. The samples in a heatmap are often ordered using hierarchical clustering for visualisation purposes, as this groups samples with similar transcriptomic profiles closely together.

5.6.2 Results

COMMUNAL was applied to the co-normalised data for a range of genes from the top 100 to the top 5,000. In the resulting optimality map (Figure 5.5) there are stable optima at K=4 from 250 genes to 1,000 genes, and at K=3 from 2,500 genes to 5,000 genes, as shown by the red and blue squares meeting. Despite the K=4 clustering assignment at 1,000 genes showing the highest mean standardized validity score of all tested clustering assignments, there were stable optima at K=3 clusters over a larger range of tested space, indicating a stronger biological signal. As such, K=3 was chosen as the optimal number of clusters in the pooled IPF dataset.

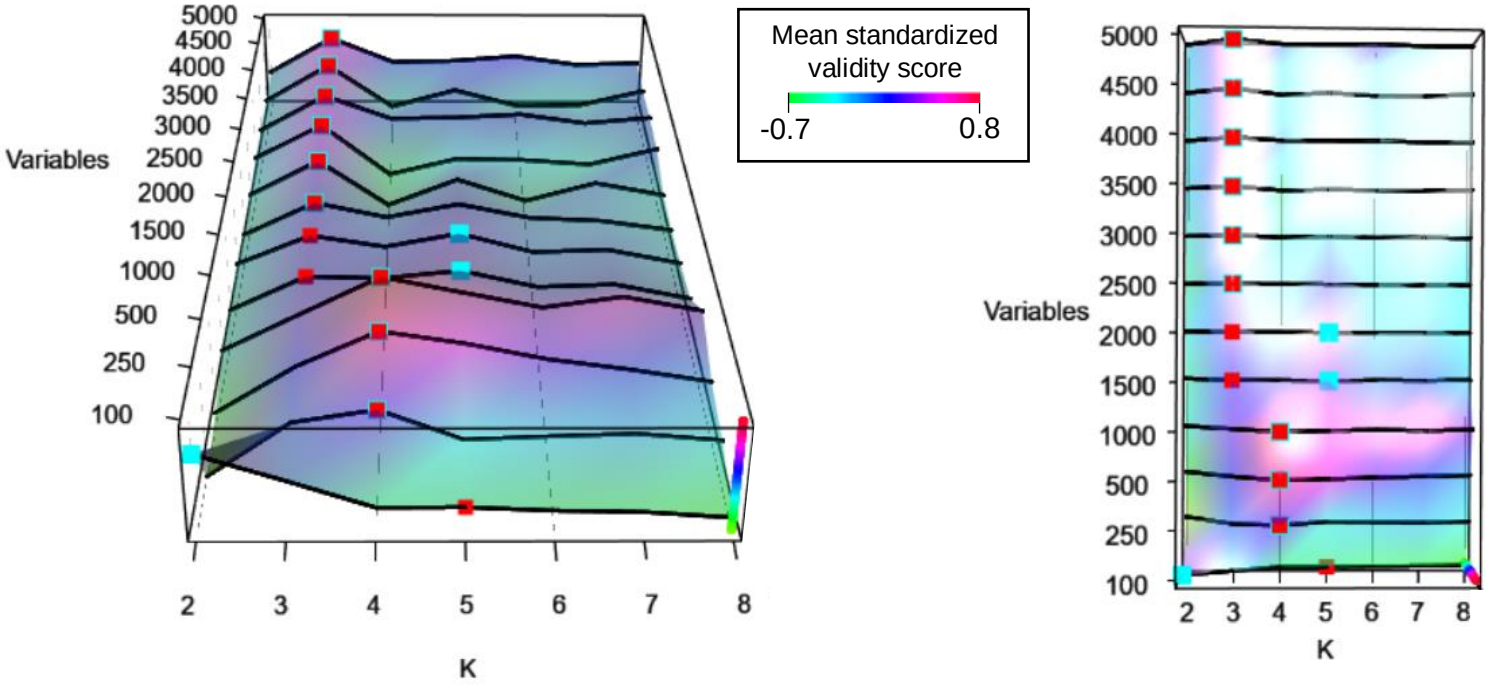


FIGURE 5.5: *The 3D optimality map produced by COMMUNAL to identify the most robust number of clusters in the co-normalised data. The map shows the mean of standardized values of each validity measure across the entire tested space. On the 3D map, blue squares indicate a potentially optimal clustering at a certain number of genes by finding the assignment where the mean combined validation metric is greatest. The absolute maximum K for any consensus subset is marked with a red square. A higher validity score indicates a better clustering assignment and stable optima are the points where the blue and red squares meet. If stable optima at K clusters are seen over most of the tested space, this indicates the presence of a strong, consistent biological signal at this number of clusters.*

The clustering at 2,500 genes (and K=3 clusters) was chosen as the optimal clustering assignment, under the assumption that using the fewest number of genes has the least amount of redundant signal. At this clustering assignment, 64 IPF cases were assigned to Cluster 1 (termed the red cluster), 95 were assigned to Cluster 2 (the blue cluster) and 37 were assigned to Cluster 3 (the yellow cluster). 24 individuals (10.4%) were not assigned into the same cluster by the two algorithms and were instead labelled 'unclustered'.

The three clusters were clearly separated in high-dimensional space (Figure 5.6) and the unclustered samples generally laid at the boundaries between the three clusters, which could represent samples that could not be perfectly assigned to a given cluster. Since the intention was to use the clustered data to create a gene-expression based cluster classifier, and classifiers trained on data with fewer errors are more robust, these uncertain samples were removed from further analysis to improve the accuracy of the classifier.

The clear separation of the clusters can also be seen when looking at a heatmap of gene expression for the clustered samples across the top 2,500 genes (Figure 5.7A). This heatmap shows that individuals who were assigned to the same cluster but originated from different studies showed differences in their gene expression profiles, particularly those in cluster 2. This suggests that some technical differences were still present in the co-normalised dataset. However, it is clear that the clustering assignment has not simply recapitulated the groups of subjects by study as Clusters 2 and 3 contain individuals from all three studies, while cluster 1 contained individuals from both GSE38958 and GSE33566.

When the IPF samples in the heatmap were ordered using hierarchical clustering (Figure 5.7B), the separation of the three clusters was mostly maintained, showing that the clusters were robust to another, independent method of clustering to the two used by COMMUNAL. This also showed that in terms of gene expression, Clusters 1 and 3 (the red and yellow clusters) were the most similar and Cluster 2 was most different to the other two.

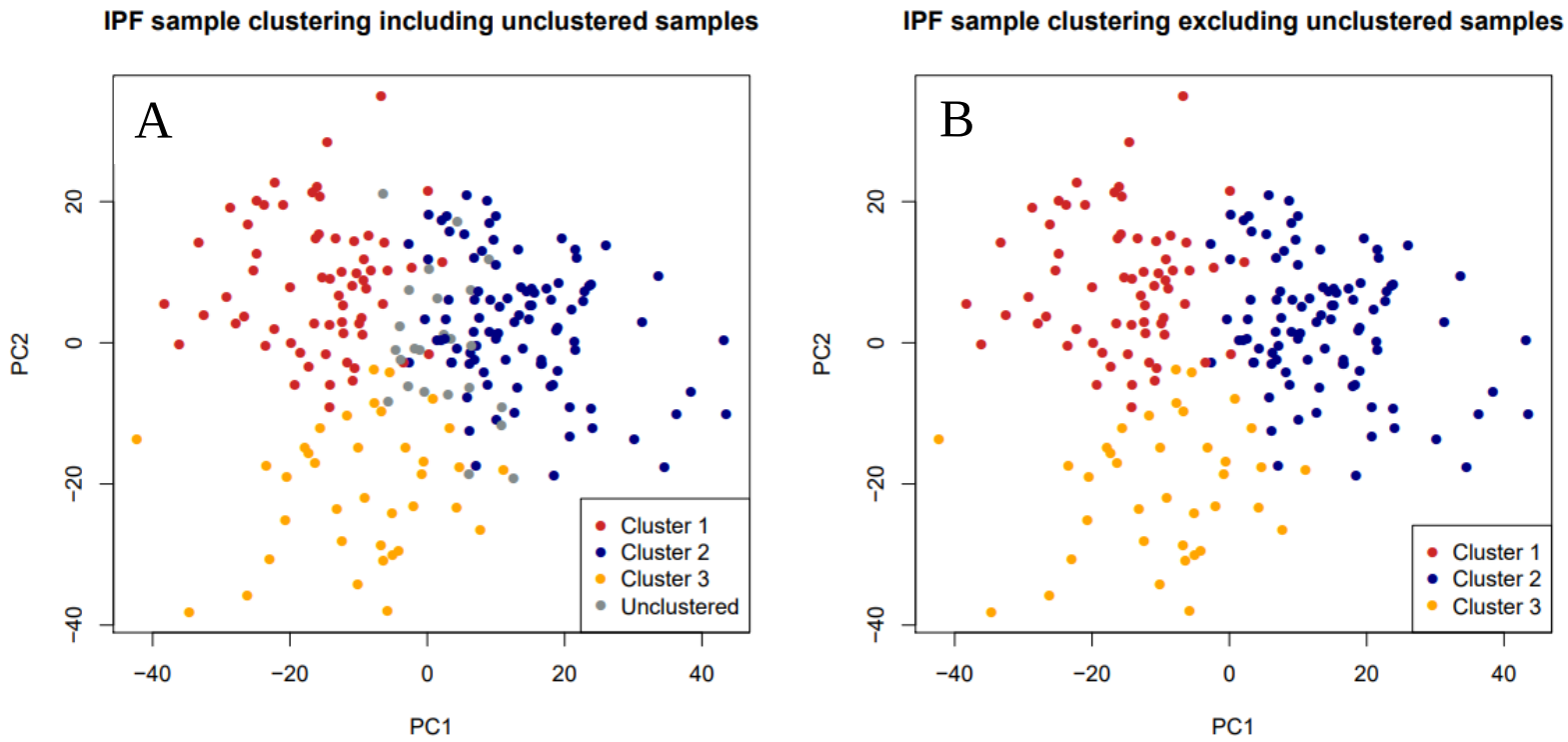


FIGURE 5.6: *Plots of the first two principal components of the co-normalised gene expression data, both with (A) and without (B) the 10.4% of samples which were unclustered, showing that the clusters were clearly separated in high-dimensional space.*

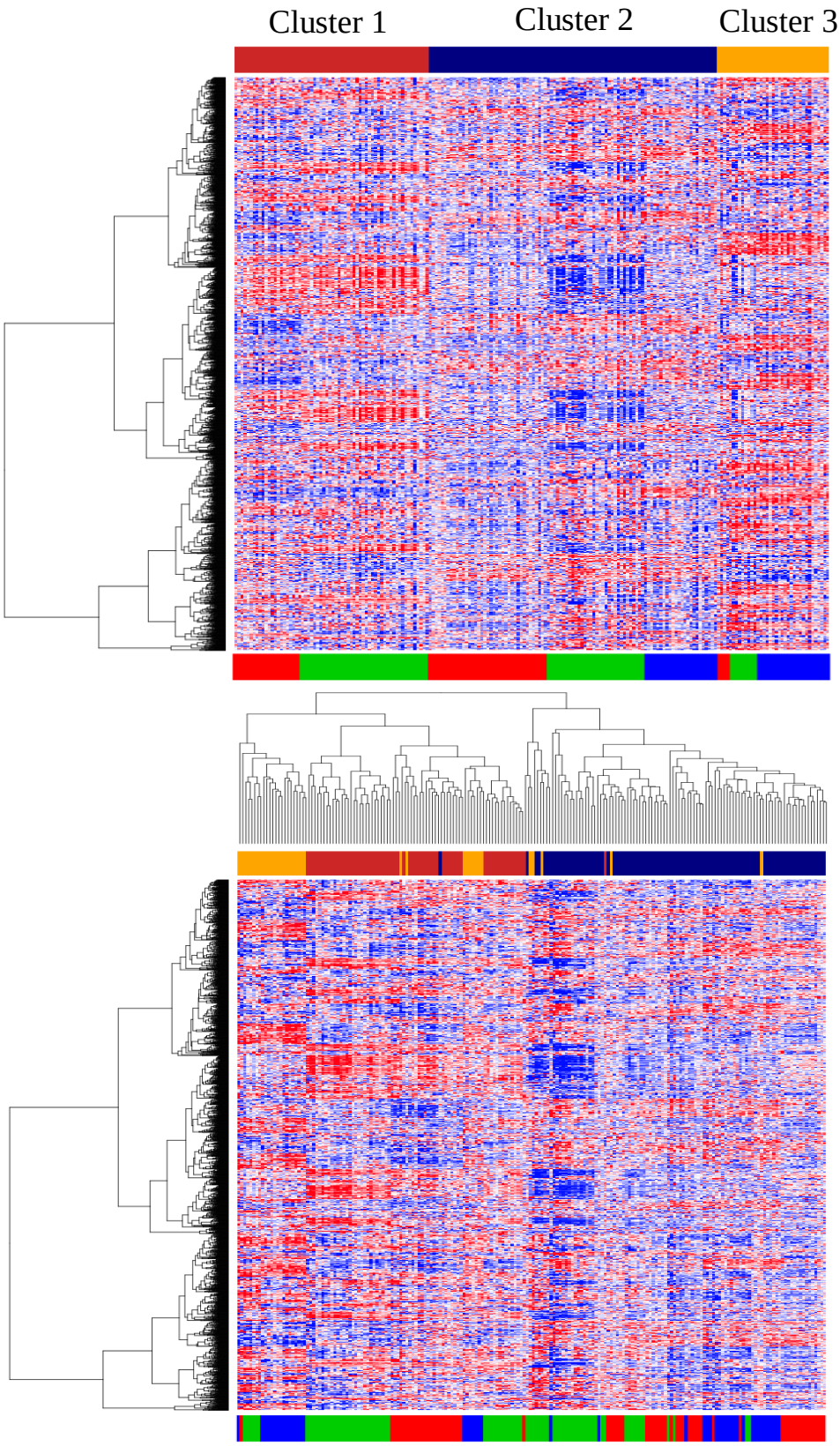


FIGURE 5.7: *Heatmaps of gene expression for the clustered samples (x-axis) across the top 2,500 genes (y-axis), without hierarchical clustering of the samples (A) and with hierarchical clustering of the samples (B). Blue inside the heatmap indicates low expression and red indicates high expression. In both plots, the genes have been hierarchically clustered for presentation purposes, the bar above the plot shows the cluster that subject was assigned in to (red = cluster 1, blue = cluster 2 and yellow = cluster 3) and the bar below the plot indicates which original study the subject was in (red = GSE38958, green = GSE33566 and blue = GSE93606).*

5.7 Comparison of clinical traits across clusters

5.7.1 Methods

The next step was to characterize the clusters by comparing the clinical and demographic traits of the individuals with IPF in each cluster. This was done for each trait that was reported in at least one discovery cohort and one validation cohort. Histograms were plotted for each continuous variable and stratified by cluster to assess whether each variable represented a normal or skewed distribution. The statistical significance of the phenotypic differences across clusters was evaluated for all studies combined using a chi-square test for count data, an analysis of variance to compare means for non-skewed continuous data and a Kruskal-Wallis rank-sum test[207] to compare medians for skewed continuous data. For traits in the form of time-to-event data, Kaplan-Meier curves (Section 2.3) were used to approximate and visualise the survival function for these variables. Further, Cox PH models (Section 2.3) were fit with cluster as the sole independent variable and the time to the event as the response variable. Scaled Schoenfeld residuals (Section 2.3) were used to evaluate these models for breaches of the PH assumption.

5.7.2 Results

The following traits were evaluated: age, sex, ancestry (subjects were classed as having European or non-European ancestry), smoking status (ex-smokers and current smokers were defined as 'ever smokers' and compared to non-smokers), predicted FVC, predicted D_{LCO}, predicted FEV_1, GAP index and genotype for the *MUC5B* promoter variant rs35705950.

One study (GSE93606) reported lung function data at multiple time points as well as patient survival information in the form of time-to-event data, including right-censored observations (which is when an individual drops out of the study or the study ends before the event of interest occurs). In this study, only baseline measurements of the lung function variables from GSE93606 were used. The proportion of subjects in GSE93606 who were observed to have died during the study was compared across clusters and a Cox PH model was used to evaluate differences in survival time across clusters.

Table 5.6 shows the summary statistics for the subjects in each cluster by study, as well as for all studies combined. As there were no individuals from study GSE93606 who were assigned to the red cluster, smoking history, predicted FVC and survival information could not be assessed for the individuals in that cluster. However, with all studies combined, statistically significant differences in predicted D_{LCO} were observed across clusters (P=0.009). In all three studies, individuals in the blue cluster had the greatest median predicted D_{LCO} (indicating relatively good lung function), whilst the subjects in the yellow cluster had the lowest. Additionally, there was a significant difference in average score from the GAP index for IPF mortality (P=0.006), with those in the red cluster having the greatest GAP score (indicating those predicted to be at a higher risk of mortality) and those in the blue cluster having the

lowest average GAP score. Individuals in the blue cluster had a higher average predicted FVC and FEV_1 than those in the yellow cluster, although these differences were not statistically significant.

Additionally, there was a statistically significant difference in survival across clusters, with death observed for 25% of subjects in the blue cluster and 67% of subjects in the yellow cluster (P=0.009) during the nearly 3-year follow-up of study GSE93606. A Kaplan-Meier plot (Figure 5.8) showed that survival over time was consistently poorer over time for those in the yellow cluster compared with those in the blue cluster. The median survival time for subjects in the yellow cluster was approximately 1 year, whilst the median survival time for those in the blue cluster was greater than 33 months, though this could not be directly calculated as more than 50% of subjects were still known to be alive at the end of the study.

The hazard ratio between the blue cluster and yellow cluster from a Cox PH model was 3.59 (95% CI: [1.40, 9.19], P=0.008), which meant that at any follow-up time individuals in the yellow cluster were estimated to be 3.59 times as likely to die as individuals in the blue cluster. There was no evidence that the PH assumption had been broken in this model (Additional Figure A.5.2).

The phenotypic differences that were observed across clusters indicated that on average, the blue cluster contained the healthiest individuals that were at a relatively low risk of mortality, whilst the red and yellow clusters contained less healthy, higher risk individuals. This could be a significant finding; as the clustering was performed independently of clinical data, yet significant differences in lung function and mortality were observed between clusters, these clusters may be representative of distinct and clinically relevant endotypes of IPF.

TABLE 5.6: *Comparison of clinical and demographic traits of clustered subjects from each study, as well as when all studies are combined. Data are presented as count (percentage), mean (standard deviation [sd]) or median (interquartile range [IQR]). NA = data not available, FVC=Forced vital capacity, D_{LCO} = Diffusing capacity for carbon monoxide, FEV_1 = Forced expiratory volume in one second, GAP index = Gender, Age and Physiology index for IPF mortality[193], MUC5B genotype = genotype for the MUC5B promoter polymorphism rs35705950. ǂ: P-value for count data is from a chi-square test, test comparing means is analysis of variance and test comparing medians is the Kruskal-Wallis log rank test. P-values less than 0.05 are highlighted in bold.*

Phenotypic trait	**GSE38958 (n=65)**			**GSE33566 (n=83)**			**GSE93606 (n=48)**			**All studies combined (n=196)**				
	Red cluster	**Blue cluster**	**Yellow cluster**	**Red cluster**	**Blue cluster**	**Yellow cluster**	**Red cluster**	**Blue cluster**	**Yellow cluster**	**Red cluster**	**Blue cluster**	**Yellow cluster**	**P-value**ǂ	**Total n used**
n subjects in cluster	22	39	4	42	32	9	0	24	24	64	95	37		
Age (years) (mean, sd)	70 .0 (6.3)	68.3 (7.9)	64.0 (2.7)	66.7 (9.8)	67.0 (14.1)	67.0 (12.1)	-	64.8 (5.9)	70.3 (8.8)	67.8 (8.9)	66.9 (10.2)	68.8 (9.4)	0.592	188
Male (%)	20 (91.0%)	30 (77.0%)	4 (100%)	32 (76.2%)	21 (65.6%)	3 (33.3%)	-	15 (62.5%)	16 (66.7%)	52 (81.3%)	66 (69.5%)	23 (62.2%)	0.091	196
European ancestry (%)	17 (81.0%)	29 (82.9%)	3 (75.0%)	NA	NA	NA	-	NA	NA	17 (81.0%)	29 (82.9%)	3 (75.0%)	0.883	60
Ever smoker (%)	NA	NA	NA	NA	NA	NA	-	15 (62.5%)	18 (78.3%)	-	15 (62.5%)	18 (78.3%)	0.389	47
Death observed during study (%)	NA	NA	NA	NA	NA	NA	-	6 (25%)	16 (66.7%)	-	6 (25%)	16 (66.7%)	**0.009**	48
FVC % predicted (median, IQR)	59.5 (19.5)	65.0 (24.0)	51.5 (7.8)	77.0 (36.0)	66.0 (46.0)	73.0 (17.5)	-	71.5 (27.7)	60.8 (24.1)	63 (35.0)	70.5 (30.1)	60.1 (23.4)	0.342	154
D_{LCO} % predicted (median, IQR)	34.5 (17.5)	49.0 (21.0)	28.5 (21.0)	65.0 (37.0)	66.0 (40.0)	30.0 (30.0)	-	38.1 (17.1)	36.6 (15.9)	35.0 (30.0)	45.0 (29.2)	34.4 (17.3)	**0.009**	133
FEV_1 % predicted (median, IQR)	NA	NA	NA	NA	NA	NA	-	74.9 (23.1)	65.4 (22.7)	-	74.9 (23.1)	65.4 (22.7)	0.216	48
GAP index (mean, sd)	5·3 (1·3)	3·9 (1·3)	4·5 (1·3)	4·3 (1·5)	4·1 (1·6)	4·3 (3·1)	-	3·7 (1·8)	4·4 (1·6)	4·9 (1·4)	3·9 (1·5)	4·4 (1·7)	**0·006**	132
MUC5B genotype: GG (%)	NA	NA	NA	5 (29·4%)	6 (28·6%)	3 (60·0%)	-	5 (26·3%)	11 (50·0%)	5 (29·4%)	11 (27·5%)	14 (51·9%)	0·230	84
MUC5B genotype: GT (%)	NA	NA	NA	10 (58·8%)	14 (66·7%)	2 (40·0%)	-	12 (63·2%)	8 (36·4%)	10 (58·8%)	26 (65·0%)	10 (37·0%)		
MUC5B genotype: TT (%)	NA	NA	NA	2 (11·8%)	1 (4·8%)	0 (0%)	-	2 (10·5%)	3 (13·6%)	2 (11·8%)	3 (7·5%)	3 (11·1%)		

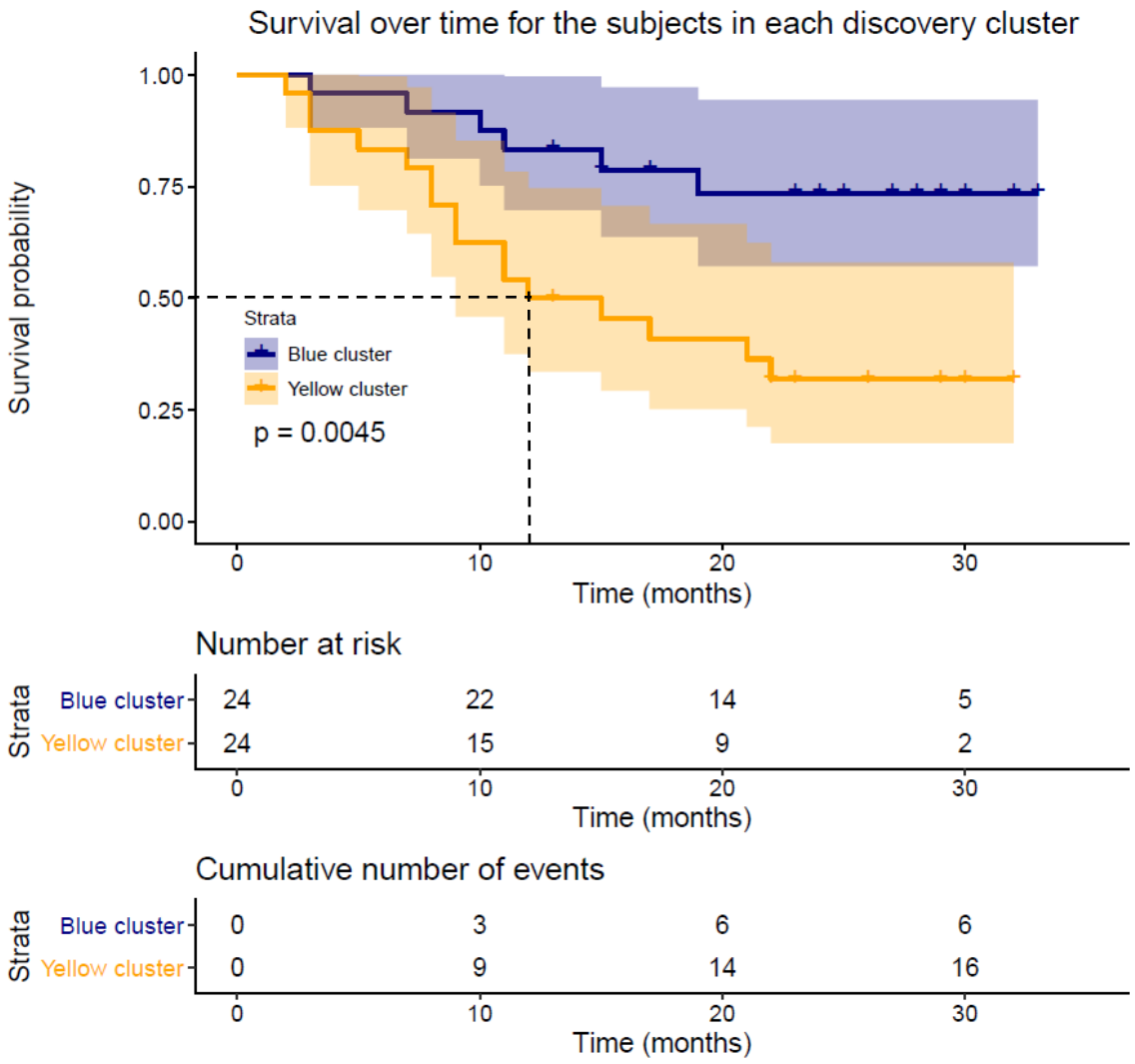


FIGURE 5.8: *Kaplan-Meier curves and corresponding 95% confidence intervals showing survival over time for the subjects from study GSE93606, stratified by the cluster which they were assigned to in this study. The dashed line indicates the median survival time for those in the yellow cluster. The p-value shown on the plot is from a log-rank test testing the two curves for equality.*

5.8 Gene enrichment analysis

5.8.1 Methods

With the clusters having been characterized using clinical information, further characterization using gene enrichment analysis was performed to investigate the biological mechanisms that could be driving the observed differences in lung function and survival between clusters. Enrichment analysis is a method that is often used to search a list of genes to identify classes of genes that are statistically overrepresented. Classes of interest often include biological processes or biological pathways. In enrichment analysis, biological processes consist of broad terms for the purpose of a gene (such as immune response), whilst the biological pathways are more specific and denote that a gene is known to be featured along a particular pathway (such as the proinsulin C-peptide signalling pathway).

First, each of the 2,500 genes used in the optimal COMMUNAL clustering assignment were assigned to the cluster in which its expression was most different to its expression in the other two clusters, as this suggests that that gene is contributing to the identity of that cluster. This was achieved by performing three ANOVA tests for each gene (one for each cluster), each comparing the expression of that gene in subjects within a given cluster against the expression of subjects in both other clusters. Each gene was then assigned to the cluster in which it had the lowest ANOVA p-value. One benefit of this approach is that the ANOVA tests allow for filtering based on statistical significance; a nominal p-value significance threshold of 0.05 was introduced and genes whose lowest ANOVA p-value was greater than this threshold were removed. The rationale for the introduction of this filtering step was that removing genes that were not associated with any cluster would reduce noise and strengthen the gene enrichment analysis for each cluster. The threshold for statistical significance was kept at a nominal level as a correction for all 7,500 ANOVA tests would have likely left too few genes assigned to each cluster to successfully perform the enrichment analysis.

Then, gene enrichment analysis was performed separately on the three resulting gene lists using the R package ‘metabaser’. This was used to search databases of gene ontology terms for statistically overrepresented biological processes and biological pathways. At the time that the analysis was performed, there were 17,552 biological processes and 12,222 biological pathways in the database accessed by metabaser. In gene enrichment analysis, overrepresented terms are given an enrichment score, which is similar to a weighted Kolmogorov-Smirnov statistic[208]. The statistical significance of the term is then determined by comparing the enrichment score to a null distribution. However, as a large number of gene ontology terms are being compared, a correction for multiple testing must be implemented. For instance, metabaser reports ‘q-values’, which are p-values that have been adjusted for multiple tests using the false-discovery rate.

Gene ontology terms with q-value < 0.05 were deemed statistically significant in this analysis. The most significantly enriched biological pathways for each cluster were visualised using Sankey diagrams to

show which genes corresponded to which enriched pathways. Additionally, the gene lists of each cluster were searched for the presence of the nearest gene for any of the 14 variants that were genome-wide significant in the largest GWAS meta-analysis of IPF susceptibility to-date[209]. The 14 genes were as follows: *AKAP13*, *ATP11A*, *DEPTOR*, *DPP9*, *DSP*, *FAM13A*, *LRRC34*, *IVD*, *KIF15*, *MAD1L1*, *MAPT*, *MUC5B*, *TERC* and *TERT*. Following this, enrichment analysis was performed on the genes of each cluster to investigate whether those genes were statistically overconnected (in terms of direct gene regulation) to any of the IPF-associated genes listed above. If the genes that were assigned to a particular cluster were found to be overconnected to one or more of the IPF-associated genes (say the exact number of overconnected IPF-associated genes is N), then a hypergeometric test was performed to approximate the statistical significance of the finding that N out of the 14 IPF-associated genes were present within the list of overconnected genes for that cluster.

5.8.2 Results

Each of the 2,500 genes used in the optimal COMMUNAL cluster classification were assigned into the cluster in which their expression was most different to all other clusters. 814 genes were assigned to the red cluster, 866 to the blue cluster and 820 to the yellow cluster. Genes whose lowest ANOVA p-value was greater than 0.05 were then removed, leaving 769 genes in the red cluster, 839 in the blue cluster and 784 in the yellow cluster. Gene enrichment was then performed on each cluster separately.

Red cluster

Several *biological processes* were significantly enriched in the red cluster, as shown in Table 5.7. The most significantly enriched terms were related to electron transport and cellular respiration. There were 32 *biological pathways* that were significantly enriched for the genes in the red cluster. The 20 most significantly enriched pathways for this cluster are shown in Figure 5.9, which is a Sankey diagram that shows which genes from the red cluster correspond to each of the pathways. The 20 most significantly enriched pathways included cell adhesion extracellular matrix remodelling, which is relevant to IPF as the disease is a result of deposition of extracellular matrix within the lung parenchyma[210]. In addition, the most significantly enriched pathways for the red cluster included two pathways related to TGF-β signalling (TGF-β signalling via kinase cascades in breast cancer and development TGF-β receptor signalling). The TGF-β signalling pathway is a well-known driver of fibrosis[211,212,212,213,213] and so these findings could support the idea that the red cluster represents a genuine endotype of IPF. Other enriched pathways included those related to adipocytes (cells that are specialised in storing energy as fat) and metabolism.

TABLE 5.7: *Significantly enriched (q-value <0.05) biological processes for the 769 genes assigned to the red cluster.*

Biological process	**Enrichment score**	**p-value**	**q-value**
Mitochondrial ATP synthesis coupled electron transport	7.18	1.0×10^{-7}	7.8×10^{-4}
ATP synthesis coupled electron transport	7.12	1.2×10^{-7}	7.8×10^{-4}
Respiratory electron transport chain	6.88	1.4×10^{-7}	7.8×10^{-4}
Cellular respiration	5.95	1.3×10^{-6}	0.005
Oxidative phosphorylation	5.84	4.0×10^{-6}	0.012
Electron transport chain	5.56	4.3×10^{-6}	0.012
Homeostasis of number of cells	5.12	1.1×10^{-5}	0.024
Homeostatic process	4.54	1.7×10^{-5}	0.032

None of the 14 genes suspected to be associated with IPF susceptibility were assigned to the red cluster, nor were they statistically overconnected to the genes that were assigned to this cluster.

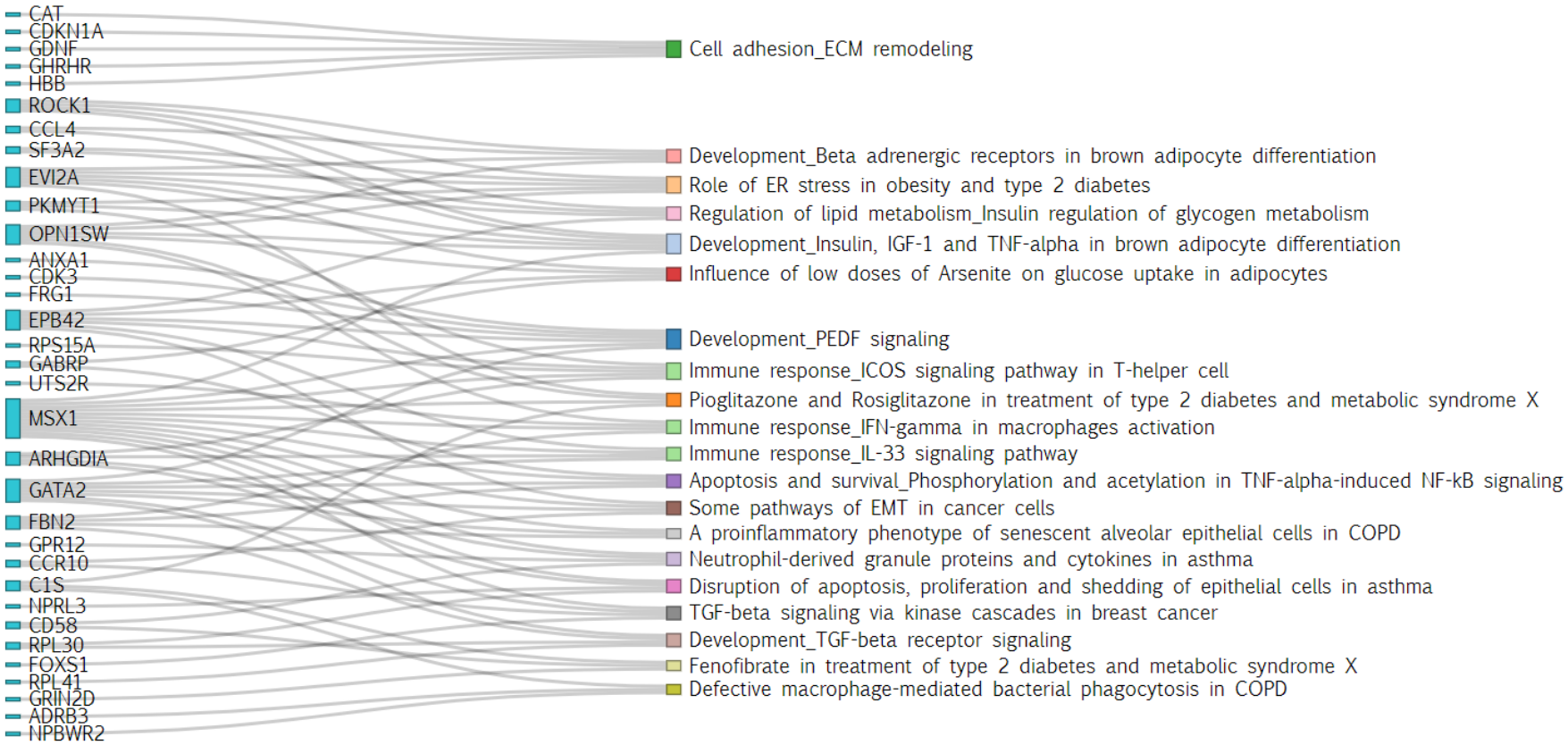


FIGURE 5.9: *A Sankey diagram for the red cluster showing which genes correspond to the 20 most significantly enriched biological pathways.*

Blue cluster

427 *biological processes* were significantly enriched in the blue cluster (q-value <0.05). The top 20 most significantly enriched processes are shown in Table 4.8. The top processes were highly significant, with q-values less than 5×10^{-15}. The 20 most significantly enriched *biological pathways* for the blue cluster (Figure 5.10) included biological pathways relating to gene regulation, DNA repair, cell cycle and apoptosis. As discussed in Section 4.4, apoptosis has been previously implicated in IPF development and apoptosis-based therapies for IPF have been proposed. Another notable significantly enriched biological pathway for the blue cluster was 'HIF-1 targets', a pathway related to hypoxia. Recent findings have shown that hypoxia induces the proliferation of pulmonary fibroblasts[214] and so it is plausible that dysregulation of the HIF-1 targets pathway could represent a causal mechanism for the development of IPF, thus supporting the hypothesis that the blue cluster could represent an endotype of the disease.

TABLE 5.8: *The 20 most significantly enriched (q-value <0.05) biological processes for the 839 genes assigned to the blue cluster.*

Biological process	Enrichment score	p-value	q-value
Cell activation	12.78	2.2×10^{-27}	3.7×10^{-24}
Immune system process	11.33	1.7×10^{-25}	1.4×10^{-21}
Leukocyte activation	11.76	2.4×10^{-23}	1.2×10^{-19}
Immune response	9.83	6.0×10^{-19}	2.5×10^{-15}
Regulation of immune system process	9.75	1.5×10^{-18}	4.9×10^{-15}
Regulated exocytosis	8.90	2.5×10^{-14}	6.9×10^{-11}
Response to stimulus	7.30	1.3×10^{-13}	3.1×10^{-10}
Defence response	8.16	1.6×10^{-13}	3.2×10^{-10}
Multi-organism process	7.74	1.9×10^{-13}	3.5×10^{-10}
Lymphocyte activation	8.73	4.5×10^{-13}	7.5×10^{-10}
Translational initiation	9.72	6.4×10^{-13}	9.1×10^{-10}
Symbiotic process	8.24	6.6×10^{-13}	9.1×10^{-10}
Interspecies interaction between organisms	8.02	1.6×10^{-12}	2.1×10^{-9}
Peptide metabolic process	8.31	1.9×10^{-12}	2.1×10^{-9}
Exocytosis	8.06	1.9×10^{-12}	2.1×10^{-9}
Peptide biosynthetic process	8.43	2.9×10^{-12}	2.9×10^{-9}
Translation	8.46	3.2×10^{-12}	3.1×10^{-9}
Regulation of biological quality	7.14	3.8×10^{-12}	3.5×10^{-9}
Myeloid leukocyte activation	8.09	4.1×10^{-12}	3.6×10^{-9}
Regulation of multicellular organismal process	7.20	5.0×10^{-12}	4.0×10^{-9}

The IPF-associated gene *FAM13A* was one of the genes that was assigned to the blue cluster, though it did not belong to any of the top 20 significantly enriched biological pathways. Additionally, the genes

in this cluster were statistically overconnected to five other IPF-associated genes. These were: *AKAP13*, *DSP*, *LRRC34*, *MAPT* and *TERT*. The hypergeometric p-value was calculated to be 0.020, indicating that it is significant that five IPF-associated genes were overconnected to the genes in this cluster and this is more than would be expected due to chance.

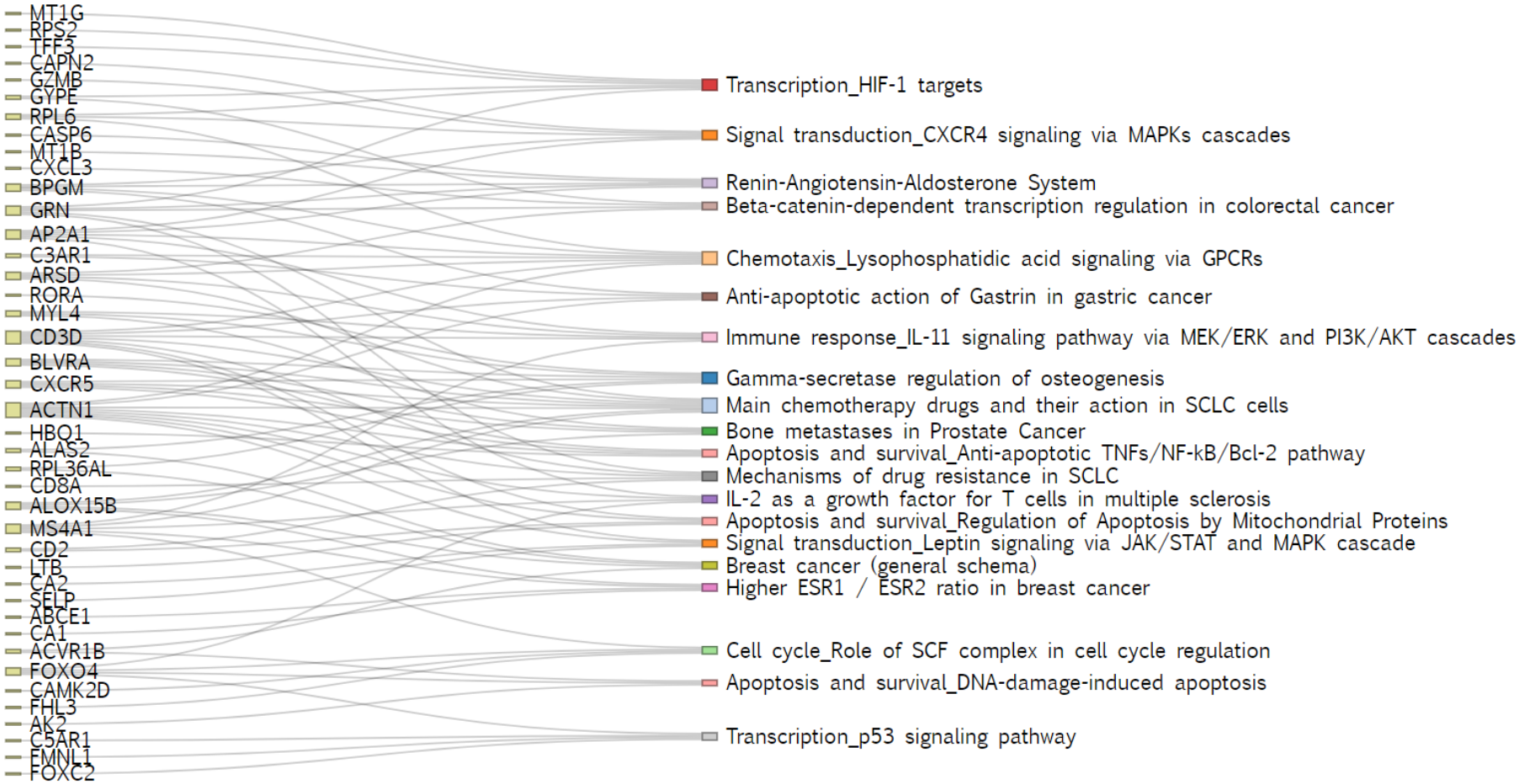


FIGURE 5.10: *A Sankey diagram for the blue cluster showing which genes correspond to the 20 most significantly enriched biological pathways.*

Yellow cluster

952 *biological processes* were significantly enriched in the yellow cluster. The most highly significant processes were related to the immune system response (Table 5.9). These were very highly significant, with q-values lower than 1×10^{-40}. The significantly enriched *biological pathways* for this cluster (Figure 5.11) included immune system response pathways related to IL-6 signalling and IL-3 signalling. The role of the immune response in IPF has been historically difficult to elucidate, however all stages of fibrosis are accompanied by innate and adaptive immune responses[215] and recent findings suggest that immune dysregulation is a key driver of disease pathophysiology[216]. The finding that genes involved in immune response pathways are dysregulated in the individuals in the yellow cluster could therefore support the idea that this cluster represents a genuine endotype of the disease.

TABLE 4.9: *The 20 most significantly enriched (q-value <0.05) biological processes for the 784 genes assigned to the yellow cluster.*

Biological process	Enrichment score	p-value	q-value
Cell activation	20.78	1.3×10^{-60}	1.5×10^{-56}
Immune response	19.53	1.8×10^{-60}	1.5×10^{-56}
Leukocyte activation	20.87	3.3×10^{-59}	1.8×10^{-55}
Immune system process	18.04	1.6×10^{-57}	6.6×10^{-54}
Immune effector process	19.19	1.2×10^{-52}	4.0×10^{-49}
Myeloid leukocyte activation	20.63	1.7×10^{-52}	4.7×10^{-49}
Leukocyte activation involved in immune response	20.07	9.2×10^{-51}	2.2×10^{-47}
Cell activation involved in immune response	19.98	1.9×10^{-50}	3.9×10^{-47}
Neutrophil activation	20.19	1.0×10^{-48}	1.9×10^{-45}
Granulocyte activation	20.02	3.5×10^{-48}	5.7×10^{-45}
Neutrophil activation involved in immune response	19.55	4.0×10^{-46}	6.1×10^{-43}
Leukocyte degranulation	19.42	5.0×10^{-46}	6.8×10^{-43}
Neutrophil degranulation	19.43	1.3×10^{-45}	1.7×10^{-42}
Myeloid cell activation involved in immune response	19.21	1.5×10^{-45}	1.8×10^{-42}
Neutrophil mediated immunity	19.23	3.6×10^{-45}	3.9×10^{-42}
Myeloid leukocyte mediated immunity	18.99	1.1×10^{-44}	1.1×10^{-41}
Leukocyte mediated immunity	17.11	4.3×10^{-43}	4.2×10^{-40}
Secretion by cell	16.63	3.9×10^{-41}	3.5×10^{-38}
Export from cell	16.50	5.9×10^{-41}	5.2×10^{-38}
Defence response	15.95	1.2×10^{-40}	1.0×10^{-37}

None of the 14 IPF-associated genes were found in the gene list for the yellow cluster. However, 4 of these genes were found to be statistically overconnected to the genes in this cluster. These were as follows: *DSP*, *MAD1L1*, *MAPT* and *TERT*. The statistical significance of this was approximated to be

P=0.008 using a hypergeometric test, again indicating that this was significantly more than would be expected due to chance.

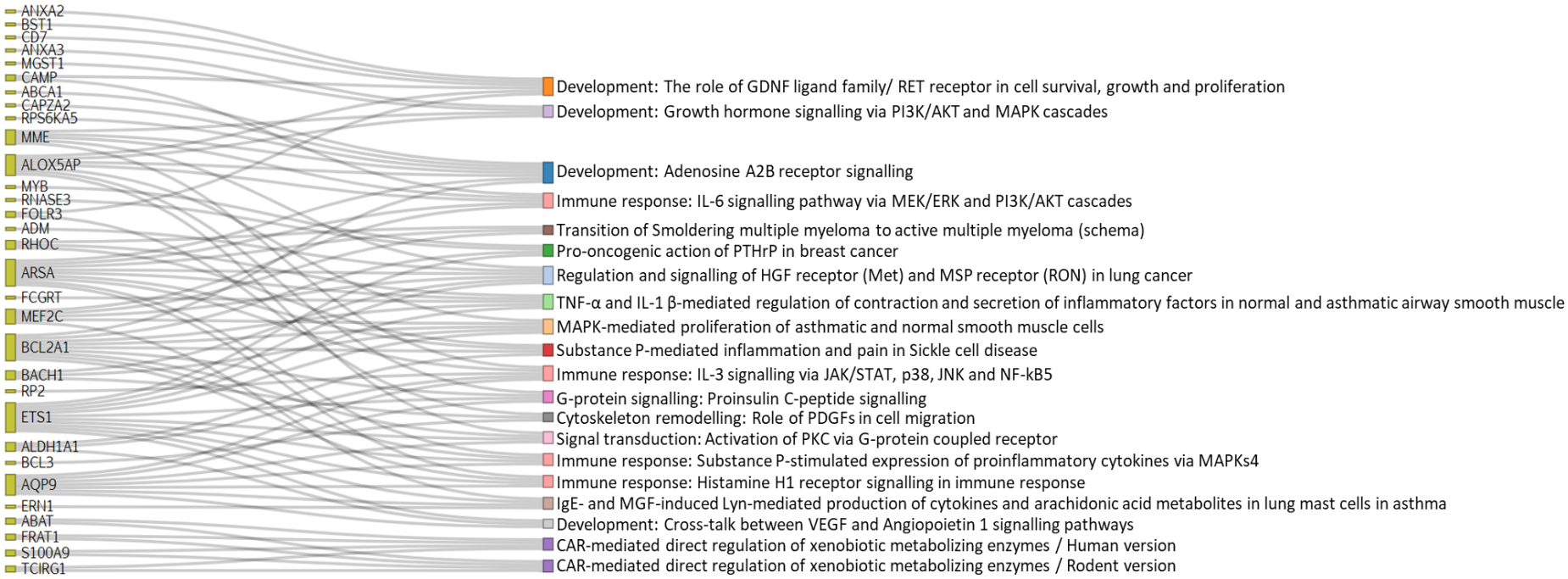


FIGURE 5.11: *A Sankey diagram for the yellow cluster showing which genes correspond to the 20 most significantly enriched biological pathways.*

5.9 Development of gene-expression based classifier

5.9.1 Methods

Following the gene enrichment analysis, a gene expression-based classifier with the ability to assign new IPF cases to one of the three clusters was designed. Classification is a method of supervised machine learning which uses a correctly labelled training dataset to predict which category new observations belong in.

This classifier was designed following the approach described by Sweeney et al. in their cluster analysis of bacterial sepsis[94]. The classifier does not use absolute levels of gene expression to make predictions, but instead utilizes relative gene expression between individuals in a cohort. This means that the classifier can be applied to a group of disease cases from the same study without first requiring the removal of technical effects, as the (scaled) relative gene expression between those individuals will be the same regardless of the presence of technical effects. This allowed for the use of datasets that did not contain data for healthy controls.

Classification accuracy was the most important feature of the classifier. However, the ability to assign individuals based on as few genes as possible was beneficial for three reasons. First, trimming the classifier would have reduced the risk of overfitting, which is a common problem in machine learning. Second, a classifier that requires the measurement of fewer genes whilst maintaining classification accuracy would be more cost-effective in a clinical setting as a prognostic biomarker. Finally, as validation studies may not have measured all 2,500 genes used in the original clustering, designing the classifier to use as few genes as possible lowered the probability of encountering unmeasured genes, and thus being unable to fully utilize the classifier.

Making predictions with the classifier was a two-stage process. First, each subject was given a classification score for each cluster. Then, these scores were used to fit a multinomial logistic regression model with the ability to predict the most likely cluster assignment for new samples.

To determine the optimal genes to include in the classifier for the IPF data, an R function called ‘classifiergenes’ was produced (see Appendix C for R code). This function was applied to each cluster separately and performed a greedy forward search to determine an optimal combination of genes to differentiate between subjects in that cluster vs all other clusters. The function required two input arguments: the first was an $i \times j$ gene expression matrix where each column represented a subject and each row represented a gene. The second argument was a numerical vector that indicated which columns contained the subjects that were assigned into the cluster under question.

The function performed an iterative algorithm, which chose the best gene at each iteration to add to the classifier. The algorithm began by fitting a receiver operating characteristic (ROC) curve to the gene expression data for each gene to assess how well each gene could discern between the subjects in that

particular cluster against subjects in all other clusters. This was evaluated using the area under the ROC curve (AUC), which ranges from 0.5 to 1. The gene with the greatest AUC indicated the gene that had the best combination of sensitivity and specificity when differentiating between subjects in that cluster against all other clusters. Thus, this gene should have been included in the cluster classifier. This optimal gene was then added to a list of over-expressed genes or under-expressed genes, based on whether that gene was more highly expressed in subjects from that cluster compared to the average expression across all subjects. Following that, a classification score was calculated for each remaining gene and for each subject. This score was calculated as the geometric mean of the over-expressed genes for subject *j* minus the geometric mean of the under-expressed genes for subject *j*, if gene *i* was included in the calculation. ROC curves were then fit to the classification scores for each gene, and the gene whose inclusion led to the greatest improvement in AUC was selected and added to the over/under-expressed gene list. This was then repeated for all remaining genes until no additional gene could have improved the AUC.

At each iteration, classifiergenes output the name of the optimal gene to include in the classifier as well as the AUC of the classifier for that cluster once that optimal gene had been added. The gene expression data for the over-expressed and under-expressed genes were then saved (separately) to the local R environment by the classfiergenes function. Once the function had been applied to all K clusters, the optimal genes to include in the classifier were known, as well as whether each gene was over-expressed or under-expressed in that cluster.

Using the same formula as described above, K classification scores were then calculated for each subject using only the optimal genes in each cluster. These scores were then mean centred around zero and scaled to reflect a Z-score (i.e. standard deviation equal to 1). Ideally, individuals that belonged to a certain cluster should have had a high classification Z-score for that cluster and low classification Z-scores for the other clusters.

The classification Z-scores were then used to fit a multinomial logistic regression model, with cluster as the independent categorical variable and the Z-scores from each cluster as the dependent variables. This model had the ability to take data from new individuals and predict which cluster they were each most likely to belong in, using only expression data from the optimal genes in the classifier. The probability of individual *i* belonging to cluster C was calculated using the following equation:

$$\Pr(Y_i = C) = \frac{e^{\beta_C \cdot X_i}}{\sum_{k=1}^{K} e^{\beta_k \cdot X_i}} \tag{5.6}$$

Where β_C denotes the model coefficients from the multinomial logistic regression model for cluster C, X_i is the vector of classification Z-scores for subject *i* and there is a total of K clusters. One cluster must be used as a reference class in the model and as such the model coefficients for this cluster were all equal to zero.

5.9.2 Results

The classifiergenes function was applied to the pooled, co-normalised gene expression data for each cluster separately, using all 196 IPF cases that were successfully clustered in the discovery analysis. To reduce the computational burden, only the genes that were assigned into each cluster prior to the gene enrichment analysis (Section 5.8) were used. The resulting optimal 23-gene classifier is shown in Table 5.10.

TABLE 5.10: *The 23 genes in the optimal classifier. 'Up genes' refer to genes that were more highly expressed in the subjects for that cluster compared to the mean expression across all subjects, and 'down genes' refer to genes that were less highly expressed in the subjects in that cluster.*

Cluster 1 (red)		**Cluster 2 (blue)**		**Cluster 3 (yellow)**	
Up genes	**Down genes**	**Up genes**	**Down genes**	**Up genes**	**Down genes**
KCNK15	*RPF1*	*NOP58*		*CA4*	
SORBS1		*PSMA5*		*BCL2A1*	
HBB		*RASGRP1*		*UGCG*	
EIF4G1		*IFI30*		*FPR2*	
		HLA-DRA			
		ATM			
		ECHDC2			
		EXOSC8			
		BLVRA			
		PSMD11			
		SLC38A1			
		MRPL41			
		PPIA			
		AES			

Classification scores were calculated for each cluster by calculating the geometric mean of the expression for the up genes minus the geometric mean of the expression of the down genes. For example, the cluster 1 classification scores were calculated as: $(KCNK15 \times SORBS1 \times HBB \times EIF4G1)^{\frac{1}{4}} - RPF1$. These scores were then mean centred and scaled to reflect Z-scores (Figure 5.12). As desired, the Z-scores based on the genes from each particular cluster were greatest on average for the subjects assigned to that cluster. This was especially effective for the Z-scores based on the Cluster 2 (blue cluster) and Cluster 3 (yellow cluster) classifier genes, where there was minimal overlap in the classification Z-score distributions between the individuals in those clusters and the individuals in the other two clusters.

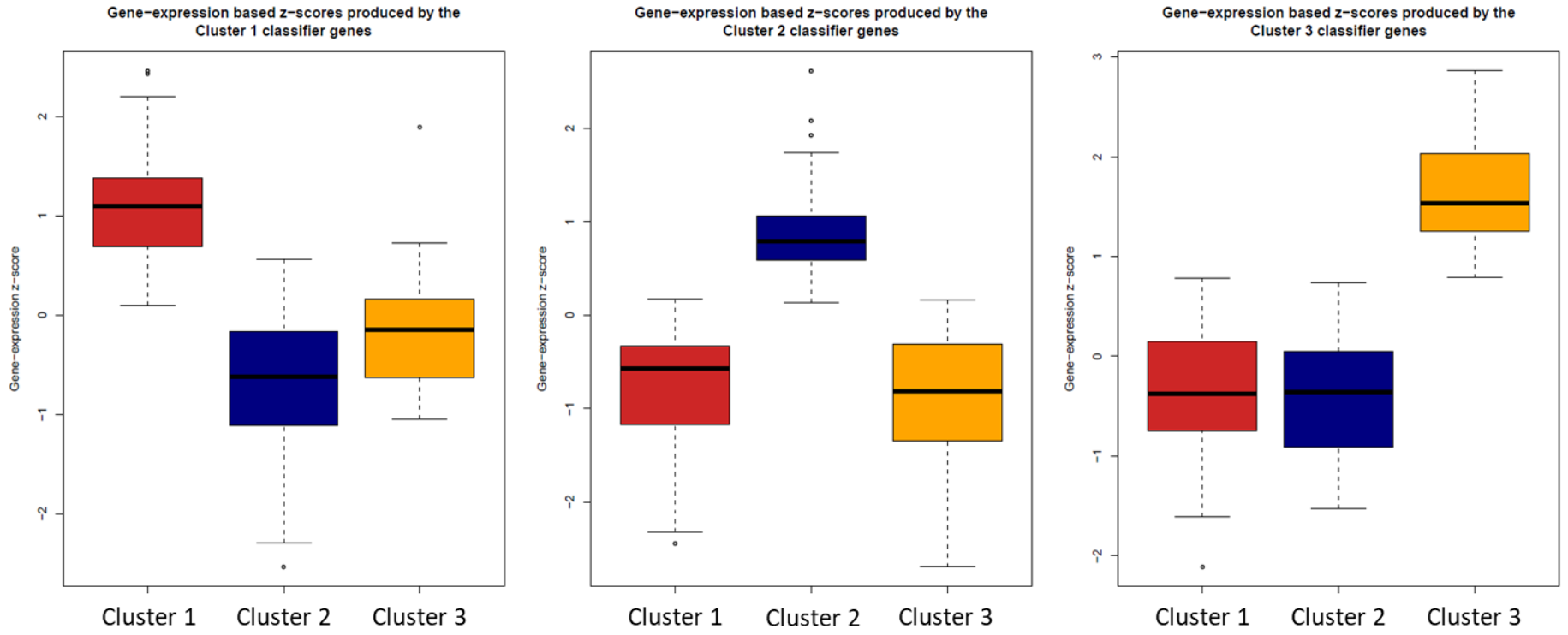


FIGURE 5.12: *The distribution of classification Z-scores for each cluster. The x-axis represents subjects, stratified by the cluster they were assigned to.*

A multinomial logistic regression model was then fit with cluster as the independent variable and the three types of classification Z-scores as the dependent variables (Table 5.11). Cluster 1 was chosen as the reference cluster and so the coefficients for this cluster were all zero and are not shown in Table 5.11. The coefficients for this model were quite large, which may indicate that the model was overfit to the data. The feasibility of classifiers with fewer genes that may be less overfit to the data is explored in Section 5.11.

TABLE 5.11: *Coefficients of the multinomial logistic regression fit to the classification scores for clusters 2 and 3. The coefficients for the reference cluster (cluster 1) are all zero and have been omitted.*

	Intercept	Cluster 1 score	Cluster 2 score	Cluster 3 score
Cluster 2	-21.22	-36.99	258.09	21.18
Cluster 3	-79.98	-93.07	-29.62	107.79

The following example demonstrates how this model may be used to classify new subjects to one of the three clusters. Say subject *i* had the classification Z-scores $Z_1 = 0.5$, $Z_2 = 0$ and $Z_3 = 2$.

First, we calculate the denominator from Equation 5.6, which will be the same for all clusters:

$$\sum_{k=1}^{K} e^{\beta_k \cdot X_i} = e^0 + e^{-21.2+(0.5\times-37.0)+(0\times258.1)+(2\times21.2)} + e^{-80.0+(0.5\times-93.1)+(0\times-29.6)+(2\times107.8)}$$

$$= e^0 + e^{2.7} + e^{89.1}$$

Then, the probability of this subject belonging to each cluster may be calculated as follows:

$$\Pr(Y_i = 1) = \frac{e^{\beta_1 \cdot X_i}}{\sum_{k=1}^{K} e^{\beta_k \cdot X_i}} = \frac{e^0}{e^0 + e^{2.7} + e^{89.1}} \approx 0$$

$$\Pr(Y_i = 2) = \frac{e^{\beta_2 \cdot X_i}}{\sum_{k=1}^{K} e^{\beta_k \cdot X_i}} = \frac{e^{-21.2+(0.5\times-37.0)+(0\times258.1)+(2\times21.2)}}{e^0 + e^{2.7} + e^{89.1}} = \frac{e^{2.7}}{e^0 + e^{2.7} + e^{89.1}} \approx 0$$

$$\Pr(Y_i = 3) = \frac{e^{\beta_3 \cdot X_i}}{\sum_{k=1}^{K} e^{\beta_k \cdot X_i}} = \frac{e^{-80.0+(0.5\times-93.1)+(0\times-29.6)+(2\times107.8)}}{e^0 + e^{2.7} + e^{89.1}} = \frac{e^{89.1}}{e^0 + e^{2.7} + e^{89.1}} \approx 1$$

The model predicts that subject *i* would almost certainly belong to Cluster 3.

This model was used to re-assign each of the 196 discovery subjects to a cluster. It correctly assigned 100% of subjects (Table 5.12), which means that the 23 gene classifier was able to completely recapitulate the original clustering that had used expression data from 2,500 genes. As the classifier was shown to be highly accurate, it was deemed suitable to apply it to the additional independent validation datasets.

TABLE 5.12: *A two-way table comparing 'True' assignment of individuals from the discovery analysis (determined using COMMUNAL with 2,500 genes) to the reassignment of these individuals using the 23 gene classifier.*

		True cluster		
		Cluster 1	Cluster 2	Cluster 3
Predicted cluster	Cluster 1	64	0	0
	Cluster 2	0	95	0
	Cluster 3	0	0	37

5.10 Validation of classifier in independent datasets

5.10.1 Methods

The classifier was designed to use a minimal number of genes for the reasons described in Section 5.9.1. However, using a small number of genes in the classifier meant that the individual contribution of each gene was greater than the contribution of a single gene in a classifier that considered many genes. As a result, applying the classifier to a dataset with even a single required gene missing could lead to a large decrease in classification accuracy. Although, if the classifier were to be used as a clinical tool, the assay would be designed in such a way that all necessary genes could be measured in all patients. Therefore, in the validation stage the classifier was applied only to datasets that had measured all of the genes that are used in the classifier as this would provide the most accurate representation of the classifier's performance in a potential clinical setting.

Any validation datasets that did not contain data from all 23 genes in the classifier were excluded. In addition, all data from non-IPF cases were removed. The gene expression data from the remaining studies and individuals in each study were then quantile normalised and put on the $\log_2$ scale. The cluster classifier was applied to each dataset separately, producing each subject a classification score for each cluster. These classification scores were then fed into the multinomial logistic regression model (Table 5.11), which assigned each IPF case to a cluster. Clinical and demographic traits were then compared across clusters for the individuals in each validation study separately, as well as across clusters with all validation studies combined. This was performed using the same methods as described in Section 5.7.1.

5.10.2 Results

There were four sets of gene expression data from blood that were reserved for validation, with GEO accession codes GSE133298, GSE132607, GSE27957 and GSE28042. GSE133298 was the only dataset that did contain information for all 23 genes in the cluster classifier (13 of the 23 were present in the dataset) and was excluded from the analysis. The data from GSE132607 comes from a study (unpublished as of April 2022) that aimed to develop a predictor of FVC progression by studying gene expression differences in 74 IPF cases over time. The data from GSE27957 and GSE28042 both originate from the same study[185], where the data in GSE27957 (n=45 IPF subjects) was used in discovery and the data in GSE28042 (n=75 IPF subjects) was used as independent validation data. This

study was described previously (Section 5.1.2). In brief, the authors developed a 52-gene signature that had the ability to successfully predict transplant-free survival in patients with IPF.

Summary statistics for the IPF cases in each of the three cohorts are shown in Table 5.13. In all three cohorts, the average individual was in their late sixties, male and of European ancestry. The individuals in each study were similar in terms of the lung function measures FVC and D_{LCO}. Survival data was available for the individuals in GSE27957 and GSE28042, where less than half of the IPF cases in each study were observed to have died during the 3.5 year follow-up period (Additional Figure A.5.3).

TABLE 5.13: *Summary statistics for the IPF subjects in each of the three cohorts that were used in the validation stage of this study. Data are presented as percentage or mean (standard deviation [sd]). FVC = forced vital capacity, D_{LCO} = diffusing capacity of lung for carbon monoxide.*

Phenotypic trait	**GSE132607** (n=74)	**GSE27957** (n=45)	**GSE28042** (n=75)
Age (years) (mean, sd)	66.6 (7.6)	67.1 (8.2)	68.9 (8.1)
Sex (% male)	70.3%	88.9%	69.3%
Ancestry (% European)	94.6%	82.2%	97.3%
FVC % predicted (mean, sd)	69.7 (18.4)	60.6 (14.3)	65.4 (16.7)
D_{LCO} % predicted (mean, sd)	45.6 (15.4)	43.4 (17.7)	48.9 (18.6)
Mortality (% death observed during study)	Unknown	37.8%	32.0%

The gene expression-based cluster classifier was applied to the IPF cases in each cohort and assigned each individual to a cluster (Table 5.14). Phenotypic traits were compared across clusters and the results for all validation studies combined are shown in Table 5.15. Comparisons of phenotypic traits for each validation study separately can be found in the appendix (Additional Table B.5.3). Again, there were significant differences in mortality between clusters, with death observed for 22% of subjects in the blue cluster, 44% of subjects in the red cluster and 63% of subjects in the yellow cluster (P=0.003). Additionally, there was a significant difference in the proportion of ever smokers across clusters, with the blue cluster having the lowest proportion of ex/current smokers and the yellow cluster having the greatest proportion. As in the discovery stage, those in the blue cluster had the greatest average D_{LCO} and lowest GAP score, although these variables were not significantly different across clusters in the validation stage (P=0.274 and P=0.377 respectively).

TABLE 5.14: *The number of IPF subjects from each validation study that were assigned into each of the three clusters.*

	Cluster 1 (red)	**Cluster 2 (blue)**	**Cluster 3 (yellow)**	**Total in study**
GSE132607	19	35	20	74
GSE27957	14	26	5	45
GSE28042	25	39	11	75
Total in cluster	58	100	36	194

TABLE 5.15: *Comparison of clinical and demographic traits across clusters for all validation studies combined. Data are presented as count (percentage), mean (standard deviation [sd]) or median (interquartile range [IQR]). NA = data not available, FVC=Forced vital capacity, D_{LCO} = Diffusing capacity for carbon monoxide, GAP index = Gender, Age and Physiology index for IPF mortality[193], FEV_1 = Forced expiratory volume in one second, MUC5B genotype = genotype for the MUC5B promoter polymorphism rs35705950. Significant P-values (P < 0.05) are highlighted in bold.*

	All validation studies combined (n=194)					
	Cluster 1 (red)	**Cluster 2 (blue)**	**Cluster 3 (yellow)**	**P-value**	**Total n used**	**No. of datasets**
Total subjects in cluster	58	100	36			
Phenotypic Trait						
Age (mean, sd)	68.6 (8.1)	67.7 (7.7)	65.8 (8.6)	0.242	194	3
Male (%)	44 (75.9%)	70 (70.0%)	30 (83.3%)	0.276	194	3
European Ancestry (%)	56 (96.6%)	91 (91.0%)	33 (91.7%)	0.412	194	3
D_{LCO} % predicted (median, IQR)	43.0 (25.1)	46.0 (22.2)	43.3 (24.8)	0.274	194	3
GAP index (mean, sd)	4.3 (1.6)	3.9 (1.5)	4.1 (1.5)	0.377	193	3
Death observed during study (%)	17 (43.6%)	14 (21.5%)	10 (62.5%)	**0.003**	120	2
FEV_1 (median, IQR)	73.5 (21.7)	74.0 (23.8)	81.8 (12.1)	0.804	75	1
Ever smoker (%)	12 (63.2%)	19 (54.3%)	18 (90%)	**0.025**	74	1
MUC5B genotype: GG (%)	3 (17.6%)	6 (18.8%)	3 (20.0%)	0.922	64	1
MUC5B genotype: GT (%)	13 (76.5%)	25 (78.1%)	12 (80.0%)			
MUC5B genotype: TT (%)	1 (5.9%)	1 (3.1%)	0 (0%)			

In terms of survival over time (Figure 5.13), individuals in the blue cluster fared the best and those in the yellow cluster fared significantly worse, which was consistent with the findings from the discovery stage. Individuals in the red cluster also had consistently poorer survival than those in the blue cluster. This too is consistent with the discovery stage findings: whilst survival information was not directly available for the individuals in cluster 1 in the discovery stage, their significantly low average D_{LCO} and high average GAP score was suggestive of poor survival. The median survival time was at approximately 17 months for those in the yellow cluster and 28 months for those in the red cluster. Again, median survival could not be directly calculated for those in the blue cluster as the survival probability for this group never dropped below 0.5, so the median survival time for this group must be greater than 45 months.

A Cox PH model was fit to the survival data, again with the blue cluster as the reference group (Table 5.16). There was no evidence that the Cox model had broken the PH assumption (Additional Figure A.5.4). There were significant differences in survival between both clusters and the blue cluster, with an estimated hazard ratio of 2.89 for the red cluster and 4.23 in the yellow cluster. This means that at any follow-up time, those in the red cluster were estimated to be 2.89 times as likely to die as those in the blue cluster, whilst those in the yellow were 4.23 times as likely to die as those in the blue cluster. Importantly, these results showed that the cluster classifier was able to assign new independent individuals with IPF in such a way that the significant differences in survival that were observed in the discovery stage had been recaptured, therefore validating the classifier.

However, the difference in survival time between the red and yellow clusters was not statistically significant, with a hazard ratio of 1.47 (95% CI [0.67, 3.22], P=0.341), using the red cluster as the reference cluster. This may have been due to a lack of statistical power, particularly as these were the two clusters with the fewest individuals, or it may indicate that these two clusters are not representative of real, clinically distinct endotypes of IPF.

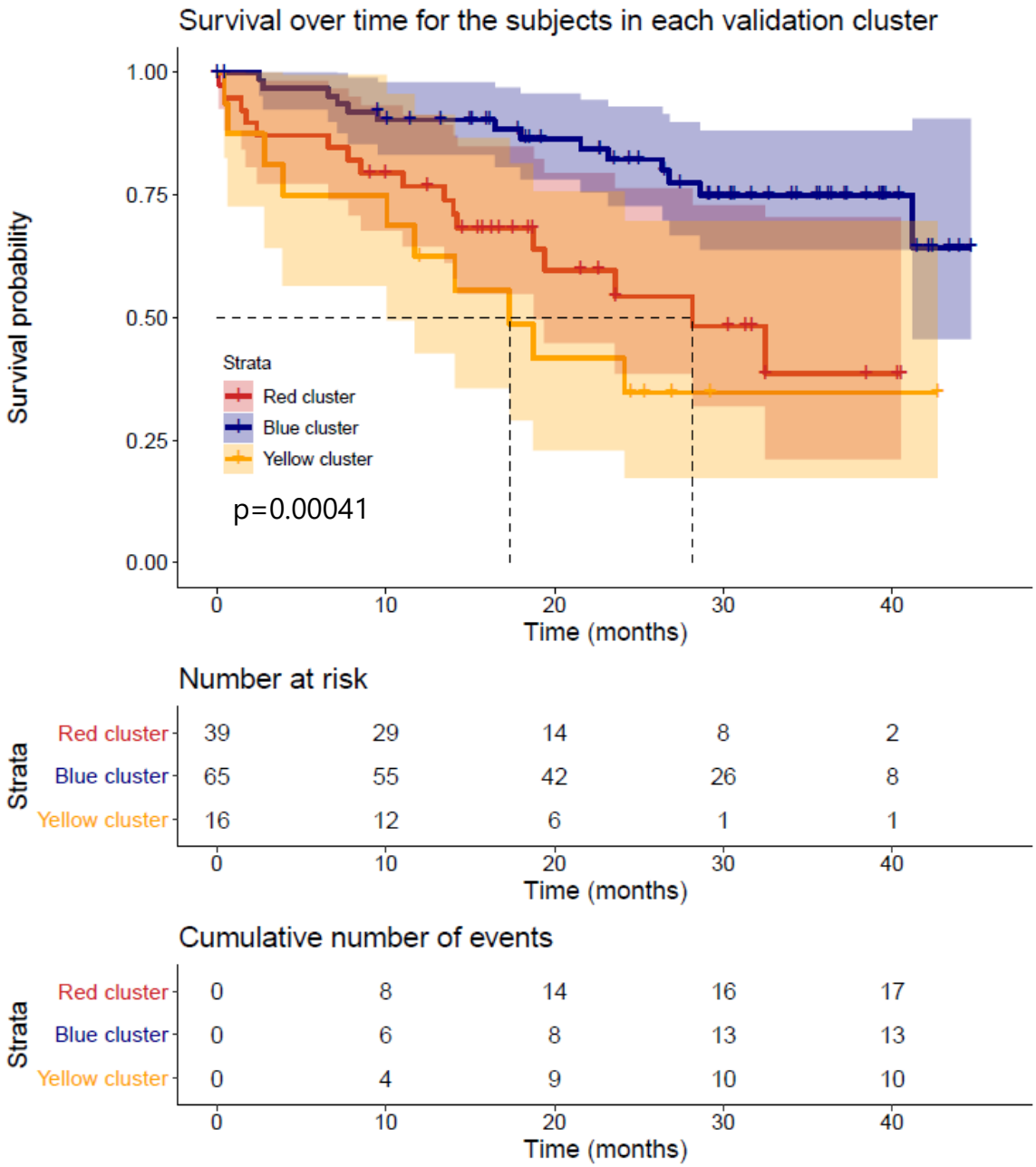


FIGURE 5.13: *A Kaplan-Meier plot showing survival over time for the validation subjects in each cluster, as well as tables showing the number of subjects at risk and the cumulative number of events at each 10-month period. The p-value shown on the plot is from a log-rank test testing the three curves for equality. Median survival in each cluster is shown by dotted lines, where possible.*

TABLE 5.16: *Summary statistics from the Cox proportional-hazards model that was fit to the survival data from the validation studies. The reference group in the model was the blue cluster.*

Cluster	Hazard ratio	95% CI	P-value
Red cluster	2.89	1.41, 5.93	**0.004**
Yellow cluster	4.23	1.87, 9.60	**0.001**

5.11 The feasibility of reduced gene classifiers

5.11.1 Methods

A classifier that requires the measurement of fewer genes could be preferable to one that requires a greater number of genes to be measured (Section 5.9.1). Additionally, the full 23-gene cluster classifier may be overfit to the data, as indicated by the large coefficients in the multinomial logistic regression model (Table 5.11). Overfitting occurs in supervised machine learning when the model is learning from a training dataset but detects noise and treats these random fluctuations as part of the true underlying distribution. This presents an issue when the overfit model is used to make predictions using additional data, as the new datasets will not contain the same patterns of noise and so the same concepts learnt by the model do not apply, leading to a loss of prediction accuracy and reduced generalisability.

If the iterative classifiergenes function (Section 5.9.1) were to be stopped once a certain AUC threshold had been reached, the resulting classifier would consist of fewer genes and would perform less well in reassignment of discovery subjects, but may improve the generalisability of the classifiers and their ability to assign validation subjects in such a way as to recreate the clusters observed in the discovery analysis. AUC thresholds of 0.99, 0.98 and 0.95 were implemented into the classifiergenes function and classification scores using these genes were calculated using the same method as described in Section 5.9.1. Multinomial logistic regression models were then fit to these classification scores to create new, reduced-gene classifiers.

The reduced classifiers were then used to reassign the IPF cases in the discovery and validation datasets to a cluster. The feasibility of these classifiers was evaluated by their ability to detect statistically significant differences in phenotypic traits between the newly clustered validation subjects, as this evaluates the classifier's ability to recreate the same clusters that were observed in the discovery analysis. For this reason, the traits that were found to be statistically significantly different across clusters in the discovery analysis (survival, D_{LCO} and GAP index) were of particular interest. Phenotypic traits were compared using the same methods as described in Section 5.7.1.

5.11.2 Results

For each cluster, the classifiergenes function was stopped once the AUC had met or exceeded 0.99 (reduced classifier 1), 0.98 (reduced classifier 2) and 0.95 (reduced classifier 3). Reducing the AUC threshold led to a decrease in the number of genes used in each classifier (Table 5.17), with 13 genes in reduced classifier 1, 7 genes in reduced classifier 2 and 4 genes in reduced classifier 3. The coefficients of the corresponding multinomial logistic regression model for each classifier are shown in Table 5.18. The coefficients in these models are lower than those found when using the full 23 gene classifier, which could indicate that these models are less overfit to the data. Reduced classifiers 1, 2 and 3 correctly reassigned 99.0%, 94.9% and 92.9% of discovery subjects respectively (Table 5.19).

TABLE 5.17: *The genes used in the full 23 gene cluster classifier and the genes included in the reduced classifiers that were produced when AUC thresholds were implemented when applying the classifiergenes function to each cluster. 'Up genes' refer to genes that were more highly expressed in the subjects for that cluster compared to the mean expression across all subjects, and 'down genes' refer to genes that were less highly expressed in the subjects in that cluster. AUC = area under curve.*

Red cluster			**Blue cluster**			**Yellow cluster**		
Up genes	**Down genes**	**AUC**	**Up genes**	**Down genes**	**AUC**	**Up genes**	**Down genes**	**AUC**
KCNK15		0.962	*NOP58*		0.935	*CA4*		0.960
			PSMA5		0.968			
	RPF1	0.984	*RASGRP1*		0.980	*BCL2A1*		0.989
SORBS1		0.984	*IFI30*		0.988	*UGCG*		0.998
HBB		0.991	*HLA-DRA*		0.989			
			ATM		0.992			
EIF4G1		0.992	*ECHDC2*		0.993	*FPR2*		1
			EXOSC8		0.995			
			BLVRA		0.996			
			PSMD11		0.998			
			SLC38A1		0.998			
			MRPL41		0.998			
			PPIA		0.999			
			AES		0.999			

Reduced classifier 3 (AUC ≥ 0.95)

Reduced classifier 2 (AUC ≥ 0.98)

Reduced classifier 1 (AUC ≥ 0.99)

Full 23 gene classifier

TABLE 5.18: *Coefficients of the multinomial logistic regression models fit using classification scores from the genes in the reduced classifiers. In each case, the red cluster is the reference cluster and as such the coefficients for this cluster are all zero and have been omitted.*

	Cluster	**Intercept**	**Red cluster score**	**Blue cluster score**	**Yellow cluster score**
Full 23 gene classifier	Blue	-21.22	-36.99	258.09	21.18
	Yellow	-79.98	-93.07	-29.62	107.79
Reduced classifier 1	Blue	3.12	-9.75	8.87	1.66
	Yellow	-16.6	-11.92	-3.15	29.42
Reduced classifier 2	Blue	1.59	-5.71	3.50	0.05
	Yellow	-3.37	-3.66	-0.40	7.48
Reduced classifier 3	Blue	1.06	-5.32	4.33	-1.35
	Yellow	-0.64	-3.13	0.78	3.17

TABLE 5.19: *Two-way tables comparing 'true' assignment of subjects from the discovery analysis (determined using COMMUNAL with 2,500 genes) to the reassignment of these subjects using the reduced gene classifiers.*

		True cluster		
		Red cluster	Blue cluster	Yellow cluster
Reduced classifier 1 predicted cluster	Red cluster	63	1	0
	Blue cluster	1	94	0
	Yellow cluster	0	0	37
Reduced classifier 2 predicted cluster	Red cluster	61	1	0
	Blue cluster	2	92	2
	Yellow cluster	1	2	33
Reduced classifier 3 predicted cluster	Red cluster	59	2	3
	Blue cluster	3	92	3
	Yellow cluster	2	1	31

The IPF cases from the validation studies were reassigned to clusters using the reduced gene classifiers and phenotypic traits were compared across clusters (Table 5.20). For all classifiers, there were statistically significant ($P<0.05$) differences across clusters in the proportion of those who were observed to have died during their study, with those in the blue cluster being the least likely to die. Furthermore, reduced classifier 1 (the 13 gene classifier) produced the most significant difference in survival, with a p-value of 0.001, which was more statistically significant than the difference found when using the full 23 gene classifier. Additionally, the difference in D_{LCO} between clusters when using reduced classifier 1 was trending towards statistical significance ($P=0.069$), with those in the blue cluster having the highest median D_{LCO}. The difference in D_{LCO} across clusters was not close to statistical significance for any other classifier, including the full 23-gene classifier. There was not a significant difference in average GAP index score across clusters under any of the classifiers. Overall, these results suggested that the 13 gene classifier was the best at recapturing the clusters that were found in the discovery analysis, where significant differences in survival and D_{LCO} were observed between clusters.

TABLE 5.20: *Comparison of phenotypic traits across clusters when all validation subjects are clustered using the full 23 gene and the reduced gene classifiers. Data are presented as count (percentage), mean (standard deviation [sd]) or median (interquartile range [IQR]). FVC=Forced vital capacity, D_{LCO} = Diffusing capacity for carbon monoxide, GAP index = Gender, Age and Physiology index for IPF mortality[193], FEV_1 = Forced expiratory volume in one second. Significant P-values (P < 0.05) are highlighted in bold.*

Trait	Total n used	No. of datasets	Full 23 gene classifier				Reduced classifier 1 (13 genes)			
			Red cluster (n=58)	Blue cluster (n=100)	Yellow cluster (n=36)	P-value	Red cluster (n=52)	Blue cluster (n=101)	Yellow cluster (n=41)	P-value
Age (mean, sd)	194	3	68.6 (8.1)	67.7 (7.7)	65.8 (8.6)	0.242	67.1 (8.1)	68.5 (7.6)	66.2 (8.6)	0.239
Male (%)	194	3	44 (75.9%)	70 (70.0%)	30 (83.3%)	0.276	38 (73.1%)	72 (71.3%)	34 (82.9%)	0.347
European Ancestry (%)	194	3	56 (96.6%)	91 (91.0%)	33 (91.7%)	0.412	51 (98.1%)	91 (90.1%)	38 (92.7%)	0.196
D_{LCO} % predicted (median, IQR)	194	3	43.0 (25.1)	46.0 (22.2)	43.3 (24.8)	0.274	42.1 (26.4)	48.2 (21.1)	43.4 (20.3)	0.069
FVC % predicted (median, IQR)	193	3	64.0 (25.6)	65.0 (23.8)	63.4 (17.0)	0.841	64.3 (23.6)	65.0 (24.3)	63.1 (15.3)	0.467
GAP index (mean, sd)	193	3	4.3 (1.6)	3.9 (1.5)	4.1 (1.5)	0.377	4·1 (1·6)	4·0 (1·5)	4·3 (1·5)	0·753
Death observed during study (%)	120	2	17 (43.6%)	14 (21.5%)	10 (62.5%)	**0.003**	16 (48.5%)	13 (19.7%)	12 (57.1%)	**0.001**
FEV_1 (median, IQR)	75	1	73.5 (21.7)	74.0 (23.8)	81.8 (12.1)	0.804	74.8 (21.7)	75.2 (22.2)	75.4 (17.7)	0.913
Ever smoker (%)	74	1	12 (63.2%)	19 (54.3%)	18 (90%)	**0.025**	11 (57.9%)	21 (60.0%)	17 (85.0%)	0.114

TABLE 5.20 (continued): *Comparison of phenotypic traits across clusters when all validation subjects are clustered using the full 23 gene and the reduced gene classifiers. Data are presented as count (percentage), mean (standard deviation [SD]) or median (interquartile range [IQR]). FVC=Forced vital capacity, D_{LCO} = Diffusing capacity for carbon monoxide, GAP index = Gender, Age and Physiology index for IPF mortality*[193]*, FEV_1 = Forced expiratory volume in one second. Significant P-values (P < 0.05) are highlighted in bold.*

Trait	Total n used	No. of datasets	Reduced classifier 2 (7 genes)				Reduced classifier 3 (4 genes)			
			Red cluster (n=56)	Blue cluster (n=103)	Yellow cluster (n=35)	P-value	Red cluster (n=62)	Blue cluster (n=100)	Yellow cluster (n=32)	P-value
Age (mean, sd)	194	3	68.2 (9.0)	67.5 (7.3)	67.0 (8.4)	0.768	68.9 (8.5)	68.0 (7.6)	66.7 (8.2)	0.560
Male (%)	194	3	40 (71.4%)	75 (72.8%)	29 (82.9%)	0.428	47 (75.8%)	73 (73.0%)	24 (75.0%)	0.919
European Ancestry (%)	194	3	55 (98.2%)	94 (91.3%)	31 (88.6%)	0.153	57 (91.9%)	92 (92.0%)	31 (96.9%)	0.619
D_{LCO} % predicted (median, IQR)	194	3	44.8 (24.5)	47.2 (25.0)	43.6 (15.8)	0.495	42.8 (23.3)	47.3 (24.3)	44.7 (22.3)	0.461
FVC % predicted (median, IQR)	193	3	63.3 (26.3)	65 (24.0)	63.4 (14.3)	0.417	62.3 (24.3)	69 (22.1)	62.8 (20.1)	**0.037**
GAP index (mean, sd)	193	3	4.2 (1.7)	4.0 (1.5)	4.3 (1.4)	0.971	4.3 (1.6)	3.9 (1.4)	4.1 (1.6)	0.429
Death observed during study (%)	120	2	18 (43.9%)	13 (21.3%)	10 (55.6%)	**0.007**	21 (46.7%)	14 (23.7%)	6 (37.5%)	**0.048**
FEV_1 (median, IQR)	75	1	76.0 (20.8)	73.3 (23.7)	75.4 (17.7)	0.995	76.0 (21.9)	80.0 (22.6)	70.5 (16.8)	0.111
Ever smoker (%)	74	1	8 (53.3%)	28 (66.7%)	13 (76.5%)	0.384	14 (82.4%)	24 (58.5%)	11 (68.8%)	0.212

Kaplan-Meier plots were used to visualise survival over time for the validation subjects in each cluster when using the reduced-gene classifiers (Figure 5.14). In all cases, those in the blue cluster fared the best over time and were most different in terms of survival to the subjects the other two clusters. As the number of genes in the classifier was reduced the survival in the yellow and red clusters became more similar, with the survival curves crossing multiple times when clusters were assigned using reduced classifiers 2 and 3.

Both reduced classifier 1 and reduced classifier 2 were able to produce clusters with similar survival over time to the full 23 gene classifier (Figure 5.13) and that were reflective of the survival shown in the blue and yellow clusters in the discovery analysis (Figure 5.8). Reduced classifier 1 produced the clusters with the greatest differences in survival, as indicated by the p-value for the log-rank test on the plot. Conversely, reduced classifier 3 produced clusters with the least difference in survival between groups. The Kaplan-Meier curves for this classifier were not as reflective of the clusters shown in the discovery analysis, particularly for the yellow cluster which had a median survival time of approximately 1 year in discovery, yet could not be calculated when using reduced classifier 3 as the survival probability for the yellow cluster never dropped below 0.5.

Cox PH models were fit to the survival data under each clustering assignment (Table 5.21). Significant differences between the blue cluster and both other clusters were found using the full classifier, reduced classifier 1 and reduced classifier 2. It was reduced classifier 1 that was shown to be the best at distinguishing between the low-risk patients in the blue cluster and the high-risk patients in the red and yellow clusters. As this information could potentially be used to predict survival in IPF, reduced classifiers 1 and 2 may each be feasible as a prognostic biomarker, with reduced classifier 1 being the preferred choice.

Importantly, the findings from this section suggest that despite using fewer genes, reduced classifier 1 is superior to the full 23 gene classifier in its ability to assign IPF subjects in such a way as to create clusters with significant differences in survival and lung function between groups.

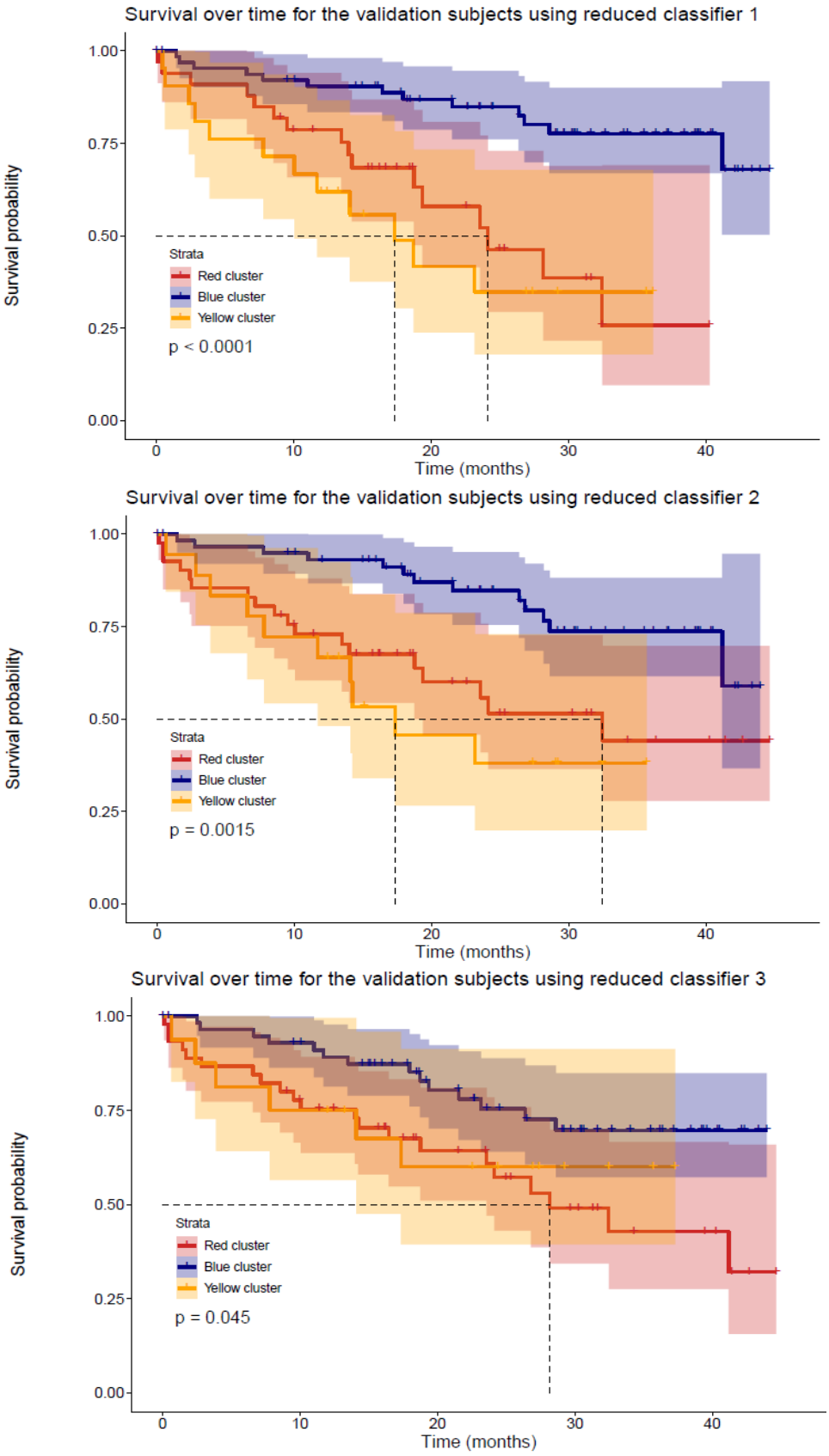


FIGURE 5.14: *Kaplan-Meier plots showing survival over time for the clustered validation subjects when using the reduced gene classifiers. The p-value shown on the plot is from a log-rank test testing the three curves for equality. Median survival in each cluster is shown by dotted lines, where possible.*

TABLE 5.21: *Summary statistics from the Cox proportional hazards models fit to the survival data from the validation studies when subjects are assigned using the full classifier and each reduced gene classifier. In all cases, the blue cluster is the reference cluster and as such the coefficients for this group are all zero and have been omitted.*

Classifier	Cluster	Hazard Ratio	95% Confidence Interval	P-value
Full 23 gene classifier	Red	2.89	(1.41, 5.93)	**0.004**
	Yellow	4.23	(1.87, 9.60)	**0.001**
Reduced classifier 1	Red	3.80	(1.78, 8.12)	**0.001**
	Yellow	5.05	(2.24, 11.35)	**9.1×10^{-5}**
Reduced classifier 2	Red	2.73	(1.33, 5.59)	**0.006**
	Yellow	3.80	(1.66, 8.74)	**0.002**
Reduced classifier 3	Red	2.30	(1.17, 4.53)	**0.016**
	Yellow	1.88	(0.72, 4.89)	0.198

5.12 Characterisation of the genes used in the classifier

The genes used in the classifier may have been previously implicated in research of pulmonary fibrosis or IPF. For each gene used in the 13 gene classifier, a search of PubMed was performed to systematically quantify the extent to which that gene has been linked to pulmonary fibrosis in the past. Each search term contained the acronym of the gene and the term 'pulmonary fibrosis'. Advanced search settings were applied so that the full gene acronym must have appeared within the main text of the publication, whilst 'pulmonary fibrosis' could have appeared in any field (e.g., main text, title or abstract).

The full name of each gene is shown in Table 5.22 along with a brief summary of its function, found using the GeneCards online resource[217]. The number of publications returned in each PubMed search is also shown. Only two genes produced relevant search results: *HBB* and *ATM*. *HBB* was mentioned in a single publication[218], which described a proteomic analysis that aimed to discover non-invasive protein biomarkers that could be used to diagnose IPF. In this analysis, the protein encoded by *HBB* was found to be one of the three most significantly upregulated proteins when comparing the peripheral blood of 60 IPF cases to that of 60 healthy controls.

ATM featured in seven publications. Mutations in the *ATM* gene result in a rare autosomal recessive disorder called Ataxia-telangiectasia which can lead to interstitial lung disease and pulmonary fibrosis[219]. As such, the publications produced by the PubMed search for this gene consisted of studies that investigated Ataxia-telangiectasia. *ATM* was not mentioned in connection with idiopathic PF. However, the protein encoded by *ATM* is thought to control the cell-cycle checkpoint signalling pathways that are required for cell response to DNA damage and for genome stability. As discussed in Section 1.1.2, cellular senescence driven through DNA damage is an important factor in the development of IPF as so it is plausible that changes in expression of *ATM* between clusters could reflect a causal disease mechanism. In addition, recent genetic studies have implicated processes involved in cell proliferation in IPF susceptibility[220], which also increases the plausibility of a causal link between dysregulation of *ATM* and the development of IPF.

None of the other genes used in the classifier have been specifically implicated in IPF, although some are involved in broader processes that have been previously linked to IPF or pulmonary fibrosis. This included genes related to the immune response[221,222,222] (*HLA-DRA*, *IFI30*, *RASGRP1*), metabolism[223,224,224] (*HBB*, *UGCG*), signalling[225,226] (*IFI30*, *RASGRP1*, *ATM*) and cell cycle[227,228] (*PSMA5*, *ATM*).

TABLE 5.22: *The full names and summaries of the genes in the 13 gene cluster classifier, as well as the number of papers found on PubMed that contained the name of that gene plus the term 'pulmonary fibrosis'.*

Cluster	Gene acronym	Full gene name	Summary	No. of papers
Red	*KCNK15*	Potassium channel subfamily K member 15	Encodes a member of the superfamily of potassium channel proteins.	0
	RPF1	Ribosome Production Factor 1 Homolog	Encodes a protein which may be required for ribosome biogenesis.	0
	SORBS1	Sorbin and SH3 domain-containing protein 1	Encodes a protein which functions in the signalling and stimulation of insulin.	0
	HBB	Hemoglobin Subunit Beta	Involved in oxygen transport from the lung to the various peripheral tissues. Among its related pathways are folate metabolism.	1
Blue	*NOP58*	Nucleolar protein 58	Related to pathways involved in rRNA processing in the nucleus and metabolism of proteins.	0
	PSMA5	Proteasome 20S Subunit Alpha 5	A component of the 20S core proteasome complex, which plays numerous essential roles within the cell by associating with different regulatory particles.	0
	RASGRP1	RAS guanyl-releasing protein 1	Activates the Erk/MAP kinase cascade and regulates T-cells and B-cells development, homeostasis and differentiation.	0
	IFI30	Gamma-interferon-inducible lysosomal thiol reductase	The protein encoded by this gene can reduce protein disulphide bonds. Among its related pathways are interferon gamma signalling and the innate immune system.	0
	HLA-DRA	HLA class II histocompatibility antigen, DR alpha chain	Plays a central role in the immune system by presenting peptides derived from extracellular proteins.	0
	ATM	Ataxia telangiectasia mutated	Encodes a protein that is thought to be one of the master controllers of cell cycle checkpoint signalling pathways that are required for cell response to DNA damage and for genome stability.	7
Yellow	*CA4*	Carbonic Anhydrase 4	Encodes part of a large family of zinc metalloenzymes that catalyse the reversible hydration of carbon dioxide and participate in a variety of biological processes, including respiration and calcification.	0
	BCL2A1	B-cell lymphoma 2-related protein A1	Encodes a member of the BCL-2 protein family. The proteins of this family form hetero- or homodimers and act as anti- and pro-apoptotic regulators that are involved in a wide variety of cellular activities such as embryonic development, homeostasis and tumorigenesis.	0
	UGCG	Ceramide glucosyltransferase	Encodes an enzyme that catalyses the first glycosylation step in the biosynthesis of glycosphingolipids, which are essential components of membrane microdomains that mediate membrane trafficking and signal transduction.	0

5.13 Comparison of the classifier to another transcriptomic biomarker for IPF

As discussed in Section 5.1.3, Herazo-Maya et al. identified 52 genes that were associated with transplant-free survival in a cohort of IPF cases[185]. In a large multi-centre validation study[186], the authors applied a method called The Scoring Algorithm for Molecular Subphenotypes (SAMS) to gene expression data (from the 52 genes) to classify the IPF cases in each cohort as either high-risk or low-risk (in terms of mortality or requiring a transplant). The results of the validation study showed that their method was successful and that the high-risk group in each centre was significantly more likely to die or require a transplant than those in the low-risk group.

The gene-expression based cluster classifier has been shown to be able to assign the individuals with IPF that are least likely to die into the blue cluster and those who are most likely to die into the two remaining clusters and so this could potentially also be used as a prognostic biomarker. As the optimal classifier uses gene expression data from only 13 genes, it would likely be more cost-effective than Herazo-Maya et al.'s prognostic tool, which uses data from 52 genes. In this section, the two approaches are compared in terms of their ability to predict mortality in IPF.

5.13.1 Methods

First, the names of the 52 genes used in Herazo-Maya et al.'s prognostic tool were compared to the names of the 13 genes used in our classifier (as well as any aliases) to check whether there were any common genes between the two methods. This was also done for the 23 genes in the full classifier.

Then, each of the IPF cases in the two validation studies for which survival data was available, GSE27957 (n=45) and GSE28042 (n=75), were classed as either 'high risk' or 'low risk' by applying Herazo-Maya et al.'s method SAMS. 7 of the 52 genes were expected to be more highly expressed in high risk cases than low risk cases ('up genes'). Likewise, the remaining 45 genes were expected to be less highly expressed in high risk cases than low risk cases ('down' genes). SAMS was implemented as follows:

1. For each gene, the geometric mean of the expression for that gene across all subjects was calculated. This value represents the average level of expression for that gene across the whole cohort. It was then subtracted from the gene expression of that gene for each subject so that positive values represented subjects that had increased expression of that gene compared to the average and negative values represented subjects that had decreased expression compared to the average.
2. For each subject, the proportion of the 7 'up genes' that were overexpressed was calculated. Similarly, the proportion of the 45 'down genes' that were less highly expressed than average was calculated. So, if a subject had 4 'up genes' that were greater than the average and 30 'down genes' that were lower than the average, these proportions would have been 0.571 and 0.667 respectively.

3. For each subject, the sum of the geometric mean-normalised expression data was summed up for the 'up genes' that were more highly expressed than average. Then the sum of the geometric mean-normalised expression data was summed up for the 'down genes' that were less highly expressed than average. So, for example, for the subject above who had 4 of the 7 'up genes' that were more highly expressed than the average, say with expression values 0.185, 0.553, 0.123 and 1.003 for these four genes, the sum would have been 1.864. The sum for the 'down genes' must always be negative, for example say that this sum for the subject above was -7.645.
4. The proportion of the 'up genes' calculated in step 2 was multiplied by the sum for the 'up genes' calculated in step 3 to produce the 'up score' for each subject. So, for the example subject above, their up score would have been 0.571×1.864 = 1.064. A 'down score' for each subject was also calculated by multiplying their proportion of down genes by their down sum from step 3. For our example subject, this would have been 0.667× -7.645=-5.099.
5. Subjects with up scores greater than the median value and down scores lower than the median value were classed as 'high risk', while all other subjects were classed as 'low risk'.

This was done separately for each cohort and by using data from as many of the 52 genes as were measured in the datasets. These subjects were also assigned into one of the three clusters (red, blue, yellow) using the 13 gene classifier. As the red and yellow clusters were not clinically distinct in the previous analyses, but both contained individuals that were more likely to die at any follow-up time than those in the blue cluster, the individuals in both of these clusters were considered 'high risk' whilst those in the blue cluster were considered 'low risk'. Two-way tables were used to compare agreement between the two methods.

Kaplan-Meier plots were used to visualise the survival over time for the validation subjects in each risk group under each method. In both cases, the log-rank test was used to test the survival curves of each risk group for equality. Univariate Cox PH models were fit to the data with risk group as the sole covariate and time-to-death as the outcome of interest. In both cases, the low-risk group was used as the reference group. The Concordance index (C-index), the equivalent of the AUC for an ROC curve, and the p-values from the log-rank test were used to assess which method performed best at assigning the IPF subjects to the correct risk group and therefore predicting survival.

Following this, multivariate Cox PH models were used to assess whether the predictions made by each method were significant predictors of mortality in the validation datasets whilst adjusting for age, sex, ancestry, FVC and D_{LCO}. The likelihood ratio test and C-index were used to assess whether either of the two methods of risk prediction led to a significant increase in predictive ability over a Cox PH model containing only age, sex, ancestry, FVC and D_{LCO}.

5.13.2 Results

The 52 genes in Herazo-Maya et al.'s gene signature to predict outcome in IPF were as follows: *LBD1, TPST1, MCEMP1, IL1R2, HP, FLT3, S100A12, LCK, CAMK2D, NUP43, SLAMF7, LRRC39, ICOS, CD47, LBH, SH2D1A, CNOT6L, METTL8, ETS1, C2orf27A, P2RY10, TRAT1, BTN3A1, LARP4, TC2N, GPR183, MORC4, STAT4, LPAR6, CPED1, DOCK10, ARHGAP5, HLA-DPA1, BIRC3, GPR174, CD28, UTRN, CD2, HLA-DPB1, ARL4C, BTN3A3, CXCR6, DYNC2LI1, BTN3A2, ITK, SNHG1, CD96, GBP4, S1PR1, NAP1L2, KLF12, IL7R.* There were no genes in common with the 13 or 23 gene classifier, or any alisases of the genes used in these classifiers (see Additional Table B.5.4 for full list of aliases). One of the 52 genes (*SNHG1*) was missing from the dataset GSE27957 and two genes (*MCEMP1* and *CPED1*) were missing from GSE28042.

The individuals in the GSE27957 and GSE28042 cohorts were each classed as 'high risk' or 'low risk' using both methods. There was 84.4% agreement between the two methods for the individuals in GSE27957, 66.7% agreement for those in GSE28042 and 68.3% agreement overall (Table 5.23).

TABLE 5.23: *The agreement between the two methods when validation subjects were assigned to risk groups.*

GSE27957 (n=45)		Our 13 gene classifier	
		High risk	Low risk
Herazo-Maya et al.'s method	High risk	13	2
	Low risk	5	25
GSE28042 (n=75)		Our 13 gene classifier	
		High risk	Low risk
Herazo-Maya et al.'s method	High risk	17	12
	Low risk	19	27
Both datasets combined (n=120)		Our 13 gene classifier	
		High risk	Low risk
Herazo Maya et al.'s method	High risk	30	14
	Low risk	24	52

Survival over time for the subjects in each risk group according to each method (for both datasets combined) was visualised using Kaplan-Meier plots (Figure 5.15). See Additional Figures A.5.5 and A.5.6 for the survival curves for each study separately. Figure 5.15A showed that the classifier performed well at predicting survival, with the individuals in the high-risk clusters having consistently poorer survival over time than those in the blue 'low-risk' cluster and a highly significant p-value ($P<0.0001$) for the log-rank test. A univariate Cox PH model estimated that at any follow-up time, those in the high-risk clusters were 4.25 times more likely to die than those in the low-risk cluster (95% CI = [2.14, 8.46], $P=3.7\times10^{-5}$). This model had a C-index of 0.664 (95% CI= [0.590, 0.737]).

SAMS (Figure 5.15B) performed less well, with survival curves that laid closer together and a less highly significant log-rank test p-value (0.027). A univariate Cox PH model estimated that at any follow-up time, those in the high-risk group were 1.98 times as likely to die than those in the low-risk group (95% CI = [1.07, 3.68], P = 0·030) and a C-index of 0.609 (95% CI = [0.531, 0.686]).

After adjusting for age, sex, ancestry, FVC and D_{LCO}, the risk predictions made using the classifier remained statistically significant (P=0.007, Table 5.24), with a HR of 2.70 between the high-risk and low-risk clusters (95% CI= [1.32, 5.53]). This model had a C-index of 0.773 (95% CI = [0.697, 0.848]), which was greater than that of the Cox model containing only age, sex, ancestry, FVC and D_{LCO} (C-index = 0.747, 95% CI = [0.670, 0.825]). A likelihood ratio test between the two models gave a p-value of 0.005, suggesting that the predictions made by the classifier were able to significantly improve the predictive ability of the model. The multivariate Cox model containing SAMS' risk predictions had a C-index of 0.760 (95% CI = [0.684, 0.837]), which was an improvement over the Cox model containing only age, sex, ancestry, FVC and D_{LCO}. However, the likelihood ratio test p-value between these two models was not statistically significant (P=0.105).

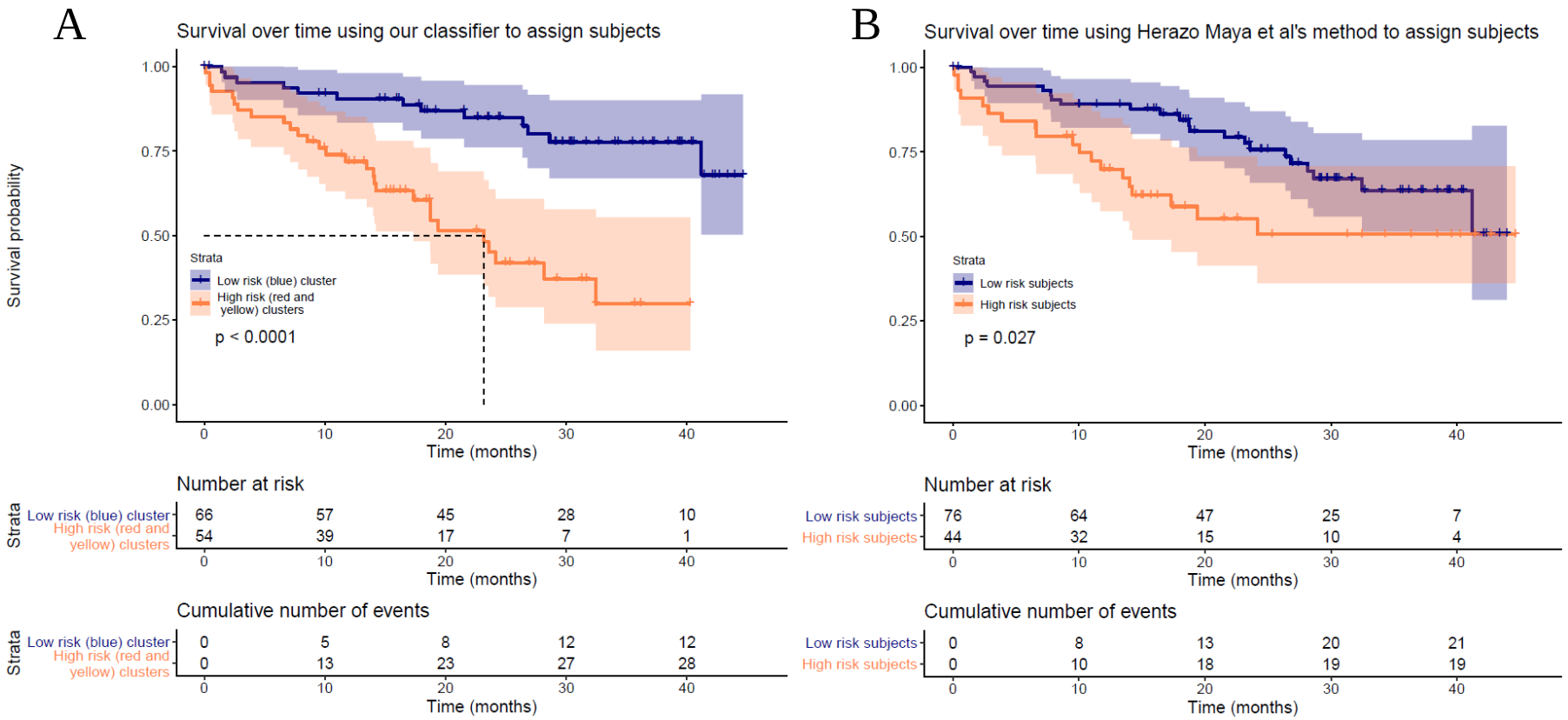


FIGURE 5.15: *Survival over time for the IPF cases in GSE27957 and GSE28042, stratified by risk group according to our 13 gene classifier (A) and Herazo-Maya et al.'s method SAMS (B). The P-value on each plot is from a log-rank test, testing the two curves for equality.*

TABLE 5.24: *Summary statistics from the Cox proportional hazards model adjusting for cluster, age, sex, ancestry, predicted forced vital capacity (FVC) and predicted diffusing capacity of the lung for carbon monoxide (D_{LCO}). OR = odds ratio, SE = standard error and CI = confidence interval. Significant p-values (P<0.05) are highlighted in bold.*

Covariate	OR	SE	P-value	95% CI
Cluster (high-risk clusters)	2.697	0.367	**0.007**	(1.315, 5.534)
Age (years)	1.006	0.020	0.748	(0.968, 1.046)
Sex (male)	5.720	0.752	**0.020**	(1.310, 24.969)
Ancestry (non-European)	1.099	0.608	0.876	(0.334, 3.619)
Predicted FVC	0.996	0.013	0.745	(0.971, 1.022)
Predicted D_{LCO}	0.967	0.013	**0.008**	(0.944, 0.991)

5.14 Application of classifiers to lung tissue datasets

The previous sections have shown that the full cluster classifier and the 13 gene classifier were both effective when applied to gene expression data that were measured from samples of whole blood from patients with IPF. However, as IPF is a lung disease it would be of value to identify whether the gene expression in the blood reflects pathology in the lungs. Whilst gene expression patterns vary across tissues and so a classifier trained on expression data from one tissue type is unlikely to be effective when applied to expression from another, expression of a gene in blood is often a significant predictor of its expression in the lung[229] and so it is possible that the classifiers could be effective when applied to data from lung tissue.

5.14.1 Available data

This part of the analysis was conducted in May 2020. By this time, clinical data had been obtained for some of the lung tissue gene expression datasets that were initially considered for the cluster analysis (Section 5.3). We had also become aware of an additional collection (with GEO accession code GSE47460) that contained data for 122 IPF subjects but did not appear in the initial search as the term 'IPF' is not mentioned in the GEO description. The updated table of available clinical traits for the collection of lung studies is shown in Table 5.25.

Both the new collection GSE47460, and GSE32537, contained samples originating from the Lung Tissue Research Consortium (LTRC). As both collections had reported the unique LTRC identification codes for each subject, neither collection need be excluded entirely as common individuals could be identified and removed from one of the collections. The 61 common individuals were removed from GSE47460 as collection GSE32537 contained more extensive clinical data for each individual.

The lung tissue samples in these studies were collected by either biopsies or transplants. As these are invasive procedures, the controls included in these studies were not necessarily 'healthy' controls (as they must have had a medical reason to justify the collection of their lung tissue sample). For instance,

the controls in the lung tissue dataset GSE47460 all went for surgery for the investigation of a nodule but no evidence of chronic lung disease was found.

TABLE 5.25: *Updated clinical and demographic traits that were reported in at least one of the lung tissue data collections, and their availability across collections. The ✔ symbol indicates that the trait was reported in that collection and the ✗ symbol indicates that the trait was not reported in that collection. FVC=Forced vital capacity, D_{LCO} = Diffusing capacity for carbon monoxide, BMI = body mass index, TLC = Total lung capacity, PT = post-transplant, FEV_1 = Forced expiratory volume in one second.*

	Collections with control subject data								**Collections without control subject data**	
	GSE10667	**GSE124685**	**GSE134692**	**GSE110147**	**GSE92592**	**GSE53845**	**GSE32537**	**GSE47460**	**GSE48149**	**GSE24988**
N IPF cases	31	10	36	22	20	40	119	122	13	44
N controls	15	3	17	11	19	8	50	15	0	0
Trait										
Age	✗	✔	✔	✔	✗	✗	✔	✔	✔	✔
Ancestry	✗	✗	✔	✗	✗	✗	✗	✗	✗	✗
Sex	✗	✔	✔	✔	✗	✔	✔	✔	✔	✔
Smoking status	✗	✗	✔	✗	✗	✗	✔	✔	✔	✗
Smoking pack years	✗	✗	✗	✗	✗	✗	✔	✗	✗	✗
St George's total score	✗	✗	✗	✗	✗	✗	✔	✗	✗	✗
FVC	✗	✔	✗	✔	✗	✗	✔	✔	✔	✔
D_{LCO}	✗	✔	✗	✔	✗	✗	✔	✔	✔	✔
BMI	✗	✗	✔	✔	✗	✗	✗	✗	✗	✔
TLC	✗	✗	✗	✔	✗	✗	✗	✗	✔	✔
PT survival	✗	✗	✗	✔	✗	✗	✗	✗	✗	✔
FEV_1	✗	✔	✗	✗	✗	✗	✗	✔	✗	✗

5.14.2 Methods

Lung tissue study cohorts were excluded if not all 13 genes in the optimal classifier were measured or if relevant clinical data (such as FVC, D_{LCO} and post-transplant survival) were not available for the individuals in that cohort. The full 23 gene classifier and the 13 gene classifier were both applied to the IPF cases from each remaining study to test whether either classifier could effectively assign IPF subjects when applied to gene expression data from lung tissue. Clinical and demographic traits were compared using the same methods as described in Section 5.7.1.

5.14.3 Results

Four studies met the criteria for inclusion in this analysis: GSE47460, GSE110147, GSE24988 and GSE32537. GSE47460 was a large collection from the LTRC containing gene expression data from whole lung samples (totalling 582 individuals, including 122 IPF cases). GSE110147 contained data from a Canadian study[230] that compared the gene expression profiles from explanted lungs of IPF patients (n=22) to those with non-specific interstitial pneumonia. The data in collection GSE24988 was from a Canadian study[231] that used gene expression from the explanted lungs of 116 pulmonary fibrosis patients (N=44 IPF cases) to investigate the relationship between pulmonary fibrosis and pulmonary hypertension. Finally, GSE32537 originates from an American study[232] that used gene expression data (collected from whole lung of 119 patients with IPF/usual interstitial pneumonia) to find transcripts that were differentially expressed compared to healthy control lungs. As most of the samples in these studies were collected from explanted lungs and transplants have a large impact on the trajectory of the disease, patient survival could not be evaluated as reliably as in the analyses of blood expression.

The 23 and 13 gene classifiers were used to assign the IPF subjects in the four datasets to a cluster and phenotypic traits were compared across clusters (Table 5.26). There was only one statistically significant ($P<0.05$) difference between clusters, which was the age of the subjects when using the full 23 gene classifier. In the absence of informative survival data, the most important variable to determine whether the classifiers were effective was D_{LCO}, which was not significant when using either classifier, but did trend toward significance when using the full 23 gene classifier ($P=0.087$). In this case, those in the yellow cluster had the lowest average D_{LCO} and those in the other two clusters had a similarly relatively high average D_{LCO}. This somewhat reflects the findings from the blood analyses, although those in the red cluster would have been expected to have a lower average D_{LCO} than those in the blue cluster. However, the 13 gene classifier did not create clusters representative of those observed in the blood analyses, with those in the yellow cluster having the highest average D_{LCO} and a P-value of 0.133.

In summary, when applied to gene expression data from lung tissue, there was little evidence that the classifiers (that were trained using transcriptomic data from blood) were able to assign patients into clinically distinguishable groups that reflected the clusters that were observed in the previous analyses.

This could be due to a lack of clinical data or be due to differing gene expression patterns between the lung and the blood, which could suggest that the pathology of the two tissue types are distinct.

TABLE 5.26: *Comparison of phenotypic traits across clusters when all subjects in the lung tissue validation cohorts are clustered using the full 23 gene classifier and reduced classifier 1. Data are presented as count (percentage), mean (standard deviation [sd]) or median (interquartile range [IQR]). FVC=Forced vital capacity, D_{LCO} = Diffusing capacity for carbon monoxide, BMI = body mass index, TLC = total lung capacity, FEV_1 = Forced expiratory volume in one second. Significant P-values (P < 0.05) are highlighted in bold.*

Trait	Total n used	No. of datasets	Full 23 gene classifier				Reduced classifier 1 (13 genes)			
			Cluster 1 (red) (n=97)	Cluster 2 (blue) (n=108)	Cluster 3 (yellow) (n=38)	P-value	Cluster 1 (red) (n=80)	Cluster 2 (blue) (n=118)	Cluster 3 (yellow) (n=45)	P-value
Age (mean, sd)	243	4	63.8 (8.2)	63.2 (8.1)	59.4 (11.2)	**0.027**	63.6 (7.1)	63.2 (9.0)	60.4 (10.4)	0.120
Male (%)	243	4	59 (60.8%)	79 (73.1%)	25 (65.8%)	0.170	53 (66.3%)	80 (67.8%)	30 (66.7%)	0.972
FVC % predicted (median, IQR)	234	4	60.0 (23.0)	64.0 (29.6)	53.9 (30.0)	0.394	58.2 (25.0)	62.9 (23.3)	60.0 (29.1)	0.813
D_{LCO} % predicted (median, IQR)	187	3	46.2 (21.3)	47.0 (24.2)	34.0 (37.4)	0.087	41.0 (23.7)	46.7 (24.7)	48.0 (33.4)	0.133
Ever smoker (%)	170	2	40 (59.7%)	50 (67.6%)	17 (58.6%)	0.545	32 (60.4%)	53 (60.9%)	22 (73.3%)	0.429
St George's score (median, IQR)	119	1	43.3 (30.6)	44.2 (41.9)	55.7 (28.0)	0.216	46.7 (34.6)	42.8 (36.8)	59.9 (50.9)	0.353
Post-transplant death observed (%)	63	2	11 (42.3%)	18 (62.1%)	5 (62.5%)	0.298	14 (51.9%)	13 (52.0%)	7 (36.4%)	0.778
BMI (median, IQR)	63	2	24.9 (6.1)	27.0 (3.9)	25.5 (6.4)	0.440	25.0 (5.8)	27.0 (3.4)	26.0 (10.3)	0.389
TLC (median, IQR)	60	2	55.5 (14.8)	61.0 (19.5)	64.5 (9.5)	0.550	56.0 (18.0)	61.0 (14.5)	59.5 (18.5)	0.861
FEV_1 (median, IQR)	59	1	74.0 (16.6)	77.0 (17.1)	70.9 (18.1)	0.441	72.3 (12.5)	76.5 (19.7)	73.8 (16.2)	0.716

5.15 Cluster analysis in multiple lung tissue datasets

Three clusters were identified when using gene expression data from whole blood, yet the classifiers were not shown to be effective at recreating these clusters when applied to data from lung tissue. Therefore, an important question stood as to whether similar clusters are present when gene expression data from a more relevant tissue type, such as lung tissue, is clustered. With sufficient levels of clinical data having been subsequently obtained from investigators, an additional cluster analysis was performed using the lung tissue datasets to address this question.

5.15.1 Methods

Gene expression data for the IPF cases in the lung tissue collections (shown in Table 5.25) that had included data for non-IPF control subjects were co-normalised using COCONUT (Section 5.5.1). Again, PCA was used to visualise and assess the efficacy of the co-normalisation. Any cohorts that were found to have not co-normalised well with the others were excluded from further analysis and the co-normalisation was repeated without this cohort. The pooled, co-normalised data was then clustered using COMMUNAL (Section 5.6.1) and the optimal clustering assignment was selected using the resulting 3D map. Phenotypic traits were compared across clusters as described in Section 5.7.1.

5.15.2 Results

Eight collections contained data for control subjects and were included in the initial co-normalisation. The collections originated from either the USA or Canada and contained a mix of microarray and RNA-seq data from a range of platforms (Table 5.27). Summary statistics for the IPF cases and the controls in each cohort are shown in Table 5.28, though there were still many missing clinical variables across the studies. Additionally, the controls tended to be younger, were less likely to smoke or have been a smoker and they were more likely to be female than the IPF subjects in the same cohorts. The differences in age across disease groups were particularly large in collections GSE134692 and GSE32537 where the controls were on average 28 and 17 years younger than the corresponding IPF cases, respectively. This could have presented a problem for the co-normalisation, as COCONUT theoretically works best when the control subjects in each study are similarly matched to the disease cases as well as being similar to the control subjects in the other studies. Despite this, no control subjects were excluded from this analysis based on their clinical variables due to the high amount of missing data across the cohorts.

TABLE 5.27: *Information on each of the lung tissue cohorts included in this analysis.*

GEO accession	Reference	Country	Platform type	Platform name	Number of gene probes	Number of unique genes
GSE10667	[28]	USA	Microarray	Agilent-014850 Whole Human Genome Microarray	43,376	19,749
GSE124685	[233]	USA	RNA-seq	Ion Torrent Proton	22,653	22,653
GSE134692	[234]	USA	RNA-seq	Illumina HiSeq 2500	15,210	15,210
GSE110147	[230]	Canada	Microarray	Affymetrix Human Gene 1.0 ST Array	33,297	23,307
GSE92592	[25]	USA	RNA-seq	Illumina HiSeq 2000	23,398	23,398
GSE53845	[235]	USA	Microarray	Agilent-014850 Whole Human Genome Microarray	41,000	19,595
GSE32537	[232]	USA	Microarray	Affymetrix Human Gene 1.0 ST Array	11,950	9,928
GSE47460	ǂ	USA	Microarray	Agilent-014850 Whole Human Genome Microarray	15,262	15,181

ǂ: GSE47460 did not contain data from one particular study but instead contained data from the Lung Tissue Research Consortium.

TABLE 5.28: *Summary statistics for the IPF and control subjects in the lung tissue cohorts. sd = standard deviation, IQR = interquartile range, FVC = Forced vital capacity, D_{LCO} = Diffusing capacity for carbon monoxide, BMI = body mass index.*

GEO accession	Disease status	Sample size	Age (years, sd)	Sex (% male)	Ever smoker (%)	FVC (sd)	D_{LCO} (IQR)	BMI (IQR)
GSE10667	IPF	31	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown
	Control	15	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown
GSE124685	IPF	10	57.0 (5.1)	100%	Unknown	58.7 (19.7)	26.0 (10.0)	Unknown
	Control	6	58.8 (11.7)	100%	Unknown	Unknown	Unknown	Unknown
GSE134692	IPF	36	62.8 (5.6)	77.8%	61.1%	Unknown	Unknown	27.7 (6.3)
	Control	17	34.0 (25.9)	52.9%	35.3%	Unknown	Unknown	23.8 (9.1)
GSE110147	IPF	22	61.5 (6.5)	77.3%	Unknown	57.4 (19.3)	37.0 (14.0)	26.5 (5.3)
	Control	11	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown
GSE92592	IPF	20	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown
	Control	19	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown
GSE53845	IPF	40	Unknown	80.0%	Unknown	Unknown	Unknown	Unknown
	Control	8	Unknown	87.5%	Unknown	Unknown	Unknown	Unknown
GSE32537	IPF	119	62.6 (8.7)	64.7%	63.1%	61.3 (17.0)	44.4 (29.7)	Unknown
	Control	50	45.6 (18.6)	46.0%	58.3%	Unknown	Unknown	Unknown
GSE47460	IPF	61	66.7 (8.2)	68.9%	67.3%	67.8 (15.9)	49.0 (24.0)	Unknown
	Control	91	63.4 (11.5)	44.0%	62.2%	94.5 (13.1)	80.0 (20.5)	Unknown

There was a total of 339 IPF cases across the eight cohorts and 5,667 common genes that were measured in all datasets. COCONUT was used to co-normalise the datasets. Before COCONUT (Figure 5.16A), the IPF cases from all eight cohorts are entirely separated in high-dimensional space due to technical differences between datasets. Whereas post-COCONUT (Figure 5.16B), the IPF cases from several of the cohorts overlapped on the plot, indicating that the co-normalisation had reduced the technical differences between those datasets. However, there was one dataset (GSE110147) that did not co-normalise well with the other studies and was still entirely separated from the other datasets in the plot. This dataset was excluded from the analysis and the COCONUT co-normalisation was repeated.

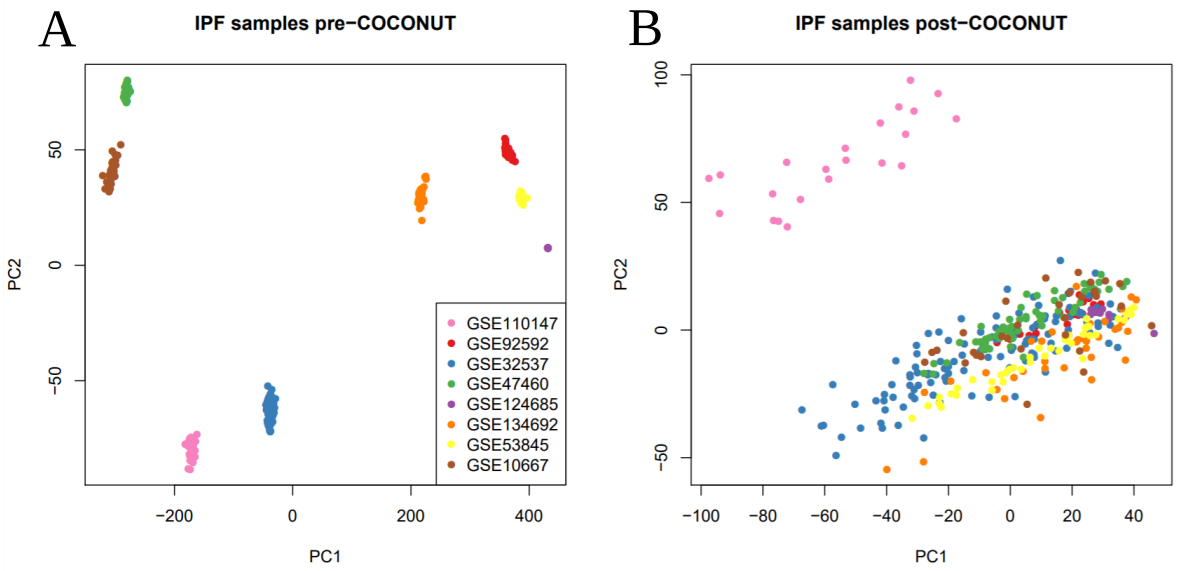


FIGURE 5.16: *Plots of the first two principal components of the gene expression data for the IPF samples from the eight lung tissue datasets, before (A) and after (B) COCONUT co-normalisation.*

With cohort GSE110147 removed, there was a total of 317 IPF subjects across the seven datasets and the number of genes measured in all datasets remained unchanged at 5,667. As before, these datasets were completely separated in gene expression space prior to the co-normalisation (Figure 5.17A). After COCONUT was applied (Figure 5.17B), the technical effects between datasets were reduced and the cohorts overlapped in gene expression space as desired. However, the co-normalisation appeared to have been imperfect. For example, the individuals from GSE124685 all laid closely together toward the top left of the plot and did not overlap at all with the individuals from GSE53845, who all laid towards the bottom of the plot. In spite of this, and as there were no further obviously outlying datasets, no additional studies were removed and the pooled data was clustered using COMMUNAL.

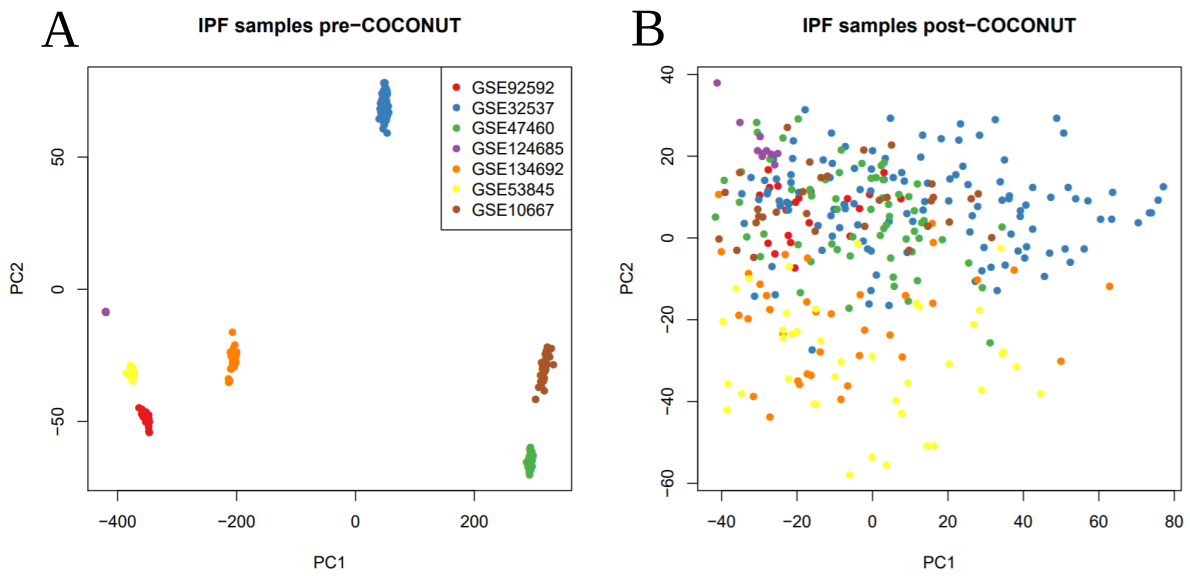


FIGURE 5.17: *Plots of the first two principal components of the gene expression data for the IPF samples from the seven remaining lung tissue datasets, before (A) and after (B) COCONUT co-normalisation.*

The resulting 3D map (Figure 5.18) was used to select the optimal clustering assignment for the data. However, there were few points on the map where the mean validity score was above 0, suggesting overall weak clustering. Additionally, there were few points where the blue and red squares met, indicating that there were not many stable optima and therefore there was no strong signal for any particular number of clusters. The greatest number of stable optima was observed for K=4 clusters and the fewest genes for which there were stable optima was using 100 genes, so this assignment was selected as the optimal assignment for the data.

Under the optimal clustering assignment (Figure 5.19A), 292 individuals (93%) were successfully clustered and clusters 1 and 2 contained individuals from many of the cohorts. However, cluster 4 only contained two individuals, both of whom were from GSE124685, which as discussed previously did not co-normalise well with the other studies. This cluster may therefore be an artefact of the poor co-normalisation. As such, these two subjects were removed along with the unclustered samples, which left three clusters (Figure 5.19B). Another issue was that the clusters were not entirely separated in high-dimensional space, particularly clusters 1 and 2 which overlapped quite substantially. Additionally, comparing Figure 5.19B to Figure 5.17B shows that cluster 3 was comprised almost entirely of individuals from GSE32537, which was presumably another artefact of poor co-normalisation.

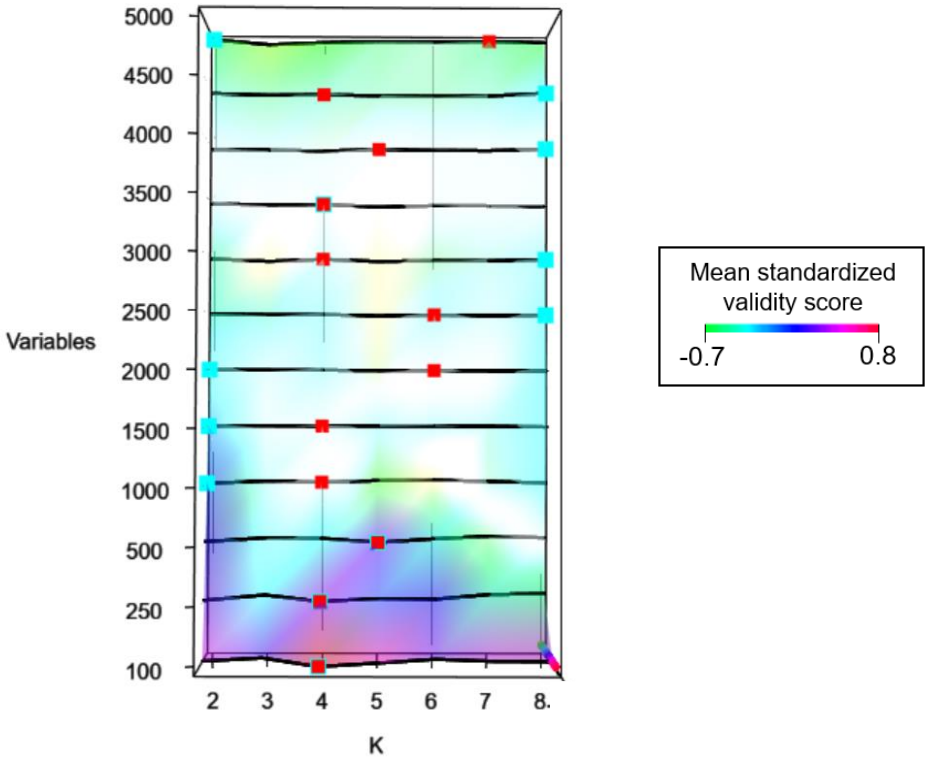


FIGURE 5.18: *The 3D map output by COMMUNAL when applied to the pooled, co-normalised data from the seven lung tissue datasets to identify the optimal cluster assignment. The map shows the mean of standardized values of each validity measure across the entire tested space. On the 3D map, blue squares indicate a potentially optimal clustering at a certain number of genes by finding the assignment where the mean combined validation metric is greatest. The absolute maximum K for any consensus subset is marked with a red square. A higher validity score indicates a better clustering assignment and stable optima are the points where the blue and red squares meet. If stable optima at K clusters are seen over most of the tested space, this indicates the presence of a strong, consistent biological signal at this number of clusters.*

The clinical and demographic traits of the clustered individuals were compared across the three clusters (Table 5.29). No significant differences in phenotypic traits were observed across clusters, though some variables may have been underpowered due to the high proportion of missing data.

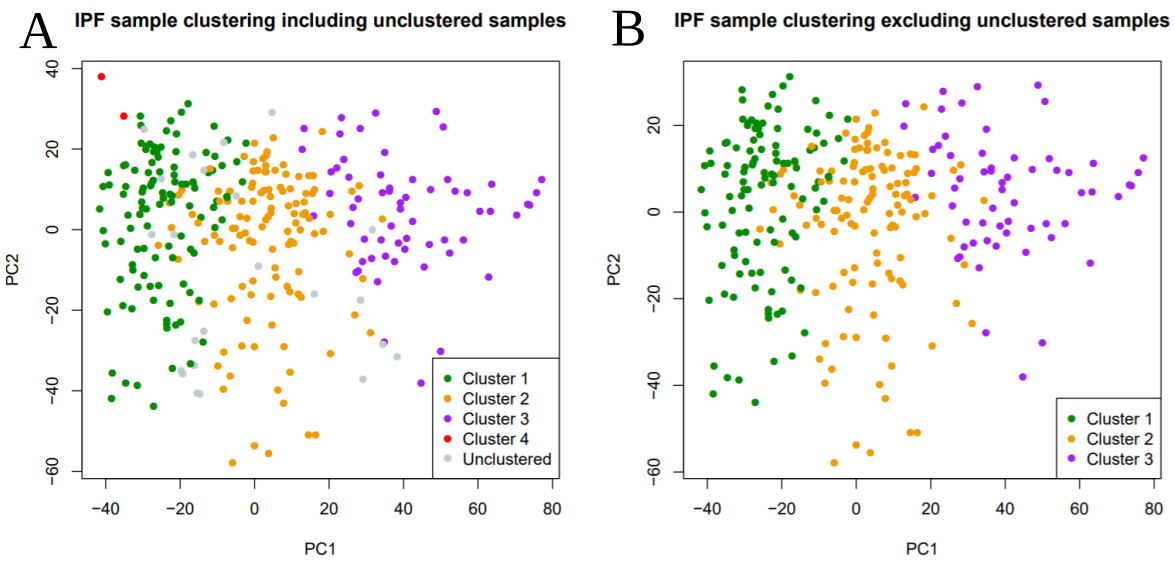


FIGURE 5.19: *Plots of the first two principal components for the clustered IPF samples from the seven remaining lung tissue datasets, before (A) and after (B) the removal of the samples in cluster 4 and the unclustered samples.*

TABLE 5.29: *Comparison of clinical and demographic traits for the clustered subjects in the lung tissue analysis. FVC = Forced vital capacity, D_{LCO} = Diffusing capacity for carbon monoxide, FEV_1 = Forced expiratory volume in one second, BMI = body mass index.*

	Total n used	**No. of datasets**	**Cluster 1 (green)** (n=111)	**Cluster 2 (orange)** (n=120)	**Cluster 3 (purple)** (n=61)	**P-value**
Phenotypic Trait						
Age (mean, sd)	214	4	62.2 (9.4)	64.8 (8.0)	63.4 (7.2)	0.170
European ancestry (%)	32	1	11 (68.8%)	10 (90.9%)	5 (100%)	0.177
Male (%)	247	5	62 (66.0%)	65 (70.7%)	40 (72.1%)	0.670
Ever smoker (%)	204	4	48 (61.5)	48 (67.6%)	35 (63.6%)	0.606
Smoking pack years (median, IQR)	114	2	20.0 (39.5)	19.5 (28.0)	14.5 (40.0)	0.897
Predicted FVC (mean, sd)	178	3	65.5 (18.1)	63.3 (16.8)	60.3 (16.8)	0.282
Predicted D_{LCO} (median, IQR)	158	3	49.8 (32.1)	43.3 (21.3)	44.9 (26.9)	0.301
Predicted FEV_1 (mean, sd)	65	2	72.9 (18.8)	72.5 (17.4)	-	0.921
St George's score (median, IQR)	115	1	49.1 (37.6)	43.0 (41.5)	43.5 (34.7)	0.724
BMI (median, IQR)	32	1	25.9 (6.3)	27.2 (6.8)	27.7 (3.4)	0.878

5.16 Cluster analysis in a single lung tissue dataset

As repeating the cluster analysis using multiple lung tissue datasets was inconclusive, perhaps due to the weak clustering as a result of the poor co-normalisation, an additional cluster analysis was performed using a single lung tissue dataset. Focusing the analysis on one dataset meant that the co-normalisation step could be circumvented, which could have improved the strength of the clustering and allowed for the identification of clinically distinct clusters despite the reduction in sample size and statistical power. In addition, using only a single cohort meant that a greater number of genes were available for inclusion in the clustering.

5.16.1 Methods

The dataset with the greatest number of IPF cases was selected as the sole dataset for inclusion in the analysis. The data for the IPF cases was clustered using COMMUNAL (Section 5.6.1) and the resulting 3D map was used to select the optimal clustering assignment. Phenotypic traits were then compared across clusters using the same approach as described previously (Section 5.7.1).

5.16.2 Results

The largest dataset was GSE47460 with a total of 122 IPF cases (including the individuals who were found to be common with GSE32537 and were excluded from the previous analyses). Age, sex, smoking status, predicted FVC, predicted D_{LCO} and predicted FEV_1 were available for the individuals in this cohort. There were 15,181 unique genes in the dataset, which was clustered using COMMUNAL (Figure 4.20). Again, there were not many stable optima across the tested space. K=2 clusters consistently displayed the highest mean standardised validity score (indicated by the blue squares) and had the greatest number of stable optima, so this was selected as the optimal number of clusters. The lowest number of genes for which there was a stable optimum at K=2 clusters was 250 genes, so this assignment was selected as the optimal assignment.

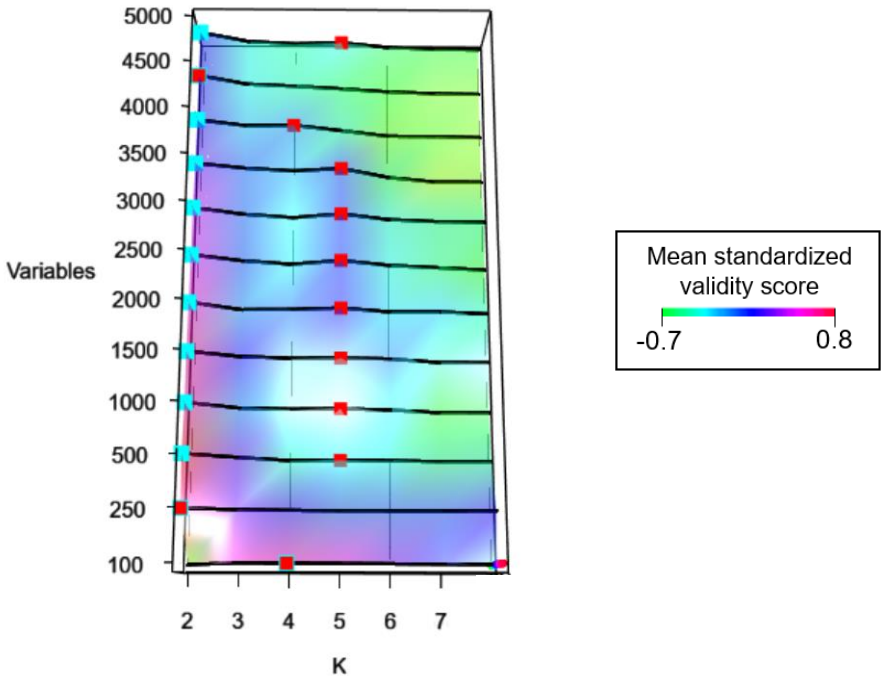


FIGURE 5.20: *The 3D map to identify the optimal cluster assignment output by COMMUNAL when applied to the data from the single lung tissue dataset GSE47460. The map shows the mean of standardized values of each validity measure across the entire tested space. On the 3D map, blue squares indicate a potentially optimal clustering at a certain number of genes by finding the assignment where the mean combined validation metric is greatest. The absolute maximum K for any consensus subset is marked with a red square. A higher validity score indicates a better clustering assignment and stable optima are the points where the blue and red squares meet. If stable optima at K clusters are seen over most of the tested space, this indicates the presence of a strong, consistent biological signal at this number of clusters.*

The two clusters were clearly separated in gene expression space (Figure 5.21), suggesting a stronger clustering than that observed when incorporating multiple lung datasets. 112 samples were successfully clustered (91.8%) and the unclustered samples were removed from the analysis. The clinical and demographic traits of the clustered subjects were compared across the two clusters (Table 5.30). As before, the results were inconclusive as there were no significant differences in clinical and demographic traits across clusters, perhaps through a lack of power due to the decreased sample size.

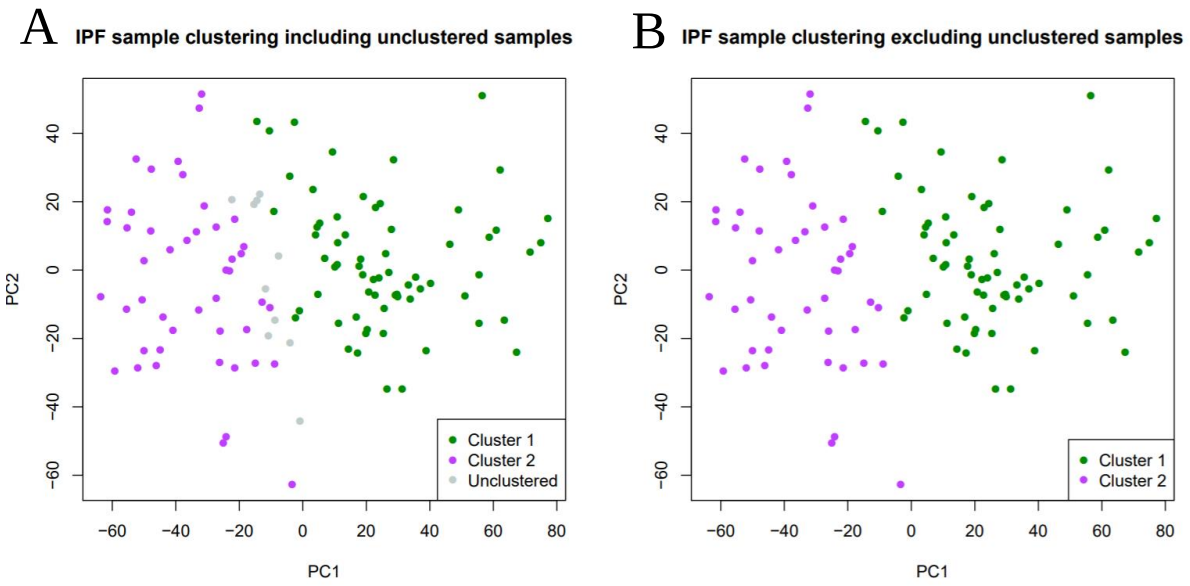


FIGURE 5.21: *Plots of the first two principal components for the clustered IPF samples from the lung tissue dataset GSE47460, before (A) and after (B) the unclustered samples were removed.*

TABLE 5.30: *Comparison of phenotypic traits across clusters for the single lung tissue dataset. FVC = Forced vital capacity, D_{LCO} = Diffusing capacity for carbon monoxide, FEV_1 = Forced expiratory volume in one second.*

Trait	Cluster 1 (green) (n=66)	Cluster 2 (purple) (n=46)	P-value	n used
Age (mean, sd)	65.0 (8.5)	64.0 (8.3)	0.534	112
Male (%)	45 (68.2%)	32 (69.6%)	1.000	112
Ever smoker (%)	38 (61.3%)	31 (67.4%)	0.431	108
FVC (mean, sd)	63.4 (18.3)	65.5 (13.9)	0.325	106
D_{LCO} (mean, sd)	46.5 (25.3)	50.0 (23.0)	0.120	101
FEV_1 (mean, sd)	69.8 (18.9)	73.2 (16.0)	0.325	106

5.17 Discussion

This chapter described cluster analyses of multiple publicly available transcriptomic datasets that were performed to identify groups of IPF patients that could be representative of distinct disease endotypes. The first of these analyses was conducted using three datasets (totalling 220 IPF cases) in which the gene expression data were measured from whole blood samples. A new method of transcriptomic data pooling (COCONUT) was utilised to reduce the technical differences between the transcriptomic datasets so that the data for the IPF cases in each cohort could be combined. Following this, another new bioinformatics method (COMMUNAL) was used to cluster the pooled data, resulting in three clusters of patients with IPF (termed the red, blue and yellow clusters). There were statistically significant differences in lung function, GAP index and survival between the clusters, with those in the blue cluster appearing to be healthier on average compared to those in the red and yellow clusters and at a lower risk of mortality.

Gene enrichment analysis was used to characterise the clusters in terms of the underlying biology. It was found that the genes that were most differently expressed in the individuals in the red cluster were significantly enriched for biological processes and biological pathways related to metabolic changes. Recent findings appear to suggest that metabolic dysregulation could be a contributing factor to fibrosis, though its role is not yet fully understood[223,224,224]. It has been previously reported that metabolic changes in IPF, such as decreased electron transport chain function, could potentially impact the functionality of lung cells and activate fibrotic responses[236]. In turn, these changes can lead to the activation of the TGF-β signalling pathway, which is a known driver of fibrosis[211,212,212,213,213]. Interestingly, the genes assigned to the red cluster were also significantly enriched for pathways related to TGF-β signalling.

Among the biological pathways that were significantly enriched for the blue cluster were pathways related to apoptosis and cell cycle. As discussed in Section 4.4, apoptosis has been previously implicated in IPF development and apoptosis-based therapies for IPF have been proposed. Additionally, genetic variants within cell cycle genes have been shown to be associated with IPF development and progression[227]. The results for this cluster could further support the idea that apoptosis and cell cycle each play an important role in the pathology of IPF.

The genes that were assigned to the yellow cluster were significantly enriched for processes related to the immune system response, as well as pathways related to the immune response, including an IL-6 signalling pathway. IL-6 has been implicated previously in IPF where it was reported that increased levels of IL-6 are related to an increased risk of death in IPF patients[237]. It is interesting that this cluster was enriched for biological pathways and processes related to the immune response as the role of the immune response in IPF has been controversial in the past, with some types of immunosuppresants being found to lead to worse clinical outcomes for some IPF patients[238]. This has led to speculation that

some immune responses in IPF are harmful whilst some are protective[221,222,222]. Perhaps IPF consists of multiple immune-driven endotypes, which is consistent with the results of the gene enrichment analysis, and these harmful and protective immune responses are each more predominant in certain endotypes. If this were the case, perhaps immunosuppressants could prove effective (and safe) in treating IPF when targeted to a specific endotype as part of a precision medicine approach.

Whilst these findings may be able to provide some mechanistic insight into the pathology of IPF, it must be acknowledged that changes in gene expression can be causal or consequential of disease. This means that it is possible that the clinically distinct groups identified in this study could be so as a result of downstream effects of having the disease and that it was the disease itself that led to the activation of pathways related to metabolic changes, apoptosis, cell cycle or the immune system response. However, the genes assigned to the blue and yellow clusters were found to be statistically overconnected (in terms of direct gene regulation) to a significant number of the 14 genes that were implicated in the largest genome-wide association study meta-analysis of IPF susceptibility to date[209]. Importantly, this could suggest that some of the differences in gene expression across the clusters reflect causal effects of the disease rather than consequential effects.

Following this, a gene expression-based cluster classifier was developed to assign additional independent IPF cases to one of the three clusters. The original classifier used data from 23 genes and was successfully validated when applied to IPF subjects from three validation datasets (totalling 194 IPF cases). The classifier accurately reassigned 100% of discovery subjects and assigned the validation subjects to clusters that displayed statistically significant differences in survival between groups that were consistent with the discovery clusters. It was estimated that at any follow-up time, those in the red cluster were 2.89 times as likely to die than those in the blue cluster (95% CI = [1.41, 5.93], P=0.004), whilst those in the yellow cluster were 4.23 times as likely to die compared to individuals in the blue cluster (95% CI = [1.87, 9.60], P=0.001). However, the difference in survival over time between the red and yellow clusters was not significant (P=0.341).

This classifier showed signs of being overfit to the discovery data, which could have affected its generalisability to additional datasets and so the feasibility of less overfit, reduced-gene classifiers was evaluated. A classifier that used 13 genes was shown to be superior to the full 23 gene classifier when applied to validation datasets, as it was able to define groups with greater differences in lung function as well as survival, therefore more closely resembling the clusters that were observed in the discovery analysis.

This 13 gene classifier was shown to have the ability to assign IPF cases in such a way as to, on average, put the individuals who are at a lower risk of death into the blue cluster and the individuals who are at a greater risk of death into the other two clusters. Therefore, it could potentially be used as a prognostic biomarker. The performance of the 13 gene classifier in predicting survival was compared to SAMS,

another transcriptomic approach to outcome prediction in IPF[186], which uses expression data from 52 genes. Both approaches were applied to the two validation cohorts for which survival information was available. The classifier was able to assign individuals so that at any follow up time, those in the 'high-risk' (red and yellow) clusters were estimated to be 4.25 times as likely to die than those in the 'low-risk' (blue) cluster (95% CI=[2.14, 8.46], $P=3.7\times10^{-5}$). SAMS performed less well at discerning between truly high-risk and low-risk individuals; those who were assigned to the high-risk group were estimated to be 1.98 times as likely to die at any particular time compared to those who were assigned to the low-risk group (95% CI=[1.07, 3.68], P=0.030). In addition, when added to a model adjusting for age, sex, ancestry, FVC and D_{LCO}, the predictions made by the classifier led to a significant improvement in predictive ability (P=0.005), whereas the predictions made by SAMS did not (P=0.105).

There are a few possible reasons that may explain why the classifier performed better than SAMS at assigning individuals into the correct risk group and thus predicting survival in IPF. Firstly, different approaches to cluster assignment were used. Using the classifier, every subject is given a score for each cluster, which are fed into a multinomial model that considers all three scores and predicts the most likely cluster assignment for that individual. Whereas in SAMS, individuals with an up score greater than the median value and a down score lower than the median value are classified as high risk. This means that an individual can have the greatest up score of the entire cohort but if their down score is just slightly above the median, they will be classed as low risk. The converse is also true for an individual with an extreme down score and an up score just above the median. These examples should highlight that there is a greater risk of misclassification using this method where truly high-risk individuals could be classed as low-risk. This is backed up by the fact that the classifier assigned 54 individuals into the high-risk clusters while SAMS only classed 44 individuals as high-risk.

Secondly, the 52 genes that are used in their method were originally included as they were found to be associated with transplant-free survival. As such, deaths and lung transplants were both treated as events of interest in the time-to-event analysis within their study. Whereas in the analyses described in this thesis chapter, deaths were the only outcome of interest. This difference in methodology may have affected the efficacy of SAMS in our analysis.

Lastly, some of the 52 genes in the SAMS gene signature were not measured in the validation cohorts, which will have also influenced the efficacy of SAMS. However, as this was only 1 of 52 genes (1.9%) in the first cohort and 2 out of 52 genes (3.8%) in the second cohort, this is unlikely to have had a considerable impact on the risk group assignments. Overall, the results in this section suggested that the classifier may be a feasible cost-effective alternative to the current best transcriptomic prognostic biomarker in IPF.

The blood expression-based classifiers were then applied to additional transcriptomic datasets from IPF cases where the gene expression was measured from whole lung tissue. This was conducted to investigate whether the pathology in the lung reflected that found in the blood, which would have added some credibility to the hypothesis that the three clusters were representative of three distinct endotypes of IPF. However, there was no evidence that the classifiers were able to assign subjects in such a way as to recreate the clinically distinct clusters observed in the discovery analysis when applied to the lung tissue data. This may have been because only a limited amount of clinical data was available to discern the resulting clusters or may suggest that the pathology in the lung was not reflected by gene expression markers the blood.

Following this, the cluster analysis was repeated using multiple lung tissue datasets to see whether the results would be concordant with those observed in blood. However, issues with the co-normalisation step led to poor clustering, which in turn led to clusters that were not clinically distinct. The clustering was repeated using only the single largest lung tissue dataset (totalling 122 IPF subjects) so that the co-normalisation step was not required. Two clusters of IPF cases were found in this analysis, though again there were no significant differences in clinical traits between the two groups. It was unclear from the results of these analyses whether clinically significant clusters were not found due to a lack of statistical power (as a result of the high proportion of missing clinical data in the first analysis and the reduced sample size of the second), or because the clusters were not representative of distinct pathophysiological states of the disease.

One of the main strengths of the first analysis described in this chapter was that the utilization of COCONUT allowed for three datasets to be combined, resulting in a relatively large discovery sample size of 220 IPF subjects, which increased the statistical power of the analysis and possibly allowed for more of the heterogeneity of the disease to be captured than if only one dataset were analysed. When also considering the three additional cohorts of independent IPF cases that were included in the validation stage, this study was one of the largest transcriptomic studies in IPF to date with a total of 416 IPF cases.

Another strength of these analyses was the application of COMMUNAL, which considered two different clustering algorithms (K-means clustering and PAM) and tested five validity measures over a range of genes. This allowed for the selection of an optimal clustering assignment, in terms of the most robust number of clusters in the dataset as well as the number of genes for which there was the greatest ratio of signal (informative genes) to noise (uninformative genes). These factors meant that this clustering was more reliable and more likely to be reproducible than the standard approach, which would have been to apply one clustering algorithm and test one validity measure.

Previous studies on this topic have often split IPF cases based on clinical variables or definitions before comparing levels of gene expression between those groups[183,184,184]. However, this approach is subject

to misclassification errors which could result in an individual being assigned to the incorrect group. For example, an individual in one of these studies could have been classed as having 'mild IPF' based on their lung function during the study but then have gone on to experience a sudden rapid progression of the disease and death shortly after the study, while another subject who was classed as having 'severe IPF' could have gone on to live for several years after the study. These misclassifications will have weakened the results of the study as those individuals' gene expression profiles were compared as part of the incorrect groups. On the contrary, the studies in this chapter did not use clinical data or definitions prior to cluster assignment and so were not reliant on these clinical variables or definitions. Furthermore, statistically significant differences in lung function, GAP index and survival were found between the clustered subjects in the first study despite clinical data not having been used in the clustering. This shows that clinically distinct groups of IPF patients can be identified through the clustering of gene expression data and could suggest that those groups may represent distinct endotypes of IPF with clinical relevance. Of course, a limitation of both approaches is that they are dependent on the presence of accurate and extensive phenotype data. In addition, it cannot be concluded unequivocally that the differences in survival across clusters were the result of differing IPF pathology between subjects, as other co-morbidities (such as age or heart disease) may have also varied across clusters. Whilst age was reported for nearly all subjects in the analysis and there were no significant differences in age observed across clusters, the limited phenotype data for other co-morbidities meant that the influence of unmeasured factors cannot be ruled out. Therefore, these clusters should be confirmed in a prospective study that considers a wide range of relevant phenotypes before conclusions regarding the existence of disease endotypes can be made.

In Sweeney et al.[94], each gene used in their clustering was assigned to the cluster in which its expression was most different to the expression in the other clusters, as this suggests that this gene contributes to the identity of that cluster. A similar approach was taken in this study, though with the improvement that ANOVA tests were introduced which allowed for the significance of the association between each gene and each cluster to be calculated. This allowed for the genes that were not at least nominally significantly associated to the cluster to which it was assigned to be excluded prior to the gene enrichment analysis. The inclusion of these uninformative genes would have weakened the enrichment analysis and so excluding them provided an additional level of confidence to the results from this section.

However, there were several limitations to these analyses. The objective of this chapter was to identify groups of patients that could represent distinct endotypes of IPF through cluster analysis. However, the red and yellow clusters could not be distinguished in terms of clinical traits. It is possible that these clusters were in fact clinically distinct, but that a difference between the groups was not detected in this analysis. A possible reason for this is that the analysis relied on the use of publicly available data and as a result some clinical variables (e.g. ancestry and FVC) were relatively underpowered in the

discovery analysis due to not being reported in all three studies. Alternatively, it is possible that the red and yellow clusters could both represent the same endotype of IPF. In the subsequent cluster analysis of a single lung tissue dataset, two clusters were found which could support this theory. However, as those clusters were not clinically distinct, the conclusions that can be made from that analysis are limited. This is an additional reason why a prospective study on this topic would be beneficial, as a well powered prospective study with comprehensive clinical data collection for all participants could help to clinically distinguish the clusters.

There are also some limitations to the gene expression data that should be recognised. Firstly, in the cluster analysis of whole blood datasets, not all genes were measured in each discovery stage study, which meant that many potentially informative genes could not be included in the analysis. This restricted the total number of genes in the analysis to approximately 9,000, even though the study with greatest coverage measured over 20,000. Additionally, due to the way that probes were mapped to genes, it is possible that the detection of different transcripts was used to represent the expression for the same genes across studies. Whilst expression for transcripts that represent the same gene are generally assumed to be correlated, this is not always the case[239]. In addition, as gene expression is tissue-specific, the results from the blood analysis may not be generalizable to other tissue types, such as lung. Finally, due to the types of data that were available (microarray and bulk RNA-seq), it was not possible to investigate whether there were any cell-type-specific effects across clusters. Future studies of this type could benefit from including single-cell RNA-seq data to address this. Finally, there are limitations to the use of blood as the tissue type in a transcriptomic analysis. For example, blood is a mixture of cell types that can change greatly in response to stimuli such as an infection, which will impact transcript abundance[240]. This could add another layer of variability between datasets from different studies and could have reduced the efficacy of the data co-normalisation.

Another weakness of these analyses is that COCONUT assumes that the healthy controls across the different studies came from the same statistical distribution and so all differences between healthy controls across studies must have been due to non-biological variation. This is a strong assumption as any large differences in confounding factors (such as age, sex and ethnicity) between the groups of healthy controls would have restricted the efficacy of the co-normalisation. However, the healthy controls in each of the three discovery blood datasets appeared similar in terms of the available phenotypic traits (age and sex) and the co-normalisation appeared to work well as the data for the control subjects from the three cohorts overlapped considerably in high-dimensional space post-co-normalization. However, the red cluster did not contain any individuals from the dataset with GEO accession code GSE93606. This could suggest that the co-normalisation was imperfect and meant that the discovery subjects in the red cluster could not be assessed for the important clinical variables that were reported in only that dataset, including: survival over time, smoking status and FEV_1. Importantly, this may explain why there were three clusters in this analysis but only two that were clinically distinct;

it is possible that the red and yellow clusters were both representative of the same endotype of IPF and it was an imperfect co-normalisation that led to the individuals in these groups being clustered separately.

It was more evident in the lung tissue analyses that the controls did not fully account for the technical differences between datasets. There could be several reasons for this. Firstly, the lung tissue studies tended to have fewer control subjects than the whole blood studies, perhaps partly because the collection of a lung tissue sample (which requires a procedure such as a biopsy or a transplant) is more invasive than the collection of a whole blood sample (done through phlebotomy or venepuncture) and thus lung tissue samples are more difficult to obtain. This would have affected the accuracy of the co-normalisation as the correction factors for each study would be less reliable. Further, as lung tissue samples are harder to obtain than blood samples, it stands to reason that the authors from the lung tissue studies were less particular over their control subjects in regard to recruiting those that were well matched to the disease cases. If the control subjects from the lung studies were indeed less well matched to the disease cases, they would likely be less well matched to the controls from the other studies, which would have been detrimental to the co-normalisation. One approach to combat this issue would have been to remove controls that were clearly poorly matched to the IPF cases, such as those much younger than a typical IPF patient. However, there was such a large amount of unreported clinical data for the lung tissue studies, particularly for the control subjects, that this could not have been done for the majority of the included datasets and doing so for only a proportion of the cohorts may have meant that the remaining cohorts would have co-normalised less well.

As mentioned previously, the 13 gene classifier could have potential use as a clinical biomarker to predict IPF patient survival. However, several important factors must be considered before this is done. First, more extensive validation in additional independent cohorts is needed to further characterise the clusters and gain additional support for the classifier's ability to predict survival in IPF. Second, the classifier should only be introduced as a prognostic tool if it could improve the outcome for the patients by leading to a change in the clinical management of the disease. For example, perhaps those in a particular cluster may react more favourably to one of the two existing IPF drugs nintedanib and pirfenidone, or perhaps those in the high-risk clusters should be prioritised for lung transplants over those in the low risk blue cluster. Further work must be done to assess the possible benefit to treatment that the classifier could provide as a biomarker.

Third, in its current form, the classifier can only work effectively when applied to transcriptomic data from a whole cohort of individuals with IPF. It was designed in this way so that non-biological variation within the data would not need to be accounted for and removed prior to the cluster assignment, which allowed for the inclusion of studies that had not considered healthy control subjects. However, this is not ideal as it would hinder the clinical applicability of the classifier. A way to standardise the procedure

so that a single individual with IPF can be assigned to a cluster would be preferable and should be developed in the future.

Fourth, there are some factors that may interfere with the efficacy of the classifier, such as drug effects or the timing of the blood sample collection in the disease course. Factors such as this should be investigated before the classifier can be introduced to a wide range of patients with IPF. Fifth, the cost-effectiveness of the classifier must be evaluated. There is a precedent for a clinically cost-effective transcriptomic biomarker: MammaPrint, the 70-gene signature that is used to predict whether an individual's breast cancer tumours are likely to spread[241]. MammaPrint is cost-effective as it can prevent low-risk patients from receiving chemotherapy unnecessarily. As the classifier requires measurement of only 13 genes, it may cost less per patient to implement than MammaPrint, but the cost must be weighed up against the benefit that the classifier can provide to assess whether clinical use is feasible.

The cohorts that were used in these analyses were checked for common subjects across datasets. This included checking for the origin of the samples or whether the study participants had been given a unique reference code from their study. Those who were found to have featured in multiple cohorts were removed from all except one in an effort to minimise any bias in the results. However, the possibility of subject overlap between datasets cannot be completely ruled out. Still, it is unlikely that there were significant levels of subject overlap and so any bias in the results should not have had a substantial effect on the findings of these analyses.

A further weakness of these analyses is that each participating cohort of IPF patients was subject to survival bias, as only individuals who survived long enough to enrol into each study could have contributed their transcriptomic data to it. It is therefore likely that there were some individuals with IPF that would have been included in one of these studies, had they not died before their enrolment or prior to the collection of their tissue/cell sample. Their absence would have meant that less heterogeneity of IPF was captured in these analyses and as a result meant that the clusters found in this chapter were less likely to represent the entire range of endotypes of IPF. This survival bias would have had a greater effect on the lung tissue studies than the blood studies, as these required the use of explanted lung tissue/biopsies and the wait for these procedures likely would have been longer than the wait for the collection of the blood samples. This may have been a contributing factor as to why clinically distinct clusters were not identified using the transcriptomic data from lung tissue.

If the clusters identified in this chapter do truly represent endotypes of IPF, it may be worth speculating about the nature of these endotypes. A traditional discrete endotype model assumes that all individuals with IPF should fit the description for just one endotype and that there should be little heterogeneity between the individuals with a particular endotype. However, this approach is more consistent with disorders that have rare, high impact genetic and environmental exposures. Endotypes of a complex disease such as IPF, which has many known common genetic and environmental exposures, would

likely behave under a more complex model, such as the palette model described by McCarthy[242]. The palette model assumes that there are a range of key pathophysiological traits and processes (termed 'component pathways') which all contribute to the risk of disease, and every single person lies somewhere on a spectrum of disease risk, with their exact position defined by the sum of their genetic risk and history of environmental exposure to each component pathway. The results of the gene enrichment analysis in this chapter could implicate metabolic changes, cell cycle, apoptosis and the immune system response as being among the component pathways for IPF. Under the palette model, an individual with IPF does not have to be neatly defined as having a particular endotype, as they may exhibit biology suggestive of multiple endotypes. This would be consistent with the findings of the cluster analysis of whole blood datasets as there were roughly 10% of subjects who could not be placed into a cluster because the two clustering algorithms disagreed on their cluster assignment.

To conclude, the results from this chapter suggest that there are at least two clinically distinct groups of individuals with IPF that can be identified through clustering transcriptomic data. The three clusters identified were defined using expression from groups of genes that were significantly enriched for many different biological pathways and processes. Therefore, these clusters could be representative of distinct pathophysiological states of IPF and could suggest the existence of multiple endotypes of IPF. If so, these findings would implicate metabolic changes, cell cycle, apoptosis and the immune response as the dominant pathways underlying these endotypes. However, the existence of these endotypes should be confirmed through additional follow-up studies. Additionally, a classifier with the ability to assign individuals with IPF to one of the clusters was developed. With further development, this classifier could be a useful tool in outcome prediction and patient stratification in IPF.

Chapter 6 – Discussion

This thesis describes a series of analyses that were conducted to improve the understanding of the pathogenesis of idiopathic pulmonary fibrosis. There were two primary aims: i) to define the genetic determinants of age-of-onset of IPF and ii) to identify endotypes of IPF using gene expression data. This chapter will summarise the main findings of these analyses and discuss how they have made original contributions to the field of IPF research. Following this will be a discussion of the strengths and limitations of these analyses, as well as speculation on future work that could potentially follow this research.

6.1 Summary of thesis

Two different approaches were used to investigate the genetic determinants of the age-of-onset of IPF. First, genome-wide association studies were performed with the objective to identify common genetic variants (MAF>1%) that were significantly associated with the age-of-onset of IPF (Chapter 3). Following this, a gene-based collapsing analysis was conducted to investigate the possible role of rare genetic variants (MAF<1%) in the age-of-onset of IPF (Chapter 4). These analyses were the first genetic studies to investigate the age-of-onset phenotype in IPF. In these analyses, the age-at-diagnosis of IPF was used as a proxy for the age-of-onset of IPF. When the age-at-diagnosis of IPF was unavailable, the age-at-enrolment into a study was used as a proxy for this, as long as all individuals in that study were recruited within six months of their initial IPF diagnosis.

Two GWAS were performed in Chapter 3, using genetic data from 465 individuals with IPF from the PROFILE study, 210 from the Trent Lung Fibrosis Study and 98 from UK Biobank. In the first GWAS, linear regression was used to model the proxy for the age-of-onset of IPF and a two-stage study design was adopted. The discovery analysis (stage 1) was performed in the PROFILE cohort and suggestive signals of association were followed-up in the remaining two cohorts (stage 2). The results of stages 1 and 2 were then meta-analysed. In the second GWAS, improvements were made to the methodology and the design of the study by modelling the proxy for the age-of-onset of IPF using time-to-event analysis methods and implementing a 3-way GWAS meta-analysis study design with internal validation criteria (only variants that were nominally significant in each cohort and had a consistent direction of effects across all cohorts were included in the meta-analysis). There were no genetic variants that reached genome-wide significance ($P_{meta}<5\times10^{-8}$) in either of these analyses, but there were five independent genetic signals in the 3-way GWAS meta-analysis that met the internal validation criteria and reached suggestive statistical significance ($P_{meta}<5\times10^{-6}$). Signal refinement and functional follow-up was performed for these suggestively significant signals but none could be robustly linked to any genes using the resources available, which limited the biological interpretation of the signals.

In Chapter 4, whole-genome sequencing data (from 493 individuals with IPF from the PROFILE cohort) were utilised to identify genes in which an aggregated excess of rare variants was associated with the age-of-onset of IPF. Two different statistical methods were used to collapse and test genetic variants at the gene level: the Morris-Zeggini approach, a type of burden test, was used in the first analysis and SKAT, a type of non-burden test, was used in the second analysis. Initial results in the first analysis showed two genes (*IGF2BP2* and *RELT*) that reached the threshold for study-wide significance ($P<2.6\times10^{-6}$), but sensitivity analyses showed that these results were largely being driven through the presence of a single individual whose age-at-enrolment into PROFILE was much lower than the average. After the removal of this individual, there were no genes that reached the threshold for study-wide significance under the primary model or in any of the models in subsequent sensitivity analyses. Likewise, there were no genes that reached the threshold for study-wide significance in the second analysis using SKAT.

Chapter 5 described a series of transcriptomic analyses that were performed to identify endotypes of IPF. In the first analysis in this chapter, three publicly available whole blood gene expression datasets (220 IPF cases total) were co-normalised and clustered to identify groups of individuals with IPF that could represent disease endotypes. Three clusters of patients were identified, with significant differences in lung function and survival between clusters. The clusters were used to build a gene expression-based cluster classifier, which was then validated using three additional cohorts of individuals with IPF (194 IPF cases total). With a total of 414 IPF cases across both the discovery and validation stages, this was one of the largest gene expression studies in IPF to-date.

As IPF is a lung disease, the transcriptome of the lung may be more informative about IPF pathogenesis than the transcriptome of the blood. As such, analyses were also performed using publicly available whole lung transcriptomic datasets. First, the whole blood-trained cluster classifiers were applied to four whole lung expression datasets (total 243 IPF cases) to assess whether the gene expression differences in the blood reflected pathology in the lungs. However, the resulting clusters were not significantly clinically distinct. Second, the transcriptomic cluster analysis was repeated using eight whole lung gene expression datasets (totalling 339 IPF cases). However, this analysis was inconclusive as the co-normalisation was ineffective with substantial technical differences remaining between the datasets, which led to poor clustering. Finally, a single lung expression dataset of 122 IPF cases was clustered, which circumvented the need to remove technical differences between studies through co-normalisation. Two clusters were identified in this analysis, though again the results were inconclusive as there were no significant differences in clinical traits between the two clusters, perhaps due to a lack of power as a result of the lower sample size compared to the previous analyses.

6.2 Original contributions to the field

The analyses in this thesis were the first genetic studies in IPF to consider the age-of-onset phenotype. Although no variants reached statistical significance overall, there were five potentially interesting association signals in the 3-way GWAS meta-analysis that may prove to be true positives when larger samples sizes are available. Furthermore, it appears that the genetic determinants for the age-of-onset are likely distinct to those for IPF risk; lookups for the known IPF risk-associated variants within the results of the age-of-onset GWAS showed that none were significantly associated with the age-of-onset following adjustment for multiple testing. However, it is possible that this was due to a lack of statistical power within the age-of-onset analysis. In addition, there were no common genetic variants that were exerting very large effects on the age-of-onset comparable to the large effect of rs35705950 on IPF susceptibility. This means that future genetic studies to investigate this phenotype must prioritise gaining a sufficiently large sample size to provide enough statistical power to detect more modest genetic effects for common variants.

With additional support in independent datasets and with further functional follow-up to robustly link those signals to genes, the five suggestively significant signals in the age-of-onset GWAS could implicate new genes and pathways in the development of IPF. As such, studying this phenotype further could improve the understanding of the disease pathogenesis and could potentially lead to new treatment options for patients. Therefore, future genetic studies in IPF should continue to pursue this phenotype, preferably in studies of greater sample size and statistical power.

The cluster analysis of whole blood expression datasets was performed to identify groups of IPF patients that could represent clinically distinct endotypes of the disease. In this analysis, three clusters of IPF patients were identified and there were significant differences in lung function and survival across clusters. As clinical data were not used in the clustering process itself, the finding that these clusters were clinically distinct could suggest that they are representative of distinct pathophysiological states. Furthermore, gene enrichment analysis showed that the genes that were differentially expressed in each cluster were significantly enriched for many different biological pathways and processes, including metabolic changes (red cluster), cell cycle and apoptosis (blue cluster) and the immune response (yellow cluster). As discussed in Section 5.17, these findings are consistent with previous studies that have implicated metabolic dysfunction[223,224,224], cell cycle[227] and apoptosis[171,172,172] in the pathogenesis of IPF. As such, our findings suggest that drugs that target these mechanisms may warrant further investigation when considered as potential therapies for subgroups of IPF patients. Furthermore, as the clusters may represent groups of IPF patients with different predominant disease processes, cluster assignment may be informative as to treatments that might prove effective when targeted to a specific group of patients. For example, if chronic inflammation is revealed to be a driver of IPF for the patients in a particular cluster, the use of anti-inflammatories (which are known to not be effective when considered in an unselected patient population[215]) might prove beneficial when targeted specifically towards the

individuals in that cluster. Similarly, the results of the gene enrichment analysis are consistent with the existence of immune-driven endotypes in IPF, which may suggest that immunosuppressants could prove effective (and safe) in treating IPF when targeted to a specific endotype.

The cluster analysis also led to the development of a gene expression-based classifier, which could potentially have clinical applications in IPF. Firstly, as the classifier was able to assign IPF cases to clusters that showed significant differences in survival across groups, it could be used as a prognostic biomarker to predict outcome for IPF patients. These predictions could be used by clinicians to prioritise earlier treatment (such as anti-fibrotic therapy or a lung transplant) to those who are classed as high-risk. The classifier was shown to be more accurate at distinguishing individuals at a high risk of death from those at a low risk of death than another transcriptomic prognostic biomarker for IPF[186] and could be a more cost-effective clinical tool as it requires expression from fewer genes to be measured.

Secondly, stratifying patients into groups using the classifier and matching treatments to each cluster or risk group could allow for a precision medicine approach in IPF. For example, patients assigned to a particular cluster could be prescribed either pirfenidone or nintedanib depending on which anti-fibrotic therapy has been found to be most effective for the patients in that cluster. Similarly, the classifier could allow for immunosuppressive therapy to be targeted towards individuals with a particular immune-driven endotype. Of course, these hypotheses must be tested and proven in clinical trials before they can be introduced clinically and there are some limitations to the classifier that should be addressed first (Section 6.4).

It was of interest to investigate whether there were any genetic variants associated with the clusters (as putative IPF endotypes). However, except for the *MUC5B* promoter polymorphism rs35705950, genetic data could not be directly compared across clusters as this required paired genetic and transcriptomic data, which was not available. For rs35705950, there were no significant differences in genotype across clusters, although this may have been due to a lack of power as a result of the large amount of missing data. Gene enrichment analysis was used to test whether the genes that defined each cluster were statistically overconnected (in terms of direct gene regulation) to IPF risk-associated genes[209]. The genes that were most differentially expressed in the blue and yellow clusters were each found to be overconnected to a significant number of IPF risk-associated genes, which suggested that genes and biological mechanisms related to IPF susceptibility may be important to the development of the pathophysiological states that those clusters could represent. However, more extensive approaches are needed to fully evaluate the genetic basis of these putative endotypes (Section 6.4).

6.3 Strengths and limitations

In addition to the original contributions to the understanding of IPF pathogenesis discussed in the previous section, the analyses in this thesis had some important strengths. Firstly, the methodology used in these analyses was rigorous to ensure that any findings were genuine and robust. For example, the

internal validation criteria in the age-of-onset of IPF GWAS ensured that the genetic variants were associated with the proxy for the age-of-onset in all cohorts and so reduced the possibility of signals being reported that were being driven through a strong association in only one cohort. Additionally, the sensitivity analyses in the gene-based collapsing analysis highlighted the spurious results that were the result of the age outlier and COMMUNAL was used in the transcriptomic cluster analysis to provide a more reproducible and robust clustering than a traditional clustering approach.

Another strength was the use of the PROFILE cohort, which was the largest individual study in the age-of-onset GWAS, the only study in the gene-based collapsing analysis and contributed to the discovery stage of the transcriptomic cluster analysis. The individuals in the PROFILE cohort were prospectively enrolled and newly diagnosed, which meant that their age at enrolment could be used as a proxy for their age-of-onset, thus allowing for the first genetic studies of the age-of-onset phenotype to be performed. In addition, the individuals in PROFILE were treatment naïve as the two anti-fibrotic treatments currently licensed for IPF were not licensed at the time of study recruitment. This therefore reduced the possibility that any differences in gene expression between individuals were due to treatment effects. Also, the rich clinical data that was available for this cohort allowed for clinically significant traits to be studied, such as smoking history, lung function measurements and survival over time. However, any unique characteristics of this cohort could have potentially weakened or biased the results of all the analyses in this thesis at once. For example, if the individuals in PROFILE had a better average socioeconomic status than the general public, the results in this thesis may not be as generalizable to that population.

The analyses in this thesis had some additional limitations that should be recognised. Firstly, statistical power was a significant issue in these analyses due to the relatively low sample sizes that were available. For example, there were a total of 773 IPF cases in the age-of-onset GWAS but a GWAS investigating a continuous phenotype for a polygenic disease would require approximately 4,000 individuals to have 80% power to detect a variant with an additive genetic effect that accounts for 1% of the phenotypic variance at a genome-wide significance level ($P<5\times10^{-8}$). A relatively straightforward way to increase the statistical power of these analyses would have been to incorporate data from additional independent cohorts of IPF patients into the studies and to meta-analyse the results. However, aside from those used in this thesis, there were no genetic datasets available for IPF patients whose age-of-onset (or a suitable proxy) was known. Statistical power due to sample size was also a limitation in the transcriptomic cluster analysis and this was exacerbated by there being large amounts of missing data for some clinical variables. This limitation highlights the need for additional independent omic studies in IPF, as well as the need for more detailed collection of clinical data in those studies. In the interim, it may be beneficial for future studies in IPF to identify cases using large biobanks and inferring disease phenotypes, such as the age-of-onset, using electronic health records (Section 6.4).

Another weakness of these analyses is that they were almost entirely comprised of individuals of white European ancestry and so the results may not be generalizable to individuals of other ancestries. Furthermore, most previous genetic studies of IPF have been in populations of European ancestry. Unfortunately, this is typical of current genetic studies, with over 75% of participants in published GWAS being of European ancestry[243]. This data inequality risks missing disease-associated genetic loci that are specific to other ancestries and must be addressed in future studies. In addition, many of the available genetic and transcriptomic datasets for IPF originate from randomised control trials and registries and biases in terms of sex and ethnicity in recruitment for these kind of studies have recently been noted, with females and non-white individuals being underrepresented[244]. Going forward, action must be taken to ensure that study populations are more representative of individuals with IPF in the general population, including offering registry enrolment to all patients seen in ILD clinics and introducing quotas for enrolment of racial minority participants in clinical trials.

6.4 Future work

One way to increase the power of future omic analyses in IPF could be to use electronic health records (EHRs) to identify patients with IPF and define their phenotype. For instance, if the phenotype of interest in a future study was the age-of-onset of IPF, individuals in a biobank who have IPF would be identified using primary care data (such as general practice records and International Classification of Diseases [ICD] codes), whilst primary care codes for common IPF symptoms (such as cough and dyspnoea) could be used to determine their age when they first visited the doctor after developing IPF. This age, if estimated accurately, would provide a more accurate approximation of the age-of-onset of IPF than the proxies used in this thesis. In addition, this could allow for the inclusion of individuals with IPF from different ancestral groups by utilising biobanks from around the world, which would make the results more generalizable to populations of non-European ancestry.

However, there also some disadvantages to this approach. One recent study (not yet peer-reviewed)[245] performed the largest meta-analysis of IPF risk to-date by using EHRs to define IPF in biobanks but found that effect sizes for IPF susceptibility-associated SNPs varied across studies based on how the IPF cases were ascertained. Effect sizes were 2.1 times greater on average in studies where IPF was diagnosed clinically than in studies where biobanks and EHRs were used to define IPF, suggesting that there was misclassification of IPF in biobanks. Therefore, due to the expectation of attenuated effect sizes, biobank studies to investigate IPF phenotypes will likely require greater sample sizes than studies that use clinically defined IPF in order to maintain the same level of statistical power. Moreover, as IPF is an uncommon disease, the biobanks will likely need to be very large in order to accumulate enough IPF cases to provide sufficient statistical power to the study. In addition, the strengths and limitations of using primary care codes will be dependent on the particular IPF phenotype being derived. For example, defining an age-of-onset using primary care symptom codes may prove difficult as some symptoms will be caused by co-morbidities.

As discussed in Chapter 5, a large prospective cohort study to follow-up the transcriptomic cluster analysis is now needed. Firstly, this would allow for more extensive clinical data to be recorded for the IPF patients, including clinically important traits that were not available in these analyses, such as lung function over time, patient reported outcomes and treatment history. Secondly, a larger sample size, coupled with more comprehensive clinical data, will provide greater statistical power and possibly allow for more of the disease heterogeneity to be captured than in the Chapter 5 analyses. This increase in statistical power will help to elucidate the true number of clinically distinct endotypes of IPF. Thirdly, conducting a single large prospective study would mean that the gene expression data would not need to be co-normalised prior to the clustering, thus eliminating a potential source of bias. Fourth, an important question still stands as to whether the findings from this analysis about gene expression in the blood reflects pathology in the lungs. A prospective study could allow for gene expression from a more relevant tissue type than blood (e.g. whole lung or lung fibroblasts) to be investigated and clustered.

In addition, a prospective study could allow for genetic and transcriptomic data to be collected for the same patients, which would allow for genetic associations with clusters to be investigated. For example, rather than simply studying the *MUC5B* promoter variant rs35705950, a future study could test for associations between cluster membership and all of the SNPs that have been previously identified as being associated with IPF susceptibility. Alternatively, genome-wide approaches could be used to search for genetic variants that are associated with cluster membership. For example, assigning patients into clusters using the gene expression-based classifier and then performing a GWAS to compare individuals in the low-risk (blue) cluster against individuals in the high-risk (red and yellow) clusters could be informative as to whether the clusters identified in this thesis reflect genetically driven endotypes. However, such studies would likely require thousands of subjects in order to be well-powered.

Future work could also focus on the classifier, which could be utilised in clinical trials to evaluate the efficacy of treatments in each cluster of IPF patients. For example, the efficacy of the two current antifibrotic interventions (pirfenidone and nintedanib) for IPF could be assessed stratified by cluster membership. This could reveal whether targeting each of these treatments to a particular cluster improves its efficacy. Importantly, this may not require a new clinical trial to be conducted and could potentially be performed post-hoc if a previous clinical trial for pirfenidone or nintedanib had measured gene expression for the 13 genes in the classifier at baseline.

However, before the classifier can possibly be implemented as a clinical tool in IPF, additional work to address its limitations should be conducted. As discussed in Section 5.15, this would include the development of a reference panel and a standardised procedure for the collection of expression data (so that the classifier can be applied to an individual IPF patient as opposed to an entire cohort), more

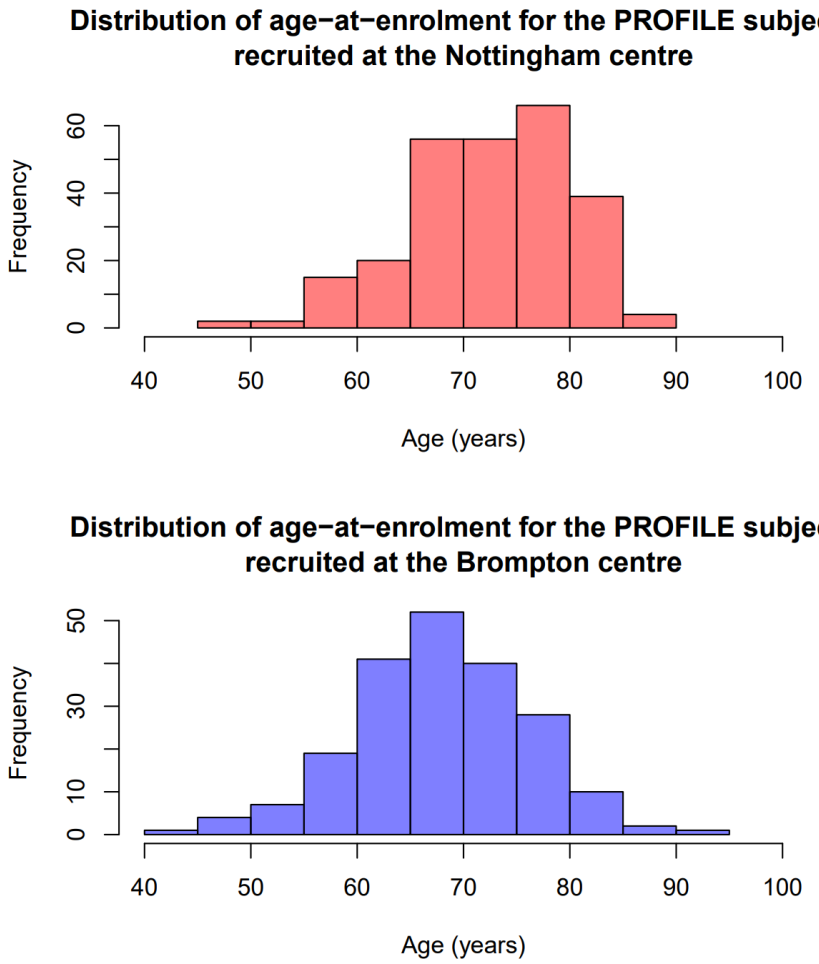
extensive validation in additional independent cohorts (to further characterise the clusters and to gain additional support for the classifier's ability to predict survival in IPF), the investigation of factors that could interfere with the classifier's efficacy (such as drug effects or the timing of the blood sample collection in the disease course) and an evaluation of the classifier's cost-effectiveness.

6.5 Conclusion

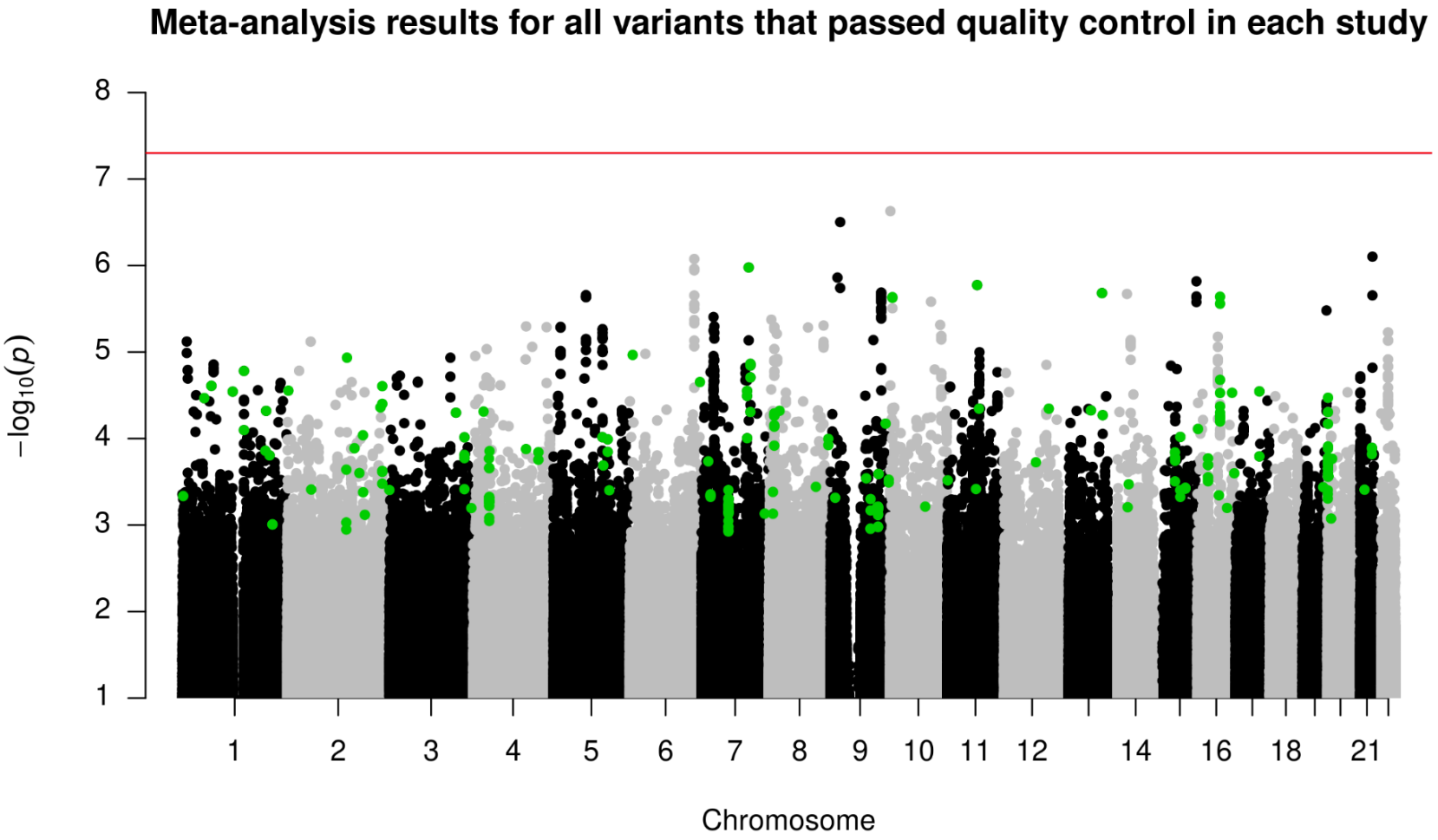
In this thesis I have contributed to the understanding of the pathogenesis of IPF by performing the first genetic analyses to study the age-of-onset of IPF, which highlighted some genes of potential interest as well as some important factors to consider when studying this phenotype. In addition, through combining and clustering multiple gene expression datasets I have identified potential endotypes of IPF and used these to develop a transcriptomic classifier capable of predicting outcome in IPF. These findings could lead to improvements in treatment for patients with IPF as well as inspiring future studies in the field of IPF research.

Appendix

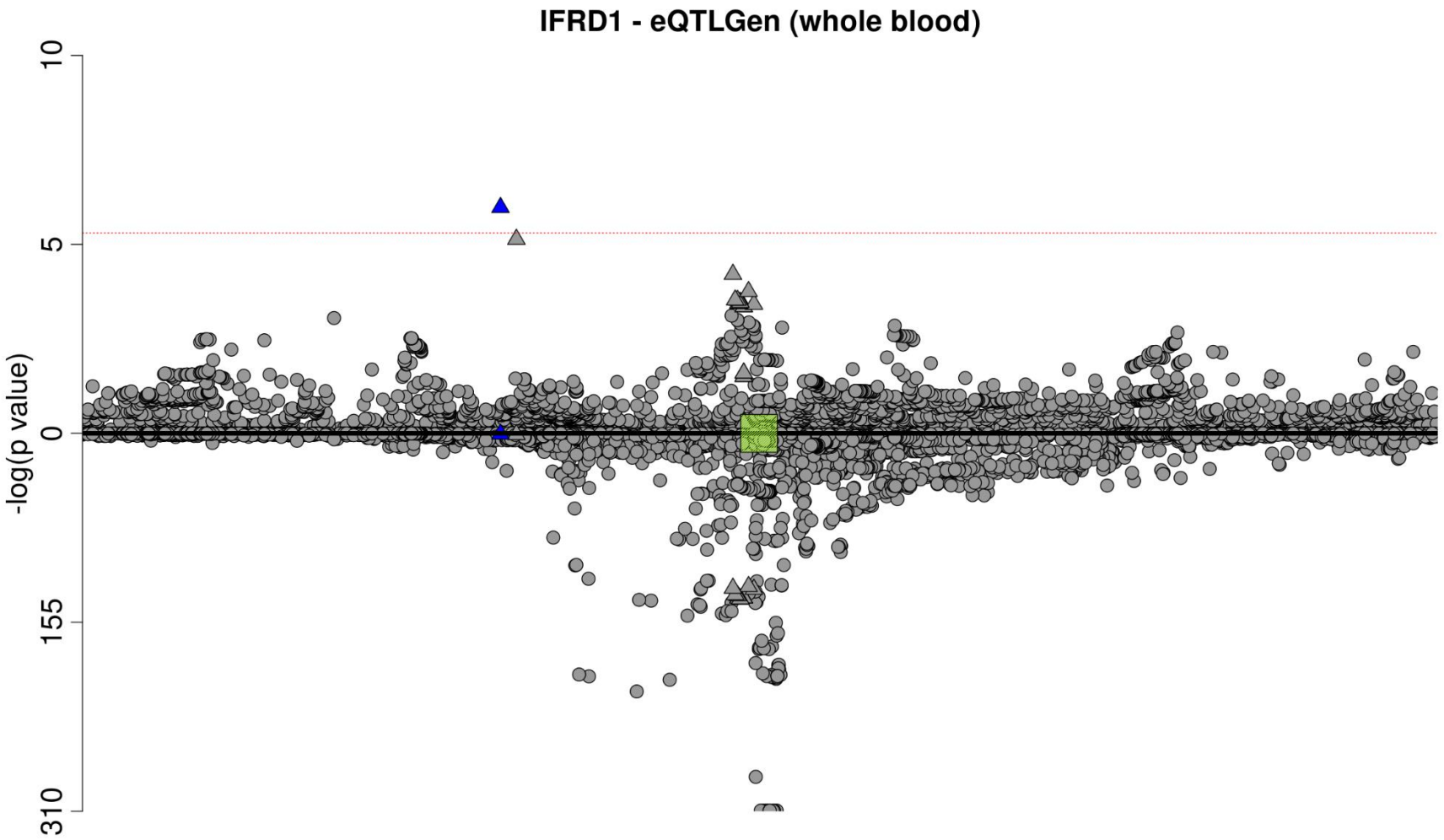
A: Additional Figures



ADDITIONAL FIGURE A.3.1: *Histograms showing the distribution of the age-at-enrolment of the PROFILE cohort, stratified by recruitment centre.*

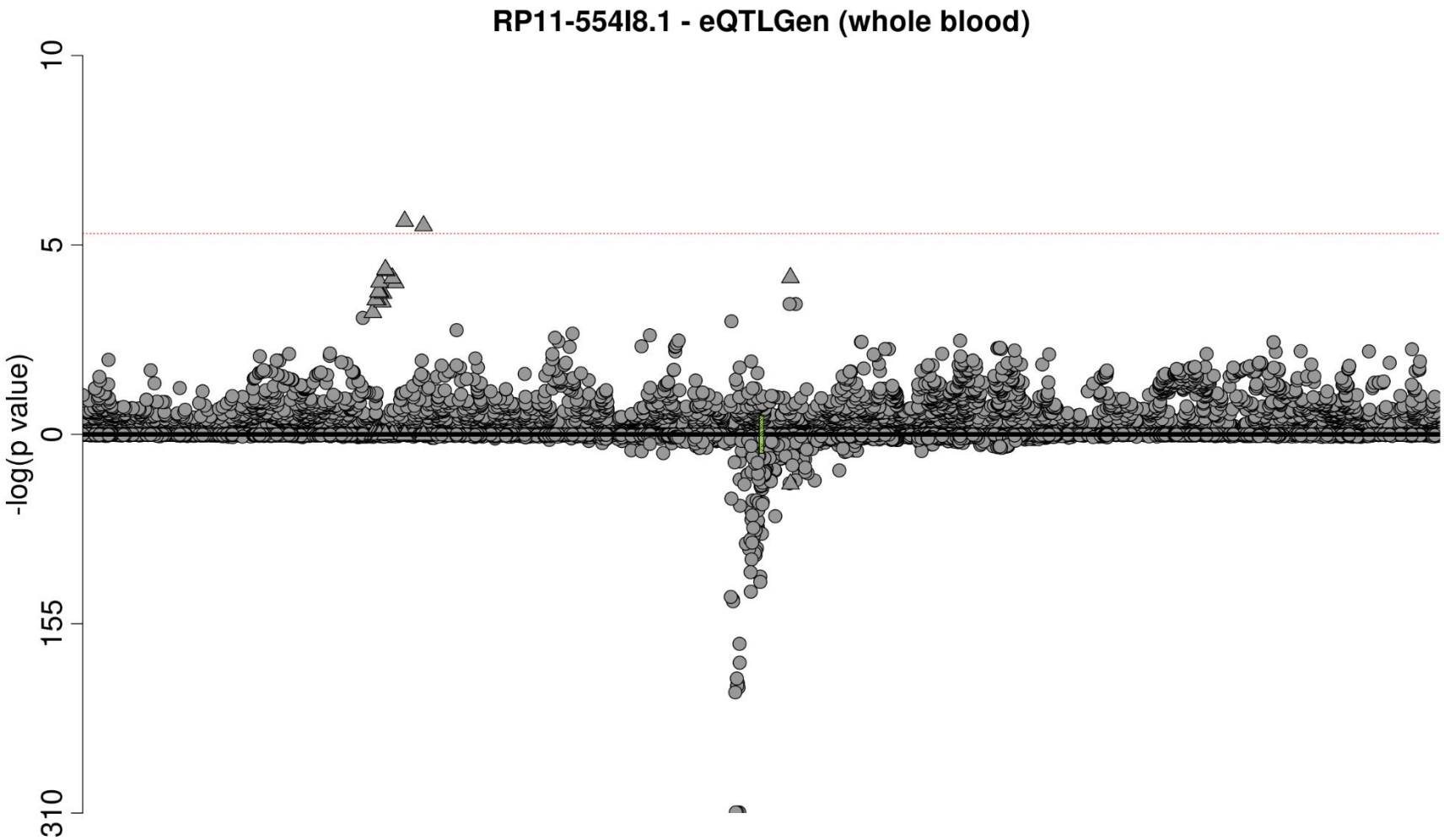


ADDITIONAL FIGURE A.3.2: *Manhattan plot showing the results of the meta-analysis for all SNPs that passed quality control in each of the three study cohorts. The SNPs that passed the internal validation procedure are highlighted green. Variants with a P-value greater than 0.1 were filtered out to reduce the computational burden of the plot. The red line indicates the threshold for genome-wide statistical significance ($P=5\times10^{-8}$).*

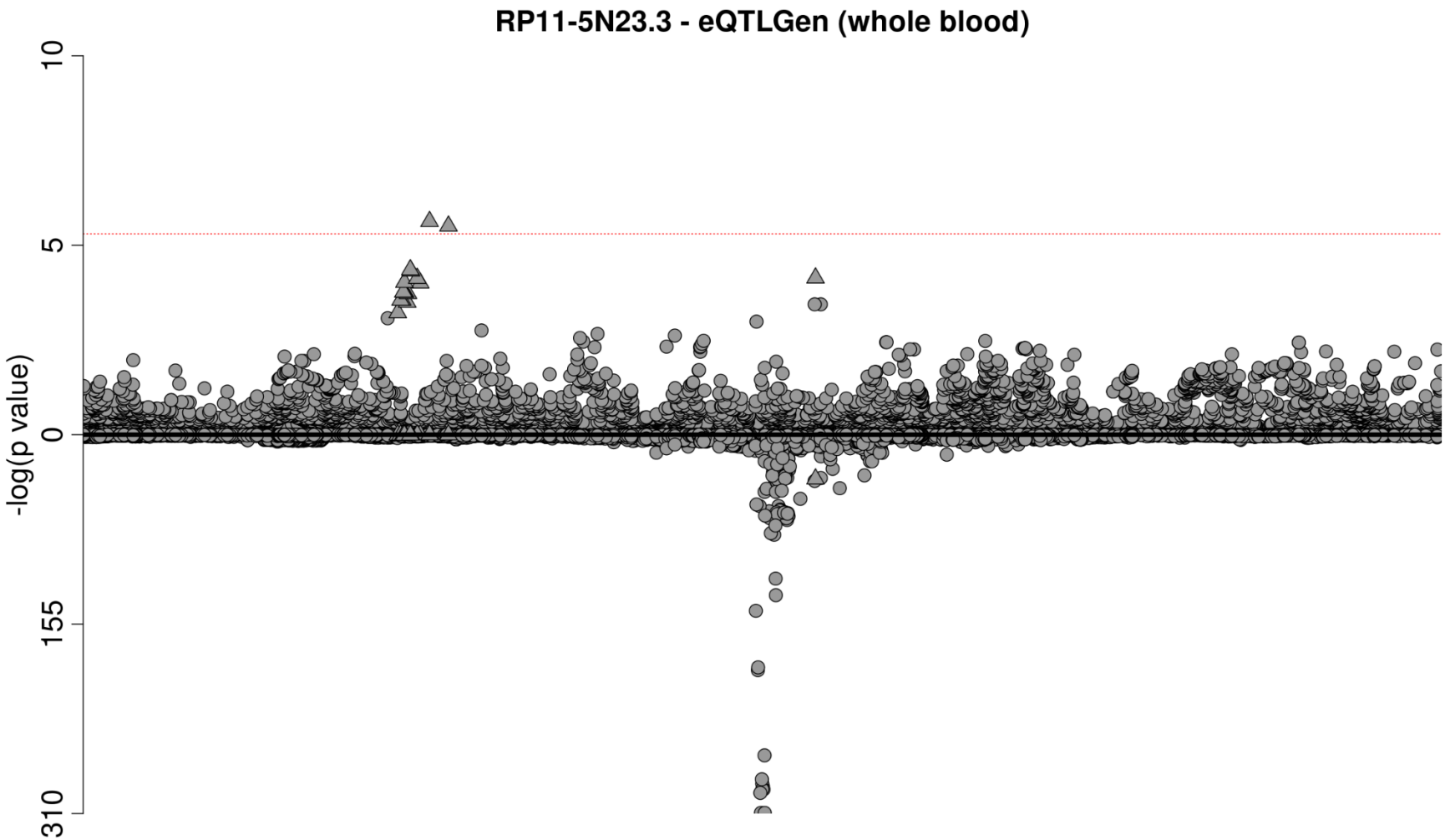


ADDITIONAL FIGURE A.3.3: *A mirror plot to jointly visualise the summary statistics from the age-at-diagnosis of IPF GWAS meta-analysis and the summary statistics from the analysis of cis-eQTLs for IFRD1. The sentinel variant (rs183759512) is coloured blue. All variants in the 95% credible set for the chromosome 7 signal in the GWAS are denoted by triangles whereas all other variants are denoted by circles. The gene region of IFRD1 is highlighted yellow.*

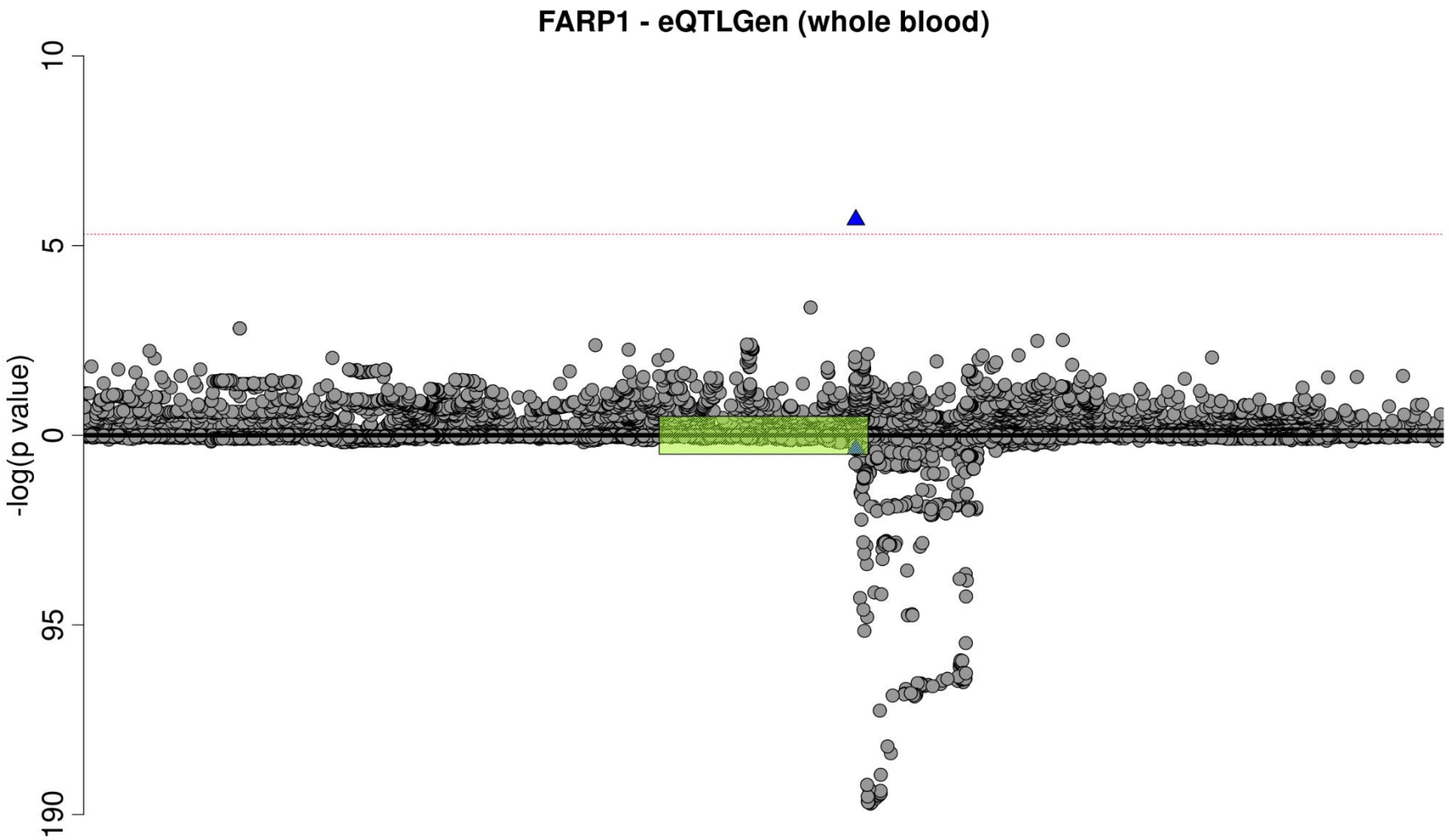
ADDITIONAL FIGURE A.3.4: *A mirror plot to jointly visualise the summary statistics from the age-at-diagnosis of IPF GWAS meta-analysis and the summary statistics from the analysis of cis-eQTLs for PRKCQ-AS1. All variants in the 95% credible set for the chromosome 10 signal in the GWAS are denoted by triangles whereas all other variants are denoted by circles. The gene region of PRKCQ-AS1 is highlighted yellow.*



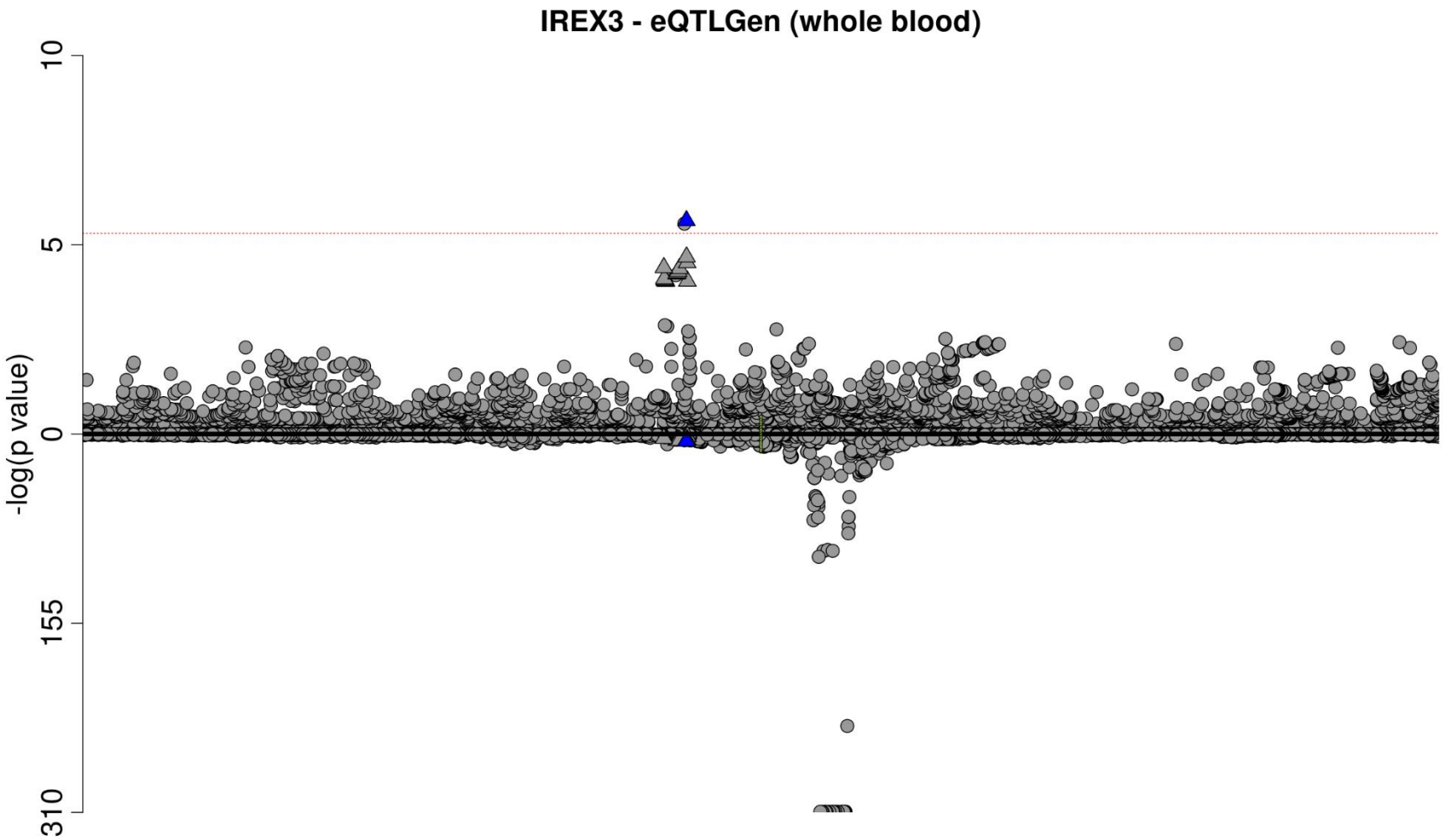
ADDITIONAL FIGURE A.3.5: *A mirror plot to jointly visualise the summary statistics from the age-at-diagnosis of IPF GWAS meta-analysis and the summary statistics from the analysis of cis-eQTLs for RP11-554I8.1. The sentinel variant (rs41295127) is coloured blue. All variants in the 95% credible set for the chromosome 10 signal in the GWAS are denoted by triangles whereas all other variants are denoted by circles. The gene region of RP11-554I8.1 is highlighted yellow.*



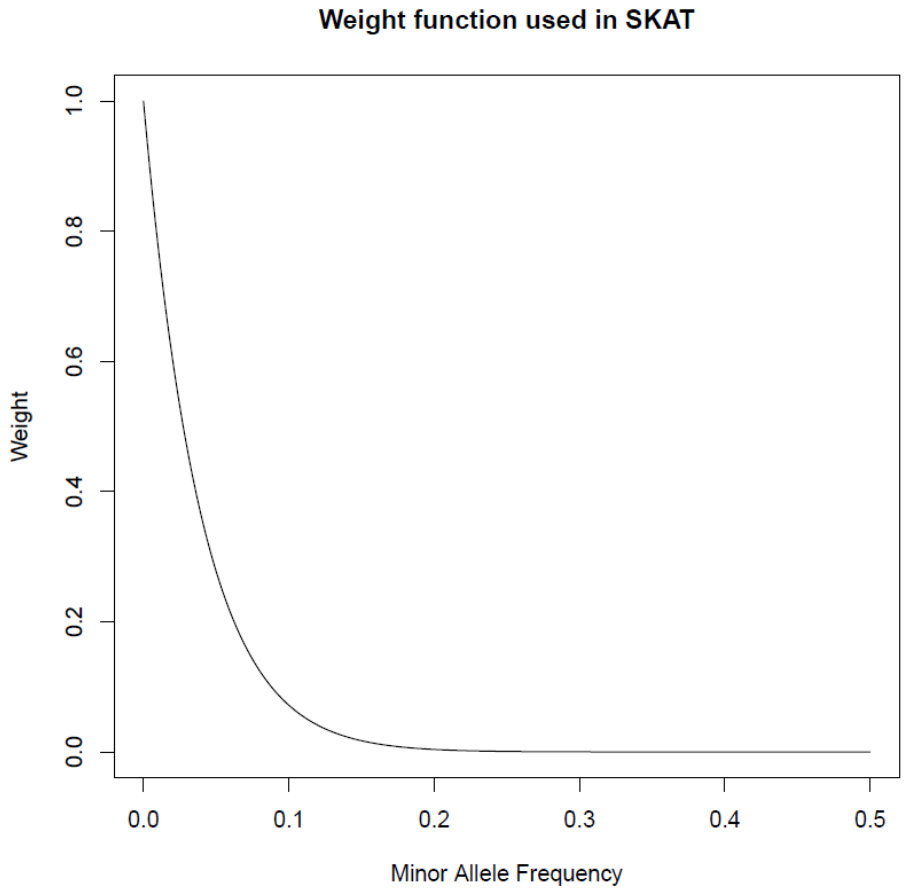
ADDITIONAL FIGURE A.3.6: *A mirror plot to jointly visualise the summary statistics from the age-at-diagnosis of IPF GWAS meta-analysis and the summary statistics from the analysis of cis-eQTLs for RP11-5N23.3. All variants in the 95% credible set for the chromosome 10 signal in the GWAS are denoted by triangles whereas all other variants are denoted by circles. The gene region of RP11-5N23.3 is highlighted yellow.*



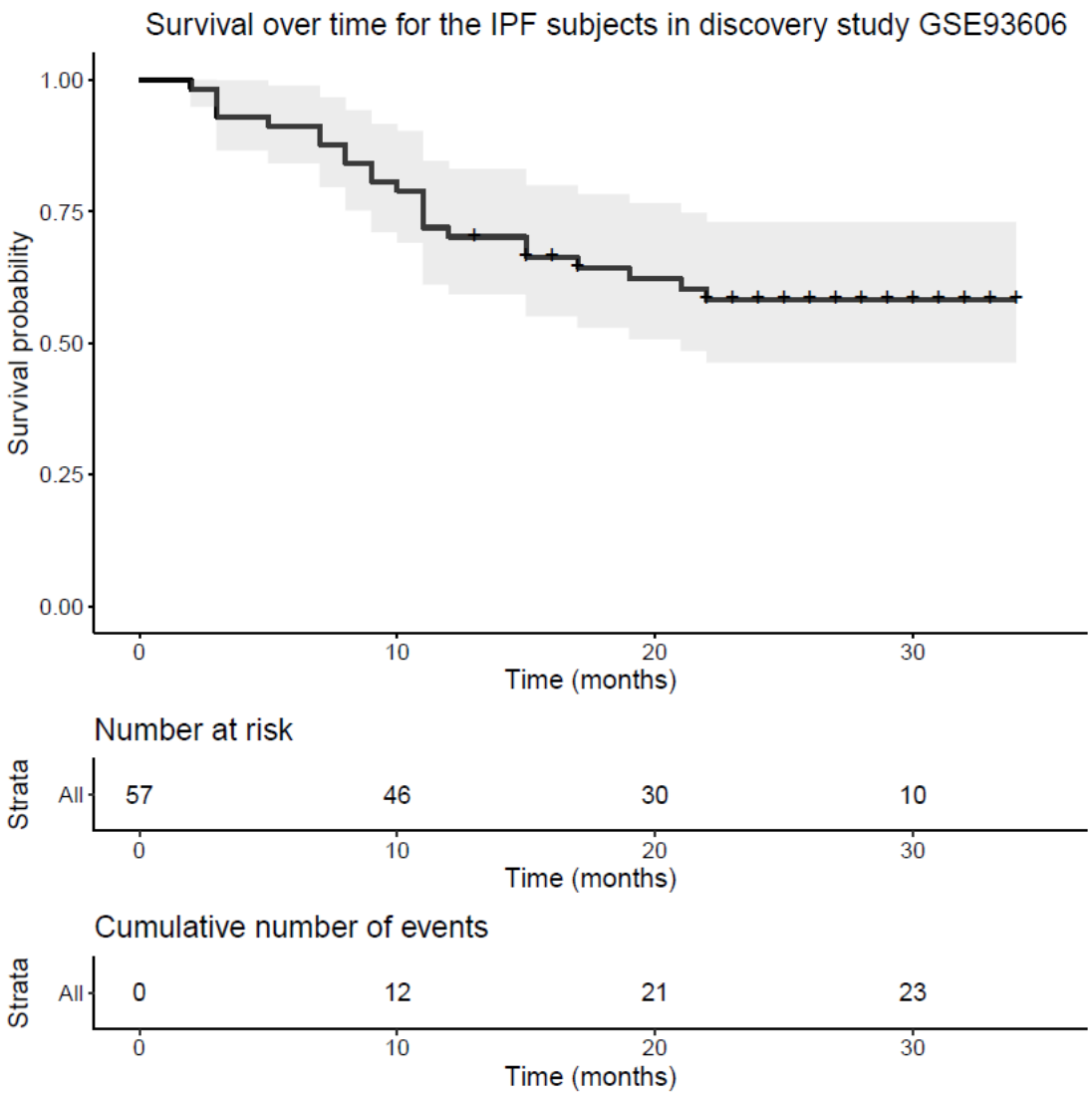
ADDITIONAL FIGURE A.3.7: *A mirror plot to jointly visualise the summary statistics from the age-at-diagnosis of IPF GWAS meta-analysis and the summary statistics from the analysis of cis-eQTLs for FARP1. The sentinel variant (rs9513422) is coloured blue. All variants in the 95% credible set for the chromosome 13 signal in the GWAS are denoted by triangles whereas all other variants are denoted by circles. The gene region of FARP1 is highlighted yellow.*



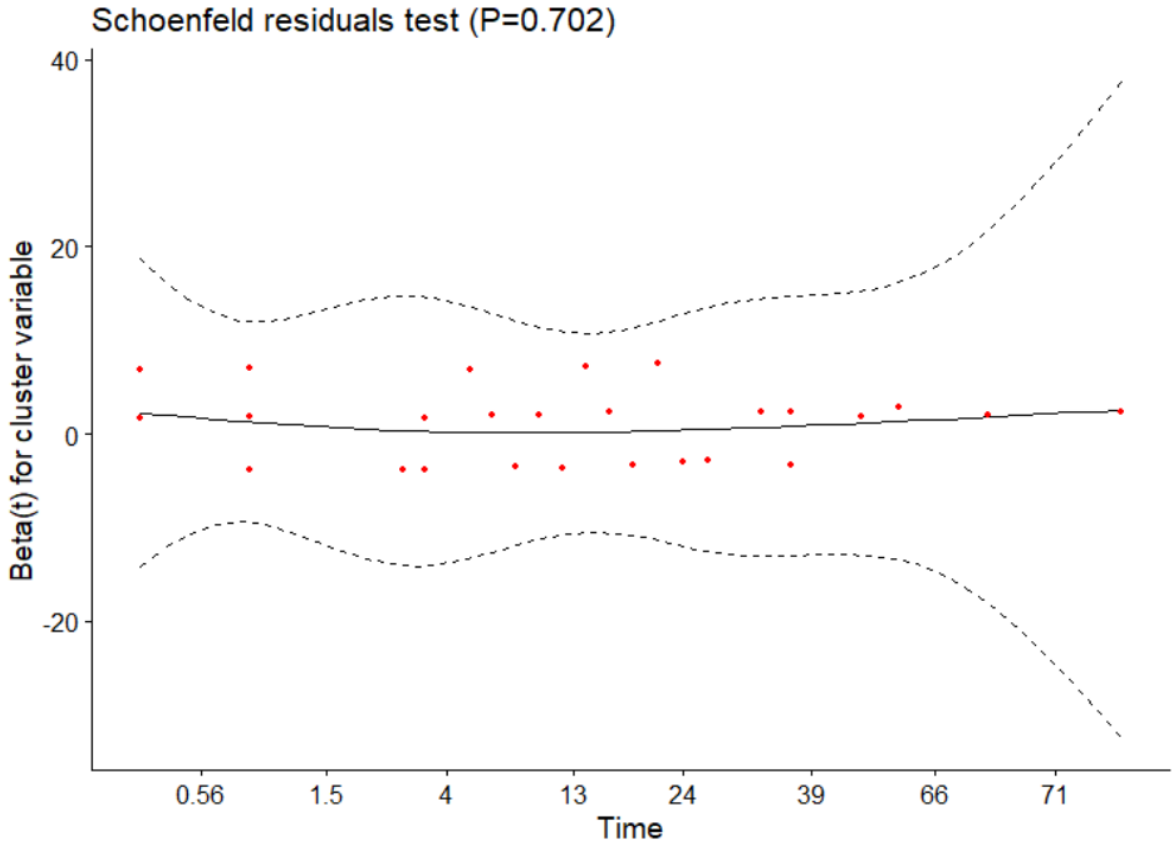
ADDITIONAL FIGURE A.3.8: *A mirror plot to jointly visualise the summary statistics from the age-at-diagnosis of IPF GWAS meta-analysis and the summary statistics from the analysis of cis-eQTLs for IRX3. The sentinel variant (rs118122250) is coloured blue. All variants in the 95% credible set for the chromosome 16 signal in the GWAS are denoted by triangles whereas all other variants are denoted by circles. The gene region of IRX3 is highlighted yellow.*



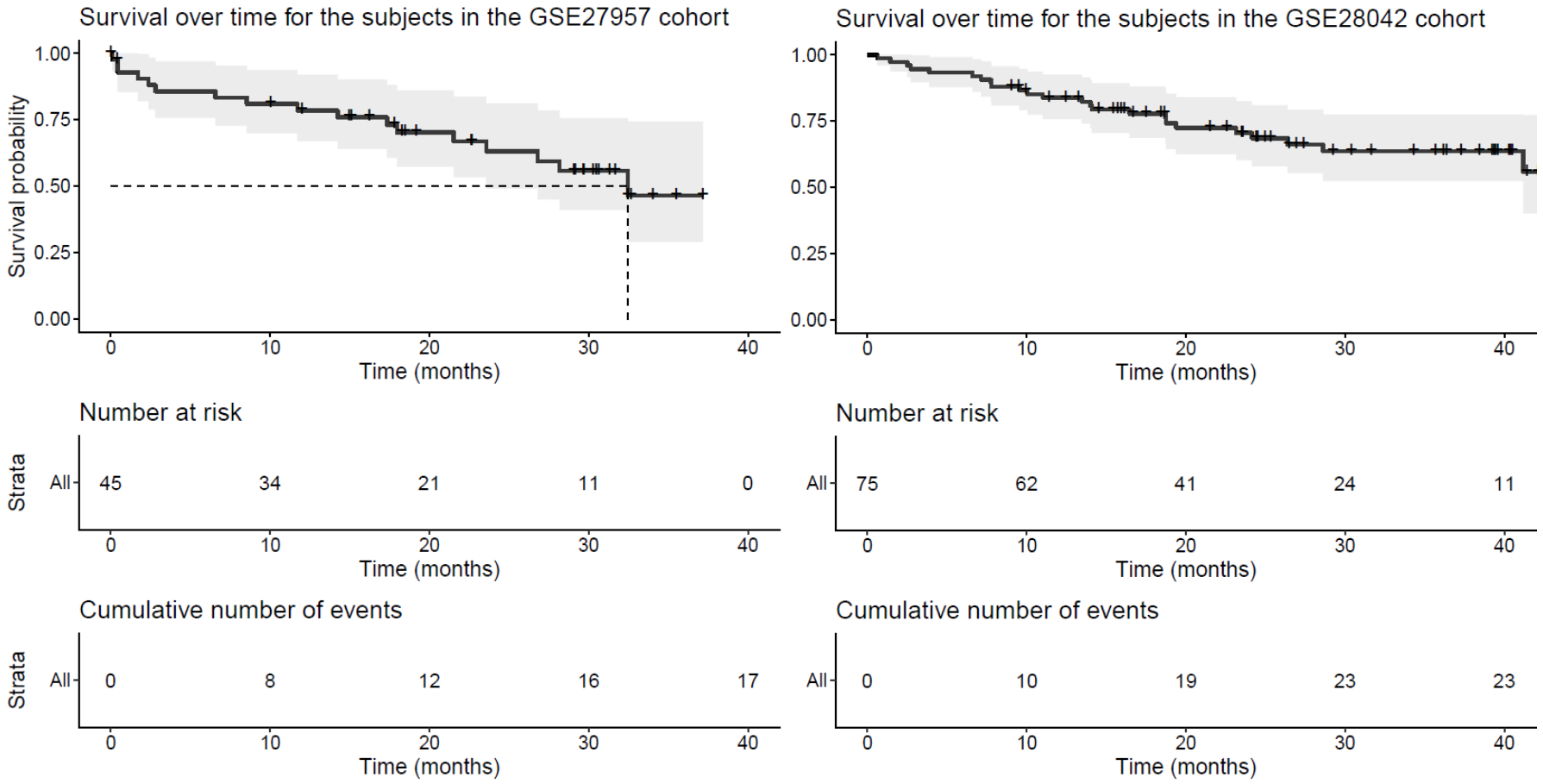
ADDITIONAL FIGURE A.4.1: *The variant weight function used by SKAT in the non-burden test analysis, which follows a Beta distribution with shape parameters 1 and 25.*



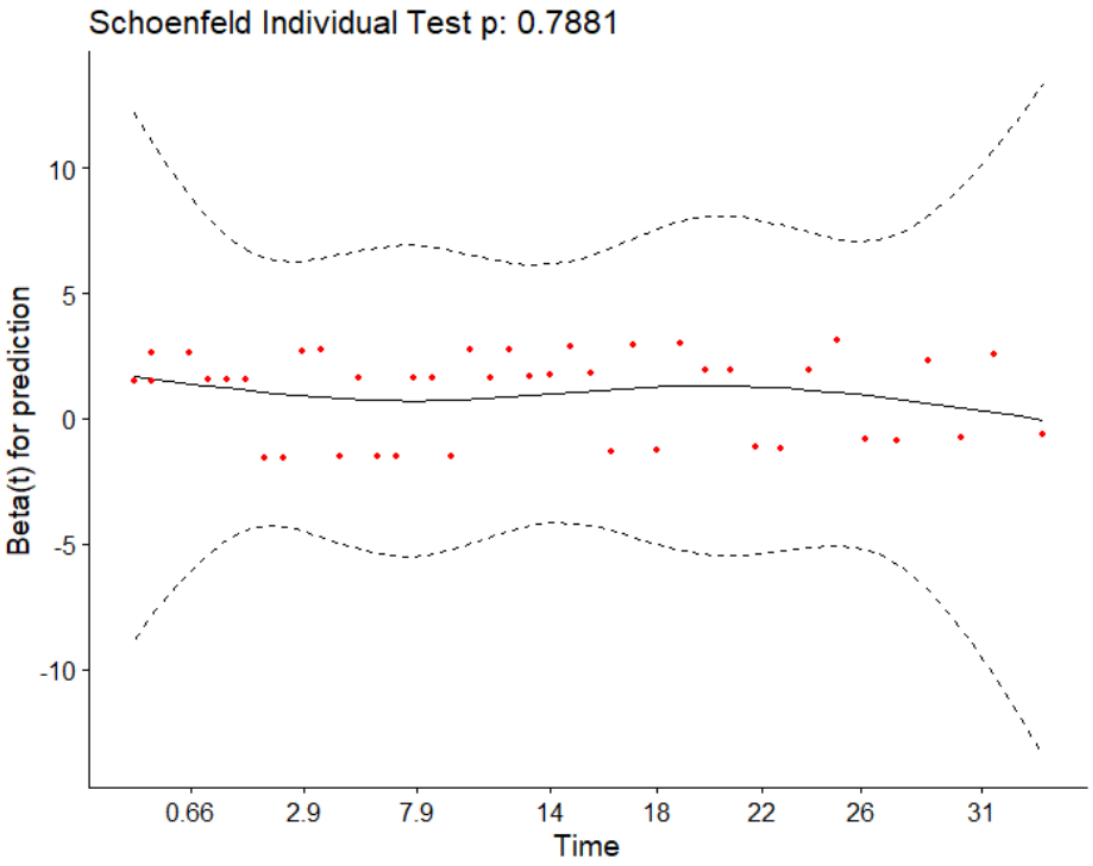
ADDITIONAL FIGURE A.5.1: *A Kaplan-Meier plot showing survival over time for the subjects in study GSE93606.*



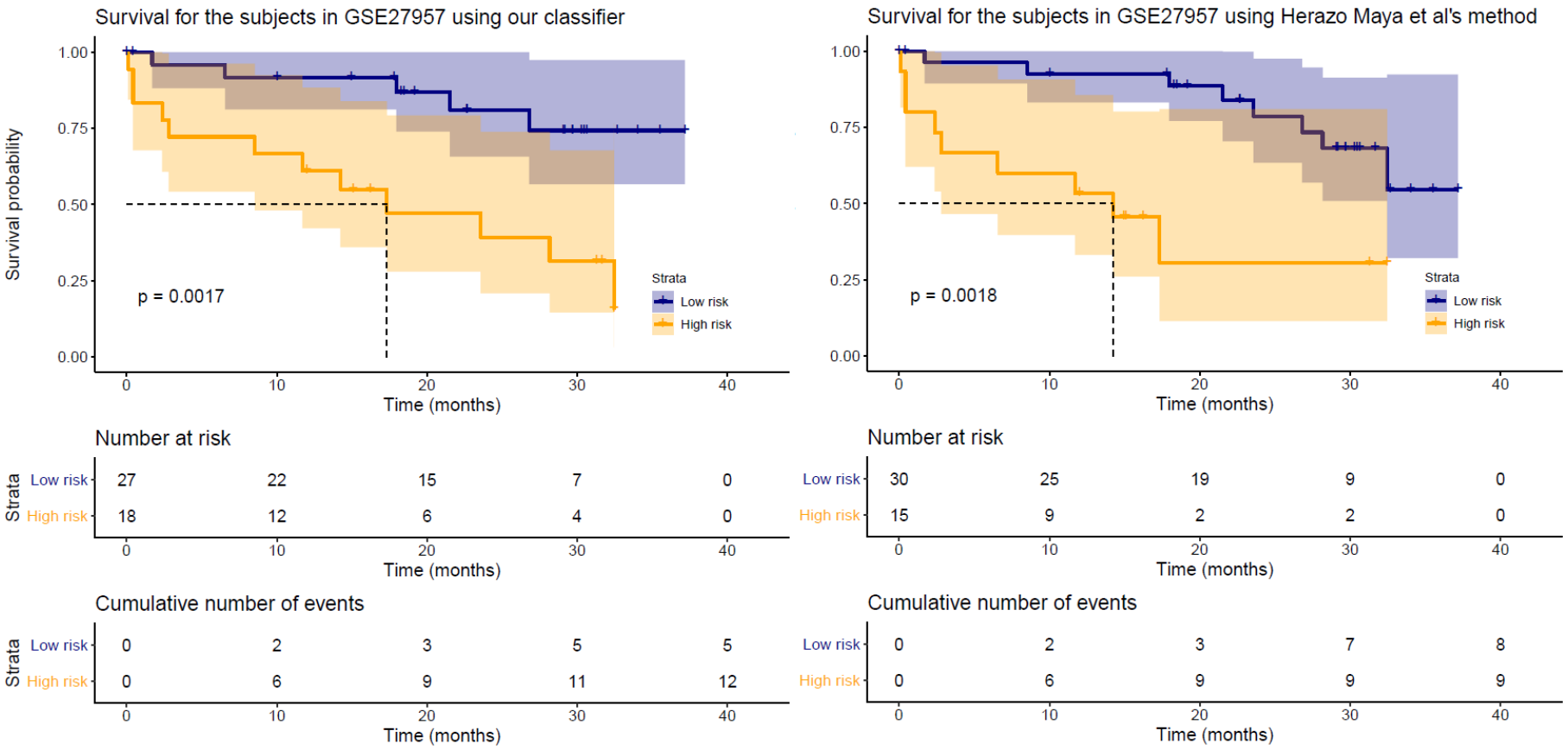
ADDITIONAL FIGURE A.5.2: *A plot of the scaled Schoenfeld residuals over time for the Cox proportional hazards model that was fit to the survival data from the discovery study GSE93606. The plot shows no trend with time and thus the proportional hazards assumption has not been broken.*



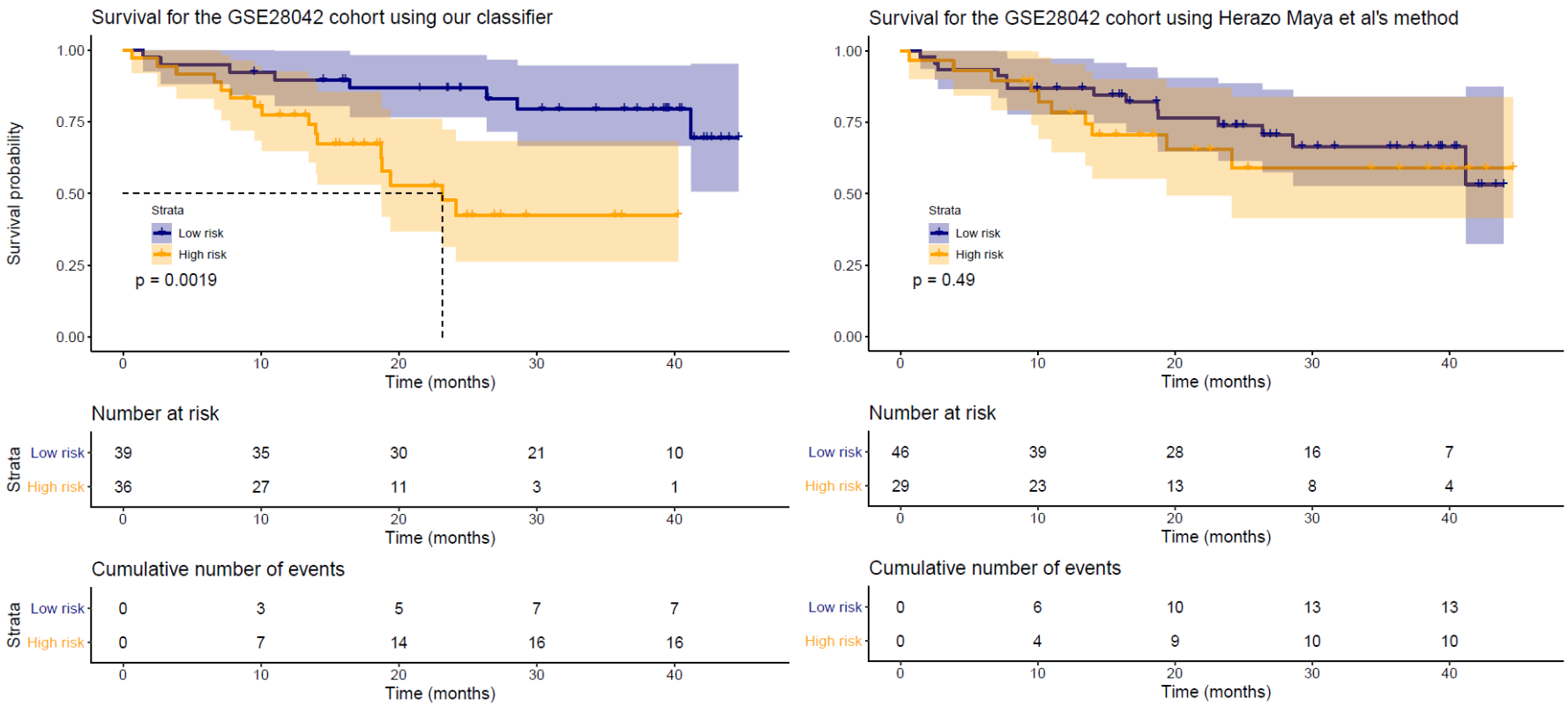
ADDITIONAL FIGURE A.5.3: *Kaplan-Meier plots showing survival over time for the subjects in the validation cohorts GSE27957 (left) and GSE28042 (right). A dotted line on the plot indicates the median survival time, if this could be calculated.* .



ADDITIONAL FIGURE A.5.4: *A plot of the scaled Schoenfeld residuals over time for the Cox proportional hazards model that was fit to the survival data from the validation studies. The plot shows no trend with time and thus there is no evidence that the proportional hazards assumption has been broken.*



ADDITIONAL FIGURE A.5.5: *Kaplan-Meier plots showing survival over time for the subjects in the validation cohort GSE27957 when subjects are assigned using our classifier (left) or Herazo Maya et al.'s SAMs method (right). A dotted line on the plot indicates the median survival time for that group and the p-value on each plot is from a log-rank test, testing the two curves for equality.*



ADDITIONAL FIGURE A.5.6: *Kaplan-Meier plots showing survival over time for the subjects in the validation cohort GSE28042 when subjects are assigned using our classifier (left) or Herazo Maya et al.'s method SAMs (right). A dotted line on the plot indicates the median survival time and the p-value on each plot is from a log-rank test, testing the two curves for equality.*

B: Additional Tables

ADDITIONAL TABLE B.3.1: *Suspected causal variants for previously reported genome-wide signals of association with IPF susceptibility.*

Chr.	Position	rsid	Locus	Reference
3	44902386	rs78238620	*KIF15*	[77]
3	169481271	rs12696304	*LRRC34/TERC*	[77]
4	89885086	rs2013701	*FAM13A*	[77]
5	1282414	rs7725218	*TERT*	[77]
5	169015479	rs116483731	*SPDL1*	[119]
6	7563232	rs2076295	*DSP*	[77]
7	1909479	rs12699415	*MAD1L1*	[77]
7	99630342	rs2897075	*7q22.1*	[77]
8	120934126	rs28513081	*DEPTOR*	[77]
11	1241221	rs35705950	*MUC5B*	[77]
13	113534984	rs9577395	*ATP11A*	[77]
15	40720542	rs59424629	*IVD*	[77]
15	86097216	rs62023891	*AKAP13*	[77]
17	44214888	rs2077551	*MAPT*	[77]
19	4717672	rs12610495	*DPP9*	[77]

ADDITIONAL TABLE B.3.2: *Summary statistics for the sentinel SNP of each signal of interest in the two-stage GWAS analysis (section 3.2). EAF = effect allele frequency.*

rsid	Chr	Position	Conditional upon	Ref allele	Effect allele	EAF	P-value	Beta	SE	R^2
rs78672887	1	11473150		C	T	0.034	1.11×10^{-6}	-0.830	0.170	0.862
rs2999900	1	13830815		T	C	0.827	4.51×10^{-5}	-0.331	0.081	0.990
rs79264639	1	64346109		C	G	0.975	2.64×10^{-5}	-0.836	0.199	0.737
rs76719272	1	156129796		C	T	0.132	3.15×10^{-5}	-0.390	0.094	0.967
rs113212335	1	234601869		G	A	0.025	2.31×10^{-5}	-0.842	0.199	0.992
rs72765831	1	244357437		A	G	0.014	4.18×10^{-5}	-1.072	0.262	0.986
rs76719272	2	28809554		T	C	0.023	2.56×10^{-5}	-0.798	0.190	0.913
rs72791696	2	29690051	rs72791696	G	A	0.013	1.7×10^{-5}	-1.146	0.267	0.857
rs12471179	2	49568428		A	C	0.324	9.89×10^{-6}	-0.286	0.065	0.972
rs74703036	2	113020207		G	T	0.012	3.22×10^{-5}	-1.180	0.284	0.879
rs7599256	2	171495421	rs192643964	G	T	0.544	6.8×10^{-6}	0.277	0.062	0.998
rs192643964	2	171819000		A	G	0.012	2.95×10^{-5}	-1.185	0.284	0.878
rs7562987	2	202815530		G	A	0.757	2.85×10^{-5}	-0.309	0.074	0.993
rs72958256	2	217676681		C	T	0.212	3.22×10^{-6}	-0.349	0.075	0.967
rs71043147	2	239509442		T	C	0.011	2.78×10^{-5}	-1.246	0.297	0.711
rs3849481	3	76448409		G	A	0.625	4.18×10^{-5}	0.257	0.063	0.977
rs798582	3	118510843		T	C	0.629	2.69×10^{-5}	0.263	0.063	0.980
rs11943143	4	24376579		A	G	0.361	2.72×10^{-5}	0.275	0.066	0.992

rs4865060	4	57035690		A	G	0.752	4.25×10^{-5}	-0.282	0.069	0.987
rs72641509	4	64265863		C	A	0.220	3.24×10^{-5}	-0.311	0.075	0.992
rs28536140	4	126144080		T	C	0.141	3.79×10^{-5}	-0.359	0.087	0.937
rs56267054	4	141953175		T	C	0.582	4.25×10^{-5}	0.260	0.064	0.964
rs72699571	4	166114257		C	T	0.037	4.38×10^{-5}	-0.677	0.166	0.995
rs79073584	4	174832314		T	C	0.038	3.30×10^{-5}	-0.659	0.159	0.985
rs112101321	4	177862046		G	T	0.031	3.86×10^{-5}	-0.734	0.178	0.966
rs1035908	5	465510		T	C	0.911	4.26×10^{-5}	0.461	0.113	0.981
rs16901464	5	11301495		A	G	0.043	1.51×10^{-5}	-0.632	0.146	0.992
rs72749864	5	54507085		T	A	0.033	6.49×10^{-6}	-0.731	0.162	0.967
rs12518082	5	103033128		T	G	0.910	1.86×10^{-5}	0.454	0.106	0.950
rs62381849	5	113452092		T	A	0.115	1.60×10^{-5}	-0.404	0.094	0.988
rs149865510	5	124640655		G	T	0.012	1.61×10^{-5}	-1.224	0.284	0.921
rs55742238	5	133022616		G	A	0.313	2.50×10^{-5}	-0.274	0.065	0.994
rs114697189	5	147442328		C	T	0.020	3.26×10^{-5}	0.905	0.218	0.920
rs72814140	5	172438977		C	T	0.012	4.77×10^{-5}	-1.154	0.284	0.931
rs185529470	6	9088333		T	C	0.015	4.27×10^{-5}	-1.033	0.252	0.744
rs189806018	6	21799534		T	C	0.019	3.09×10^{-5}	-0.931	0.224	0.820
rs114791520	6	29487974		T	C	0.011	9.48×10^{-6}	-1.316	0.297	0.906
rs113727117	7	37818692		G	A	0.054	2.30×10^{-5}	0.589	0.139	0.892
rs2392610	7	38952795		C	T	0.908	3.37×10^{-5}	-0.431	0.104	0.979
rs73364290	7	63682097		G	A	0.075	4.35×10^{-5}	-0.485	0.119	0.986
rs76259754	7	65599188		C	T	0.014	5.34×10^{-6}	-1.190	0.262	0.933
rs140795717	7	110907680		A	C	0.034	1.67×10^{-5}	-0.651	0.151	0.933
rs113419491	8	18627701		A	G	0.012	3.52×10^{-5}	-1.174	0.284	0.883
rs75681116	8	20248930		C	T	0.029	2.87×10^{-7}	-0.946	0.184	0.998
rs12155839	8	130184065		C	T	0.173	4.63×10^{-6}	-0.369	0.081	0.987
rs10974627	9	4563255		G	C	0.016	1.08×10^{-5}	-1.074	0.244	0.947
rs148740072	9	5249347	rs10974627	T	C	0.058	3.8×10^{-5}	-0.521	0.126	0.881
rs62541874	9	16840005		A	T	0.068	3.72×10^{-5}	-0.502	0.122	0.866
rs113887396	9	138670699		A	G	0.099	4.28×10^{-5}	-0.416	0.102	0.950
rs1769215	10	802445		C	T	0.300	1.16×10^{-5}	0.290	0.066	0.997
rs11252666	10	4735548		C	T	0.554	4.01×10^{-5}	0.258	0.063	0.959
rs77761367	10	25031285		G	A	0.041	1.93×10^{-5}	-0.673	0.157	0.964
rs182317201	10	133919390		C	T	0.017	5.02×10^{-6}	-1.080	0.237	0.974
rs117946291	11	38076713		T	C	0.014	2.11×10^{-5}	1.112	0.262	0.988
rs7944090	11	44513229		G	C	0.233	2.76×10^{-5}	-0.297	0.071	0.976
rs3018480	11	77208493		T	C	0.423	1.11×10^{-5}	-0.270	0.061	0.987
rs112777739	11	90510739		C	T	0.022	1.41×10^{-5}	0.923	0.213	0.908
rs74715174	12	10065926		A	T	0.039	6.81×10^{-5}	-0.705	0.157	0.950
rs113262525	12	10291177	rs74715174	T	G	0.080	4.6×10^{-6}	-0.477	0.104	0.977
rs181150706	12	29738685		T	C	0.014	2.95×10^{-5}	-1.093	0.262	0.830
rs12431370	13	23525132		C	T	0.058	4.53×10^{-5}	-0.527	0.129	0.965
rs111673280	13	66165105		G	A	0.183	2.37×10^{-5}	-0.328	0.078	0.987

rs61969491	13	106268991		T	C	0.268	1.04×10^{-5}	0.301	0.068	0.972
rs2391813	13	110767948		C	T	0.011	4.92×10^{-5}	-1.207	0.297	0.908
rs10130876	14	35983946		T	C	0.160	3.09×10^{-5}	0.346	0.083	0.979
rs60667478	14	45206340		A	G	0.058	4.72×10^{-5}	0.548	0.135	0.932
rs79552221	14	85195323		G	T	0.035	1.56×10^{-5}	-0.726	0.168	0.964
rs79577629	14	88798784		A	G	0.044	2.50×10^{-5}	-0.625	0.148	0.983
rs41317306	14	96864374		T	G	0.022	3.16×10^{-5}	-0.885	0.213	0.991
rs11633299	15	22768701		T	C	0.522	1.89×10^{-5}	0.258	0.060	0.805
rs143547312	15	77136340		C	G	0.015	4.60×10^{-5}	-1.028	0.252	0.972
rs61459715	15	80777127		G	A	0.016	1.67×10^{-6}	-1.169	0.244	0.883
rs72755373	15	91421010		G	C	0.011	2.20×10^{-5}	1.261	0.297	0.831
rs76137938	15	94190256		C	T	0.020	2.79×10^{-5}	0.913	0.218	0.985
rs55752757	16	3112715		C	G	0.181	4.69×10^{-5}	-0.333	0.082	0.893
rs72784561	16	65795865		A	G	0.338	1.71×10^{-5}	-0.279	0.065	0.983
rs189629869	17	13341945		T	C	0.013	1.90×10^{-5}	-1.163	0.272	0.870
rs9906660	17	64882726		C	T	0.308	4.74×10^{-5}	0.272	0.067	0.995
rs72900412	17	75777150		C	T	0.092	2.55×10^{-5}	-0.449	0.107	0.951
rs139518622	19	4524306		G	A	0.025	1.90×10^{-5}	-0.850	0.199	0.734
rs117388035	19	57260406		G	A	0.026	5.36×10^{-7}	-0.977	0.195	0.981
rs118021492	20	39722647		C	T	0.015	4.74×10^{-5}	-1.027	0.252	0.929
rs78914239	20	40275250	rs118021492	A	T	0.063	3.8×10^{-5}	-0.505	0.123	0.990
rs148565491	20	43602081		G	A	0.012	4.49×10^{-5}	-1.158	0.284	0.895
rs6025056	20	55255474		A	G	0.156	1.71×10^{-5}	0.362	0.084	0.909
rs11910170	21	18194592		G	A	0.586	3.36×10^{-5}	0.253	0.061	0.987
rs1007614	22	22935155	rs144531880	G	A	0.312	2.8×10^{-5}	0.280	0.067	0.971
rs144531880	22	23728611		C	T	0.011	3.99×10^{-5}	1.221	0.297	0.918

ADDITIONAL TABLE B.3.3: *Results from the sensitivity analysis in Section 3.3, in which the five suggestively significant results in the 3-way GWAS meta-analysis were re-meta-analysed without including a genomic control correction. None of the signals reached genome-wide significance ($P<5\times10^{-8}$) in the sensitivity analysis.*

rsid	Chr.	Position	Gene	Ref/effect allele	HR	95% CI	P-value
rs183759512	7	111708942	*DOCK4* (intron)	C / T	2.20	(1.64, 2.96)	1.5×10^{-7}
rs41295127	10	6134617	*RBM17* (intron)	A / T	2.77	(1.86, 4.09)	4.2×10^{-7}
rs3915628	11	71682613	*RNF121* (intron)	C / T	4.00	(2.36, 6.81)	2.7×10^{-7}
rs9513422	13	99083935	*FARP1* (exon)	C / T	1.38	(1.22, 1.56)	3.6×10^{-7}
rs118122250	16	54209057	Intergenic	G / A	2.20	(1.62, 2.98)	3.7×10^{-7}

ADDITIONAL TABLE B.3.4: *95% credible sets for each of the five suggestively significant signals from the GWAS meta-analysis (Section 3.3). In each credible set, variants are presented in descending order of posterior probability. MAF and P-value refer to the minor allele frequency and P-value for the variant in the meta-analysis of Section 3.3.*

rsid	Chr	Position	MAF	P-value	Posterior probability	Cumulative probability	Gene	Functional annotation	eQTL genes
rs183759512	7	111708942	4.1%	1.05E-06	31.4%	31.4%	*DOCK4*	Protein coding, processed transcript, promoter flanking region	
rs11761827	7	112051706	22.0%	6.19E-05	15.5%	46.9%	*AC004112.4*	Antisense, CTCF binding site, promoter flanking region	*IFRD1*
rs116923970	7	111732366	4.3%	7.33E-06	11.9%	58.8%	*DOCK4*	Protein coding, processed transcript, enhancer	
rs6944057	7	112074766	20.4%	1.76E-04	6.3%	65.1%	*IFRD1*	Protein coding, enhancer	*IFRD1*
rs6969907	7	112054974	21.8%	2.97E-04	4.2%	69.3%		CTCF binding site	*IFRD1*
rs58295122	7	112058508	21.8%	3.01E-04	4.1%	73.4%	*IFRD1*	Protein coding	*IFRD1*
rs7798715	7	112062523	21.8%	3.43E-04	3.7%	77.1%	*IFRD1*	Protein coding, promoter flanking region, CTCF binding site	*IFRD1*
rs7780160	7	112062683	21.8%	3.44E-04	3.7%	80.8%	*IFRD1*	Protein coding, promoter flanking region	*IFRD1*
rs13223482	7	112059772	21.7%	3.62E-04	3.5%	84.3%	*IFRD1*	Protein coding	*IFRD1*
rs4730544	7	112059842	21.7%	3.63E-04	3.5%	87.8%	*IFRD1*	Protein coding, promoter flanking region	*IFRD1*
rs12216563	7	112060989	21.7%	3.66E-04	3.5%	91.3%	*IFRD1*	Protein coding, promoter flanking region	*IFRD1*
rs7810707	7	112082651	20.4%	3.83E-04	3.3%	94.6%	*IFRD1, AC079741.2*	Protein coding, processed pseudogene	*IFRD1*
rs720639	7	112067629	21.3%	4.46E-04	2.9%	97.5%	*IFRD1*	Protein coding, enhancer	*IFRD1*
rs41295127	10	6134617	1.9%	2.32E-06	12.7%	12.7%	*RBM17*	Protein coding	
rs75827481	10	6138462	1.9%	2.36E-06	12.6%	25.3%	*RBM17*	Protein coding	
rs117413330	10	6166270	1.9%	3.12E-06	10.7%	36.0%			
rs62626316	10	6110135	2.9%	4.45E-05	7.0%	43.1%	*RP11-414H17.2*	Processed pseudogene	
rs41295051	10	6111210	2.9%	4.79E-05	6.7%	49.8%	*RP11-414H17.2*	Processed pseudogene	
rs62626324	10	6120365	3.0%	7.46E-05	5.2%	55.0%			
rs17421433	10	6124793	2.9%	9.95E-05	4.4%	59.4%		Promoter flanking region	

rs1924137	10	6099887	3.0%	1.72E-04	3.6%	63.0%	*IL2RA*	Protein coding	
rs12722492	10	6100471	3.0%	1.73E-04	3.6%	66.6%	*IL2RA*	Protein coding, promoter flanking region	
rs41294927	10	6102259	3.0%	1.75E-04	3.6%	70.1%	*IL2RA*	Protein coding	
rs41294935	10	6106609	3.0%	1.88E-04	3.4%	73.5%	*IL2RA*	Protein coding, promoter	
rs12722491	10	6101430	2.2%	9.54E-05	3.0%	76.5%	*IL2RA*	Protein coding	
rs80214570	10	5317344	4.8%	5.03E-04	2.9%	79.4%	*AKR1C7P, RP11-445P17.8*	Transcribed unprocessed pseudogene, Processed transcript, lincRNA, promoter flanking region	
rs12722498	10	6095836	2.9%	2.82E-04	2.7%	82.1%	*IL2RA*	Protein coding, CTCF binding site	
rs12722494	10	6098177	2.9%	2.84E-04	2.7%	84.8%	*IL2RA*	Protein coding	
rs142101282	10	6105544	3.0%	3.19E-04	2.6%	87.3%	*IL2RA*	Protein coding	
rs77744503	10	6706598	1.9%	7.31E-05	2.4%	89.7%			*PRKCQ-AS1, RP11-554I8.1, RP11-5N23.3*
rs76673065	10	5557435	1.7%	1.02E-04	2.3%	92.0%	*CALML3-AS1*	antisense, enhancer	
rs11256971	10	6187096	3.6%	8.63E-04	1.9%	93.9%	*RN7SKP78*	misc_RNA, Protein coding, promoter	
rs12722503	10	6091643	3.0%	6.05E-04	1.8%	95.7%	*IL2RA*	Protein coding	
rs3915628	11	71682613	1.2%	1.68E-06	100.0%	100.00%	*RNF121*	Protein coding, nonsense mediated decay, retained intron	
rs9513422	13	99083935	28.5%	2.08E-06	100.0%	100.00%	*FARP1*	Protein coding, antisense, CTCF binding site	*FARP1*
rs118122250	16	54209057	3.3%	2.30E-06	17.8%	17.8%		Intergenic variant	*IRX3*
rs78387400	16	54206097	3.3%	2.76E-06	16.0%	33.8%		Regulatory region variant, TF binding site variant	*IRX3*
rs1013170	16	54210447	15.4%	9.12E-05	10.1%	44.0%		Intergenic variant	
rs72811760	16	54210144	5.6%	2.97E-05	9.8%	53.8%		Intergenic variant	*IRX3*
rs115769282	16	54175752	4.2%	4.01E-05	5.1%	58.9%		Intergenic variant	
rs116940325	16	54209223	3.1%	2.10E-05	4.6%	63.5%		Intergenic variant	*IRX3*
rs59504288	16	54176478	4.0%	7.86E-05	3.1%	66.6%		Intergenic variant	

rs117820196	16	54198778	3.0%	4.26E-05	3.0%	69.6%	Intergenic variant	*IRX3*
rs60589529	16	54177237	4.0%	8.07E-05	3.0%	72.6%	Intergenic variant	
rs182248506	16	54177423	4.0%	8.69E-05	2.9%	75.5%	Intergenic variant	
rs4567704	16	54177413	4.0%	8.69E-05	2.9%	78.4%	Intergenic variant	
rs78644184	16	54178671	4.0%	8.95E-05	2.9%	81.3%	Intergenic variant	
rs115079931	16	54178564	4.0%	9.23E-05	2.8%	84.1%	Intergenic variant	
rs112409776	16	54196864	3.0%	5.02E-05	2.8%	86.8%	Intergenic variant	*IRX3*
rs113905985	16	54195689	3.0%	5.48E-05	2.6%	89.5%	Regulatory region variant, TF binding site variant	*IRX3*
rs76289419	16	54195273	3.0%	5.53E-05	2.6%	92.1%	Regulatory region variant	*IRX3*
rs111930423	16	54194787	3.0%	5.73E-05	2.6%	94.7%	Intergenic variant	*IRX3*
rs77489535	16	54194567	3.0%	5.88E-05	2.5%	97.2%	Intergenic variant	*IRX3*

ADDITIONAL TABLE B.5.1: *The distribution of collection sizes of the 125 collections of human gene expression data on GEO that contained the term 'IPF'.*

Number of samples	1-10	11-20	21-30	31-50	51-100	>100
Number of IPF studies on GEO	32	36	21	10	11	15

ADDITIONAL TABLE B.5.2: *Summary information for the collections on the Gene Expression omnibus with transcriptomic data for human IPF subjects and at least 30 samples.*

GEO accession	Total IPF cases	Total samples	Date public on GEO	Included controls	Total controls	PubMed ID	Technology	Platforms	Tissues	Phenotype
GSE147066	32	58	16/03/2020	Yes	26		RNA-seq	Illumina HiSeq 4000	Lung	IPF
GSE15197	8	39	24/01/2020	Yes	13	20081107	Array	Whole Human Genome Microarray 4x44K G4112F	Lung	IPF
GSE98764	6	30	10/12/2019	No			RNA-seq	Illumina HiSeq 2000	Lung	IPF
GSE121849	10	33	26/10/2019	Yes			Array	Illumina HumanMethylation450 BeadChip	Lung squamous cell carcinoma	IPF
GSE124685	10	59	11/10/2019	Yes	6	31600171	RNA-seq, microRNA	Ion Torrent Proton	Lung	Severe IPF
GSE133298	54	54	30/09/2019	No			RNA-Seq	Illumina HiSeq 4000	Blood	IPF
GSE132607	74	276	30/09/2019	No			Array	Affymetrix Human Gene Expression Array	Blood	IPF
GSE135099	5	30	31/07/2019	Yes		31603936	Array	Illumina HumanMethylation450 BeadChip, Affymetrix Human Gene Expression Array	Lung fibroblasts	IPF
GSE134692	36	80	24/07/2019	Yes	17	31423451	RNA-seq	Illumina HiSeq 2500	Lung	Severe IPF/ALI
GSE98925	13	61	30/06/2019	No			Array	Affymetrix Human Genome U133 Plus 2.0 Array	Lung fibroblasts	IPF
GSE70866	176	196	31/12/2018	Yes		30141961	Array	Agilent-028004 SurePrint G3 Human GE 8x60K Microarray	Bronchoalveolar lavage (BAL)	IPF
GSE110147	22	48	06/02/2018	Yes	11	30111332	Array	Affymetrix Human Gene 1.0 ST Array	Lung	IPF
GSE93606	57	174	14/01/2017	Yes	20	28085486	Array	Affymetrix Human Gene 1.1 ST Array	Blood and BAL	IPF
GSE92592	20	39	01/01/2017	Yes	19	28230051	RNA-seq	Illumina HiSeq 2000	Lung	IPF
GSE86618	325	540	08/12/2016	Yes		28157391	RNA-seq	Illumina NextSeq 500	Epithelial cells	IPF
GSE85268	68	68	15/08/2016	No		28157391	Array	Affymetrix Human Gene Expression Array	Blood	IPF
GSE53845	40	48	14/10/2014	Yes	8	25217476	Array	Agilent-014850 Whole Human Genome Microarray 4x44K G4112F	Lung	IPF
GSE49072	15	84	30/06/2014	Yes		23924348	Array	Affymetrix Human Genome U133A Array	Alveolar macrophages	IPF
GSE38958	70	115	30/06/2014	Yes	45	26286721	Array	Affymetrix Human Exon 1.0 ST Array	Blood	IPF
GSE45686	5	40	04/03/2014	Yes		24590289	Array	Illumina HumanHT-12 V4.0 expression beadchip	Lung fibroblasts	IPF
GSE27957	45	45	15/10/2013	No		24089408	Array	Affymetrix Human Exon 1.0 ST Array	Blood	IPF

GSE28042	75	94	15/10/2013	Yes	19	24089408	Array	Agilent-014850 Whole Human Genome Microarray	Blood	IPF
GSE32537	119	217	21/06/2013	Yes	50	23783374	Array	Affymetrix Human Gene 1.0 ST Array	Lung	IPF/UIP
GSE48149	13	53	21/06/2013	No		21360508	Array	Illumina HumanRef-8 v3.0 expression beadchip	Lung	IPF
GSE33566	93	123	01/07/2012	Yes	30	22761659	Array	Agilent-014850 Whole Human Genome Microarray 4x44K	Blood	IPF
GSE17978	12	58	19/05/2010	Yes		20451601	Array	Duke Human Operon 36k v4.0 spotted microarray	Lung fibroblasts	End-stage IPF
GSE10667	31	46	20/02/2009	Yes	15	19363140	Array	Agilent-014850 Whole Human Genome Microarray 4x44K G4112F	Lung	IPF

ADDITIONAL TABLE B.5.3: *Comparison of phenotypic traits across clusters in each validation study. Data are presented as count (percentage), mean (standard deviation) or median (interquartile range). NA = data not available, FVC=Forced vital capacity, D_{LCO} = Diffusing capacity for carbon monoxide, FEV1 = Forced expiratory volume in one second. Significant p-values (P < 0.05) are highlighted in bold.*

	GSE132607 (n=74)				**GSE27957 (n=45)**				**GSE28042 (n=75)**			
Trait	**Cluster 1**	**Cluster 2**	**Cluster 3**	**P-value**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**P-value**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**P-value**
Total in cluster	19	35	20		14	26	5		25	39	11	
Age (mean, sd)	67.6 (8.1)	66.8 (65.4)	65.4 (6.6)	0.660	68.3 (7.6)	67.6 (7.5)	61.4 (12.2)	0.248	69.6 (8.6)	68.8 (7.5)	68.5 (10.2)	0.903
Male (%)	13 (68.4%)	24 (68.6%)	15 (75.0%)	0.864	14 (100%)	21 (80.8%)	5 (100%)	0.128	17 (68.0%)	25 (64.1%)	10 (90.9%)	0.231
European Ancestry (%)	19 (100%)	33 (94.3%)	18 (90%)	0.383	13 (92.9%)	20 (76.9%)	4 (80%)	0.449	1 (4.0%)	1 (2.6%)	0 (0%)	0.789
FVC % predicted (median, IQR)	74 (27.2)	70.0 (30.9)	66.2 (22.5)	0.735	50.5 (20.8)	65.0 (19.3)	52.0 (8.0)	**0.024**	65.1 (18.6)	62.0 (25.7)	62.7 (13.8)	0.636
D_{LCO} % predicted (median, IQR)	41.6 (19.3)	45.5 (22.7)	40.2 (21.1)	0.157	38 .0 (18.8)	45.1 (23.3)	56.0 (28.0)	0.195	52.7 (31.5)	47.0 (20.1)	44.1 (20.4)	0.537
Death observed during study (%)	NA	NA	NA	-	9 (64.3%)	4 (15.4%)	4 (80%)	**0.001**	8 (32.0%)	10 (25.6%)	6 (54.5%)	0.193
Ever smoker (%)	12 (63.2%)	19 (54.3%)	18 (90%)	**0.025**	NA	NA	NA	-	NA	NA	NA	-
FEV_1 (median, IQR)	NA	NA	NA	-	NA	NA	NA	-	73.5 (21.7)	74.0 (23.8)	81.8 (12.1)	0.804

ADDITIONAL TABLE B.5.4: *A list of aliases for each the 23 genes in the full classifier.*

Gene name	Aliases
KCNK15	TASK5, K2p15.1, TASK-5, KCNK11, KCNK14
SORBS1	CAP, KIAA1296, SH3P12, SH3D5, FLJ12406, KIAA0894, Sh3p12, FLAF2, R85FL, SORB1
HBB	ECYT6
EIF4G1	P220, EIF-4G1, EIF4GI, PARK18, EIF4F, EIF4G, EIF-4G 1, EIF4G1
RPF1	BXDC5
NOP58	HSPC120, NOL5
PSMA5	ZETA, PSC5
RASGRP1	IMD64
IFI30	GILT, IP30, IFI-30, IP-30, MGC32056
HLA-DRA	HLA-DRA1
ATM	TELO1, TEL1, ATDC, AT1, ATA, ATC, ATD, ATE
ECHDC2	FLJ10948
EXOSC8	RRP43, OIP2, P9, OIP-2, CIP3, EAP2, EXOSC8, PCH1C
BLVRA	BLVR, BVR, BVRA
PSMD11	P44.5, S9, MGC3844
SLC38A1	ATA1, NAT2, SAT1, SNAT1
MRPL41	MRPL27, RPML27, BMRP, MRP-L27, MRP-L41, PIG3
PPIA	CYPA, HEL-S-69p, CYPH
AES	GRG5, TLE5, GRG, AES-1, AES-2, ESP1
CA4	CA-IV, CAIV, RP17
BCL2A1	BCL2L5, BFL1, GRS, ACC-1, ACC-2, HBPA1, ACC1, ACC2
UGCG	GCS, GLCT-1, GLCT1
FPR2	LXA4R, HM63, FMLPX, FPR2A, FPRH1, FPRH2, FPRL1, ALXR, RFP, FPR2

C: R code

classifiergenes function

```
classifiergenes <- function(cluster1genes3,clustersubjects){
clustersubjectsindicator <- rep(0,ncol(cluster1genes3))
clustersubjectsindicator[clustersubjects] <- 1
Nsubjects <- ncol(cluster1genes3)
auc <- rep(0,20)
X <- rep(0,20)
upgenes <- as.matrix(cluster1genes3[1,1:Nsubjects])
upgenes[1,1:Nsubjects] <- 1
Nupgenes <- 1
rownames(upgenes) <- "referencegene"
downgenes <- as.matrix(cluster1genes3[1,1:Nsubjects])
downgenes[1,1:Nsubjects] <- 1
Ndowngenes <- 1
rownames(downgenes) <- "referencegene"
for (i in 1:nrow(cluster1genes3)){
  ifelse(mean(as.numeric(cluster1genes3[i,clustersubjects]))>mean(as.numeric(cluster1genes3[i,])),
        cluster1genes3$up[i]<-TRUE,cluster1genes3$up[i]<-FALSE)
}

for (k in 1:20){

for (i in 1:nrow(cluster1genes3)){
  cluster1genes3$auc[i] <- 0
  ifelse(cluster1genes3$up[i]==TRUE,
#if the gene has higher expression in the cluster than the average across all clusters, then:
        roc <-
roc(clustersubjectsindicator,(as.numeric(colProds(as.matrix(upgenes)))*as.numeric(cluster1genes3[i,1:Nsubject
s]))^(1/Nupgenes)-(colProds(as.matrix(downgenes))^(1/Ndowngenes))),
# or if the gene has lower expression in the cluster than the average across all clusters, then:
        roc <- roc(clustersubjectsindicator,(as.numeric(colProds(as.matrix(upgenes)))^(1/Nupgenes))-
((colProds(as.matrix(downgenes))*as.numeric(cluster1genes3[i,1:Nsubjects]))^(1/Ndowngenes))))
  cluster1genes3$auc[i] <- auc(roc)
}

if(max(auc) < max(cluster1genes3$auc)){
    X[k] <- which(cluster1genes3$auc==max(cluster1genes3$auc))
    auc[k] <- max(cluster1genes3$auc)
    ifelse(cluster1genes3$up[X[k]]==TRUE,upgenes<-
rbind(upgenes,cluster1genes3[X[k],1:Nsubjects]),downgenes<-
rbind(downgenes,cluster1genes3[X[k],1:Nsubjects]))
    Nupgenes <- max(nrow(upgenes)-1,1)
    Ndowngenes <- max(nrow(downgenes)-1,1)
    print(c(rownames(cluster1genes3)[X[k]],k,auc[k]))
}
  else {break}
}
auc <<- auc
upgenes <<- upgenes
downgenes <<- downgenes
}

#example input
```

```
classifiergenes(cluster1genes,1:64)      #4 upgenes, 1 downgene
classifiergenes(cluster2genes,65:159) #14 upgenes, 0 downgenes
classifiergenes(cluster3genes,160:196) #4 upgenes, 0 downgenes
```

D: Publications

i. Allen, R.J., Guillen-Guio, B., Oldham, J.M., Ma, S.F., Dressen, A., Paynton, M.L., **Kraven, L.M.**, Obeidat, M.E., Li, X. *et al.*, 2020. Genome-wide association study of susceptibility to idiopathic pulmonary fibrosis. American journal of respiratory and critical care medicine, 201(5), pp.564-574.

ii. Fadista, J., **Kraven, L.M.**, Karjalainen, J., Andrews, S.J., Geller, F., Baillie, J.K., Wain, L.V., Jenkins, R.G., Feenstra, B. and the COVID-19 Host Genetics Initiative, 2021. Shared genetic etiology between idiopathic pulmonary fibrosis and COVID-19 severity. EBioMedicine, 65, p.103277.

iii. **Kraven, L.M.**, Taylor, A.R. Molyneaux, P.L., Maher, T.M., McDonough, J.E., Mura, M., Yang, I.V., *et al.*, 2022. Cluster analysis of transcriptomic datasets to identify endotypes of Idiopathic Pulmonary Fibrosis. Thorax, 10.1136/thoraxjnl-2021-218563.

iv. Allen, R.J., Guillen-Guio, B., Croot, E., **Kraven, L.M.**, Moss, S., Stewart, I., Jenkins, R.G. and Wain, L.V., 2022. Genetic overlap between idiopathic pulmonary fibrosis and COVID− 19. *European Respiratory Journal*, 10.1183/13993003.03132-2021.

v. (Pre-print) Fainberg, H.P., Oldham J.M., Molyneaux, P.L., Allen, R.J., Kraven, L.M., Fahy, W.A., Porte, J., Braybrooke, R., *et al.*, 2022. Analysis of Forced Vital Capacity (FVC) trajectories in Idiopathic Pulmonary Fibrosis (IPF) identifies four distinct clusters of disease behaviour. SSRN, 4000658.

vi. (Pre-print) Mohammadi-Nejad, A., Allen, R.J., **Kraven, L.M.**, Leavy, O.C., Jenkins R.G., Wain, L.V., Auer, D.P. and Sotiropoulos S.N., 2022. Mapping brain endotypes associated with idiopathic pulmonary fibrosis genetic risk. medRxiv, 2022.03.25.22272932.

vii. Leavy, O.C., Allen, R.J., **Kraven, L.M**., Morgan, A., Tobin, M.D., Quint, J.K., Jenkins R.G. and Wain, L.V., 2022. Using genetic information to define idiopathic pulmonary fibrosis in UK Biobank. Chest, 2022.07.027.

E: References

1. Raghu G, Remy-Jardin M, Myers JL, et al. Diagnosis of idiopathic pulmonary fibrosis. An official ATS/ERS/JRS/ALAT clinical practice guideline. *American journal of respiratory and critical care medicine* 2018; **198**(5): e44-68.

2. Gribbin J, Hubbard RB, Le Jeune I, Smith CJ, West J, Tata LJ. Incidence and mortality of idiopathic pulmonary fibrosis and sarcoidosis in the UK. *Thorax* 2006; **61**(11): 980-5.

3. Bjoraker JA, Ryu JH, Edwin MK, Myers JL, Tazelaar HD, Schroeder DR, Offord KP. Prognostic significance of histopathologic subsets in idiopathic pulmonary fibrosis. *American journal of respiratory and critical care medicine* 1998; **157**(1): 199-203.

4. Nicholson AG, Colby TV, Dubois RM, Hansell DM, Wells AU. The prognostic significance of the histologic pattern of interstitial pneumonia in patients presenting with the clinical entity of cryptogenic fibrosing alveolitis. *American journal of respiratory and critical care medicine* 2000; **162**(6): 2213-7.

5. Rudd RM, Prescott RJ, Chalmers JC, Johnston ID. British Thoracic Society Study on cryptogenic fibrosing alveolitis: response to treatment and survival. *Thorax* 2007; **62**(1): 62-6.

6. Maher TM. Pirfenidone in idiopathic pulmonary fibrosis. *Drugs of today (Barcelona, Spain: 1998)* 2010; **46**(7): 473-82.

7. Varone F, Sgalla G, Iovene B, Bruni T, Richeldi L. Nintedanib for the treatment of idiopathic pulmonary fibrosis. *Expert Opin Pharmacother* 2018; **19**(2): 167-75.

8. Strongman H, Kausar I, Maher TM. Incidence, prevalence, and survival of patients with idiopathic pulmonary fibrosis in the UK. *Adv Ther* 2018; **35**(5): 724-36.

9. Kaunisto J, Salomaa E, Hodgson U, et al. Demographics and survival of patients with idiopathic pulmonary fibrosis in the FinnishIPF registry. *ERJ open research* 2019; **5**(3).

10. Meltzer EB, Noble PW. Idiopathic pulmonary fibrosis. *Orphanet journal of rare diseases* 2008; **3**(1): 8.

11. Maher TM, Bendstrup E, Dron L, et al. Global incidence and prevalence of idiopathic pulmonary fibrosis. *Respiratory research* 2021; **22**(1): 1-10.

12. British Lung Foundation. Idiopathic pulmonary fibrosis statistics. 2019; Available at: https://www.blf.org.uk/support-for-you/idiopathic-pulmonary-fibrosis-ipf/statistics. Accessed 08/08/, 2019.

13. Ley B, Collard HR, King Jr TE. Clinical course and prediction of survival in idiopathic pulmonary fibrosis. *American journal of respiratory and critical care medicine* 2011; **183**(4): 431-40.

14. Taskar VS, Coultas DB. Is idiopathic pulmonary fibrosis an environmental disease? *Proceedings of the American Thoracic Society* 2006; **3**(4): 293-8.

15. Olson AL, Swigris JJ. Idiopathic pulmonary fibrosis: diagnosis and epidemiology. *Clin Chest Med* 2012; **33**(1): 41-50.

16. Assayag D, Morisset J, Johannson KA, Wells AU, Walsh SL. Patient gender bias on the diagnosis of idiopathic pulmonary fibrosis. *Thorax* 2020; **75**(5): 407-12.

17. Moore BB, Moore TA. Viruses in idiopathic pulmonary fibrosis. Etiology and exacerbation. *Annals of the American Thoracic Society* 2015; **12**(Supplement 2): S186-92.

18. Raghu G, Collard HR, Egan JJ, et al. An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. *American journal of respiratory and critical care medicine* 2011; **183**(6): 788-824.

19. Sheng G, Chen P, Wei Y, et al. Viral infection increases the risk of idiopathic pulmonary fibrosis: a meta-analysis. *Chest* 2020; **157**(5): 1175-87.

20. Macneal K, Schwartz DA. The genetic and environmental causes of pulmonary fibrosis. *Proceedings of the American Thoracic Society* 2012; **9**(3): 120-5.

21. Goodwin AT, Jenkins G. Molecular endotyping of pulmonary fibrosis. *Chest* 2016; **149**(1): 228-37.

22. Spagnolo P, Kropski JA, Jones MG, et al. Idiopathic pulmonary fibrosis: disease mechanisms and drug development. *Pharmacol Ther* 2021; **222**: 107798.

23. Kellogg DL, Musi N, Nambiar AM. Cellular senescence in idiopathic pulmonary fibrosis. *Current Molecular Biology Reports* 2021; **7**(3): 31-40.

24. Lehmann M, Korfei M, Mutze K, et al. Senolytic drugs target alveolar epithelial cell function and attenuate experimental lung fibrosis ex vivo. *European Respiratory Journal* 2017; **50**(2).

25. Schafer MJ, White TA, Iijima K, et al. Cellular senescence mediates fibrotic pulmonary disease. *Nature communications* 2017; **8**(1): 1-11.

26. Kropski JA, Lawson WE, Blackwell TS. Personalizing therapy in idiopathic pulmonary fibrosis: a glimpse of the future? *American Journal of Respiratory and Critical Care Medicine* 2015; .

27. Jenkins G. Endotyping idiopathic pulmonary fibrosis should improve outcomes for all patients with progressive fibrotic lung disease. *Thorax* 2015; .

28. Rosas IO, Richards TJ, Konishi K, et al. MMP1 and MMP7 as potential peripheral blood biomarkers in idiopathic pulmonary fibrosis. *PLoS Med* 2008; **5**(4): e93.

29. Song JW, Do KH, Jang SJ, Colby TV, Han S, Kim DS. Blood biomarkers MMP-7 and SP-A: predictors of outcome in idiopathic pulmonary fibrosis. *Chest* 2013; **143**(5): 1422-9.

30. Maher TM, Oballa E, Simpson JK, et al. An epithelial biomarker signature for idiopathic pulmonary fibrosis: an analysis from the multicentre PROFILE cohort study. *The Lancet Respiratory Medicine* 2017; **5**(12): 946-55.

31. Organ LA, Duggan AR, Oballa E, et al. Biomarkers of collagen synthesis predict progression in the PROFILE idiopathic pulmonary fibrosis cohort. *Respiratory research* 2019; **20**(1): 1-10.

32. Collins FS, Mansoura MK. The human genome project: revealing the shared inheritance of all humankind. *Cancer: Interdisciplinary International Journal of the American Cancer Society* 2001; **91**(S1): 221-5.

33. Van de Peer Y. Genomes: the truth is in there. *EMBO Rep* 2011; **12**(2): 93.

34. Jackson M, Marks L, May GH, Wilson JB. The genetic basis of disease. *Essays Biochem* 2018; **62**(5): 643-723.

35. Ellegren H. Microsatellites: simple sequences with complex evolution. *Nature reviews genetics* 2004; **5**(6): 435-45.

36. Fallin MD, Duggal P, Beaty TH. Genetic epidemiology and public health: the evolution from theory to technology. *Am J Epidemiol* 2016; **183**(5): 387-93.

37. Klein RJ, Zeiss C, Chew EY, et al. Complement factor H polymorphism in age-related macular degeneration. *Science* 2005; **308**(5720): 385-9.

38. Buniello A, MacArthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* 2019; **47**(D1): D1005-12.

39. Evangelou E, Warren HR, Mosen-Ansorena D, et al. Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. *Nat Genet* 2018; **50**(10): 1412-25.

40. Lee JJ, Wedow R, Okbay A, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat Genet* 2018; **50**(8): 1112-21.

41. Uffelmann E, Huang QQ, Munung NS, et al. Genome-wide association studies. *Nature Reviews Methods Primers* 2021; **1**(1): 1-21.

42. Nelson MR, Tipney H, Painter JL, et al. The support of human genetic evidence for approved drug indications. *Nat Genet* 2015; **47**(8): 856.

43. King EA, Davis JW, Degner JF. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLoS genetics* 2019; **15**(12): e1008489.

44. Okada Y, Wu D, Trynka G, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* 2014; **506**(7488): 376-81.

45. Drew L. Pharmacogenetics: the right drug for you. *Nature* 2016; **537**(7619): S60-2.

46. Lambert SA, Abraham G, Inouye M. Towards clinical utility of polygenic risk scores. *Hum Mol Genet* 2019; **28**(R2): R133-42.

47. Loos RJ. 15 years of genome-wide association studies and no signs of slowing down. *Nature Communications* 2020; **11**(1): 1-3.

48. Omenn GS, Nass SJ, Micheel CM. Evolution of translational omics: lessons learned and the path forward. 2012; .

49. Marnellos G. High-throughput SNP analysis for genetic association studies. *Curr Opin Drug Discov Devel* 2003; **6**(3): 317-21.

50. Consortium GP, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM. A global reference for human genetic variation. *Nature* 2015; **526**(7571): 68-74.

51. Palmer L, Smith GD, Burton PR. An introduction to genetic epidemiology. Policy Press; 2011.

52. Hellwege JN, Keaton JM, Giri A, Gao X, Velez Edwards DR, Edwards TL. Population stratification in genetic association studies. *Current protocols in human genetics* 2017; **95**(1): 1.22. 1,1.22. 23.

53. Bycroft C, Freeman C, Petkova D, et al. Genome-wide genetic data on~ 500,000 UK Biobank participants. *BioRxiv* 2017; : 166298.

54. Lamy P, Andersen CL, Wikman FP, Wiuf C. Genotyping and annotation of Affymetrix SNP arrays. *Nucleic Acids Res* 2006; **34**(14): e100.

55. Auton A, Brooks LD. A global reference for human genetic variation. *Nature* 2015; **526**(7571): 68-74.

56. UK10K consortium. The UK10K project identifies rare variants in health and disease. *Nature* 2015; **526**(7571): 82.

57. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016; **48**(10): 1279.

58. Schurz H, Müller SJ, Van Helden PD, Tromp G, Hoal EG, Kinnear CJ, Möller M. Evaluating the accuracy of imputation methods in a five-way admixed population. *Frontiers in genetics* 2019; **10**: 34.

59. Steele MP, Speer MC, Loyd JE, et al. Clinical and pathologic features of familial interstitial pneumonia. *American journal of respiratory and critical care medicine* 2005; **172**(9): 1146-52.

60. Bitterman PB, Rennard SI, Keogh BA, Wewers MD, Adelberg S, Crystal RG. Familial idiopathic pulmonary fibrosis. *N Engl J Med* 1986; **314**(21): 1343-7.

61. Raghu G, Hert R. Interstitial lung diseases: genetic predisposition and inherited interstitial lung diseases. Seminars in respiratory medicine; Copyright© 1993 by Thieme Medical Publishers, Inc.; 1993.

62. Hodgson U, Laitinen T, Tukiainen P. Nationwide prevalence of sporadic and familial idiopathic pulmonary fibrosis: evidence of founder effect among multiplex families in Finland. *Thorax* 2002; **57**(4): 338-42.

63. García-Sancho C, Buendía-Roldán I, Fernández-Plata MR, et al. Familial pulmonary fibrosis is the strongest risk factor for idiopathic pulmonary fibrosis. *Respir Med* 2011; **105**(12): 1902-7.

64. Fernandez BA, Fox G, Bhatia R, et al. A Newfoundland cohort of familial and sporadic idiopathic pulmonary fibrosis patients: clinical and genetic features. *Respiratory research* 2012; **13**(1): 64.

65. Tsakiri KD, Cronkhite JT, Kuan PJ, et al. Adult-onset pulmonary fibrosis caused by mutations in telomerase. *Proceedings of the National Academy of Sciences* 2007; **104**(18): 7552-7.

66. Armanios MY, Chen JJ, Cogan JD, et al. Telomerase mutations in families with idiopathic pulmonary fibrosis. *N Engl J Med* 2007; **356**(13): 1317-26.

67. Mushiroda T, Wattanapokayakit S, Takahashi A, et al. A genome-wide association study identifies an association of a common variant in TERT with susceptibility to idiopathic pulmonary fibrosis. *J Med Genet* 2008; **45**(10): 654-6.

68. Bilgili H, Białas AJ, Górski P, Piotrowski WJ. Telomere abnormalities in the pathobiology of idiopathic pulmonary fibrosis. *Journal of Clinical Medicine* 2019; **8**(8): 1232.

69. Zhang Y, Noth I, Garcia JG, Kaminski N. A variant in the promoter of MUC5B and idiopathic pulmonary fibrosis. *N Engl J Med* 2011; **364**(16): 1576-7.

70. Seibold MA, Wise AL, Speer MC, et al. A common MUC5B promoter polymorphism and pulmonary fibrosis. *N Engl J Med* 2011; **364**(16): 1503-12.

71. Fingerlin TE, Murphy E, Zhang W, et al. Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. *Nat Genet* 2013; **45**(6): 613.

72. Noth I, Zhang Y, Ma S, et al. Genetic variants associated with idiopathic pulmonary fibrosis susceptibility and mortality: a genome-wide association study. *The lancet Respiratory medicine* 2013; **1**(4): 309-17.

73. Allen RJ, Porte J, Braybrooke R, et al. Genetic variants associated with susceptibility to idiopathic pulmonary fibrosis in people of European ancestry: a genome-wide association study. *The Lancet Respiratory Medicine* 2017; **5**(11): 869-80.

74. Zhu Q, Zhang X, Zhang S, et al. Association between the MUC5B Promoter polymorphism rs35705950 and idiopathic pulmonary fibrosis: a meta-analysis and trial sequential analysis in Caucasian and Asian populations. *Medicine* 2015; **94**(43).

75. Peljto AL, Zhang Y, Fingerlin TE, et al. Association between the MUC5B promoter polymorphism and survival in patients with idiopathic pulmonary fibrosis. *JAMA* 2013; **309**(21): 2232-9.

76. Dudbridge F, Allen RJ, Sheehan NA, et al. Adjustment for index event bias in genome-wide association studies of subsequent events. *Nature communications* 2019; **10**(1): 1561.

77. Allen RJ, Guillen-Guio B, Oldham JM, et al. Genome-wide association study of susceptibility to idiopathic pulmonary fibrosis. *American journal of respiratory and critical care medicine* 2020; **201**(5): 564-74.

78. Lorenzo-Salazar JM, Ma S, Jou J, et al. Novel idiopathic pulmonary fibrosis susceptibility variants revealed by deep sequencing. *ERJ open research* 2019; **5**(2): 71.

79. Moore C, Blumhagen RZ, Yang IV, et al. Resequencing Study Confirms Host Defense and Cell Senescence Gene Variants Contribute to the Risk of Idiopathic Pulmonary Fibrosis. *American journal of respiratory and critical care medicine* 2019; (ja).

80. Dhindsa RS, Mattsson J, Nag A, et al. Identification of a missense variant in SPDL1 associated with idiopathic pulmonary fibrosis. *Communications biology* 2021; **4**(1): 1-8.

81. Fingerlin TE, Zhang W, Yang IV, et al. Genome-wide imputation study identifies novel HLA locus for pulmonary fibrosis and potential role for auto-immunity in fibrotic idiopathic interstitial pneumonia. *BMC genetics* 2016; **17**(1): 74.

82. Peljto AL, Selman M, Kim DS, et al. The MUC5B promoter polymorphism is associated with idiopathic pulmonary fibrosis in a Mexican cohort but is rare among Asian ancestries. *Chest* 2015; **147**(2): 460-4.

83. Bustin SA, Benes V, Nolan T, Pfaffl MW. Quantitative real-time RT-PCR–a perspective. *J Mol Endocrinol* 2005; **34**(3): 597-601.

84. Nickless A, Bailis JM, You Z. Control of gene expression through the nonsense-mediated RNA decay pathway. *Cell & bioscience* 2017; **7**(1): 1-12.

85. Mueller WF, Larsen LS, Garibaldi A, Hatfield GW, Hertel KJ. The silent sway of splicing by synonymous substitutions. *J Biol Chem* 2015; **290**(46): 27700-11.

86. Tseng C, Wong M, Liao W, Chen C, Lee S, Yen J, Chang S. Genetic variants in transcription factor binding sites in humans: triggered by natural selection and triggers of diseases. *International Journal of Molecular Sciences* 2021; **22**(8): 4187.

87. Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. *PLoS computational biology* 2017; **13**(5): e1005457.

88. Evans C, Hardin J, Stoebel DM. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Briefings in bioinformatics* 2018; **19**(5): 776-92.

89. Bolstad BM, Irizarry RA, Åstrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 2003; **19**(2): 185-93.

90. Han J, Chen M, Wang Y, Gong B, Zhuang T, Liang L, Qiao H. Identification of biomarkers based on differentially expressed genes in papillary thyroid carcinoma. *Scientific reports* 2018; **8**(1): 1-11.

91. Jiang K, Liu H, Xie D, Xiao Q. Differentially expressed genes ASPN, COL1A1, FN1, VCAN and MUC5AC are potential prognostic biomarkers for gastric cancer. *Oncology letters* 2019; **17**(3): 3191-202.

92. Oyelade J, Isewon I, Oladipupo F, et al. Clustering algorithms: their application to gene expression data. *Bioinformatics and Biology insights* 2016; **10**: BBI. S38316.

93. Van Dam S, Vosa U, van der Graaf A, Franke L, de Magalhaes JP. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in bioinformatics* 2018; **19**(4): 575-92.

94. Sweeney TE, Azad TD, Donato M, et al. Unsupervised analysis of transcriptomics in bacterial sepsis across multiple datasets reveals three robust clusters. *Crit Care Med* 2018; **46**(6): 915.

95. Neumark N, Cosme Jr C, Rose K, Kaminski N. The idiopathic pulmonary fibrosis cell atlas. *American Journal of Physiology-Lung Cellular and Molecular Physiology* 2020; **319**(6): L887-92.

96. Lewis CM, Knight J. Introduction to genetic association studies. *Cold Spring Harb Protoc 3: 297–306* 2012; .

97. Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, Derks EM. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International journal of methods in psychiatric research* 2018; : e1608.

98. Turner S, Armstrong LL, Bradford Y, et al. Quality control procedures for genome-wide association studies. *Current protocols in human genetics* 2011; **68**(1): 1.19. 1,1.19. 18.

99. Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, Derks EM. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International journal of methods in psychiatric research* 2018; **27**(2): e1608.

100. Chen B, Cole JW, Grond-Ginsbach C. Departure from Hardy Weinberg Equilibrium and Genotyping Error. *Frontiers in genetics* 2017; **8**: 167.

101. Bush WS, Moore JH. Genome-wide association studies. *PLoS computational biology* 2012; **8**(12): e1002822.

102. Novembre J, Johnson T, Bryc K, et al. Genes mirror geography within Europe. *Nature* 2008; **456**(7218): 98-101.

103. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006; **38**(8): 904.

104. Johnson RC, Nelson GW, Troyer JL, Lautenberger JA, Kessing BD, Winkler CA, O'Brien SJ. Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC Genomics* 2010; **11**(1): 724.

105. Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 1999; **55**(4): 997-1004.

106. Spain SL, Barrett JC. Strategies for fine-mapping complex traits. *Hum Mol Genet* 2015; **24**(R1): R111-9.

107. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association* 1958; **53**(282): 457-81.

108. Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* 1972; **34**(2): 187-202.

109. Schoenfeld D. Partial residuals for the proportional hazards regression model. *Biometrika* 1982; **69**(1): 239-41.

110. Krauss E, Gehrken G, Drakopanagiotakis F, et al. Clinical characteristics of patients with familial idiopathic pulmonary fibrosis (f-IPF). *BMC pulmonary medicine* 2019; **19**(1): 1-13.

111. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Optimal designs for two-stage genome-wide association studies. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* 2007; **31**(7): 776-88.

112. Maher TM. PROFILEing idiopathic pulmonary fibrosis: rethinking biomarker discovery. *European Respiratory Review* 2013; **22**(128): 148-52.

113. Navaratnam V, Fogarty AW, McKeever T, et al. Presence of a prothrombotic state in people with idiopathic pulmonary fibrosis: a population-based case–control study. *Thorax* 2014; **69**(3): 207-15.

114. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *Plos med* 2015; **12**(3): e1001779.

115. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W. Robust relationship inference in genome-wide association studies. *Bioinformatics* 2010; **26**(22): 2867-73.

116. Allen RJ, Porte J, Braybrooke R, et al. Genetic variants associated with susceptibility to idiopathic pulmonary fibrosis in people of European ancestry: a genome-wide association study. *The Lancet Respiratory Medicine* 2017; **5**(11): 869-80.

117. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018; **562**(7726): 203-9.

118. Pruim RJ, Welch RP, Sanna S, et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 2010; **26**(18): 2336-7.

119. Dhindsa RS, Mattsson J, Nag A, et al. Identification of a novel missense variant in SPDL1 associated with idiopathic pulmonary fibrosis. *bioRxiv* 2020; .

120. Leavy OC, Ma S, Molyneaux PL, et al. Proportion of Idiopathic pulmonary fibrosis risk explained by known common genetic loci in european populations. *American journal of respiratory and critical care medicine* 2021; **203**(6): 775-8.

121. Li X, Wong W, Lamoureux EL, Wong TY. Are linear regression techniques appropriate for analysis when the dependent (outcome) variable is not normally distributed? *Invest Ophthalmol Vis Sci* 2012; **53**(6): 3082-3.

122. Speed D, Balding DJ. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nat Genet* 2019; **51**(2): 277-84.

123. Hughey JJ, Rhoades SD, Fu DY, Bastarache L, Denny JC, Chen Q. Cox regression increases power to detect genotype-phenotype associations in genomic studies using the electronic health record. *BMC Genomics* 2019; **20**(1): 1-7.

124. Maller JB, McVean G, Byrnes J, et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet* 2012; **44**(12): 1294-301.

125. Wakefield J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *The American Journal of Human Genetics* 2007; **81**(2): 208-27.

126. Cingolani P, Platts A, Wang LL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* 2012; **6**(2): 80-92.

127. Koscielny G, An P, Carvalho-Silva D, et al. Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res* 2017; **45**(D1): D985-94.

128. Nica AC, Dermitzakis ET. Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences* 2013; **368**(1620): 20120362.

129. Võsa U, Claringbould A, Westra H, et al. Large-scale cis-and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat Genet* 2021; : 1-11.

130. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, Plagnol V. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS genetics* 2014; **10**(5): e1004383.

131. Lappi-Blanco E, Lehtonen ST, Sormunen R, Merikallio HM, Soini Y, Kaarteenaho RL. Divergence of tight and adherens junction factors in alveolar epithelium in pulmonary fibrosis. *Hum Pathol* 2013; **44**(5): 895-907.

132. Zou J, Li Y, Yu J, Dong L, Husain AN, Shen L, Weber CR. Idiopathic pulmonary fibrosis is associated with tight junction protein alterations. *Biochimica et Biophysica Acta (BBA)-Biomembranes* 2020; **1862**(5): 183205.

133. Han Y, Zhang M, Chen D, Li H, Wang X, Ma S. Downregulation of RNA binding motif protein 17 expression inhibits proliferation of hypopharyngeal carcinoma faDu cells. *Oncology Letters* 2018; **15**(4): 5680-4.

134. Cai C, Tang Y, Zhai J, Zheng C. The RING finger protein family in health and disease. *Signal Transduction and Targeted Therapy* 2022; **7**(1): 1-23.

135. Galati D, Zanotta S, Polistina GE, Coppola A, Capitelli L, Bocchino M. Circulating dendritic cells are severely decreased in idiopathic pulmonary fibrosis with a potential value for prognosis prediction. *Clinical Immunology* 2020; **215**: 108454.

136. Bellefroid EJ, Kobbe A, Gruss P, Pieler T, Gurdon JB, Papalopulu N. Xiro3 encodes a Xenopus homolog of the Drosophila Iroquois genes and functions in neural specification. *EMBO J* 1998; **17**(1): 191-203.

137. Schupp JC, Kaminski N. Toward early detection of idiopathic pulmonary fibrosis. *Toward early detection of idiopathic pulmonary fibrosis* 2019; .

138. Putman RK, Rosas IO, Hunninghake GM. Genetics and early detection in idiopathic pulmonary fibrosis. *American journal of respiratory and critical care medicine* 2014; **189**(7): 770-8.

139. Assayag D, Morisset J, Johannson KA, Wells AU, Walsh SL. Patient gender bias on the diagnosis of idiopathic pulmonary fibrosis. *Thorax* 2020; **75**(5): 407-12.

140. Oh CK, Murray LA, Molfino NA. Smoking and idiopathic pulmonary fibrosis. *Pulmonary medicine* 2012; **2012**.

141. Laudicella M, Siciliani L, Cookson R. Waiting times and socioeconomic status: evidence from England. *Soc Sci Med* 2012; **74**(9): 1331-41.

142. Sesé L, Cavalin C, Bernaudin J, Maesano IA, Nunes H. Patient Registries in Idiopathic Pulmonary Fibrosis: Don't Forget Socioeconomic Status. *American journal of respiratory and critical care medicine* 2020; **201**(8): 1014-5.

143. Dudbridge F, Allen RJ, Sheehan NA, et al. Adjustment for index event bias in genome-wide association studies of subsequent events. *Nature communications* 2019; **10**(1): 1-10.

144. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet* 2014; **46**(2): 100-6.

145. Thickett D, Voorham J, Ryan R, et al. Historical database cohort study addressing the clinical patterns prior to idiopathic pulmonary fibrosis (IPF) diagnosis in UK primary care. *BMJ open* 2020; **10**(5): e034428.

146. Maher B. Personal genomes: The case of the missing heritability. *Nature News* 2008; **456**(7218): 18-21.

147. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics* 2010; **11**(6): 415-25.

148. Gibson G. Rare and common variants: twenty arguments. *Nature Reviews Genetics* 2012; **13**(2): 135-45.

149. Gorlov IP, Gorlova OY, Frazier ML, Spitz MR, Amos CI. Evolutionary evidence of the effect of rare variants on disease etiology. *Clin Genet* 2011; **79**(3): 199-206.

150. Povysil G, Petrovski S, Hostyk J, Aggarwal V, Allen AS, Goldstein DB. Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nature Reviews Genetics* 2019; **20**(12): 747-59.

151. Pang H, Xia Y, Luo S, Huang G, Li X, Xie Z, Zhou Z. Emerging roles of rare and low-frequency genetic variants in type 1 diabetes mellitus. *J Med Genet* 2021; **58**(5): 289-96.

152. Povysil G, Chazara O, Carss KJ, et al. Assessing the role of rare genetic variation in patients with heart failure. *JAMA cardiology* 2021; **6**(4): 379-86.

153. Höglund J, Rafati N, Rask-Andersen M, Enroth S, Karlsson T, Ek WE, Johansson Å. Improved power and precision with whole genome sequencing data in genome-wide association studies of inflammatory biomarkers. *Scientific reports* 2019; **9**(1): 1-14.

154. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics* 2014; **95**(1): 5-23.

155. Dering C, Ziegler A, König IR, Hemmelmann C. Comparison of collapsing methods for the statistical analysis of rare variants. BMC proceedings; Springer; 2011.

156. Moore C, Blumhagen RZ, Yang IV, et al. Resequencing study confirms that host defense and cell senescence gene variants contribute to the risk of idiopathic pulmonary fibrosis. *American journal of respiratory and critical care medicine* 2019; **200**(2): 199-208.

157. Petrovski S, Todd JL, Durheim MT, et al. An exome sequencing study to assess the role of rare genetic variation in pulmonary fibrosis. *American journal of respiratory and critical care medicine* 2017; **196**(1): 82-93.

158. Pedersen BS, Quinlan AR. Who's who? Detecting and resolving sample anomalies in human DNA sequencing studies with peddy. *The American Journal of Human Genetics* 2017; **100**(3): 406-13.

159. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* 2015; **526**(7571): 68.

160. Koch L. Exploring human genomic diversity with gnomAD. *Nature Reviews Genetics* 2020; **21**(8): 448.

161. Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol* 2010; **34**(2): 188-93.

162. Zhan X, Hu Y, Li B, Abecasis GR, Liu DJ. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics* 2016; **32**(9): 1423-6.

163. Dering C, König IR, Ramsey LB, Relling MV, Yang W, Ziegler A. A comprehensive evaluation of collapsing methods using simulated and real data: excellent annotation of functionality and large sample sizes required. *Frontiers in genetics* 2014; **5**: 323.

164. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* 2011; **89**(1): 82-93.

165. Hasegawa T, Kojima K, Kawai Y, Misawa K, Mimori T, Nagasaki M. AP-SKAT: highly-efficient genome-wide rare variant association test. *BMC Genomics* 2016; **17**(1): 1-8.

166. Guiot J, Henket M, Bondue B, Corhay J, Louis R. IGFBP-2 a new interesting target in IPF. *IGFBP-2 a new interesting target in IPF* 2016; .

167. He J, Li X. Identification and validation of aging-related genes in idiopathic pulmonary fibrosis. *Frontiers in genetics* 2022; **13**.

168. Guiot J, Bondue B, Henket M, Corhay JL, Louis R. Raised serum levels of IGFBP-1 and IGFBP-2 in idiopathic pulmonary fibrosis. *BMC pulmonary medicine* 2016; **16**(1): 1-7.

169. Chiahsuan C, Lee J, Ravichandran R, et al. IGFBP2 protects against pulmonary fibrosis through inhibiting P21-mediated senescence. *bioRxiv* 2021; .

170. Sica GL, Zhu G, Tamada K, Liu D, Ni J, Chen L. RELT, a new member of the tumor necrosis factor receptor superfamily, is selectively expressed in hematopoietic tissues and activates transcription factor NF-κB. *Blood, The Journal of the American Society of Hematology* 2001; **97**(9): 2702-7.

171. Plataki M, Koutsopoulos AV, Darivianaki K, Delides G, Siafakas NM, Bouros D. Expression of apoptotic and antiapoptotic markers in epithelial cells in idiopathic pulmonary fibrosis. *Chest* 2005; **127**(1): 266-74.

172. Wang Q, Xie Z, Wu Q, Jin Z, Yang C, Feng J. Role of various imbalances centered on alveolar epithelial cell/fibroblast apoptosis imbalance in the pathogenesis of idiopathic pulmonary fibrosis. *Chin Med J* 2021; **134**(3): 261.

173. Du Bois RM. Strategies for treating idiopathic pulmonary fibrosis. *Nature reviews Drug discovery* 2010; **9**(2): 129-40.

174. Lee S, Emond MJ, Bamshad MJ, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *The American Journal of Human Genetics* 2012; **91**(2): 224-37.

175. Pirinen M, Donnelly P, Spencer CC. Including known covariates can reduce power to detect genetic effects in case-control studies. *Nat Genet* 2012; **44**(8): 848-51.

176. Van't Veer LJ, Dai H, Van De Vijver, Marc J, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002; **415**(6871): 530-6.

177. Buyse M, Loi S, Van't Veer L, et al. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J Natl Cancer Inst* 2006; **98**(17): 1183-92.

178. Drukker CA, Bueno-de-Mesquita JM, Retèl VP, et al. A prospective evaluation of a breast cancer prognosis signature in the observational RASTER study. *International journal of cancer* 2013; **133**(4): 929-36.

179. Cardoso F, van't Veer LJ, Bogaerts J, et al. 70-gene signature as an aid to treatment decisions in early-stage breast cancer. *N Engl J Med* 2016; **375**(8): 717-29.

180. Slodkowska EA, Ross JS. MammaPrint™ 70-gene signature: another milestone in personalized medical care for breast cancer patients. *Expert review of molecular diagnostics* 2009; **9**(5): 417-22.

181. Der S, Zhu C, Brower S, Uihlein A. Predicting prognosis of early-stage non-small cell lung cancer using the GeneFx® lung signature. *PLoS Currents* 2015; **7**.

182. Kopetz S, Tabernero J, Rosenberg R, et al. Genomic classifier ColoPrint predicts recurrence in stage II colorectal cancer patients more accurately than clinical factors. *Oncologist* 2015; **20**(2): 127-33.

183. Boon K, Bailey NW, Yang J, et al. Molecular phenotypes distinguish patients with relatively stable from progressive idiopathic pulmonary fibrosis (IPF). *PloS one* 2009; **4**(4): e5134.

184. Wang Y, Yella J, Chen J, McCormack FX, Madala SK, Jegga AG. Unsupervised gene expression analyses identify IPF-severity correlated signatures, associated genes and biomarkers. *BMC pulmonary medicine* 2017; **17**(1): 1-10.

185. Herazo-Maya JD, Noth I, Duncan SR, et al. Peripheral blood mononuclear cell gene expression profiles predict poor outcome in idiopathic pulmonary fibrosis. *Science translational medicine* 2013; **5**(205): 205ra136.

186. Herazo-Maya JD, Sun J, Molyneaux PL, et al. Validation of a 52-gene risk profile for outcome prediction in patients with idiopathic pulmonary fibrosis: an international, multicentre, cohort study. *The Lancet Respiratory Medicine* 2017; **5**(11): 857-68.

187. Ley B, Collard HR, King Jr TE. Clinical course and prediction of survival in idiopathic pulmonary fibrosis. *American journal of respiratory and critical care medicine* 2011; **183**(4): 431-40.

188. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007; **8**(1): 118-27.

189. Fare TL, Coffey EM, Dai H, et al. Effects of atmospheric ozone on microarray data quality. *Anal Chem* 2003; **75**(17): 4672-5.

190. Sweeney TE, Wong HR, Khatri P. Robust classification of bacterial and viral infections via integrated host gene expression diagnostics. *Science translational medicine* 2016; **8**(346): 346ra91.

191. Sweeney TE, Chen AC, Gevaert O. Combined Mapping of Multiple clUsteriNg ALgorithms (COMMUNAL): a robust method for selection of cluster number, K. *Scientific reports* 2015; **5**(1): 1-10.

192. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002; **30**(1): 207-10.

193. Ley B, Ryerson CJ, Vittinghoff E, et al. A multidimensional index and staging system for idiopathic pulmonary fibrosis. *Ann Intern Med* 2012; **156**(10): 684-91.

194. Huang LS, Berdyshev EV, Tran JT, et al. Sphingosine-1-phosphate lyase is an endogenous suppressor of pulmonary fibrosis: role of S1P signalling and autophagy. *Thorax* 2015; **70**(12): 1138-48.

195. Yang IV, Luna LG, Cotter J, et al. The peripheral blood transcriptome identifies the presence and extent of disease in idiopathic pulmonary fibrosis. *PloS one* 2012; **7**(6): e37708.

196. Meltzer EB, Barry WT, Yang IV, et al. Familial and sporadic idiopathic pulmonary fibrosis: making the diagnosis from peripheral blood. *BMC Genomics* 2014; **15**(1): 1-11.

197. Molyneaux PL, Willis-Owen SA, Cox MJ, et al. Host–microbial interactions in idiopathic pulmonary fibrosis. *American journal of respiratory and critical care medicine* 2017; **195**(12): 1640-50.

198. Garge NR, Page GP, Sprague AP, Gorman BS, Allison DB. Reproducible clusters from microarray research: whither? BMC bioinformatics; BioMed Central; 2005.

199. Huang Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery* 1998; **2**(3): 283-304.

200. Kashef R, Kamel MS. Cooperative clustering. *Pattern Recognit* 2010; **43**(6): 2315-29.

201. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2001; **63**(2): 411-23.

202. Mar JC, Wells CA, Quackenbush J. Defining an informativeness metric for clustering gene expression data. *Bioinformatics* 2011; **27**(8): 1094-100.

203. Brito MR, Chávez EL, Quiroz AJ, Yukich JE. Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection. *Statistics & Probability Letters* 1997; **35**(1): 33-42.

204. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 1987; **20**: 53-65.

205. Gordon AD. Classification. CRC Press; 1999.

206. Halkidi M, Batistakis Y, Vazirgiannis M. On clustering validation techniques. *J Intell Inform Syst* 2001; **17**(2): 107-45.

207. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association* 1952; **47**(260): 583-621.

208. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* 2005; **102**(43): 15545-50.

209. Allen RJ, Guillen-Guio B, Oldham JM, et al. Genome-wide association study of susceptibility to idiopathic pulmonary fibrosis. *American journal of respiratory and critical care medicine* 2020; **201**(5): 564-74.

210. Coward WR, Saini G, Jenkins G. The pathogenesis of idiopathic pulmonary fibrosis. *Therapeutic advances in respiratory disease* 2010; **4**(6): 367-88.

211. Biernacka A, Dobaczewski M, Frangogiannis NG. TGF-β signaling in fibrosis. *Growth Factors* 2011; **29**(5): 196-202.

212. Meng X, Nikolic-Paterson DJ, Lan HY. TGF-β: the master regulator of fibrosis. *Nature Reviews Nephrology* 2016; **12**(6): 325.

213. Györfi AH, Matei A, Distler JH. Targeting TGF-β signaling for the treatment of fibrosis. *Matrix biology* 2018; **68**: 8-27.

214. Senavirathna LK, Huang C, Yang X, et al. Hypoxia induces pulmonary fibroblast proliferation through NFAT signaling. *Scientific Reports* 2018; **8**(1): 1-16.

215. P Heukels, C.C Moor, J.H. von der Thüsen, M.S. Wijsenbeek, M. Kool. Inflammation and immunity in IPF pathogenesis and treatment. *Respiratory Medicine* 2019; **197**: 79-91.

216. Shenderov K, Collins SL, Powell JD, Horton MR. Immune dysregulation as a driver of idiopathic pulmonary fibrosis. *J Clin Invest* 2021; **131**(2).

217. Safran M, Dalah I, Alexander J, et al. GeneCards Version 3: the human gene integrator. *Database* 2010; **2010**.

218. Niu R, Liu Y, Zhang Y, et al. iTRAQ-based proteomics reveals novel biomarkers for idiopathic pulmonary fibrosis. *PloS one* 2017; **12**(1): e0170741.

219. Duecker R, Baer P, Eickmeier O, et al. Oxidative stress-driven pulmonary inflammation and fibrosis in a mouse model of human ataxia-telangiectasia. *Redox Biology* 2018; **14**: 645-55.

220. Zhang D, Povysil G, Kobeissy PH, et al. Rare and Common Variants in KIF15 Contribute to Genetic Risk of Idiopathic Pulmonary Fibrosis. *American Journal of Respiratory and Critical Care Medicine* 2022; (ja).

221. Adegunsoye A, Hrusch CL, Bonham CA, et al. Skewed lung CCR4 to CCR6 CD4+ T cell ratio in idiopathic pulmonary fibrosis is associated with pulmonary function. *Frontiers in immunology* 2016; **7**: 516.

222. Desai O, Winkler J, Minasyan M, Herzog EL. The role of immune and inflammatory cells in idiopathic pulmonary fibrosis. *Frontiers in medicine* 2018; **5**: 43.

223. Zhao YD, Yin L, Archer S, et al. Metabolic heterogeneity of idiopathic pulmonary fibrosis: a metabolomic study. *BMJ open respiratory research* 2017; **4**(1).

224. Bargagli E, Refini RM, d'Alessandro M, et al. Metabolic Dysregulation in Idiopathic Pulmonary Fibrosis. *International Journal of Molecular Sciences* 2020; **21**(16): 5663.

225. Hosseinzadeh A, Javad-Moosavi SA, Reiter RJ, Hemati K, Ghaznavi H, Mehrzadi S. Idiopathic pulmonary fibrosis (IPF) signaling pathways and protective roles of melatonin. *Life Sci* 2018; **201**: 17-29.

226. Froidure A, Marchal-Duval E, Homps-Legrand M, Ghanem M, Justet A, Crestani B, Mailleux A. Chaotic activation of developmental signalling pathways drives idiopathic pulmonary fibrosis. *European Respiratory Review* 2020; **29**(158).

227. Korthagen NM, Van Moorsel CH, Barlo NP, Kazemier KM, Ruven HJ, Grutters JC. Association between variations in cell cycle genes and idiopathic pulmonary fibrosis. *PLoS One* 2012; **7**(1): e30442.

228. Lv X, Liu C, Liu S, et al. The cell cycle inhibitor P21 promotes the development of pulmonary fibrosis by suppressing lung alveolar regeneration. *Acta Pharmaceutica Sinica B* 2022; **12**(2): 735-46.

229. Halloran JW, Zhu D, Qian DC, Byun J, Gorlova OY, Amos CI, Gorlov IP. Prediction of the gene expression in normal lung tissue by the gene expression in blood. *BMC medical genomics* 2015; **8**(1): 1-6.

230. Cecchini MJ, Hosein K, Howlett CJ, Joseph M, Mura M. Comprehensive gene expression profiling identifies distinct and overlapping transcriptional profiles in non-specific interstitial pneumonia and idiopathic pulmonary fibrosis. *Respiratory research* 2018; **19**(1): 1-12.

231. Mura M, Anraku M, Yun Z, et al. Gene expression profiling in the lungs of patients with pulmonary hypertension associated with pulmonary fibrosis. *Chest* 2012; **141**(3): 661-73.

232. Yang IV, Coldren CD, Leach SM, et al. Expression of cilium-associated genes defines novel molecular subtypes of idiopathic pulmonary fibrosis. *Thorax* 2013; **68**(12): 1114-21.

233. McDonough JE, Ahangari F, Li Q, et al. Transcriptional regulatory model of fibrosis progression in the human lung. *JCI insight* 2019; **4**(22).

234. Sivakumar P, Thompson JR, Ammar R, et al. RNA sequencing of transplant-stage idiopathic pulmonary fibrosis lung reveals unique pathway regulation. *ERJ open research* 2019; **5**(3).

235. DePianto DJ, Chandriani S, Abbas AR, et al. Heterogeneous gene expression signatures correspond to distinct lung pathologies and biomarkers of disease severity in idiopathic pulmonary fibrosis. *Thorax* 2015; **70**(1): 48-56.

236. Zank DC, Bueno M, Mora AL, Rojas M. Idiopathic pulmonary fibrosis: aging, mitochondrial dysfunction, and cellular bioenergetics. *Frontiers in medicine* 2018; **5**: 10.

237. Papiris SA, Tomos IP, Karakatsani A, et al. High levels of IL-6 and IL-8 characterize early-on idiopathic pulmonary fibrosis acute exacerbations. *Cytokine* 2018; **102**: 168-72.

238. Idiopathic Pulmonary Fibrosis Clinical Research Network. Prednisone, azathioprine, and N-acetylcysteine for pulmonary fibrosis. *N Engl J Med* 2012; **366**(21): 1968-77.

239. Stalteri MA, Harrison AP. Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips. *BMC Bioinformatics* 2007; **8**(1): 1-15.

240. Chaussabel D. Assessment of immune status using blood transcriptomics and potential implications for global health. Seminars in immunology; Elsevier; 2015.

241. Er Chen M, Tong KB, Malin JL. Cost-effectiveness of 70-gene MammaPrint signature in node-negative breast cancer. 2011; .

242. McCarthy MI. Painting a new picture of personalised medicine for diabetes. *Diabetologia* 2017; **60**(5): 793-9.

243. Peterson RE, Kuchenbaecker K, Walters RK, et al. Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations. *Cell* 2019; **179**(3): 589-603.

244. Jalbert A, Siafa L, Ramanakumar AV, Assayag D. Gender and racial equity in clinical research for idiopathic pulmonary fibrosis: a systematic review and meta-analysis. *European Respiratory Journal* 2022; .

245. Partanen JJ, Happola P, Zhou W, et al. Leveraging global multi-ancestry meta-analysis in the study of Idiopathic Pulmonary Fibrosis genetics. *medRxiv* 2021; .

246. Yorke J, Jones PW, Swigris JJ. Development and validity testing of an IPF-specific version of the St George's Respiratory Questionnaire. *Thorax* 2010; **65**(10): 921-6.