Contribution of complex genetic variation at the Human leukocyte antigen (HLA) and Killer immunoglobulin-like receptor (KIR) regions to Idiopathic pulmonary fibrosis (IPF) susceptibility

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

By

Megan Louise Paynton BSc MSc

Department of Health Sciences

University of Leicester

August 2021

Abstract

Contribution of complex genetic variation at the Human leukocyte antigen (HLA) and Killer immunoglobulin-like receptor (KIR) regions to Idiopathic pulmonary fibrosis (IPF) susceptibility

Megan Louise Paynton

Idiopathic Pulmonary Fibrosis (IPF) is a rare lung disease characterised by inflammation and scarring of the alveoli. Although genetic and environmental factors have been reported in IPF, the biological processes underlying IPF development remain unclear. The largest genetic risk factor for IPF is a common variant in mucin gene *MUC5B.* It is believed that IPF develops because of microinjury in the lung from for example cigarette smoke or viral infection. The Human leukocyte antigen (HLA) and Killer immunoglobulin-like receptor (KIR) molecules play a vital role in immune response against infection. HLA allele, *HLA-DQB1*06:02* has evidence for association with fibrotic idiopathic interstitial pneumonias (fIIP) (including IPF). The HLA and KIR regions harbour complex variation and although there are links between IPF and viral infection, these regions have not been studied in depth. The aims of this thesis were to investigate the contribution of complex genetic variation in these regions to IPF susceptibility.

The HLA-wide association meta-analysis identified a common novel signal near *ZNRD1ASP* as associated with IPF susceptibility. Bioinformatic investigation highlighted associations with immunity and respiratory traits and differential expression of HLA and non-HLA genes. The *HLA-DQB1*06:02* variant did not replicate in three independent IPF datasets suggesting that the originally reported association may have been driven by the inclusion of non-IPF fIIPs. The *MUC5B**HLA interaction analysis in IPF susceptibility did not present any novel signals but identified some suggestively significant signals that warrant further investigation. The KIR-wide association meta-analysis did not identify any novel signals and highlighted concerns in the KIR imputation.

Overall, this thesis did not support the previously reported association of *HLA-DQB1*06:02* with IPF susceptibility. Also, there was no evidence the regions exhibit a large genetic effect on IPF risk, however the analyses were limited by sample size and the imputation quality in the KIR region.

Acknowledgements

I would like to thank my supervisors Professor Louise Wain and Dr Ed Hollox for all their help and support over the last four years. I would like to thank everyone in the Genetic Epidemiology group at the University of Leicester for all their help, insight and kindness, especially Dr Richard Allen for his support all things IPF. I would also like to thank all the PhD students in Health Sciences who I worked alongside for providing so much support, friendship and help with anything and everything.

I would like to acknowledge all my collaborators and all the IPF patients who allowed us to use their data for this research.

Finally, I would like to thank my family, for all their love and support over the past four years. I would especially like to thank my Mum, Dad and Husband-to-be Ryan who have never stopped believing in me and given me all the love, support, and encouragement I needed throughout my studies.

Table of Contents

List of Tables:

List of Figures:

Abbreviations and Glossary:

Term	Description
95% CI	95% confidence interval – 95% probability that the true parameter sits between the interval
Allele	Form of a genetic variant
Alveoli	Air cavities in the lungs where gas exchange takes place
CNV	Copy number variation – the number of copies of a sequence of DNA
COPD	Chronic Obstructive Pulmonary Disease – an obstructive lung disease
DNA	Deoxyribonucleic acid
eQTL	Expression quantitative trait loci – a genetic variation associated with expression of a gene
FEV_1	Forced expiratory volume in 1 second – the amount of air that an individual can forcibly expire in one second
fIIP	Fibrotic idiopathic interstitial pneumonias – interstitial lung diseases of no known cause characterised by scarring

FVC	Forced vital capacity – the total volume of air that an individual can expire
Gene	A section of the genome that encodes a protein
Genotype	The pair of alleles that an individual has at a given loci
Genotyping	Process of determining an individuals genotype at a given loci
GWAS	Genome-wide association study – variants across the genome are tested for association with a phenotype
Heterozygous	For a given variant, where the allele on one chromosome is different to the allele on the other
HLA	Human Leukocyte Antigen – molecules coded for by genes in the HLA region that play a role in the immune response to infection
Homozygous	For a given variant, where the allele on one chromosome is the same as the allele on the other
ILD	Interstitial lung disease – a disease affecting lung tissue
Imputation	Estimating genotypes for variants that have not been directly genotyped

IPF	Idiopathic Pulmonary Fibrosis – a chronic lung disease characterised by inflammation and scarring
KIR	Killer immunoglobulin-like receptors – molecules coded for by genes in the KIR region that play a role in the activation and inhibition of natural killer cells
LD	Linkage disequilibrium – a correlation between alleles at different loci
Locus (plural loci)	A fixed position on the chromosome where a particular variant is located
MAC	Minor allele count – the number of alleles of a variant in a sample
MAF	Minor allele frequency – the number of alleles of a variant in a sample divided by the total number of alleles in the variant
Major allele	The most common allele in a population
Minor allele	The least common allele in a population
NK cells	Natural killer cells – cells involved in the body's immune response to infection
OR	Odds ratio – the proportion of odds of an outcome in a group compared to the odds of an outcome in another group

PEF	Peak expiratory flow – how fast a person can forcibly expire
Phenotype	An observed trait
Polygenic	Where a phenotype is associated with multiple genes
Promoter	A section of DNA that affects the expression of a gene
QC	Quality control
SNP	Single Nucleotide Polymorphism – a single base variant in the DNA

Chapter 1: Introduction

Idiopathic pulmonary fibrosis (IPF) is a progressive interstitial lung disease with limited treatment options and poor prognosis. Previous studies have demonstrated that there are multiple genetic risk factors associated with IPF susceptibility (1-6). However, the extent to which these genetic factors contribute to IPF susceptibility, and how this can progress understanding of the underlying biology of IPF and lead to new interventions, is not yet fully known.

The HLA (human leukocyte antigen) and KIR (killer immunoglobulin like cell receptor) genes encode molecules involved in the immune response to bacterial and viral infections. Infection is hypothesised to be a trigger for development of IPF (7, 8) and is believed to be a cause of exacerbation events (9-11). Both the HLA and KIR regions harbour high levels of complex variation and genetic associations of HLA alleles with other respiratory diseases have been reported (12-14). A link between the HLA and IPF has been reported in one IPF case control study (5) whilst the KIR region has not been studied in IPF susceptibility to date. The aim of my PhD is to further characterise the contribution of genetic variation of the complex immune system genes; in particular, in the Killer-cell immunoglobulin-like receptor (KIR) and the Human Leukocyte Antigen (HLA) regions to susceptibility to Idiopathic Pulmonary Fibrosis (IPF) to improve our understanding of the disease and help identify and guide new treatment options.

1.1 Introduction to genetic epidemiology:

1.1.1 Genetic Variation

Most variation found throughout populations is rare or very rare, with common variation (frequency of > 5%) being relatively uncommon. The most common form of variation is the Single Nucleotide Polymorphism (SNP). Figure 1.1 denotes a part of a chromosome with a SNP indicated in red. The least frequent allele at each SNP is known as the minor allele, the frequency of this allele is calculated across a population to give a minor allele frequency (MAF). SNPs are relatively common in the Human Genome (about every 300bp), the SNP Map Working Group (15) first identified 1.42 million SNPs in a single genome in 2001, from then the 1000 Genome Project (16) data showed that each individual can contain around 11 million SNPs. More than 400 million single-nucleotide and insertion or deletion variants after alignment with the reference genome (17). There are other less common forms of variation in the genome such as indels (small insertions and deletions), copy number variation (different number of copies of a sequence) and inversions.

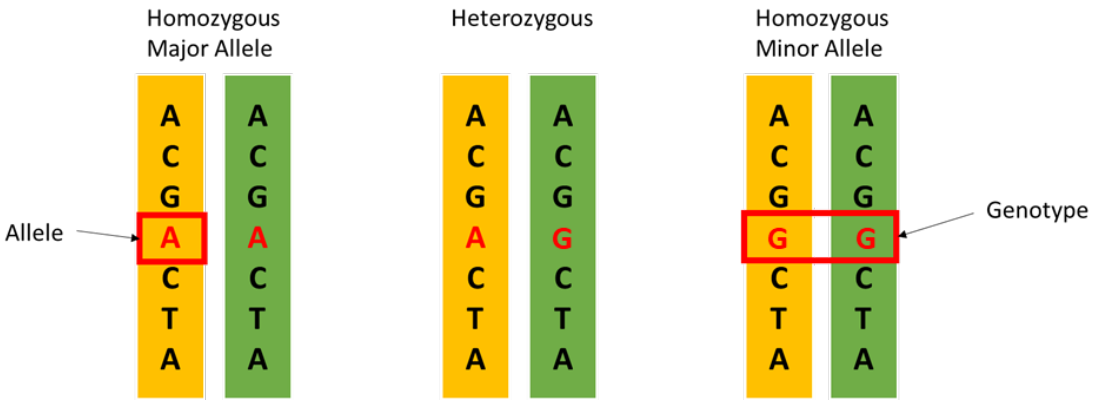


Figure 1.1: Sections of chromosome with a SNP denoted in red. The yellow chromosome has been inherited from the mother and the green chromosome has been inherited from the father. Only the positive strand is shown. An allele is the DNA nucleotide found on one chromosome and the genotype is both the alleles found across both chromosomes.

SNPs in the coding regions (exons) of genes can be described as synonymous (does not change the encoded amino acid) or non-synonymous (changes the encoded amino acid). The effect of a non-synonymous variant on protein structure and function can vary according to the specific change. For example, 'missense' is the term given to variants where the amino acid changes to another. If the amino acid change is to one with similar properties (e.g. hydrophobic) then it may not affect protein structure or function. However, if the change is to an amino acid with an opposite physical property (e.g. hydrophobic to hydrophilic), this may affect the protein product. Nonsense variants (stop loss or stop gain) can prematurely truncate the protein or produce an abnormally long protein and cause significant changes in protein function and structure.

SNPs may be found in non-coding parts of the genome and affect protein formation and expression in other ways. For example, SNPs in regulatory regions (such as promoters and enhancers) can affect levels of gene, and ultimately protein expression. Splice sites are regions which are cleaved to remove introns to produce mRNA; a variant in a splice site can cause exon skipping or read-through into introns, producing an altered protein.

1.1.2 Linkage DisequilibriumLinkage disequilibrium (LD) describes the non-random correlation of two or more SNP alleles across a general population (18) and the degree to which an allele of one SNP is associated with an allele of another SNP. Blocks of SNP alleles or haplotype

blocks (groups of SNP alleles or genes that are inherited together) are split during recombination events that occur during meiosis where portions of the chromatid arms can be exchanged (19). The extent of correlation between the two or more alleles is dependent on recombination events and population history. SNP alleles are said to be in high LD when they have not been subjected to recent recombination events and are therefore often inherited together.

LD varies across the genome but there are known hotspots where there are higher numbers of recombination events which results in areas of considerable variation (including SNPs, insertions, deletions and translocations). The HLA region exhibits lower rates of recombination compared to the rest of the genome, resulting in higher and longer-range linkage disequilibrium between alleles (20, 21).

LD is measured using R^2 which is a measure of correlation between the two SNPs. LD should be considered in association studies because correlation of SNP alleles means that the strongly associated SNP may not be the causal SNP and also it can be utilised for the imputation of missing genotypes.

1.1.3 Genetic Epidemiology

Genetic epidemiology is the study of the individual and joint contributions of genetic and environmental factors to health and disease in populations. In this thesis, the focus will be on the application of genetic epidemiology approaches to study to complex traits and diseases which are known to be polygenic (i.e. for which there are many tens, hundreds or thousands of SNPs either known or hypothesised to be associated with risk).

Genetic epidemiology studies have already yielded thousands of associations of common alleles (single nucleotide polymorphisms) with complex polygenic diseases such as Type 1 and Type 2 Diabetes Mellitus and Chronic Obstructive Pulmonary Disease (COPD), and complex traits such as height and blood pressure (22, 23). Genetic association signals can implicate genes as being important in the disease process to give new biological insight and potentially identify new therapeutic targets. Genetic association signals can also be used, individually or combined into risk scores, to improve disease prediction (24, 25).

1.1.4 Genome-wide association study design and statistical analyses:

Single nucleotide polymorphisms (SNPs) that are associated with risk of disease or other phenotypes can be identified using Genome Wide Association Studies (GWAS). Common SNPs

that have been shown to be associated with risk of complex polygenic diseases typically have individually small effects on disease risk (see Figure 1.2).

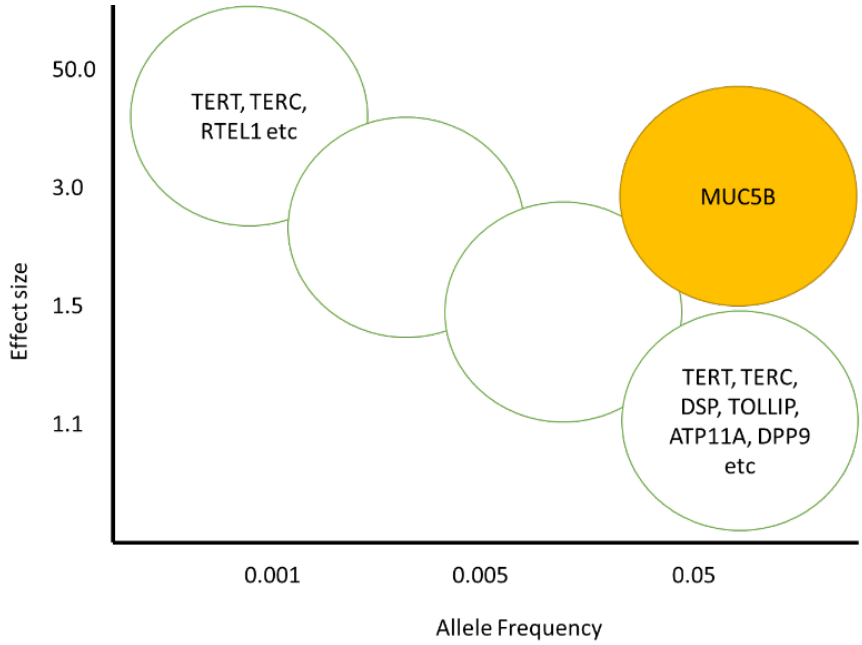


Figure 1.2: Correlation between effect size and allele frequency in IPF susceptibility. MUC5B sits outside the normal correlation with a larger effect size than expected considering its allele frequency.

GWAS test all measured variants across the genomes of a large number of individuals to determine if any of these variants are associated with the disease or trait of interest. They are considered to be a hypothesis-free approach as they are not limited to specific genomic regions. In an association analysis, an additive genetic model, a dominant genetic model or a recessive model can be assumed (Figure 1.3). In an association analysis, the outcome tested is a disease or trait and the risk of these may be affected by an exposure (such as a SNP allele). For a SNP with alleles A and G, (assuming that allele "A" confers an increased risk) the dominant model would assume AA and AG genotypes would both have the same increased risk of the outcome compared to GG. The recessive model would assume two copies of the risk allele are required to increase risk; therefore AA would increase the risk of outcome relative to AG and GG. Finally, the additive model would assume a linear effect with no risk alleles (GG) having the smallest risk effect, two risk alelles (AA) having the largest risk effect and one allele (GA) in between. In association studies, it is common to use additive models because there is sufficient power to detect both additive and dominant effects.

The most commonly used statistical tests for association in GWAS are logistic and linear regression (see below) (19). These methods are computationally efficient and enable inclusion of covariates in the association model. Risk of disease, such as IPF, is usually studied using a binary case-control analysis with logistic regression. Where quantitative traits, such as lung function, are the outcome, linear regression is used (see below).

The equation for a linear regression is as follows:

$$Y_i = \beta_0 + \beta_1 G_i + \beta_2 cov1_i + \beta_3 cov2_i + \cdots$$

Where Y_i is the phenotype for individual i, β_0 is the intercept, G_i is the genotype of the SNP (coded as 0, 1 or 2 for each copy of the allele of interest for a biallelic SNP under an additive genetic model), β_1 is the change in the phenotype for each copy of the effect allele and cov are covariates for individual i. For quantitative traits, the intercept is reported as effect size.

Case-control analyses are tested using logistic regression. Logit is the log odds of the probability of an individual (i) being a case or control (p_i). For case-control analyses the log odds is reported.

$$logit(p_i) = \beta_0 + \beta_1 G_i + \beta_2 cov1_i + \beta_3 cov2_i + \cdots$$

Logistic and linear regression can also be used to examine the effects of statistical interactions between different SNPs, and between SNPs and environmental factors, on outcome, by the inclusion of an interaction term. See section 1.4 “Chapter 3”.

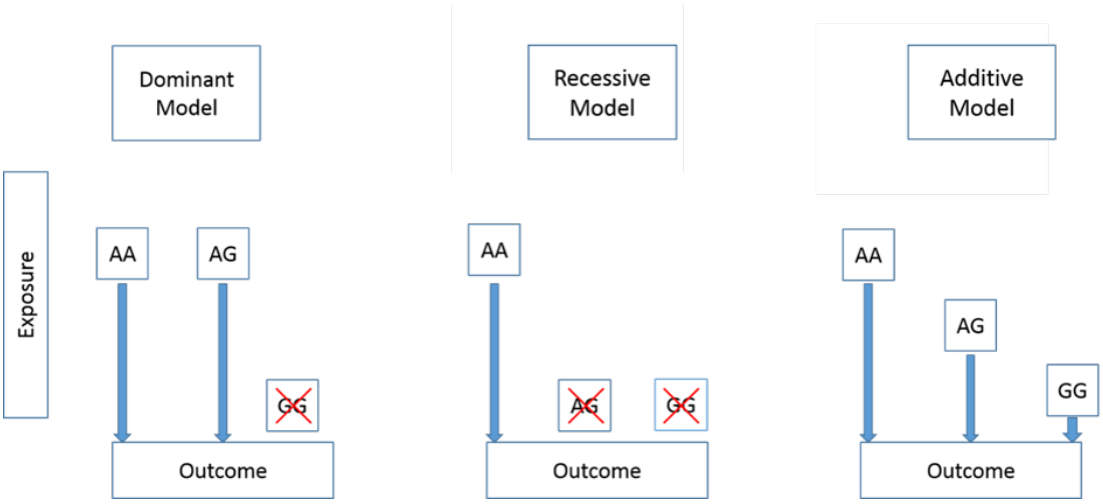


Figure 1.3: Genetic models for association analyses. "Outcome" indicates a degree of increased risk (in the case of a binary trait, for example, risk of disease) or increase in value (in the case of a quantitative trait, for example, increase of X mmHg for blood pressure) and "exposure" is the SNP. The length of the blue arrow is indicative of the risk of the outcome associated with the exposure genotype.

Replication of results is important for any scientific experiment to minimise reporting of false positives. Replication studies for GWAS should ideally be larger than the original study size to account for over estimation of the effect size in discovery (so-called Winner's curse bias) (19). In a study design that includes separate discovery and replication stages (Figure 1.4) signals that reach a genome-wide significance threshold of $P<5x10^{-8}$ in the discovery stage are selected for replication in a larger data set where a Bonferroni corrected (for multiple tests) significance level is defined based on the number of SNPs for which replication is sought.

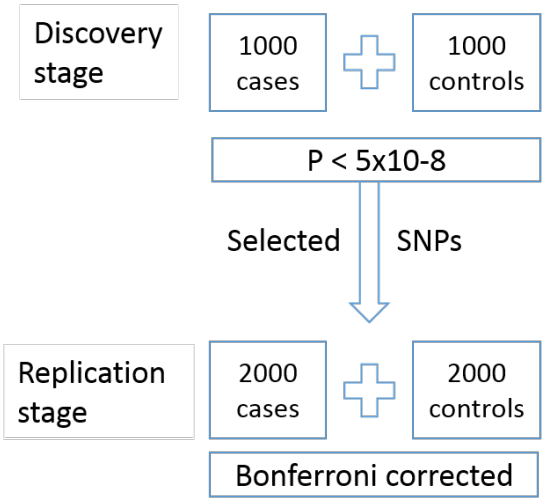


Figure 1.4: Discovery and replication study design shortlists signals from the discovery phase (at $P<5x10^{-8}$) and tests these signals for replication in an independent data set (replication stage). For example, if 10 SNPs are selected at discovery stage, they would be tested at a Bonferroni corrected threshold of 0.005 at replication stage.

A widely used alternative study design includes two stages as shown in Figure 1.5. The first stage (stage 1) identifies independent suggestive associations (for example, with $P< 5x10^{-5}$), which are then tested for association in an independent data set (stage 2). Signals that meet the genome-wide significant threshold of $P < 5x10^{-8}$ following meta-analysis of the stage 1 and stage 2 results are then reported. This 2-stage design is a helpful strategy when stage 1 is statistically underpowered (i.e. smaller than optimal sample size) as it allows for a more lenient P value threshold to be applied to identify signals for follow-up but still retains the requirement to meet a pre-determined genome-wide significant threshold overall.

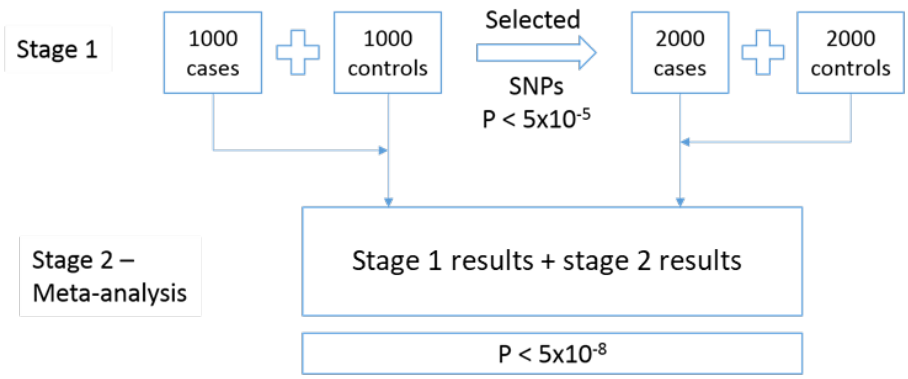


Figure 1.5: Two stage study design identifies signals at a suggestive threshold in Stage 1 (e.g. P<5x10^{-5}), and then meta-analyses the stage 1 results with an independent data set (stage 2) to identify signals which reach a genome-wide significant threshold of P<5x10^{-8}.

Two assumptions can be made when meta-analysing the results from independent datasets; fixed-effects or random-effects. Random-effects meta-analyses assume that the genetic variants have different effect sizes in each of the independent datasets. Fixed-effects assume that the genetic variants have the same effect sizes in all the datasets and this is the most widely used method of meta-analysis in genome-wide association studies.

The two-stage study design is of benefit when the studies that contribute to stage 2 do not have genome-wide data available, this strategy enables identification of a small number of SNPs to be followed-up by direct genotyping assays in those studies. Thus increasing the overall sample size to enable true positive signals to meet stringent significance thresholds. However, as many more studies now have genome-wide data, it is more logical to maximise power for all genome-wide variants by combining all available data. In doing this there is a balance that must be made between increasing discovery sample size vs ensuring the robustness of signals through replication. Replication can be sought externally through additional independent datasets or an internal validation approach can be applied to ensure the signals are not entirely driven by one contributing study (2). Alternatively, corroborative evidence can be sought to support a signal, for example, associations with related phenotypes (26).

1.1.5 Multiple testing problem in genome wide association studies

In a GWAS, thousands or millions of individual SNP association tests are performed in a single analysis which increases the likelihood that there will be a large number of statistically significant (P<0.05) results due to chance. For example, using a P-value threshold of P<0.05 in

a genetic study of 1,000,000 variants, one would expect 5,000 false positives (if the null hypothesis was true). When association testing is conducted across a particular locus, for example the human leukocyte antigen (HLA) region, the number of independent tests should be calculated to define an appropriate multiple testing threshold. This can be done using methods such as Bonferroni correction (27). The Bonferroni correction for multiple testing is deduced by dividing alpha (significance level, usually alpha=0.05) by the number of tests being performed (27), for example if the number of tests is 1,000,000, only SNPs with P values of < $5x10^{-8}$ will be reported as significantly associated with the trait. A limitation of the Bonferroni correction is that it is assuming that all of the individual tests are independent of each other. However, not all SNPs are independent of each other as some will be in linkage disequilibrium meaning that the Bonferroni correction may be over conservative and could yield a high rate of false negatives. $P<5x10^{-8}$ is a widely used P value threshold in GWAS and can be interpreted as being based on an assumption of one million independent tests. As studies are becoming larger and whole genome sequencing is becoming more commonly used, correcting for this multiple testing is becoming even more vital as it is possible that genome wide significance ($P<5x10^{-8}$) may not be an appropriate correction for the large number of variants being tested. As whole genome sequencing studies typically measure more lower frequency (MAF 1-5%) and rare (MAF <1%) variants, it is possible that many more than 1 million independent SNPs are being tested in a study, therefore a threshold of $P<5x10^{-9}$ has been recommended based on whole genome sequencing experiments in European, Asian or admixed ancestry populations, and $P<1x10^{-9}$ recommended for African samples (28).

1.1.6 Power in Genetic Association Studies

The power (ability to identify a true positive association) of a GWAS can be affected by many different factors. The number of cases in an analysis (the larger the sample size the better the power), the ratio of cases to controls, the minor allele frequency (MAF, the smaller the MAF the lower the power), the effect size (the larger the effect size the better the power) and the chosen alpha all affect the power of an analysis. Power is important as it improves the ability to identify a true association in association studies and it helps to understand that a lack of findings in an analysis may be due to low power and not because there is no association. 80% power is often used as indicative which means there is an 80% chance a true association will be identified in the analysis.

1.1.7 Genotype Imputation

Imputation of unmeasured genotypes is now commonly used to increase genome coverage (the number of SNPs that can be analysed in a study), increase power by enabling meta-

analysis of multiple data sets, and correct genotyping errors. Phasing is required to be undertaken before imputation and is the process by which alleles of directly genotyped SNP alleles are mapped to either the maternal or paternal chromosomes. Imputation is the process in which haplotype reference panels (panels of haplotypes derived from large sequencing projects) (for example Haplotype Reference Consortium panel (29)) are used to infer alleles for known SNPs that have not been directly genotyped (30, 31). Imputation uses LD (which varies by population) to infer SNP alleles in an individual's genome dependent on which genotype they possess at correlated loci to create imputed genotypes (30, 31). The quality of imputation is dependent on the quality and size of the imputation panel and the number and choice of SNPs directly genotyped in the data set being imputed. Increasing ancestry diversity of imputation reference panels has been shown to improve imputation quality (29, 32). Imputation uncertainty and quality can be captured in the imputation quality measures or correlation scores and this uncertainty can be used to either filter out poorly imputed variants or can be factored into downstream genetic analyses of the variants.

1.1.8 Ancestry and Cryptic Relatedness

Allele frequency correlates with ancestry (or geographical distance), and this is an important confounder and needs to be corrected for in association studies, especially when the geographical distances are large. In some cases, there are SNPs associated with both disease and ancestry which may have resulted through evolutionary selection in those populations (for example skin pigmentation) (33). Including these SNPs in an analysis could lead to an increase in the false positive rate because, for example, a SNP that has been reported to be associated with a disease in an association analysis could actually be associated with the population or the population differences of the individuals in the study (confounding). Fine scale population structure (for example throughout Europe or even within the UK) can also contribute to allele frequency differences which could confound genetic studies. Principal Component Analysis (PCA) is used to infer continuous axes of genetic variation (eigenvectors/Principal Components) which can describe as much variability between individuals as possible while reducing the data to a small number of dimensions. Because PCA can describe the variability between groups of individuals it can be used to determine the genetic distances and relatedness between populations. Ten principal components are most often used in GWAS but the optimum number of principal components can be determined by plotting the cumulative proportion of variation explained by the different numbers of principal components (scree plot). The curve will flatten at the point at which the best number of principal components sits. Recently, methodological approaches have been developed which account for relatedness

in association studies, an example approach is SIAGE which allows you to account for sample relatedness (34).

1.2 Idiopathic Pulmonary Fibrosis

1.2.1 Biological processes and clinical features in Idiopathic Pulmonary Fibrosis

Idiopathic pulmonary fibrosis (IPF) is a form of chronic, progressive, fibrosing interstitial pneumonia characterised by scarring and inflammation around the alveolar wall which stiffens the lungs and reduces their capacity and elasticity and impairs gas exchange. The biological processes behind the development and progression of IPF are still incompletely understood but disease initiation is thought to be caused by an abnormal wound healing response to injury (for example, smoking or viral infection) in susceptible individuals (35, 36). There is no cure for IPF and treatment options are limited. The treatment associated with the most improvement in patients is a lung transplant, but this is not often a viable option since many individuals diagnosed with IPF are elderly. Two drugs have been licenced for use in IPF; Pirfenidone (37) and Nintedanib (38, 39). These drugs are anti-fibrotic and slow down the rate of scarring on the lungs, but they cannot reverse it (37-39) and they are known to have quite significant side effects and, as such, are often not well tolerated. NICE guidelines therefore suggest that Pirfenidone and Nintedanib should only be prescribed in relatively early stage disease (individuals with a Forced Vital Capacity (FVC) of 50-80%) and the medication must be stopped if their FVC continues to decline at more than 10% a year (40, 41). IPF is a complex disease, and the risk is affected by multiple environmental and genetic factors. The risk of IPF increases with age (42) and males are also at higher risk. A history of smoking was found to be associated with an increased risk of IPF(43). Another risk factor for IPF is telomere length. Short telomere lengths have been found in individuals with familial pulmonary fibrosis caused by rare loss-of-function telomerase mutations (44, 45). Telomerase dysfunction is increasingly being recognised to be a risk factor in IPF (1-4, 6).

1.2.2 Epidemiology

There are significant challenges in the diagnosis of IPF, including misdiagnosis and changing diagnostic guidelines meaning that identifying the true incidence rate of IPF is difficult. Recently, an incidence rate of between 2.85 (narrow case definition) and 8.65 (broad case definition) for every 100,000 patient years has been published for the UK (42). The narrow case definition included individuals with hospital data codes incorporating a diagnosis of idiopathic fibrosing alveolitis, cryptogenic fibrosing alveolitis, idiopathic fibrosing alveolitis NOS, usual interstitial pneumonitis, and idiopathic pulmonary fibrosis and the broad case

definition included three additional diagnoses: diffuse pulmonary fibrosis, Pulmonary fibrosis, and Hamman–Rich syndrome (42). This rate varies across the UK and has increased by 78% from the year 2000 to 2012 (42). The incidence of IPF increases with age with most individuals being diagnosed between the ages of 75 and 90 years old (42). The incidence of IPF is also much higher in males (42), although a recent study suggests this may be a diagnostic bias (46). Only around 50% of individuals survive past 3 years with survival rates of 34% and 19% for 5 and 10 years respectively (42).

1.2.3 Genetic associations

Pulmonary Fibrosis has been shown to have a genetic component and has been proven to segregate through families (Familial Pulmonary Fibrosis [FPF]).

Early studies identified a genetic component of IPF susceptibility in twins, for example Peabody and Hayes in 1950 (47) and Javaheri *et all* in 1980 who identified a pair of monozygotic twins (who had lived apart during their lifetime) had both developed IPF (48).

More recently, further linkage and genome wide association studies (GWAS) were undertaken to identify common variation in IPF risk since it is known that there is a genomic factor to IPF but it was not a Mendelian disease. The first genome wide association study (GWAS) on susceptibility to IPF was undertaken in 2008 on 159 cases and 934 controls from Japanese ancestry with follow-up in a 83 cases and 535 controls from Biobank Japan (6). In this first study, a variant was identified in the *TERT* gene suggesting there may be a role of telomere maintenance in the development of IPF (6).

In 2011, 83 subjects with familial interstitial pneumonia, 492 subjects with IPF and 322 controls were analysed (49). In this analysis, a common SNP in the promoter region of *MUC5B* was identified at a frequency of 34% of IPF cases and only 9% in controls (49). This was the first study to identify the *MUC5B* SNP and suggest that dysregulated MUC5B expression could have a role in IPF pathogenesis. The *MUC5B* signal is described further below in section 1.2.4.

In 2013, 542 European IPF cases and 542 European controls from Chicago were analysed with replication in 544 cases and 687 controls and then further follow up in 324 cases and 702 controls (3). Five variants were identified in this GWAS: rs35705950 in *MUC5B*, three variants in *TOLLIP* and one SNP in *SPPL2C* (3). rs35705950 was found in the promoter region of *MUC5B* which encodes a mucin protein which is found in the airways and traps foreign bodies to stop them entering the lungs (3).

Additionally in 2013, 1,616 fibrotic idiopathic interstitial pneumonia (fIIP) cases and 4,683 controls from Colorado (of European ancestry) were analysed with replication in 876 cases and 1,890 controls (4). In this analysis, 11 loci were identified as associated with IPF susceptibility, these include *3q26, 7q22, 15q14-15, 17q21, DSP, DPP9, FAM13A, OBFC1, ATTP11A, TERT and MUC5B* (4) which further implicated the role of telomere maintenance in IPF and also suggested a role of cell-cell adhesion (*DSP, DPP9*). This study was repeated in 2016 incorporating HLA imputation (HIBAG – described below in Section 1.3.1 HLA Imputation strategies (50)) which identified a classical HLA allele (*HLA-DQB1*06:02*) as associated with fIIP susceptibility (described further in Chapter 3) (5).

In 2017, 602 European cases and 3,366 controls from UK Biobank were studied with replication in 2,158 IPF cases and 5,195 controls (from the Chicago and Colorado datasets as above) (2). In this study, variants in *DSP* and *MUC5B* were replicated (in independent datasets, i.e. not including the Chicago and Colorado datasets) and a novel signal in *AKAP13* was identified (2). *AKAP13* plays a role in a profibrotic signalling pathway, implicating cell signalling in IPF development and pathogenesis (2).

In total, the five IPF susceptibility association studies identified 17 genome-wide significant signals. In 2020, the largest GWAS of IPF susceptibility was undertaken in 2,668 IPF cases and 8,591 controls with replication in 1,456 IPF cases and 11,874 controls (1). In this study, 11 of the 17 previously identified signals (see above) were confirmed and 3 novel signals were identified in *DEPTOR*, *MAD1L1* and *KIF15* (1) (see table 1.1). *DEPTOR* is involved in cell signalling and *MAD1L1* and *KIF15* are involved in mitotic spindle assembly (1).

Missing heritability is the gap between heritability estimates from genetic associations from association studies and heritability estimates from twin studies (51). A challenge in missing heritability is that GWAS may not be powerful enough to detect the many variants associated with a trait with small genetic effects (52). Another challenge is that perhaps genotyping arrays are not efficiently capturing the rare genetic variants that may explain some of the missing heritability (52). Polygenic risk scores suggest that there are still thousands of unreported genetic variants associated with susceptibility to IPF (1). The *MUC5B* genetic variant explains 5.9%-9.4% disease liability and the remaining 13 non-*MUC5B* SNPs (not including *HLA-DQB1* and *SPDL1* SNPs) (53). In total, the current SNPs are only explaining around 12.4% of the disease in the general population (53).

In 2021, 752 IPF cases (541 PROFILE cases and 272 UK Biobank cases) and 119,055 controls (from UK Biobank) were studied with replication in 1,028 IPF cases and 196,986 controls from

FinnGen (54). In this study, a novel, rare missense variant in *SPDL1* was identified which suggests a role for mitotic checkpoint signalling during cell division in IPF (54).

Table 1.1: SNP alleles associated with risk of IPF.

Nearest Gene	**SNP/risk allele**	**Odds ratio (95% Confidence interval)**	**P-value**	**Reference**
7q22.1	rs6963345/T	1.30 (1.21-1.38)	$3.10x10^{-14}$	(4)
AKAP13	rs62025270/A	1.27 (1.18-1.36)	$1.27x10^{-10}$	(2)
ATPIIA	rs1278769/G	0.77 (0.71-0.83)	$1.34x10^{-10}$	(4)
DEPTOR	rs28513081/G	0.82 (0.76-0.87)	$1.20x10^{-9}$	(1)
DPP9	rs12610495/G	1.31 (1.22-1.42)	$2.92x10^{-12}$	(4)
DSP	rs2076295/G	1.46 (1.37-1.56)	$2.79x10^{-30}$	(4)
FAM13A	rs2609255/T	0.78 (0.74-0.84)	$3.30x10^{-13}$	(4)
KIF15	rs78238620/A	1.58 (1.37-1.83)	$5.12x10^{-10}$	(1)
HLA-DQB1	06:02	1.34 (1.18, 1.52)	$6.1x10^{-8}$	(5)
IVD	rs2034650/G	0.77 (0.71-0.82)	$7.30x10^{-16}$	(4)
MAD1L1	rs12699415/A	1.28 (1.19-1.37)	$7.15x10^{-13}$	(1)
MAPT	rs1981997/C	0.71 (0.65-0.87)	$2.83x10^{-16}$	(4)
MUC5B	rs35705950/T	4.84 (4.37-5.36)	$1.18x10^{-203}$	(49)
SPDL1	NM_017785.5:g.169588475 G > A p.Arg20Gln	2.87 (2.03-4.07)	$2.4x10^{-7}$	(54)
TERC	rs1881984/G	1.31 (1.21-1.40)	$7.09x10^{-13}$	(4)
TERT	rs2736100/A	0.72 (0.67-0.77)	$1.54x10^{-20}$	(3, 4, 6)

1.2.4 *MUC5B* variant

The variant with the largest effect size reported in GWAS of IPF susceptibility is a common SNP in the promoter region of the gene *MUC5B* (49, 55). The *MUC5B* risk allele is at a higher

frequency in IPF cases compared to controls (present in 30-35% of IPF cases compared to 11% in the general population) and is associated with a five-fold increased odds of IPF (1, 56). A recent study has shown that the *MUC5B* SNP explains 5.9-9.4% of risk of IPF susceptibility (54). The *MUC5B* risk allele results in an increased expression of the *MUC5B* gene which encodes for a protein called Mucin 5 subtype B which is a salivary mucin found in the tracheobronchial tract (57). *MUC5B* is a mucous secreting/gel forming mucin found in the tracheobronchial tract. Research in mouse models suggests that an increased expression of the *MUC5B* gene acts on IPF pathogenicity because the cells and bacteria (from the inflammatory response) get stuck in the excess mucin and are therefore present in the lungs for longer (58), this in turn promotes an abnormal healing response. A study in 2013 analysed the association between the *MUC5B* risk allele and interstitial lung abnormalities which found that for each copy of the *MUC5B* risk allele, the odds of interstitial lung abnormalities was increased by 2.8 times (59). This study suggests that there is a role for the *MUC5B* SNP in the wider group of interstitial lung diseases.

Although the *MUC5B* variant is associated with an unusually large (for a complex polygenic disease) increased risk of IPF, there are IPF cases without the *MUC5B* variant and there are healthy controls with the *MUC5B* variant. In IPF cases without the *MUC5B* SNP, there may be alternative genetic variants which affect their IPF risk. To date, analyses investigating a potential interaction with MUC5B have not yet been published.

1.2.5 Role of Viruses in Idiopathic Pulmonary Fibrosis (IPF)

Viral infection is hypothesised to be a trigger for IPF; infection of viruses such as herpes viruses have previously been linked to IPF development (7, 8, 60-63). In 2003 a study into Herpesvirus in IPF lungs found evidence of cytomegalovirus (CMV), Epstien-Barr virus (EBV), human herpesvirus 7 (HHP-7) and human herpesvirus 8 (HHP-8) in 32/33 IPF cases (97%) and only in 9/25 (36%) of controls suggesting herpesvirus could be a driving factor of IPF pathogenicity (60). Additionally, a study in 2014 used lung biopsies from 21 IPF cases and 21 age matched controls to identify any differences in Herpesvirus Saimiri DNA (64). Herpesvirus Saimiri DNA was found in 21/21 IPF cases but was not found in any of the control lung biopsies suggesting an association between Herpesvirus Saimiri and IPF (64). Studies in mice with bleomycin induced pulmonary fibrosis showed that those who received bleomycin and murine gamma herpesvirus (closely related to human herpes virus) had, histologically, higher fibrosis and inflammation scores and also more collagen than those who received only bleomycin(65). Latent virus in mice was also shown to predispose the lung to develop pulmonary fibrosis upon another exposure (in the case of the mice, bleomycin) (65). Similarly murine gamma

herpesvirus can also induce exacerbations in mice with existing fibrosis (65). In a recent meta-analysis of 20 IPF case-control studies (with 1,287 participants) identified that all studied viruses (apart from human herpesvirus 6) including Epstein-Barr virus (EBV), cytomegalovirus (CMV), human herpesvirus 7 (HHP-7) and human herpesvirus 8 (HHP-8) was associated with a significantly increased risk of IPF but was not associated with exacerbation of IPF (66).

1.3 Immune system genes

Two regions of the genome that are enriched for genes involved in the immune response to viral and bacterial infection are the Major Histocompatibility Complex (MHC) region and the Killer-cell immunoglobulin-like receptor (KIR) region. The genes in both of these regions are extremely polymorphic resulting in variation that cannot be appropriately captured using standard SNP imputation.

1.3.1 Human Leukocyte Antigen (HLA) region

In humans the MHC region is known as the Human Leukocyte Antigen (HLA) region. The HLA molecules have a significant role in antigen presentation and are required to display peptide fragments from pathogens on the surface of a T cell (67). The HLA region is located at 6p21.31 (chr6:28,477,797-33,448,354 on assembly GRCh37, according to the Genome Reference Consortium(68)). The genes within this region are split into three main classes; I, II and III (Figure 1.6, table 1.2). The HLA region is one of the most polymorphic regions in the Human genome, but it is known to have lower recombination rates compared to the rest of the genome (20). This results in large recombination blocks of SNPs in high LD, creating difficulty in identifying causal SNPs in association analyses.

HLA alleles were historically typed and named serologically (microlymphocytoxicity, i.e. by the use of anti-HLA antibodies) (69) using a system whereby names contain up to 7 sections (Figure 1.7 and table 1.3).The microlymphocytotoxicity assay used anti-HLA antibodies to detect mainly class I HLA gene alleles (the assay was not sensitive enough to identify class II molecules (70)) and molecular DNA typing was also used to type classical class I and class II HLA alleles more sensitively (70). Due to the polymorphic and polygenic nature of the HLA region, long-read sequencing (for example Pacific Bioscience's [PacBio] single molecule real time [SMRT] (71) and Oxford Nanopore sequencing (72)) is currently the gold standard for class I and II typing. Long-read sequencing involves the sequencing of 10,000-100,000 base pairs which enables improved sequencing of structural variation. SMRT long-ready sequencing by PacBio is currently being utilised by Anthony Nolan for transplants after it was identified that utilising this sequencing provided better HLA matching, resulting in improved survival

(73). Long-read sequencing however remains prohibitively expensive to undertake on a large scale; genome-wide genotyping arrays are a cheaper way to measure SNPs across the genome and can be used to impute type I and type II HLA gene alleles and amino acid alleles (74). HLA molecules can be grouped dependant on their historical serological epitope (e.g. HLA-C1, HLA-Bw4), these are defined based on differences in amino acid positions 77 and 80 (table 1.4). The HLA alleles that belong to each serological epitope can be found in table 1.4.

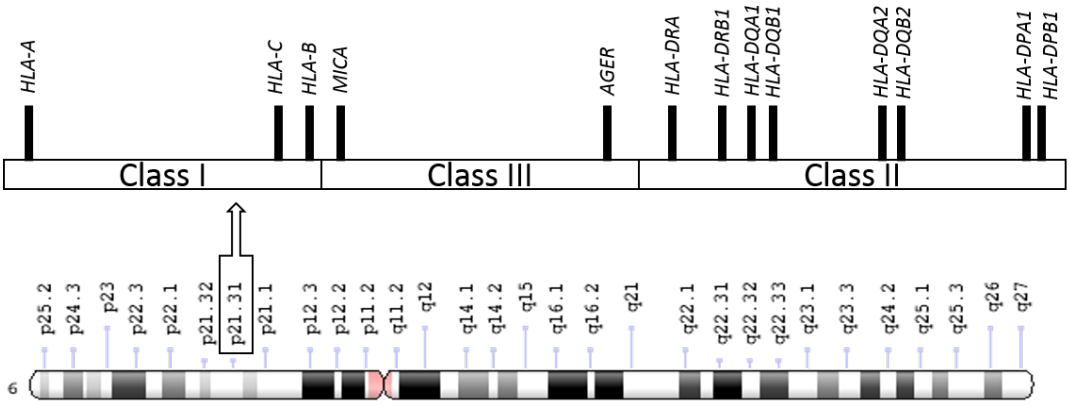


Figure 1.6: *Architecture of the HLA classes and genes in the human genome. Adapted from* (75).

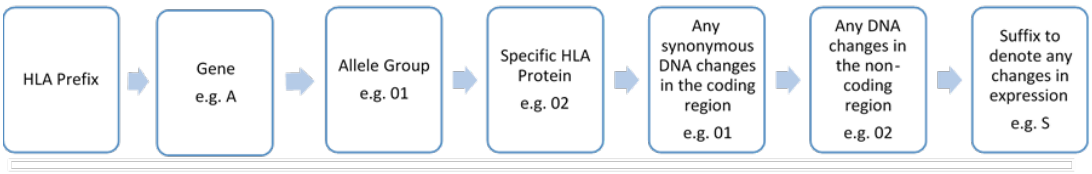


Figure 1.7: *Denotes the process in which a HLA allele is named, adapted from (76). The example given would be written as* HLA-A*01:02:01:02:S*. The S suffix denotes that the encoded protein is secreted (Table 1.2).*

Table 1.2: *HLA Class I, II and III genes and their properties* (77).

Gene Name	**Properties**	**Gene Name**	**Properties**
Class I			
HLA-A	Class I α chain	*HLA-N*	Pseudogene
HLA-B	Class I α chain	*HLA-P*	Pseudogene
HLA-C	Class I α chain	*HLA-S*	Pseudogene
HLA-E	Class I α chain paralogue	*HLA-T*	Pseudogene
HLA-F	Class I α chain paralogue	*HLA-U*	Pseudogene

HLA-G	Class I α chain paralogue	*HLA-V*	Pseudogene
HLA-H	Pseudogene	*HLA-W*	Pseudogene
HLA-J	Pseudogene	*HLA-X*	Pseudogene
HLA-K	Pseudogene	*HLA-Y*	Pseudogene
HLA-L	Pseudogene	*HLA-Z*	Pseudogene
Class II			
HLA-DRA	DR α chain	*HLA-DQA2*	DQ α chain-related sequence (not expressed)
HLA-DRB1	DR β1 chain	*HLA-DQB2*	DQ β chain-related sequence (not expressed)
HLA-DRB2	Pseudogene	*HLA-DQB3*	DQ β chain-related sequence (not expressed)
HLA-DRB3	DR β3 chain	*HLA-DOA*	DO α chain
HLA-DRB4	DR β4 chain	*HLA-DOB*	DO β chain
HLA-DRB5	DR β5 chain	*HLA-DMA*	DM α chain
HLA-DRB6	Pseudogene	*HLA-DMB*	DM β chain
HLA-DRB7	Pseudogene	*HLA-DPA1*	DP α chain
HLA-DRB8	Pseudogene	*HLA-DPB1*	DP β chain
HLA-DRB9	Pseudogene	*HLA-DPA2*	Pseudogene
HLA-DQA1	DQ α chain	*HLA-DPB2*	Pseudogene
HLA-DQB1	DQ β chain	*HLA-DPA3*	Pseudogene
Class III			
TAP1	ATP Binding Cassette (ABC) transporter	*MICAB*	Class I Chain related gene
TAP2	ATP Binding Cassette (ABC) transporter	*MICC*	Pseudogene
PSMB9	Proteasome related sequence	*MICD*	Pseudogene
PSMB8	Proteasome related sequence	*MICE*	Pseudogene
MICA	Class I Chain related gene	*AGER*	Advanced Glycosylation End-Product Specific Receptor

Table 1.3: The suffixes available for HLA nomenclature and their meanings. Information from (76).

Suffix	Meaning
A	Aberrant – some uncertainty to whether a protein is expressed.
C	Cytoplasm – protein is found here instead of cell surface.
L	Low cell surface expression.
Q	Questionable expression.
S	Soluble - the molecule is secreted but not expressed on the cell surface.

Table 1.4: Amino acid position differences in serological epitopes and which HLA alleles carry the epitope (78).

Serological epitope	**Amino acid position**		**HLA Alleles**
	77	**80**	B*05, B*51:02, B*51:03, B*13, B*17, B*27, B*37, B*38, B*44, B*47, B*49, B*51, B*52, B*53, B*57, B*58, B*59, B*63, B*77
HLA-Bw4	Asn	Ile	
	Asn	Thr	
	Asp	Thr	
	Ser	Ile	
HLA-Bw6	Gly	Asn	B*07, B*07:03, B*08, B*14, B*18, B*22, B*27:08, B*35, B*39, B*40, B*41, B*42, B*45, B*46, B*48, B*50, B*54, B*55, B*56, B*60, B*61, B*62, B*64, B*65, B*67, B*70, B*71, B*72, B*73, B*75, B*76, B*78, B*81, B*82
	Ser	Asn	
HLA-C1	-	Asn	C*01:02, C*01:03, C*03:02, C*03:03, C*03:41, C*07:01, C*07:02, C*03:04, C*07:05, C*07:06, C*08:01, C*08:02, C*08:03, C*12:03, C*12:06, C*12:21, C*12:22, C*14:02, C*14:03, C*16:01, C*16:03, C*16:41
HLA-C2	-	Lys	C*02:21, C*02:22, C*02:23, C*02:24, C*04:01, C*05:01, C*06:02, C*07:07, C*12:05, C*12:41, C*12:42, C*15:02, C*15:03, C*15:04, C*15:51, C*15:52, C*16:02, C*17:01, C*17:02, C*18:01, C*18:02

Class I and Class II HLA molecules

Class I HLA molecules can be found spanning the membrane of almost every cell in an organism and are recognised by cytotoxic CD8+ T cells (79). CD8+ T cells are a type of T Lymphocyte that express T cell receptors and a CD8 dimeric co-receptor which allows the cells to recognise peptides presented by HLA Class I molecules. There are three class I α-chain (also known as "classical") genes in humans; *HLA-A, HLA-B* and *HLA-C* (table 1.2) (77). There are also additional class I genes which are paralogues of the classical class I genes, they have been termed HLA class Ib genes, like classical class I genes, they code for cell surface molecules (*HLA-E, HLA-F* and *HLA-G*) (table 1.2) (77).

Class II HLA molecules are restricted to only specific immune system cells such as macrophages and lymphocytes and are recognised by CD4+ T cells (79). CD4+ T cells are also known as T helper cells and aid immune response through the use of cytokines. There are three main pairs of classical HLA Class II α and β chain genes, named *HLA-DR, HLA-DP* and *HLA-DQ* (77) (table 1.2). The HLA-DR cluster also contains extra β-chains which can pair with the HLA-DRα chains creating several different HLA-DR genes (77) (table 1.2). Many in the HLA-DR cluster are pseudogenes and are not expressed such as *HLA-DRB2* and *HLA-DRB6* (table 1.2). There are a

few other less known HLA class II genes called *HLA-DOA/DOB* and *HLA-DMA/DMB*, these are not as variable as the class II classical HLA genes (they do not code for as many alleles and proteins as classical genes). Class III HLA molecules contain genes which code for a variety of different proteins such as *MHC Class I Polypeptide-Related Sequence A* (*MICA,* stress induced self-antigen), *Advanced Glycosylation End-Product Specific Receptor (AGER*, cell surface receptor) and *Transporter 1, ATP Binding Cassette Subfamily B Member* (*TAP1)* and *TAP2* which are involved in the processing for antigen presentation by HLA molecules (80).

Polymorphism of genes in the HLA region

Polymorphism in the HLA region means that some HLA class I and class II genes can encode more than 2000 alleles (with many alleles at a relatively high frequency in the population, for example HLA-A*01:01 is found in around 35% of individuals in European populations) (table 1.5 and 1.6). This means that each individual can possess a set of HLA molecules which have different, extended ranges of peptide binding specificities (81).

Table 1.5: *Number of alleles and proteins expressed for HLA Class I genes (accessed 11th July 2021)* (81).

HLA Gene	***A***	***B***	***C***	***E***	***F***	***G***
Alleles	6,766	7,967	6,621	271	45	82
Proteins	4,064	4,962	3,831	110	6	22

Table 1.6: *Number of alleles and proteins expressed for HLA Class II genes (accessed 11th July 2021)* (81).

HLA Gene	***DRA***	***DRB***	***DQA1***	***DQB1***	***DPA1***	***DPA2***	***DPB1***	***DPB2***
Alleles	29	3,701	306	1,997	258	5	1,749	6
Proteins	2	2,557	143	1,303	107	0	1,106	0

HLA Imputation strategies

The HLA region is one of the most complex and polymorphic in the human genome. When the HLA region was first mapped in 2004, it was found that around 22% of the expressed HLA genes had higher numbers of SNPs than the average number across the rest of the genome (82). With the release of 1000 genomes project in 2010, there was further evidence that the HLA region had significantly higher variation than the rest of the genome (around 9 SNPs per KB compared to around 5 SNPs per KB which was average across the genome (16). Studying linkage disequilibrium (LD) in the HLA region has been rare because there are such stark differences between individuals, ancestries and gene alleles however, the HLA region appears to exhibit long-range LD blocks (83, 84). This makes interpretation of association study findings

complex in terms of identification of causal variants. To overcome this there are well characterised HLA gene alleles and amino acid changes that can be imputed using SNP data to "fine-map" associations to potentially causal gene alleles and their encoded molecules.

There are reference panels that include HLA alleles and amino acid changes (as well as dense coverage of the HLA-region SNPs) which can be used to impute the HLA region from relatively sparse SNP array data. There are methods available which utilise these reference panels including, SNP2HLA, HLA*IMP and HIBAG, which will be described below.

SNP2HLA is an imputation method which utilised a reference panel produced from 5,225 unrelated individuals from the Type one Diabetes Genomics Consortium (T1DGC) (85). In these individuals, 7,135 SNPs across the HLA region were genotyped with the Illumina Immunochip array and classical HLA genes were typed (HLA-A, -B, -C, -DQA1, -DQB1, -DPA1, -DRB1) at a four digit resolution using the Illumina GoldenGate platform (see section 1.3.1: Nomenclature) (85). Using the EMBL-EBI immunogenetics database(86), binary markers were defined for SNPs, amino acids and 2- and 4-digit classical HLA alleles. For multi-allelic positions (i.e. amino acid changes and HLA alleles), the binary markers were defined based on the presence or absence of each amino acid/allele (85).

HLA*IMP (74) is an imputation method which utilised a reference dataset created from 2,420 samples from the 1958 Birth Cohort (also known as the 1958 National Child Development Study) typed using Illumina 1.2M and Affymetrix genome-wide human SNP array 6.0) and 92 HapMap CEU samples (87). HLA genotypes were inferred using classical typing techniques and converted into SNP haplotypes using PHASE (haplotype estimator) (88). The total reference data consisted of 5,024 haplotypes containing data on 7,733 SNPs across HLA-A, -B, -C, -DQA1, - DQB1 and -DRB1 (74).

HIBAG (50, 89) is an imputation method utilised samples from HapMap (87), 1958 Birth Cohort and HLARES to create a reference panel for HLA imputation. The HapMap dataset was created by combining SNPs genotyped using several different methods including Affymetrix, Illumina and Perlegen, the HLA data was then derived by combining the genotypes and sequence data (53). The HLARES and 1958 birth cohort HLA data was generated using typing methodologies (including sequence specific primer methodology) (50).

SNP2HLA, HLA*IMP and HIBAG were compared in a population of 3,265 individuals from the Vanderbilt DNA databank (90). Concordance (with sequenced HLA results) and call rate were compared between the datasets for each HLA gene allele imputed (90). SNP2HLA had the

highest concordance rate (0.975) and call rate (1.00) and predicted the highest number of alleles (210 alleles) in European ancestries, of the three tools (90). This effect was also seen in the separate HLA genes where SNP2HLA had the highest call rate across all genes and also the highest (or joint highest) concordance rate across all genes apart from HLA-A and HLA-B where HIBAG was slightly higher (HLA-A, 0.983 and 0.986 respectively and HLA-B, 0.969 and 0.978 respectively) (90).

HLA imputation of the data sets to be used in this thesis using SNP2HLA will be described in more detail in Chapter 2.

Biological importance of HLA

HLA class I molecules allow T-cells to detect host cells which are infected with a harmful microorganism (such as a virus) (91). When a cell is infected, antigen presentation by HLA class I molecules is triggered. Abnormal peptides in the cytoplasm of the cell are transported to the endoplasmic reticulum by TAP proteins (encoded by class III HLA genes) where they are incorporated into HLA class I molecules as they are synthesised (80). Then, they are exported to the plasma membrane of the cell within vesicles. The abnormal peptides are then displayed by the HLA class I molecules on the cell for T-cell recognition.

HLA class II molecules allow T-cells to detect when specific immune cells, termed antigen presenting cells (APCs) have ingested infectious microorganisms (91). When phagocytic APCs such as macrophages engulf microorganisms, they will partially digest them to produce peptide fragments. After the HLA class II molecules are produced in the endoplasmic reticulum, they can bind the peptide fragments (91). The compound is transported to the membrane where the T-cell recognises the foreign peptide fragment and binds to it, initiating an immune response. Macrophages (or other APCs) are activated to kill the microorganisms which they have engulfed and B cells are activated to produce antibodies to remove free microorganisms. Other immune cells such as natural killer cells can also be activated to destroy infected host cells (see section 1.3.2). In cells that have not been infected by a microorganism, HLA molecules present self-peptides on the cell surface to which T-cells would not normally react.

The HLA region has been implicated as having a key function in hundreds of different phenotypes and diseases (22) (table 1.7). The most significant associations have been identified in conditions in which autoimmunity and inflammation play a part (22). Autoimmunity is the process of an immune response of an organism targeting its own healthy

cells and inflammation is part of the body's immune response to harmful stimuli, characterised by the movement of white blood cells to the infected or damaged area followed by tissue-healing processes. Chronic inflammation can lead to the development of several diseases including Rheumatoid Arthritis. Table 1.7 shows the top most significant associations identified in the HLA region as reported in GWAS catalog (22) (accessed 29th November 2018). Many are autoimmune disorders or phenotypes in which inflammation is known to play a part in progression (such as Ankylosing Spondylitis).

Table 1.7: A selection of the most significantly associated traits reported for the three classical HLA class I genes and two classical class II genes ($P < 5x10^{-8}$); identified using GWAS catalog (accessed 29th November 2020) (22).

HLA Gene	Reported Traits	References
HLA-A	Birdshot Chorioretinopathy	(92)
	Nasopharyngeal Carcinoma	(93, 94)
	Vitiligo	(95)
	Ankylosing Spondylitis	(96)
	Autism Spectrum Disorder	(97)
HLA-B	Psoriasis	(98-100)
	Ankylosing Spondylitis	(96, 101)
	Psoriatic Arthritis	(99)
	Idiopathic Membranous Nephropathy	(102)
	Graves' Disease	(103)
HLA-C	Psoriasis	(98-100, 104-107)
	Ulcerative colitis	(108)
	Crohns Disease	(109, 110)
	Vitiligo	(111)
	Atopic Dermatitis	(112)
HLA-DQB1	Rheumatoid Arthritis	(113)
	Juvenile idiopathic arthritis	(114)
	Ulcerative Colitis	(109)
	Sjögren's syndrome	(115)
	Systemic Lupus Erythematosus	(116)
HLA-DRB1	Rheumatoid arthritis	(113, 117-126)
	Multiple Sclerosis	(127-132)
	Type 1 Diabetes	(123, 133, 134)
	Ulcerative Colitis	(108, 109, 135, 136)
	Sjögren's syndrome	(115, 137, 138)

1.3.2 Killer Immunoglobulin like-receptor (KIR)

Natural killer (NK) cells are cytotoxic lymphocytes that play a vital part in the innate immune response to virally infected host cells (139). NK cells detect the foreign viral proteins presented

by human leukocyte antigen (HLA) molecules on the surface of infected cells which triggers secretion of cytokines (such as IFNγ and TNFα) to initiate cell death and enhance the immune response (Figure 1.8). The activation of NK cells is controlled by a balance between activating receptors and inhibitory receptors on their surface (including killer immunoglobulin like cell receptors [KIR]) (140-142). The KIRs are encoded by genes in a region on chromosome 19 between the base pairs of 50,900,000-58,617,616 (19q13.4) on genome build 37. KIRs are a family of tautological (genetically diverse and polymorphic) receptors encoded by more than 1,110 alleles of 17 genes identified to date (78, 143) (table 1.8) along with many haplotypes made up of KIR gene CNVs (table 1.7). KIRs bind with HLA class I ligands to either activate or inhibit NK cell response, depending on the receptor-ligand combination (table 1.4, table 1.8). KIRs preferentially attack cells with cells that have down-regulated (fewer) HLA Class I molecules on the cell's surface (141) (Figure 1.8).

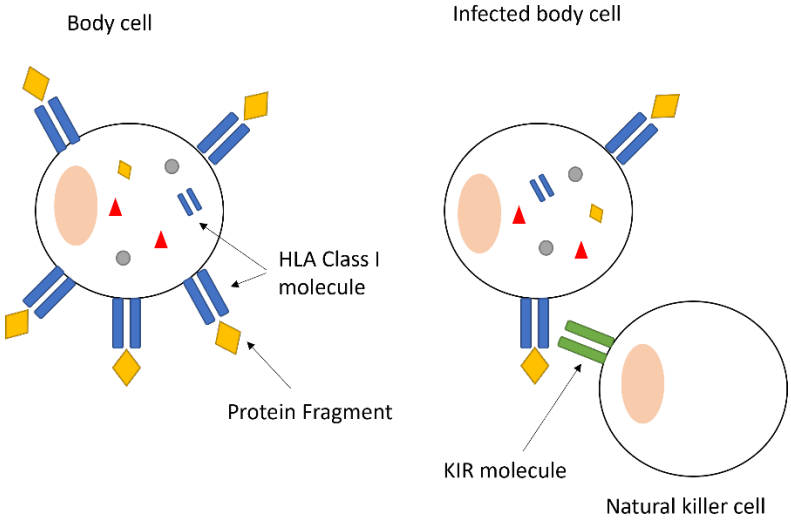


Figure 1.8: Image presenting how KIR molecules interact with HLA molecules on body cells. KIR molecules preferentially attack body cells with fewer HLA Class I molecules on the cell's surface.

KIR molecules are not only presented on natural killer cells, they are also presented on natural killer T (NKT) cells (144). Additionally, KIR molecules are also known to be presented on CD8 memory T-cells (145). It is believed that activating KIRs on CD8+ T cells may boost T cell responses and inhibiting KIRs on CD8+ T cells may affect T-cell receptor signalling and T cell responses (146).

KIR genes are named based on the structure of the molecule they encode, the first number following the KIR acronym corresponds to the number of IG-like domains in the KIR molecule. The D denotes “domain” and the letter following corresponds to the molecules properties; L denotes “long” cytoplasmic tail, S denotes “short” cytoplasmic tail and P denotes “pseudogene” (78, 147). The final number is the gene number encoding the protein. Two similar genes can have the same number and are distinguished between by a letter, e.g. KIR2DL5A and KIR2DL5B.

The KIR region genes are present on two types of haplotypes, A and B (Figure 1.9). The A haplotype is characterised by stable copy number of the KIR genes (with genes being present as one copy only or missing) while the B haplotype has more extensive copy number variation of the genes, with each gene varying in copy number from 0 to 2. The A haplotype is comprised of seven KIR genes and two pseudogenes (Figure 1.9). Six of the seven KIR genes are inhibitory, the final gene (KIR2DS4) has been deactivated by a 22bp frameshift mutation in around 75% of A haplotypes (148, 149); when functional it is an activating gene. The B haplotype is made up of varying numbers of all known activating and inhibiting KIR genes (Figure 1.9). Haplotypes are split into telomeric and centromeric regions of KIR genes which exhibit copy number variation (CNVs) (Figure 1.9). A and B haplotypes are named depending on the centromeric and telomeric groups (for example, haplotype A will have centromeric group A and telomeric group A) (see Figure 1.9). The KIR haplotypes and what CNVs they are comprised of can be seen in supplementary table 1.1.

Table 1.8: Names and descriptions of each activating or inhibiting KIR gene (and their respective HLA ligand) with the number of alleles and proteins ((78, 150).

Gene name	Description	Activating or inhibiting?	No. alleles	N. Proteins	N. Nulls	HLA Ligand
KIR2DL1	killer cell immunoglobulin-like receptor, two domains, long cytoplasmic tail, 1	Inhibiting	111	36	2	HLA-C2
KIR2DL2	killer cell immunoglobulin-like receptor, two domains, long cytoplasmic tail, 2	Inhibiting	34	15	0	HLA-C1, HLA-C2, HLA-B*46:01, HLA-B*73:01
KIR2DL3	killer cell immunoglobulin-like receptor, two domains, long cytoplasmic tail, 3	Inhibiting	64	35	1	HLA-C1, HLA-C2, HLA-B*46:01, HLA-B*73:01
KIR2DL4	killer cell immunoglobulin-like receptor, two domains, long cytoplasmic tail, 4	Inhibiting	107	54	0	HLA-G
KIR2DL5A/B	killer cell immunoglobulin-like receptor, two domains, long cytoplasmic tail, 5A/5B	Inhibiting	57	24	0	Unknown
KIR2DS1	killer cell immunoglobulin-like receptor, two domains, short cytoplasmic tail, 1	Activating	16	8	0	HLA-C2
KIR2DS2	killer cell immunoglobulin-like receptor, two domains, short cytoplasmic tail, 2	Activating	24	9	0	HLA-C1
KIR2DS3	killer cell immunoglobulin-like receptor, two domains, short cytoplasmic tail, 3	Activating	16	7	1	Unknown
KIR2DS4	killer cell immunoglobulin-like receptor, two domains, short cytoplasmic tail, 4	Activating	37	18	0	HLA-C*05:01, HLA-A*11:02, HLA-C*16:01
KIR2DS5	killer cell immunoglobulin-like receptor, two domains, short cytoplasmic tail, 5	Activating	24	17	0	Unknown
KIR2DP1	killer cell immunoglobulin-like receptor, two domains, pseudogene 1	Pseudogene	40	0	0	N/A
KIR3DL1	killer cell immunoglobulin-like receptor, three domains, long cytoplasmic tail, 1	Inhibiting	183	92	3	HLA-A, HLA-Bw4

KIR3DL2	killer cell immunoglobulin-like receptor, three domains, long cytoplasmic tail, 2	Inhibiting	164	4	1	HLA-A*03, HLA-A*11
KIR3DL3	killer cell immunoglobulin-like receptor, three domains, long cytoplasmic tail, 3	Inhibiting	165	92	1	Unknown
KIR3DS1	killer cell immunoglobulin-like receptor, three domains, short cytoplasmic tail, 1	Activating	39	22	1	Unknown
KIR3DP1	killer cell immunoglobulin-like receptor, three domains, pseudogene, 1	Pseudogene	29	0	0	N/A

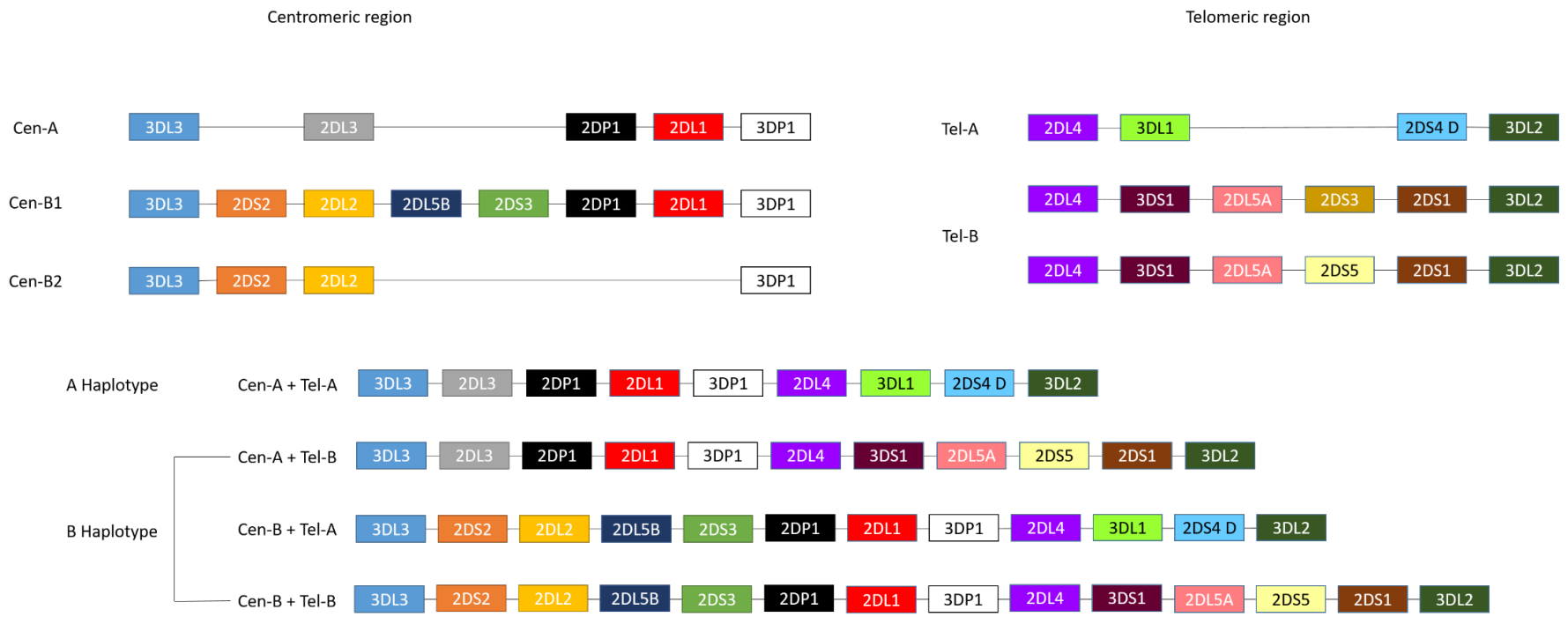


Figure 1.9: Diagram of the organisation of KIR genes in A and B haplotypes (adapted from (151)). Each gene is assigned a different colour and haplotypes are made up of centromeric and telomeric regions.

KIR region Imputation strategies

Complex genetic variation in the KIR region requires bespoke imputation to appropriately capture the variation across the region (without directly genotyping or typing using next generation sequencing) because SNP panels cannot infer CNVs or haplotypes. This is because SNP panels are not dense enough and do not provide enough information to decipher larger structural variation. Long-read sequencing is currently the gold standard for typing KIR haplotypes as it enables typing of larger structural variation however this is time-consuming and expensive, imputation provides a cheaper way to infer KIR variation on large scale. KIR*IMP uses a UK reference panel created from 698 individuals from the UK DNA banking network (DBN) which contains individuals who were selected for having atopic dermatitis or asthma (152). The copy-number variation in the KIR loci had previously been typed using qPCR (136). The imputation strategy was validated using a panel from a total of 1,338 unrelated German and Norwegian cases and healthy controls from a study of primary sclerosing cholangitis where the KIR gene copy numbers had been typed using qPCR (as above). KIR*IMP had an accuracy of over 90% for each of the KIR gene copy number variants and 87.1% for the KIR haplotypes and performed better than using conventional imputation panels (variants highly correlated with a haplotype or gene) and better than other bespoke imputation methods designed for the HLA region (HLA*IMP and HIBAG) (152). The imputation of the KIR locus in IPF case-control datasets will be described in more detail in Chapter 5.

KIR variation in disease

The KIR region, and copy number variation (CNVs) in the KIR region, has not been extensively studied in relationship to disease risk and associations are limited in number because the complex variation is hard to infer and the CNVs are not well tagged by SNP variation. Some KIR variation and KIR:HLA interactions have been associated with some viral infection phenotypes, disease susceptibility and progression and response to treatment.

Activating KIR genes in general have been suggested to affect the outcome of different viral infections including human immunodeficiency virus (HIV), cytomegalovirus (CMV) and hepatitis C (150, 153-155). HIV-infected individuals who were homozygous for *KIR3DS1 (*i.e. carried two B haplotypes containing at least one copy of the *KIR3DS1* gene) progressed more slowly to AIDS if they also carried HLA-Bw4 with an isoleucine at position 80 (HLA-Bw4^{80I}) (155). HLA-Bw4^{80I} individuals with increasing copy number variation in *KIR3DS1* have been shown to have a lower HIV viral set load (156). The activation of natural killer (NK) cells by activating KIRs and their respective HLA ligands has been suggested to play a role in controlling NK cell response to cancer cells (since NK cells expressing activating KIRs are less responsive in

the presence of self-HLA antigens) and therefore inhibit the body's ability to produce an appropriate response to cancer cells (150). Variation of individual KIR genes including *KIR2DS1* have been associated with disease progression (breast cancer, (157)), antibody treatments (imatinib in chronic myeloid leukaemia (CML) (158)) and progression free survival (after stem cell transplants in CML (159)) in cancer. Some KIR genes have been suggested to be associated with inflammatory and autoimmune disorders including systemic lupus erythematosus (SLE) (*KIR2DS1* (160)), psoriasis vulgaris (*KIR2DS1* (161)) and multiple sclerosis (*KIR3DS1* (162)).

Genetic variation of the KIR region has not been studied in lung disease. There have, however, been studies into the role of natural killer cells in lung infection and disease (163). Natural killer cell function can be impaired by smoking and therefore differences in natural killer cells in lung function and disease needs to take this into account. (164-167). There have been studies that show that natural killer cell activity in peripheral blood mononuclear cells (PBMCs) is enhanced in asthmatics (168, 169) and this has also been seen in an animal model of allergic airway sensitisation (170). Additionally, acute exacerbations in asthmatic children was associated with an increased frequency of natural killer cells in PBMCs (171). Natural killer cell function has also been suggested to be impaired in COPD (172). In COPD and asthma, most exacerbations are caused by respiratory infections (173, 174) and therefore the ability of natural killer cells to protect against infection may suggest a role of these cells in respiratory disorders and offer novel therapeutic targets. In IPF, expression of NKG2D (a receptor on natural killer cells which binds to self-antigens) was reduced on natural killer cells in IPF cases compared to healthy controls (175-177). Models of bleomycin-induced pulmonary fibrosis showed that a lack of natural killer-cell recruitment (and subsequent interferon gamma (IFN-γ) release) stopped advanced fibrosis, suggesting that IFN- γ release from NK cells could have a role in regulating pulmonary fibrosis (166, 167)). Also, the ability of natural killer cells to protect against infection may in turn limit lung inflammation and subsequent fibrosis. Additionally, there is a suggestive role of Invariant NKT (iNKT) cells in lung disease including chronic obstructive pulmonary disease (COPD) (178, 179). A study in 2014 identified that the frequency of activated iNKT cells was increased in COPD patients (179) and a study in 2016 used mouse models to study the link between increased frequency of iNKT cells and pulmonary emphysema, mucus production, and pulmonary fibrosis (178). There has also been work into the suggestive role of T-cells in lung disease. In 1995, Finkelstein *et al* identified that T-lymphocytes were the most common cells in the inflammatory response in COPD and emphysema patients (180). Further to this, in 1998, Saetta *et al* noted that the number of CD8+ T cells was directly related to the extent of airflow limitation (181).

1.4 Outline and aims of thesis

The aim of this thesis is to assess the contribution of complex genetic variation in immune system genes to Idiopathic Pulmonary Fibrosis (IPF) susceptibility. This study may identify individual variants, gene alleles or copy number variation that affect an individual's likelihood of developing IPF. In addition to providing new information about how variation in immunity contributes to risk of IPF, this study could drive precision medicine approaches to treatment through identification of subtypes of disease for which immune response dysregulation might be a key driver. This will be the largest study of the role of HLA region variation, and the first study of KIR gene region variation, in IPF susceptibility.

Chapter 2, entitled *"The imputation of the Human Leukocyte Antigen (HLA) region in four Idiopathic Pulmonary Fibrosis (IPF) datasets"* describes the method used to impute the HLA region in four IPF case-control datasets. The four datasets, genotype quality control and phasing methods which were used in the following chapters are described. The imputed and quality-controlled datasets were then analysed for association with IPF risk in chapters 3 and 4.

Chapter 3, entitled *"HLA-wide association analyses of Idiopathic Pulmonary Fibrosis susceptibility in European populations"* outlines an association study of HLA variation with IPF susceptibility. First, a HLA-wide association of study of 612 IPF cases and 3,366 controls was conducted and replication of signals was sought in a further 2,015 IPF cases and 5,193 controls. Secondly, to maximise the sample size used to discover signals, thereby increasing statistical power, a meta-analysis of all available datasets (1,905 IPF cases and 13,876 controls) was undertaken. The likely functional mechanism underlying novel associations were evaluated using gene expression data and phenome-wide association studies.

Chapter 4, entitled *"SNP-SNP interaction analyses of variants in the HLA region and the MUC5B risk allele in IPF susceptibility"* tests the hypothesis that the contribution of immune gene variation to IPF susceptibility may be dependent on *MUC5B* risk allele carrier status. 612 IPF cases and 3,366 controls were included in a HLA*MUC5B interaction association discovery analysis and replication was sought in 2,308 cases and 14,683 controls. Where interaction signals were identified, the effects of those signals were then analysed independently for association with IPF risk in *MUC5B* risk allele carriers and individuals without *MUC5B* risk alleles.

Chapter 5, entitled *"The imputation of variation within the Killer Immunoglobulin like Cell Receptor (KIR) region in four Idiopathic Pulmonary Fibrosis (IPF) datasets"* describes the

methods used to impute the KIR region in the four IPF datasets. The imputed and quality-controlled datasets were then analysed in chapter 6.

Chapter 6, entitled *"KIR-wide association analysis of Idiopathic Pulmonary Fibrosis susceptibility in four Idiopathic Pulmonary Fibrosis (IPF) datasets",* describes the association analyses of KIR gene copy number variation and haplotypes with IPF susceptibility. The chapter describes the association testing and meta-analysis of four IPF case control studies.

Chapter 7 evaluates the work undertaken in this thesis. The positives and negatives of the approaches and methodology that have been used in the thesis are described, and how these affect the conclusions drawn from the results. There is discussion around the possible implications of this work in the clinical setting and how this will affect future patients. Finally, potential future work for this area of research is outlined.

Chapter 2: The imputation of the Human Leukocyte Antigen (HLA) region in four Idiopathic Pulmonary Fibrosis (IPF) datasets

2.1 Introduction:

The human leukocyte antigen (HLA) region plays a major role in the initiation of immune responses to bacterial or viral infection. Since viral infection has been suggested to play a part in Idiopathic Pulmonary Fibrosis (IPF) development and progression (1-6) genetic variation in the HLA region could provide insight into the biological processes that underlie IPF. The genetic variation in the HLA region is complex and SNP imputation panels cannot adequately capture this variation. Bespoke HLA imputation panels have been developed which enable the imputation of classical HLA alleles and amino acid alleles across the HLA region. Using the T1DGC panel for HLA-specific imputation (see chapter 1, section 1.4.1) enables measurement of HLA alleles and amino acid alleles in *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DPA*, *HLA-DPB*, *HLA-DQA*, *HLA-DQB*, and *HLA-DRB* (7). However, the coverage of SNPs over the HLA region is limited when using directly genotyped variants alone (7). New genome-wide SNP imputation panels such as the Haplotype Reference Consortium (HRC) panel enable high confidence imputation of additional SNPs within the HLA region. Combining the SNPs imputed using the HRC panel with the SNPs, amino acids and gene alleles imputed using the HLA panel will yield a more comprehensive coverage of variation across the HLA region for association testing in Chapter 3 and Chapter 4.

This chapter describes the imputation strategy and the quality control of the HLA-imputed IPF datasets that will be analysed in chapters 3 and 4.

2.2 Summary of Idiopathic Pulmonary Fibrosis (IPF) datasets:

Four IPF case-control datasets were imputed to the HLA imputation panel; UK (8), UUS (9), Colorado (10) and Chicago (11). All four datasets were comprised of unrelated individuals of European ancestry (inferred by principal component analysis (see below) for all datasets completed for a previous study (9)) with IPF diagnosed using the American Thoracic Society and European Respiratory Society guidelines as described previously (12-14). Individuals overlapping between datasets had been identified in previous analyses and removed prior to this analysis (15). All studies had appropriate institutional board or ethics approval (8, 10, 11).

2.2.1 UK IPF dataset:

The UK IPF dataset was comprised of 612 IPF cases from centres around the UK and 3,366 controls from UK Biobank (table 2.1) (8). The controls were matched for age, sex and smoking

status (ever vs never smokers) to the cases (8). All the cases and 2,135 controls were genotyped using UK Biobank Affymetrix Axiom array and 1,231 controls and all the cases were genotyped using UK. BiLEVE Affymetrix Axiom array (8).

Table 2.1: Demographics of the UK IPF dataset.

Demographic	**Cases**	**Controls**
Number	612	3,366
Sex Male (%)	433 (71%)	2356 (70%)
Age (mean (sd))	70 (8.4)[a]	65 (5.5)

[a] Calculated on 602 cases.

2.2.2 Colorado dataset:

The Colorado dataset was comprised of 1,515 fibrotic idiopathic interstitial pneumonia (fIIP) cases and 4,683 controls that were matched on a similar identical by state (IBS) estimate (table 2.2) (10, 16). The controls were selected at random from a large database of anonymous genotyped individuals (genotyped at Centre d'Etude du Polymorphism Humain [CEPH]). All individuals were of self-reported non-Hispanic white ancestry (later confirmed by genotyping in (9)). All samples were genotyped using the Illumina Human 660W Quad BeadChip array. 77% of cases were classified as IPF with the remaining 23% of cases classified as non-specific interstitial pneumonia (NSIP, 6%), cryptogenic organizing pneumonia (COP, <1%), respiratory bronchiolitis-associated interstitial lung disease (RB-ILD, <1%), desquamative interstitial pneumonia (DIP, <1%) or were unclassified interstitial pneumonias (15%) (10). The disease phenotype percentages were provided however the exact disease sub-classification of each case sample was not available.

Table 2.2: Demographics of the Colorado dataset.

Demographic	**Cases**	**Controls**
Number	1,515	4,683
Sex Male (%)	1,118 (69%)	2,261 (48%)
Age (mean (sd))	65.5 (9.5)	NA[a]

[a] Age of controls was not known.

2.2.3 Chicago IPF dataset:

The Chicago dataset was comprised of 541 unrelated IPF cases and 542 unrelated controls matched by similarities in their first four principal components (11). Sex was not provided for 73 individuals in this dataset and therefore they were removed from future analyses (because sex was to be included as a covariate). In total, 500 cases and 510 controls remained for

analyses (table 2.3). All samples were genotyped using the Affymetrix Genome-Wide Human SNP 6.0 Array.

Table 2.3: Demographics of the Chicago IPF dataset.

Demographic	Cases	Controls
Number	500	510
Sex Male (%)	380 (76%)	241 (47%)
Age (mean)	68	63[a]

[a] Age was only available for 103 controls.

2.2.4 UK, USA and Spain (UUS) dataset

The UK, USA and Spain (UUS) dataset was comprised of 793 unrelated IPF cases and 10,000 unrelated controls (9). The IPF cases were selected from 9 centres from the UK, USA and Spain and the controls were selected from UK Biobank and matched for a similar age, sex and smoking status distribution (15) (table 2.4). The IPF cases and controls were of European ancestry confirmed by principal component analysis(1). The controls and the UK and USA IPF cases (754 of the 793 cases) were genotyped using the Affymetrix Axiom UK Biobank array and the Spanish IPF cases (39 of the 793 cases) were genotyped using the Axiom Spain Biobank array.

Table 2.4: Demographics of the UUS IPF dataset.

Demographic	Cases	Controls
Number	793	10,000
Sex Male (%)	584 (73.6%)	7210 (72.1%)
Age (mean (sd))	69 (9.1)	58 (7.8)

2.3 Methods: Imputation of the HLA region:

HLA alleles, amino acid alleles and SNPs (variation in the HLA region is described in Chapter 1: Introduction section 1.2.1 and an example of the relationship between the variants can be seen in figure 2.1) were imputed using the type one diabetes genomics consortium (T1DGC) and haplotype reference consortium (HRC) panels on chromosome 6 between the base pairs of 28,477,797 and 33,448,354 (on Genome Reference Consortium Human Build 37 (GRCh37) from the IPD/HLA database from EBI (17)). A diagram showing the architecture of the HLA region with a breakdown of the key genes can be seen in Chapter 1: Introduction, section 1.2.1.

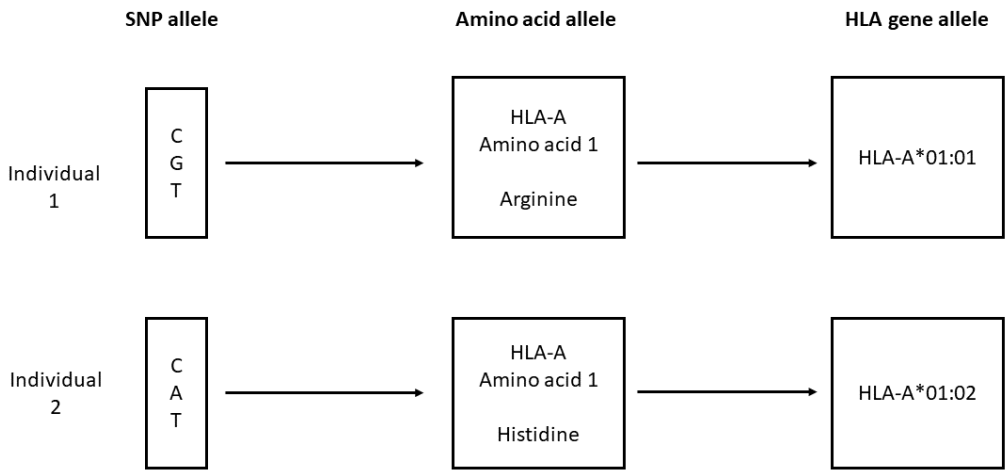


Figure 2.1: Simplified example of the relationship between SNP alleles, amino acid alleles and gene alleles in the HLA region.

2.3.1 Principal Component Analysis:

Principal component analysis (PCA) (described in Chapter 1: Introduction, section 1.1.8) was undertaken to confirm that all samples in each dataset were of European ancestry and to produce principal components (PCs) for use in the model as covariates. The smartpca program (from the EIGENSOFT package V7.2.1 (18)) was used to run principal component analysis on each of the four datasets using genome-wide genotyped SNPs from the datasets (UK=43,145 SNPs, Chicago=65,429 SNPs, Colorado=78,806 SNPs and UUS=56,786 SNPs).

2.3.2 Imputation to the Haplotype Reference Consortium panel:

Phasing and imputation to the Haplotype Reference Consortium (HRC) 1.1 panel (19) was completed on all IPF data sets for another study by Dr Allen (15) using the Michigan imputation server (https://imputationserver.sph.umich.edu/) (20).

2.3.3 Phasing:

As part of this thesis each of the four datasets were phased separately. Phasing is the process of assigning alleles to the maternal and paternal chromosomes and is an essential first step for genotype imputation. In these analyses, ShapeIt (v 2.837) was used to phase only SNPs that passed pre-imputation quality control steps as follows. SNPs were required to be present on (Axiom UK BiLEVE and Axiom UK Biobank, UK dataset only), required to have a call rate of more than 95%, a minor allele frequency of more than 1% and a Hardy Weinberg Equilibrium P > $1x10^{-6}$ to be phased and then imputed.

2.3.4 HLA allele and amino acid imputation:

Following phasing, SNPs, amino acids and HLA gene alleles in the HLA region were imputed with the T1DGC panel (7) using IMPUTE2 (V2.3.2) between the base pairs of 28,477,797 and 33,448,354 (Genome Reference Consortium Human Build 37 (GRCh37). The T1DGC panel incorporates the genetic data from 5,255 individuals and provides haplotypes for imputation of up to 424 HLA alleles and 1,276 amino acid alleles for *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB*, *HLA-DPA*, *HLA-DPB*, *HLA-DQA* and *HLA-DQB* (table 2.5).

Table 2.5: Number of alleles and amino acid alleles in each HLA gene imputed using the HLA specific imputation.

HLA Gene	**Number of imputed HLA alleles**	**Number of imputed amino acid alleles (across 399 sites)**
HLA-A	70	213
HLA-B	129	312
HLA-C	47	182
HLA-DRB	64	256
HLA-DPA	11	27
HLA-DPB	66	50
HLA-DQA	14	71
HLA-DQB	64	165

2.3.5 Merging of HRC panel imputed SNPs and HLA panel imputed SNPs:

SNPs imputed using the HRC panel and SNPs imputed using the HLA imputation panel were first matched by rsid or position and allele and duplicate SNPs were identified and removed from the HLA imputed data set (all HLA alleles, amino acid alleles and multi-allelic SNPs were retained).

2.3.6 Evaluating the use of HRC-imputed variants to improve imputation of HLA gene, amino acid and SNP alleles

The HLA imputation (for the HLA gene alleles and amino acid alleles) pipeline included only SNPs that were phased and had been directly genotyped. Inclusion of robustly imputed SNPs from the HRC imputation as input to the HLA imputation pipeline could improve overall imputation quality for the region. To test this, SNPs from the HRC imputation that had an imputation quality >0.98 were phased and combined with directly genotyped SNPs. The HLA imputation qualities of the SNPs, HLA gene alleles and amino acids with and without HRC-imputed variants as input were then compared for the UK dataset (using T-test).

2.3.7 Post-imputation quality control of HLA and HRC imputed variants:

Rare variants (minor allele frequency [MAF] < 0.01) and variants with a poor imputation quality (imputation quality < 0.4 from either panel) were removed from the datasets. HLA gene allele and amino acid frequencies were calculated per allele or amino acid (i.e. the number of chromosomes carrying HLA allele/amino acid divided by total number of chromosomes).

2.4 Results: Principal Component Analysis:

Genome-wide genotyped SNPs from all four datasets were used to calculate principal components which were plotted along with HapMap ancestral principal components. Individuals with non-European ancestry had already been removed from the datasets (8, 9) but this principal component analysis was used to derive the principal components to be used in the model and re-confirmed that each dataset was comprised of European individuals (supplementary figures 1-4).

2.5 Results: HLA region imputation pipeline:

The bespoke HLA imputation enables the imputation of classical HLA gene alleles and amino acid alleles across the HLA region. The classical HLA genes and amino acids were imputed with a binary code with each allele (gene or amino acid) at a given locus coded separately. An example of a HLA-A allele and amino acid position for one individual can be seen in table 2.6. This example shows that the individual has 2 copies of HLA-A*01:01 (and no copies of HLA-A*01:02 and HLA-A*01:03) and 2 copies of an Isoleucine at HLA-A amino acid -22 (and no copies of a deletion or Valine at amino acid position 22).

Table 2.6: Example of an imputed HLA-A allele and amino acid position for a single individual from the UK IPF dataset, imputed using the T1DGC HLA panel (7) on IMPUTE 2 (v2.3.2) (V=valine, x=deletion, I=isoleucine, P = present, A = absent). AlleleA and AlleleB were alternative and reference alleles. HLA-A has 78 amino acid positions, this table shows only a single amino acid position to provide an example.

RSID	Position	AlleleA	AlleleB	genotype
HLA_A_0101	29911991	P	A	1/1
HLA_A_0102	29911991	P	A	0/0
HLA_A_0103	29911991	P	A	0/0
AA_A_-22_V	29910338	P	A	1/1
AA_A_-22_x	29910338	P	A	0/0
AA_A_-22_I	29910338	P	A	0/0

The T1DGC panel is comprised of 12,306 SNPs (along with the HLA gene alleles and amino acid alleles) and the HRC panel (19) enables the imputation of 83,866 SNPs. There were some SNPs (411) that were present in both the HRC panel and the bespoke HLA imputation panel. In this case, when the SNPs were merged they were excluded from the bespoke HLA panel (an example of how these variants were split between the two reference panels can be seen in table 2.7).

Table 2.7: Number and type of variants imputed across the HLA region in the UK IPF dataset, split by panel.

Variant Type:	**HRC Panel**	**T1DGC Panel**
SNPs	77,762	12,306
HLA alleles	0	424
Amino acid alleles	0	1,276

2.5.1 Evaluating the use of HRC-imputed variants to improve imputation of HLA gene alleles, amino acid alleles and SNP alleles

To test whether inclusion of well-measured HRC-imputed SNPs improved imputation of HLA variants, 6,909 directly genotyped and 45,937 well-imputed (imputation quality >0.98) SNPs were used to impute HLA SNPs, gene and amino acid alleles in the UK dataset.

When comparing the use of directly genotyped SNPs only vs additional inclusion of HRC imputed SNPs there were some observable differences (87% of variants had imputation quality > 0.98 with only directly genotyped SNPs vs 91% when including HRC imputed SNPs) (figure 2.2). However, there was no significant difference between the means of the imputation qualities (directly genotyped vs imputed=0.995 and 0.996 respectively, P-value=0.2). Similarly, when comparing the imputation qualities of the HLA alleles and amino acid alleles together imputed using directly genotyped vs HRC imputed SNPs, there was no significant difference (imputation quality means of directly genotyped vs imputed=0.866 and 0.867 respectively, P-value=0.93). There was also no significant difference between the imputation qualities when looking at only amino acid alleles (imputation quality means of directly genotyped vs imputed=0.90 and 0.90 respectively, P=0.37) and only HLA gene alleles (imputation quality means of directly genotyped vs imputed=0.77 and 0.799 respectively, P=0.2) (figure 2.2). Because there was no significant difference in imputation qualities, the HLA

gene alleles, amino acid changes and SNPs in the HLA region in the four IPF datasets were imputed using directly genotyped SNPs only.

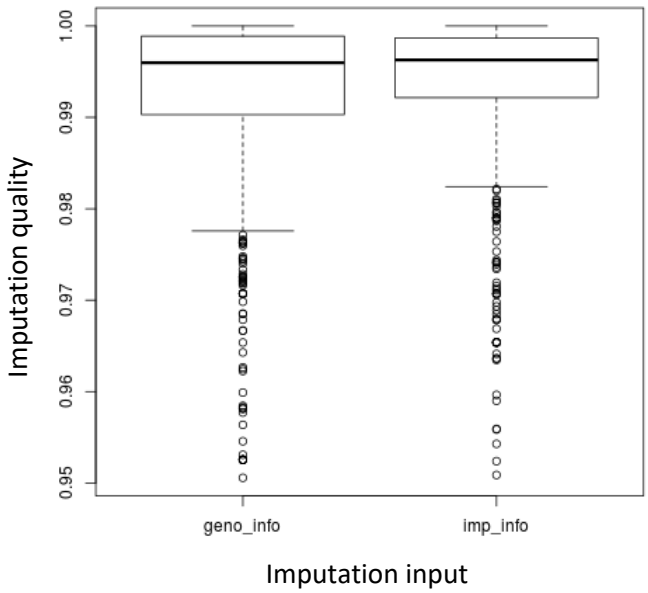


Figure 2.2: Comparison of imputation qualities of all variants imputed to the HLA imputation panel using either directly genotyped SNPs as the input (geno_info) or well imputed SNPs (imputation quality > 0.98) from the HRC imputation panel as the input (imp_info).

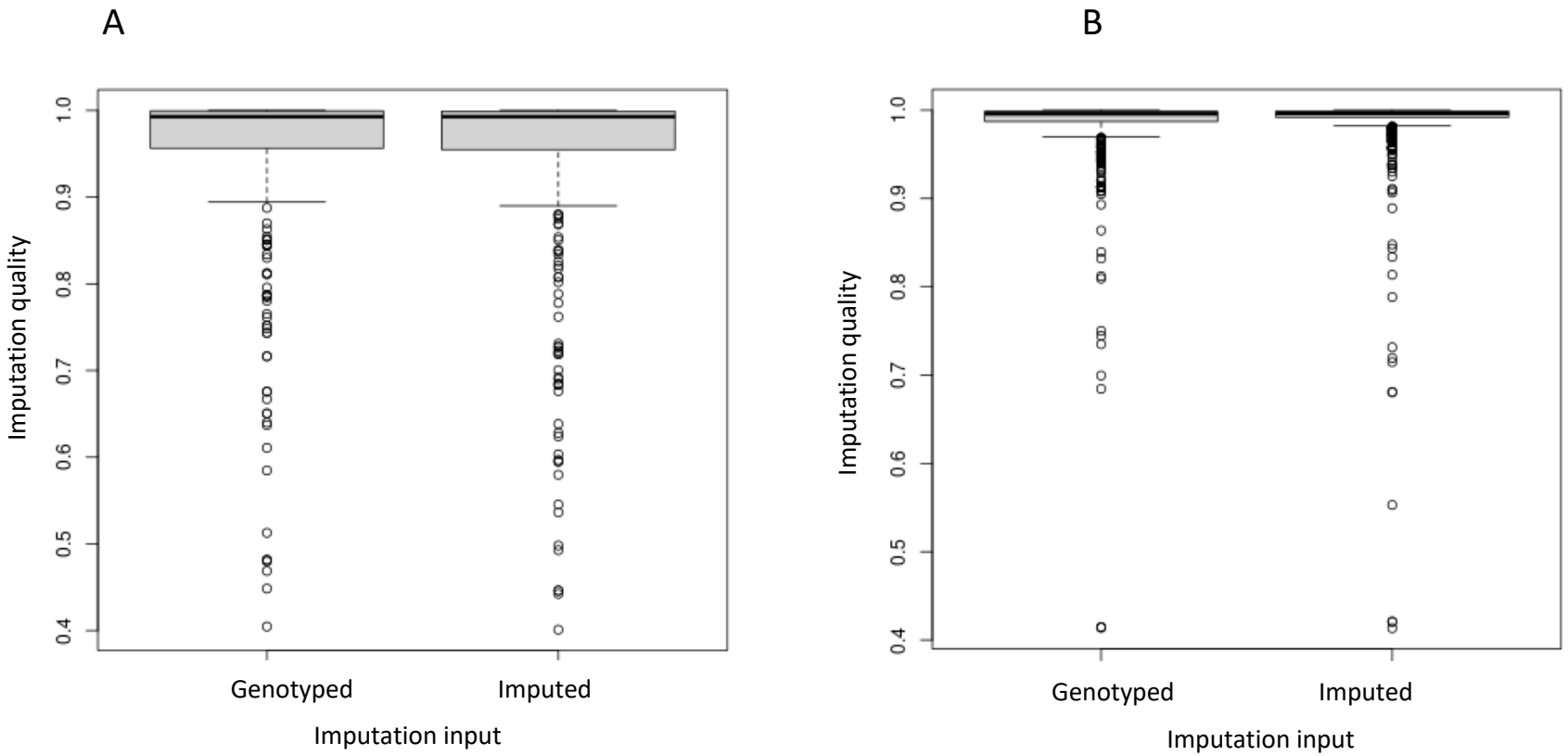


Figure 2.3: Comparison of imputation qualities of HLA alleles (A) and amino acid alleles (B) imputed to the HLA imputation panel using either directly genotyped SNPs as the input (genotyped) or well imputed SNPs (imputation quality > 0.98) from the HRC imputation panel as the input (imputed).

2.6 Results: HLA imputation of IPF datasets:

HLA gene alleles, amino acid alleles and SNPs were imputed in 3,420 cases and 18,559 controls of European ancestry across four datasets (supplementary figures 1-4) using directly genotyped SNPs. Between 86,000-88,000 variants were imputed in each dataset in total. Around 60% of the imputed variants were removed in the quality control step (frequency and imputation quality filters) which left between 34,000-37,000 variants for analyses in the four datasets. A vast majority (99%) of the variants removed were SNPs and the remaining 1% was made up of 596 (UK), 533 (Colorado), 529 (Chicago) and 544 (UUS) rare HLA gene alleles and amino acid alleles (table 2.9). The UK and UUS datasets had considerably more SNPs genotyped across the HLA region (these were both genotyped using the Affymetrix Axiom UK Biobank array) compared to the Chicago and Colorado datasets (genotyped using the Affymetrix Genome-Wide Human SNP 6.0 Array and Illumina Human 660W Quad BeadChip, respectively) which could explain the higher number of variants removed for poor imputation quality (table 2.8) and the differences in imputation quality distributions (figure 2.6). Of the variants that passed the quality control filters, there was a mean imputation quality of 0.98 for the UK, UUS and Colorado datasets and 0.97 for the Chicago dataset (figure 2.3). The distributions of allele frequencies were similar across all four datasets (figure 2.7).

Table 2.8: Demographics and results of the HLA imputation across the UK, Colorado, Chicago and UUS datasets.

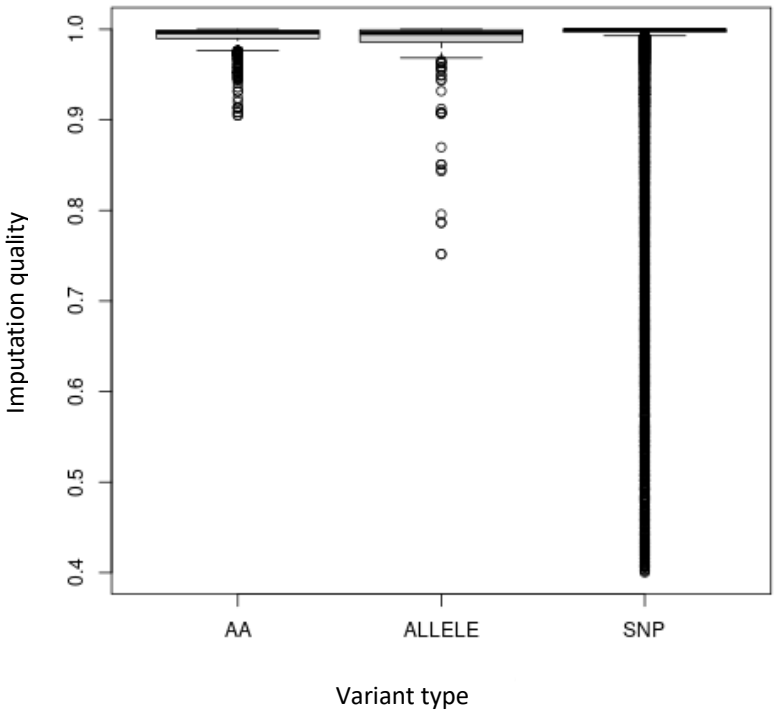
Dataset	**Directly genotyped SNPs used for imputation**	**Total variants imputed**	**Variant exclusions**		**Total quality-controlled variants**	**Mean allele frequency**	**Mean imputation quality**
			Allele frequency < 0.01	Imputation quality < 0.4			
UK	6,909	87,355	50,224	388	36,743	0.18	0.98
Colorado	1,495	86,646	49,573	2,168	34,905	0.17	0.98
Chicago	1,012	86,647	49,376	2,248	35,023	0.17	0.97
UUS	6,256	87,441	49,925	551	36,965	0.17	0.98

Table 2.9: Number of HLA gene alleles, amino acid alleles and SNPs at an allele frequency less than 1% (and removed from the quality-controlled datasets).

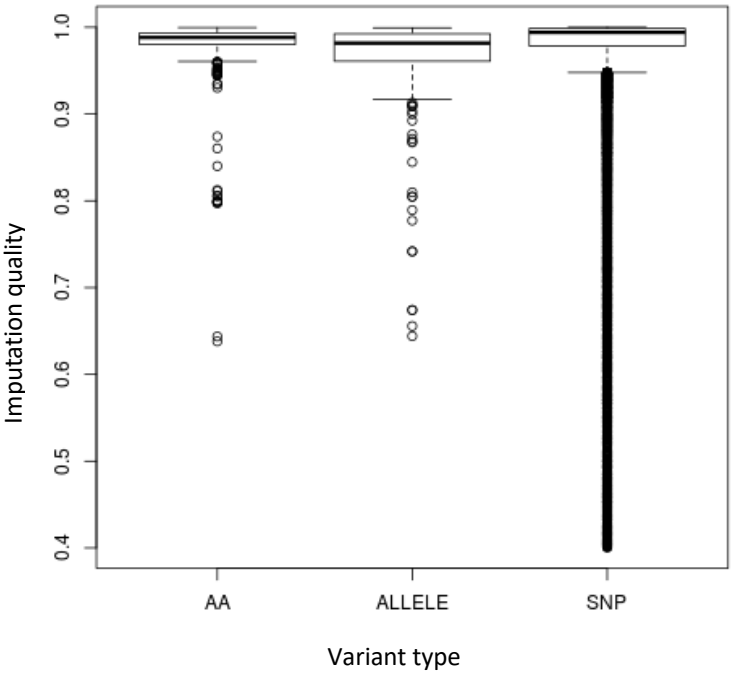
Dataset	**HLA gene alleles**	**Amino acid alleles**	**SNPs**
UK	256	340	49,628
Colorado	245	288	49,040
Chicago	243	286	48,847
UUS	252	292	49,381

Across the three types of variants there were different ranges of allele frequencies (can be seen in figure 2.4). The mean allele frequency (across the four datasets) was 9% for the HLA gene alleles, 26% for the amino acid alleles and 17% for SNPs (figure 2.4). The ranges and mean allele frequencies across the variant types were similar across all four datasets. The UK and UUS datasets appear to cluster at a higher quality compared to the Colorado and Chicago datasets, this could be because the two groups are on different genotyping panels (UK/UUS used UK Biobank Affymetrix array, UK BiLEVE Affymetrix array and Axiom Spain Biobank array. Chicago used Affymetric Genome-Wide Human SNP 6.0 Array and Colorado used Illumina Human 660W quad BeadChip array). A comparison of allele frequencies between the HLA imputation panel and the imputed HLA variation showed strong correlation across all four datasets (figure 2.5). Association testing in subsequent chapters would include multiple independent datasets enabling identification and exclusion of spurious signals driven by imputation errors in individual datasets. Therefore, no variant exclusions were made based of these comparisons.

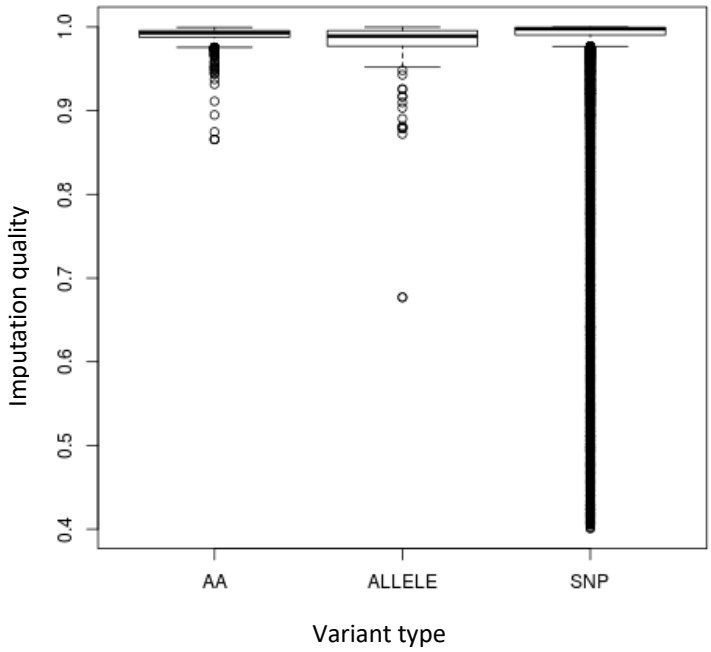
UK dataset



Chicago dataset



Colorado dataset



UUS dataset

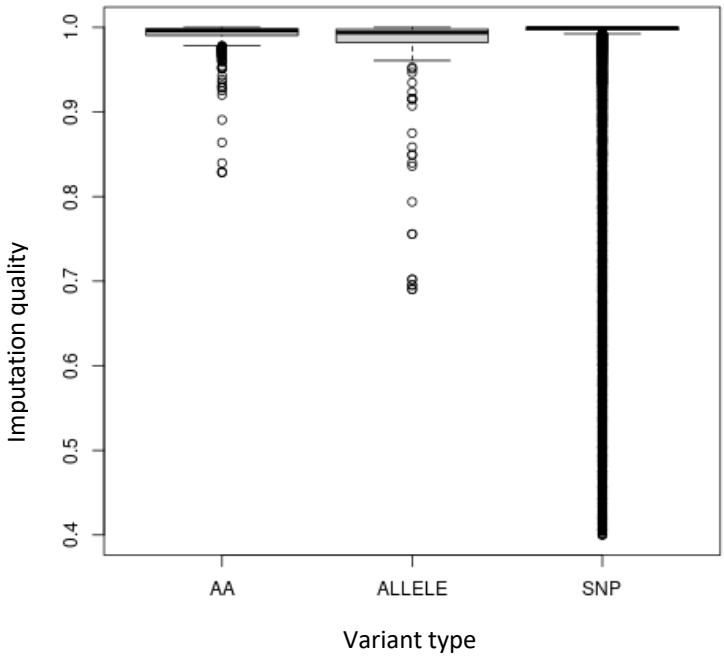


Figure 2.4: Imputation quality of variants in IPF susceptibility in the UUS IPF dataset, split by variant type (AA= HLA amino acid alleles, ALLELES=HLA alleles).

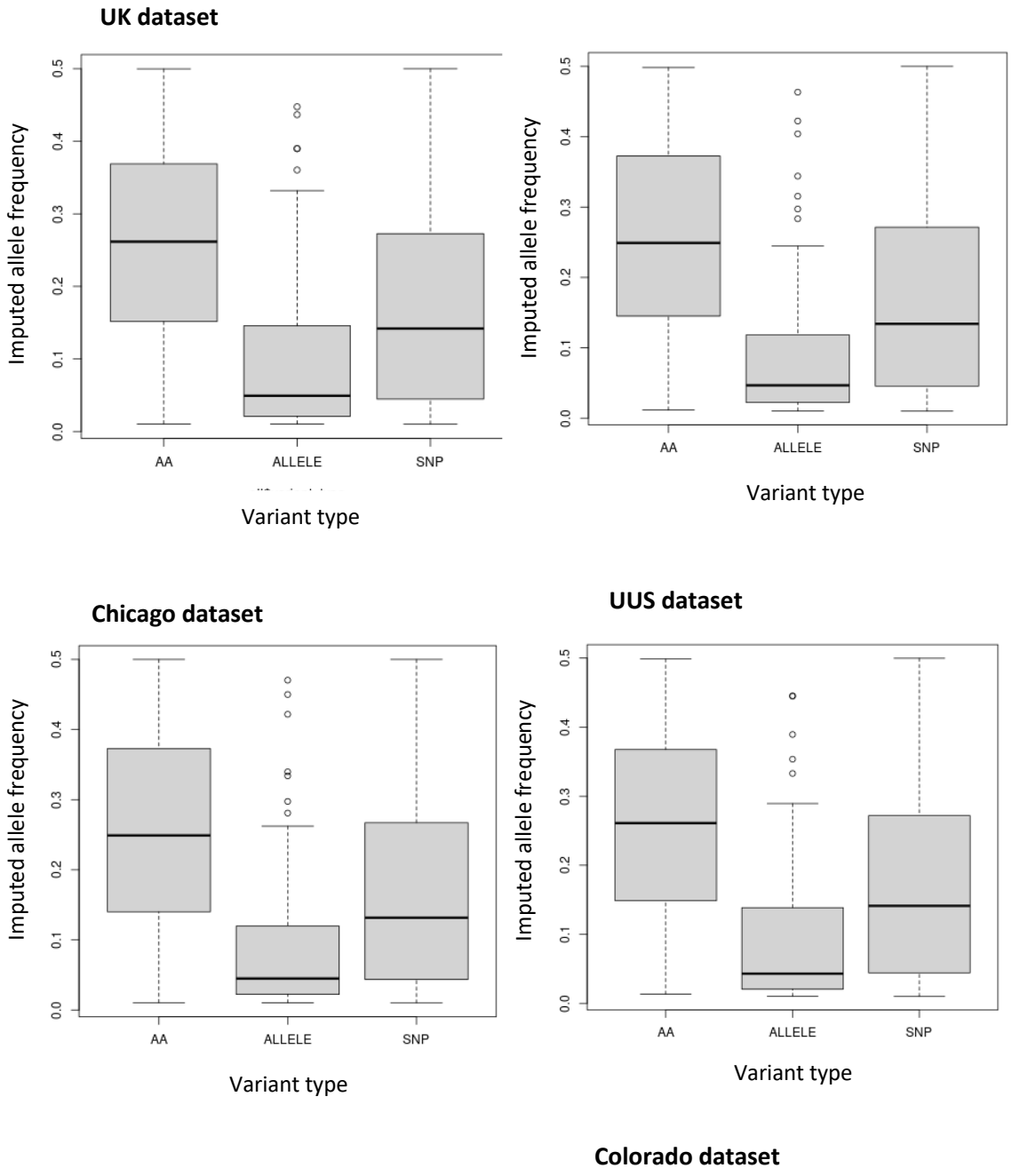
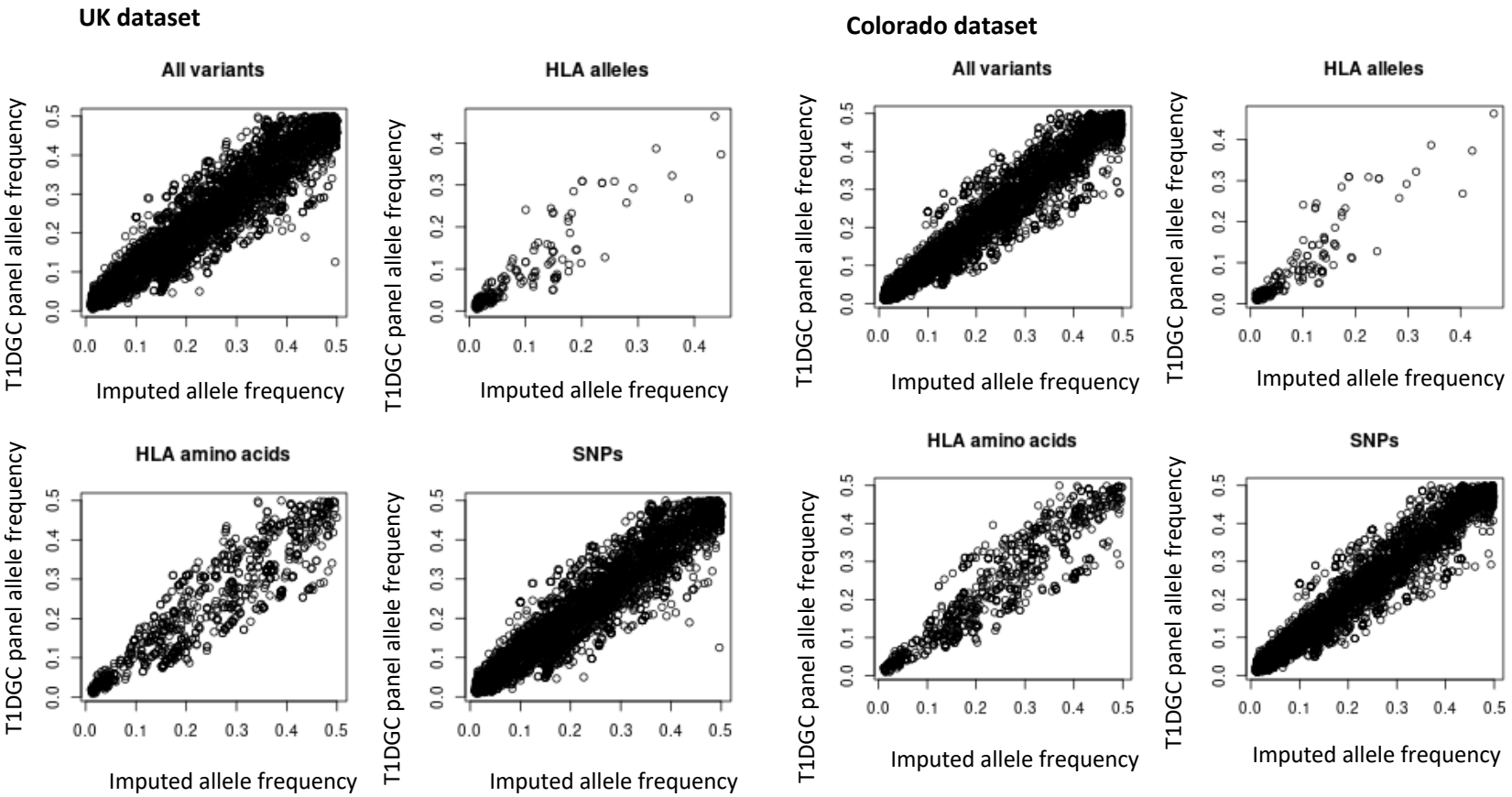


Figure 2.5: Boxplot of the allele frequencies of variants in the UUS IPF dataset, split by variant type (AA= HLA amino acid alleles, ALLELE= HLA alleles).


UK dataset
Colorado dataset
All variants
HLA alleles
HLA amino acids
SNPs
T1DGC panel allele frequency
Imputed allele frequency
0.0
0.1
0.2
0.3
0.4
0.5

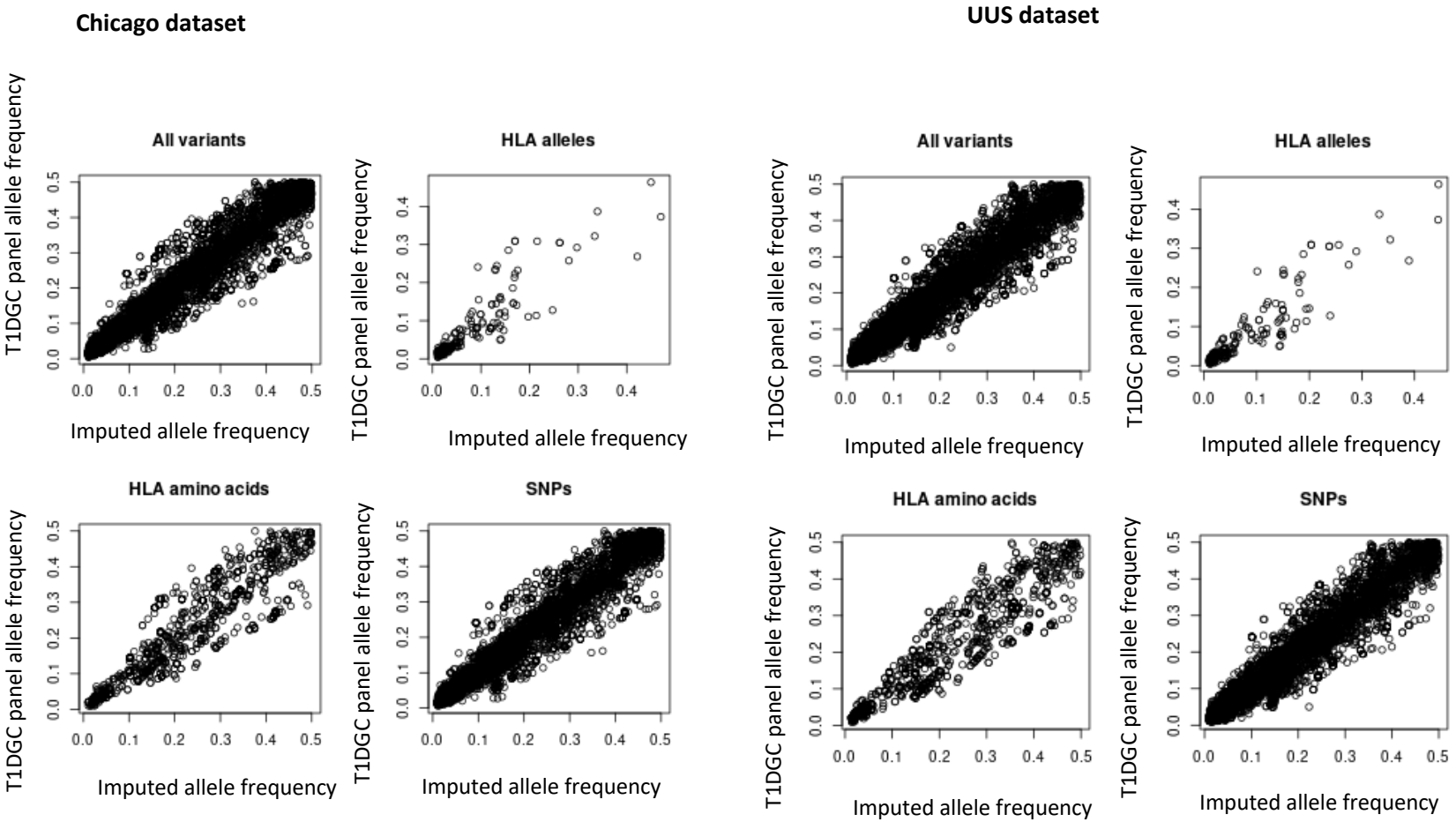


Figure 2.6: Comparison of the minor allele frequencies of variants in the T1DGC panel (Y axis) and the allele frequencies from the imputed UUS dataset (X axis).

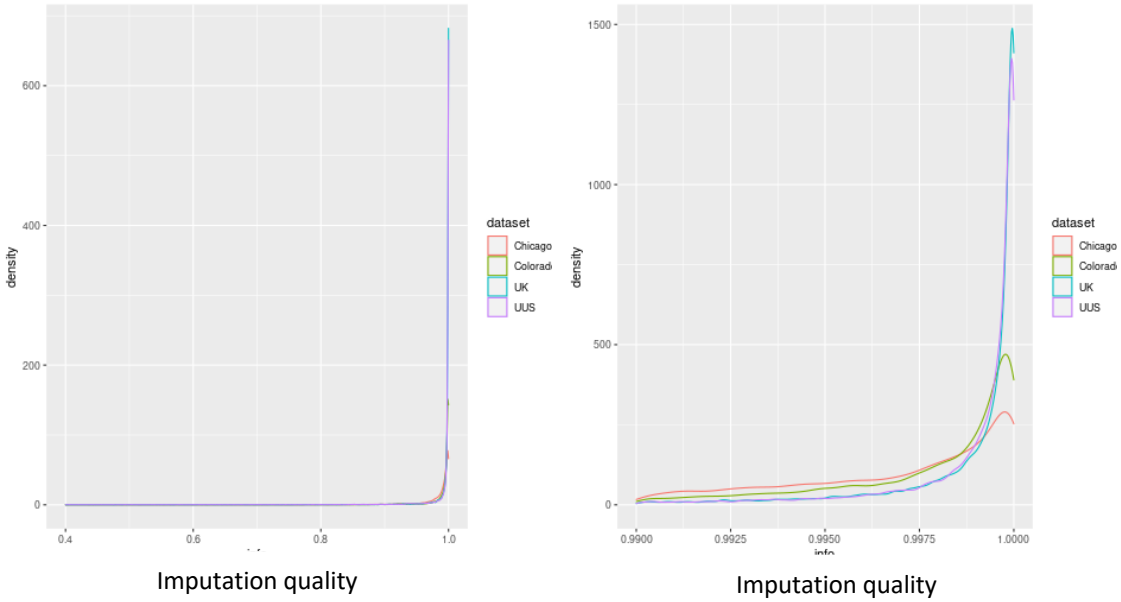


Figure 2.7: Density plots of imputation quality for all four datasets across the HLA region (left=all qualities, right= qualities >0.9).

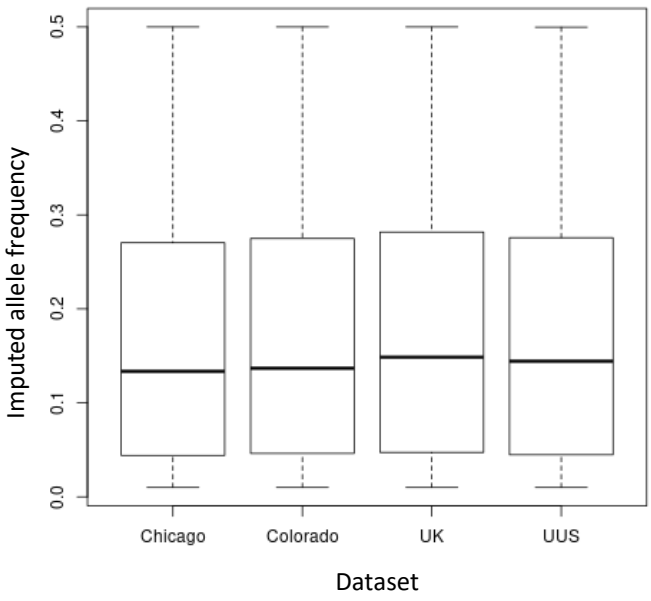


Figure 2.8: Boxplot of imputed allele frequency of all four datasets across the HLA region.

Genes across the HLA region are said to be highly polymorphic, as this may be advantageous to the response to bacterial and viral infections. The data from the UK IPF dataset confirms a lack of homozygotes, table 2.10 shows that homozygotes were rare across HLA-A*01, *02, *03 and *11.

Table 2.10: Table of genotypes of the imputed HLA-A genes in the UK IPF dataset (P=presence and A=absence of gene allele in the individual).

Genotype	HLA-A*01:01	HLA-A*01:02	HLA-A*01:03	HLA-A*02:01	HLA-A*02:02	HLA-A*02:03	HLA-A*02:05	HLA-A*02*06	HLA-A*02:07	HLA-A*02:11	HLA-A*02:16	HLA-A*02:17	HLA-A*03:01	HLA-A*03:02	HLA-A*11:01	HLA-A*11:02	HLA-A*11:03	HLA-A*11:05
A/A	2623	3940	3976	2050	3973	3966	3917	3963	3978	3972	3978	3973	2903	3934	3523	3978	3978	3978
P/A	1157	0	0	1572	2	0	54	12	0	0	0	2	943	0	441	0	0	0
P/P	157	0	0	311	0	0	0	0	0	0	0	0	87	0	14	0	0	0

2.7 Discussion:

This chapter outlined the imputation approach that was used to infer SNPs, amino acid alleles and classical HLA gene alleles in the IPF datasets to be used in chapters three and four. Directly genotyped SNPs were used to impute the HLA region in the UK, UUS, Chicago and Colorado datasets with the Haplotype Reference Consortium SNP panel and the T1DGC HLA panel. This approach utilised the most up to date SNP panel and a HLA region specific panel, which provided a high coverage of SNPs across the HLA region and allowed the incorporation of classical HLA alleles and amino acid alleles into these data. This allows the fine mapping of genetic associations to a particular HLA isoform as described in the following chapters. This chapter described an investigation to determine if the inclusion of HRC-imputed SNPs improved the imputation of the HLA gene alleles and amino acid alleles in these four IPF datasets. It was found that there was no significant improvement of the mean imputation qualities across the HLA gene alleles and amino acid alleles. Therefore, I decided to only use directly genotyped SNPs for the bespoke HLA imputation. This could be because the genotyped SNPs provided sufficient information to effectively impute the HLA gene alleles and amino acid alleles. Also, including the imputed SNPs could be introducing unnecessary noise, meaning they are not providing any more useful information for the imputation.

The main three HLA class I molecules (HLA-A, -B and –C) and the main HLA class II molecules (HLA-DRB, -DQA1, -DQB, -DPA1 and –DPB1) were all well represented using this method with an average imputation quality of 0.97 for HLA gene alleles and 0.99 for amino acid alleles (figure 2.3). These molecules are highly polymorphic (shown in our data, table 2.10) and typically include clinically relevant alleles and amino acid alleles for example *HLA-B*27* (ankylosing spondylitis (21)) and *HLA-DQA1*05:01* and HLA-DQB1 amino acid 57 (type one diabetes (22, 23)). Non-classical HLA genes (HLA-E, -F and G from class I and HLA-DRA, -DQA2, DPA2 and DPB2 from class II) were not represented using this imputation method. Non-classical HLA genes are not thought to be as polymorphic as the classical HLA genes (see chapter 1, section 1.3.1) and therefore it has been suggested that any variation found across these genes may be appropriately captured using SNP imputation alone (7) however these genes have been implicated in autoimmune disorders previously and therefore may be relevant to disease processes. Alleles and amino acid polymorphisms that may be of importance to future analyses of IPF were well captured, for example the *HLA-DQB1*06:02* allele previously associated with IPF risk (16) and the HLA-DQB1*06:02 amino acid 57 polymorphism previously associated with lung function (24) have average imputation qualities of 0.99.

A limitation of the quality control approach is the allele frequency exclusion threshold of 1% which removed half the variants in the analysis, although 99% of these were SNPs. It could be suggested that using an allele frequency cut off is not appropriate for the HLA gene alleles and the amino acid alleles (because they were multi-allelic). Over 240 gene alleles imputed were removed due to low frequency (table 2.9, the same gene alleles were removed in both the cases and controls), the mean frequency of those excluded was 0.1% (and only 0.008% in the IPF cases) therefore there would not be sufficient power to detect associations in these alleles in subsequent analyses. The allele frequency cut off was used across all the variant types to reduce the chance of spurious results since rare alleles are often not well imputed (182) or genotyped and there was not sufficient power to effectively test the associations with the rare variants. The allele frequency threshold of 0.01 could be reduced or removed to include more variants in the analyses. However, this would increase the type one error rate and there would not be sufficient power to detect any associations in these rare variants in the following analyses.

Imputation quality thresholds is used to remove poorly imputed variants that could cause spurious findings in the analysis. In this analysis, a threshold of 0.3 was used, thresholds of 0.3 and 0.5 have been commonly used in analyses of common variation (2, 5, 183) and higher thresholds are used for rarer variants. For example, in the 2017 paper by Allen et al (2) in which an imputation quality threshold of 0.5 was used for common variants (frequency ≥ 1%) and a threshold of 0.8 was used for rare variants (frequency < 1%). The imputation quality metric could have been applied differently, for example a higher imputation quality threshold for example to 0.5 as would provide higher confidence in the imputed SNP alleles, however this would have reduced the number of variants available for analysis.

This chapter outlined the method for imputing the HLA region across all four IPF datasets. The datasets with quality-controlled HLA gene alleles, amino acid alleles and SNPs will be utilised for association testing in the following two chapters.

Chapter 3: HLA-wide association analyses of Idiopathic Pulmonary Fibrosis susceptibility in European populations

3.1 Introduction:

During this chapter I will study the role of genetic variation in immune system genes; in particular, genes in the Human Leukocyte Antigen (HLA) region in susceptibility to Idiopathic Pulmonary Fibrosis (IPF). The HLA genes encode molecules involved in the immune response to bacterial and viral infections. Infection is hypothesised to be a trigger for development of respiratory disease (7) and is known to be a major cause of exacerbation events (9-11). Infection of viruses such as herpes viruses (7, 8, 60-63) have previously been linked to IPF and could play a role in its pathogenesis. However, there is conflicting evidence on the matter with some studies identifying no difference of viral load between healthy controls and IPF cases (184, 185). The HLA region has also been linked to response to herpes infections, for example HLA-DRB1 and HLA-DQB1 have been implicated in a GWAS of Epstein-Barr virus infection in Hispanic populations (186) and HLA-B has been implicated in susceptibility to Shingles infection (187). The HLA region is highly polymorphic, it has been shown to harbour around nine SNPs per kilobase compared to around three which is average across the rest of the genome (16).The HLA region therefore requires specific imputation techniques to fine map the variation to a specific functional HLA allele. Genetic studies have provided evidence of a link between the HLA region and respiratory traits (188) and a HLA gene allele has been implicated in susceptibility to fibrotic idiopathic interstitial pneumonia (fIIP) (5). fIIP as a group encompasses several different interstitial pneumonias including IPF, rheumatoid arthritis associated interstitial lung disease (RA-ILD), desquamative interstitial pneumonia, cryptogenic organising pneumonia, respiratory bronchiolitis-associated interstitial lung disease. Although they all share a fibrotic phenotype, these diseases have radiological and histopathological differences in their presentation which could suggest there are some differences in their underlying processes.

Previous analyses for IPF susceptibility have utilised either only standard SNP imputation methods (2-4) or specific HLA imputation methods to fine map signals (5). By using both these panels I will infer a high coverage of SNPs in the HLA region and also infer HLA alleles and amino acid changes to test for association with IPF susceptibility.

The aims of this chapter were to; utilise imputation of SNPs, amino acid changes and HLA gene alleles to identify new IPF susceptibility signals in the HLA region and replicate the *HLA-DQB1*06:02* signal previously identified in the Colorado dataset (5).

This chapter first describes a HLA-wide discovery association analysis in 612 IPF cases and 3,366 controls (UK dataset) with replication in 2,015 IPF and fIIP cases and 5,193 controls (Colorado and Chicago datasets) (see chapter 2, sections 2.2.1, 2.2.2 and 2.2.3). To maximise the power to identify novel signals in these datasets, a 3-way meta-analysis was then undertaken in 2,769 IPF and fIIP cases and 8,591 controls (UK, Chicago and Colorado datasets – UUS was not available at the time). When evaluating the results of the 3-way meta-analysis, there was a notable difference between the Colorado datasets and the other IPF datasets. This was perhaps due to subtle phenotype differences (IPF in UK and Chicago datasets and fIIP in Colorado datasets). To address this, a 3-way meta-analysis was undertaken in 1,905 IPF cases and 13,876 controls (UK, UUS and Chicago datasets).

3.2 Methods:

3.2.1 Datasets:

The phasing, imputation and quality control of the UK, Colorado, UUS and Chicago datasets is outlined in chapter 2. Briefly, the UK dataset is comprised of 612 IPF cases and 3,366 controls and has 36,743 well imputed (allele frequency >1%, imputation quality >0.4) variants for analysis. The Colorado dataset is comprised of 1,515 fibrotic idiopathic interstitial pneumonia cases and 4,683 controls with 34,905 well imputed (allele frequency >1%, imputation quality >0.4) variants for analysis. Finally, the Chicago dataset has 500 IPF cases and 510 controls with 35,023 well imputed (allele frequency >1%, imputation quality >0.4) variants for analysis. The UUS dataset is comprised of 793 IPF cases and 10,000 controls with 35,809 well imputed variants for analysis.

3.2.2 Testing the association between variants in the HLA region and susceptibility to IPF:

A HLA region-wide association analysis of IPF susceptibility was conducted using SNPtest (v2.5.2) in the UK, Colorado and Chicago datasets assuming an additive model. The HLA alleles and amino acid changes were also modelled assuming an additive model (e.g. present vs absent and not as a multi-allelic variant). Ten principal components (to adjust for fine-scale population structure) and sex were included as covariates. Manhattan plots were created for

each analyses using qqman covering the whole of the HLA region (189) and independent variants were visualised on region plots using Locuszoom (190) in Python.

3.2.3 Defining significance thresholds and statistical significance:

A Bonferroni corrected significance threshold was determined for the HLA-wide association analysis using the number of independent SNPs, amino acid changes and alleles in this region. LD Prune from Plink v1.9 was utilised to identify the number of independent signals in the region. Windows of 50 variants were analysed and a variant was removed from a pair if the r^2 was greater than 0.2. For the replication of signals, a Bonferroni corrected threshold was calculated using the number of independent variants to be replicated (i.e. 0.05/number of variants in replication).

Signals that passed these predefined thresholds were selected as significant. If none passed Bonferroni corrected threshold, a threshold of $P<5x10^{-3}$ was used to identify suggestive signals. Those that passed these significance thresholds were identified as independent from one another by excluding those with an r^2 of more than 0.2 with the lead variant.

3.2.4 Signal Characterisation:

To refine the association signals to include only variants that were the most likely to be causal, 95% credible sets were calculated (i.e. a set in which there is 95% confidence the causal variant is in – under the assumption there is one causal variant and it is measured). Posterior probabilities were calculated from approximate bayes factors using the Wakefield formula (191) (Wakefield prior set at 0.4) for variants within 1Mb and that were in linkage disequilibrium with the lead variant ($r^2 > 0.2$). The 95% credible set was produced by adding variants to the set until the sum of their posterior probabilities was equal to or greater than 0.95.

Lead variants of signals (and those in the credible set) identified in the analysis and variants in high LD ($r^2>0.8$) (SNPs or HLA alleles) were investigated in using Phenoscanner (192) to identify associations with respiratory, autoimmune, inflammatory or immunity phenotypes. Signals identified in the association analyses were said to be associated with a phenotype if it met a $P<5x10^{-8}$ threshold in or Phenoscanner. GTEx consortium (tissue sample sizes from 4-706) was used to identify if the signals were associated with the expression of any genes in 49 tissues around the body (193). The colocalisation of association and eQTL signals in the lung were tested by studying the linkage disequilibrium between the lead SNPs of the two signals.

3.2.5 Meta-analysis of IPF susceptibility study design:

SNPs, HLA alleles and amino acid changes (hereafter to be known as variants) in each dataset were tested for association with IPF susceptibility and those that passed quality control were included in the meta-analysis of IPF susceptibility. A fixed-effects weighted meta-analysis was performed on these data to provide a weighted *P*-value, beta and standard error for each variant. Variants were required to be present in at least two studies to be included in the analysis. For significance, variants were also required to be in the same direction of effect and have $P < 0.05$ in all studies.

3.2.6 HLA amino acid joint regression data and study design:

As well as the logistic regression analysis (described in 3.2.2, a joint regression analysis was undertaken to test the effects of each amino acid at multi-allelic sites (e.g. in HLA amino acid alleles). 1,276 amino acid changes at 399 sites were imputed in all four IPF datasets (see chapter 2) across all HLA genes (table 3.1). The number of amino acid alleles imputed at one position varies from 1-31, a typical amino acid site can be seen in table 3.2.

Table 3.1: Number of amino acids sites and alleles in each HLA gene.

Gene	**Number of amino acid alleles**	**Number of amino acid sites**	**Number of amino acid sites with ≥2 alleles**	**Number of amino acid sites with ≥3 alleles**
HLA-A	213	78	70	11
HLA-B	312	77	61	14
HLA-C	182	64	41	8
HLA-DPA	27	15	4	0
HLA-DPB	50	24	12	9
HLA-DQA	71	36	20	8
HLA-DQB	165	54	42	38
HLA-DRB	256	51	39	28

Joint logistic regression model was used to test for the effect of each amino acid at the position simultaneously, for example if there were three amino acids at one position (AA1, AA2 and AA3) they would be tested for association with the phenotype (IPF susceptibility) as follows:

$$Phenotype \sim AA1 + AA2 + AA3 + Sex + PCs$$

Only single amino acid alleles were taken across all the sites (for example from table 3.1, F, T, S and Y would be used but FT, FS and FY were excluded as this was appropriately captured when

the separate amino acid alleles were tested). Rare amino acid alleles (frequency < 1%) were removed from each IPF dataset and only sites with two or more alleles (after rare alleles were removed) were retained for the joint regression analysis.

In order to reduce collinearity in the model, two sets of amino acids were tested, all amino acids after quality control (set 1) and a smaller set with the most frequent amino acid allele removed at each loci (loci with 3 or more alleles) (set 2). The frequency of each amino acid was calculated using Plink and the amino acid with the highest frequency at each position was removed. The most common amino acid was only removed at positions with three or more amino acids. Collinearity is the correlation between variables (that they express a linear relationship) in the model, collinearity in the model can cause inaccuracies in the estimates, confidence intervals and association testing (194).

The joint regression analyses (using the above model) was run across both amino acid sets in all four IPF datasets using R.

The Bonferroni corrected significance threshold was calculated using the number of amino acid sites in each analysis in each dataset. The results from the joint regression from each individual dataset were combined using a fixed-effects weighted meta-analysis on R. Amino acids were required to pass the Bonferroni corrected significance threshold (as described above) and be nominally significant (P<0.05) in at least two studies.

3.3 Results: Dataset quality control.

3.3.1 Defining significance thresholds:

Of a total 79,485 variants analysed, 17,338 variants were identified as independent using LD prune from Plink v1.9. A Bonferroni corrected threshold of $P<2.8x10^{-6}$ was used in subsequent analyses as this has been corrected for the number of independent signals across the region.

3.4 Results: HLA-wide association analyses of IPF susceptibility: discovery in UK IPF dataset and replication in Chicago and Colorado datasets.

A HLA-wide discovery association analysis of IPF susceptibility was conducted in the UK IPF dataset with significant signals analysed in the Colorado and Chicago datasets for replication.

3.4.1 HLA-wide association of IPF susceptibility: discovery in the UK IPF dataset:

The UK IPF dataset was comprised of 612 IPF cases and 3,366 controls of European ancestry. 36,743 well imputed variants across the HLA region (6:28,477,797 – 6:33,448,354) were tested

for an association with IPF susceptibility. There was no inflation of the test statistic (figure 3.1, λ=1.02). No variants passed the significance threshold of $P<2.8x10^{-6}$ for association with IPF susceptibility (figure 3.2). Twelve independent variants passed the suggestive significance threshold of $P<5x10^{-3}$ (Green variants in Table 3.2 show all variants that passed the suggestive significance threshold). Seven signals identified in this analysis were found in HLA Class I and the rest reside in Class III. Table 3.3 shows that four variants in this analysis were tagging at least one class I classical HLA gene allele (at an r^2 of at least 0.2) but the alleles were not directly associated with IPF susceptibility in this analysis.

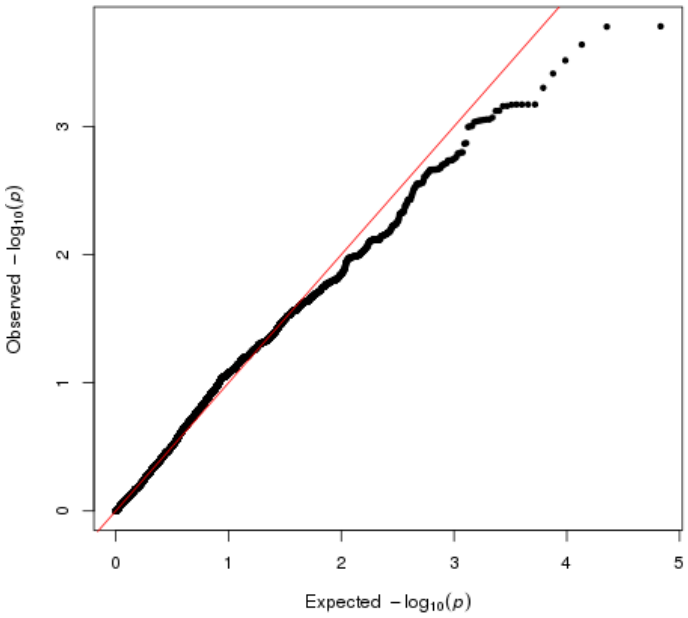


Figure 3.1: qq plot of the test statistics of an association analysis of IPF susceptibility across the HLA region (6:28,477,797 – 6:33,448,354) in the UK dataset (λ=1.02).

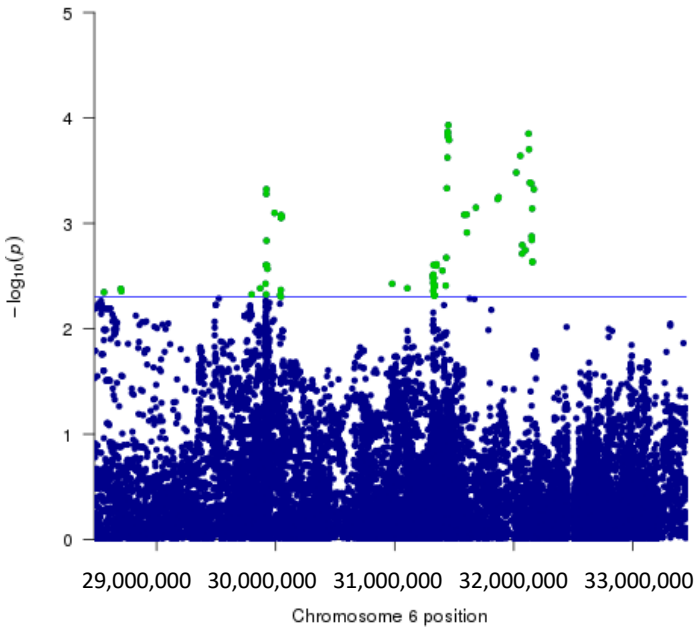


Figure 3.2: Manhattan plot of the HLA region for IPF susceptibility in the UK IPF dataset (the green variants are all the variants that passed the suggestive significance threshold). Blue line is suggestive significance threshold of $P<5x10^{-3}$).

Table 3.2: Independent signals ($P<5x10^{-3}$) in a HLA-wide association analyses of IPF susceptibility in the UK IPF dataset.

rsid	Gene	Position – build 37 (HLA Class)	Coded/ non-coded allele	Imp qual	Coded Allele Freq	OR (95% CI)	P-value
rs112417354	*HCG26* (-10382bp) / *MICB* (+11487bp)	31450567 (Class III)	G/A	1.00	0.11	1.58 (1.25-1.99)	$1.17x10^{-4}$
rs9267812	*PPT2*	32128394 (Class III)	C/T	1.00	0.13	1.45 (1.19-1.76)	$1.99x10^{-4}$
rs9260570	*HLA-A* (-8477bp) / *HCG4P4* (+849bp)	29922138 (Class I)	G/A	0.58	0.03	2.76 (1.56-4.89)	$4.74x10^{-4}$
rs3132684	*ZNRD1-AS1*	29990708 (Class I)	A/G	1.00	0.32	1.29 (1.11-1.50)	$7.97x10^{-4}$
rs190452998	*BAG6*	31606528 (Class III)	G/A	0.50	0.02	3.66 (1.67– 8.04)	$1.22x10^{-3}$
rs41268902	*TNXB*	32071008 (Class III)	G/A	0.97	0.11	1.44 (1.15-1.80)	$1.61x10^{-3}$
rs2442752	*HLA-S* (-1500bp) / *LOC101929072* (+10302bp)	31351764 (Class III)	T/C	1.00	0.42	1.28 (1.09-1.50)	$2.52x10^{-3}$
rs9278834	*MUC22*	30978884 (Class I)	T/A	1.00	0.16	1.32 (1.09-1.58)	$3.74x10^{-3}$
rs9263726	*PSORS1C1/ PSORS1C2*	31106499 (Class I)	G/A	1.00	0.14	1.34 (1.10-1.63)	$4.12x10^{-3}$
rs116612059	*RPSAP2*	28699440 (Class I)	C/T	0.95	0.01	2.27 (1.30-3.93)	$4.17x10^{-3}$
rs709053	*HLA-B*	31324077 (Class I)	C/G	1.00	0.38	1.27 (1.08-1.50)	$4.38x10^{-3}$
rs17179108	*HLA-G*	29798642 (Class I)	T/C	0.97	0.09	1.39 (1.11-1.75)	$4.72x10^{-3}$

Table 3.3: Table of HLA gene alleles correlated with independent variants in the association analyses of IPF in the UK IPF dataset. r^2 <0.2 = not correlated, r^2 = 0.2-0.5 weakly correlated, r^2 = 0.5-0.8 = correlated, r^2 >0.8 = strongly correlated.

Lead variant rsid	**HLA Gene Allele**	**r^2**
rs2442752	*HLA-B*08:01*	0.24
	*HLA-B*44*	0.29
rs9260570	*HLA-A*29:02*	0.50
	*HLA-A*32:01*	0.44
rs9263726	*HLA-C*05:01*	0.29

3.4.2 HLA-wide association of IPF susceptibility: replication in the Chicago and Colorado datasets:

This analysis utilised the Chicago and the Colorado datasets and the HRC and HLA-specific imputation in an attempt to replicate the suggestive signals identified in the discovery analyses in the UK IPF dataset (Table 3.2). The two datasets were meta-analysed in order to increase power to replicate the suggestive novel findings.

The twelve independent variants that reached a threshold of $P<5x10^{-3}$ for association with IPF susceptibility in the UK data set were tested for replication for significance in IPF susceptibility in 2,015 IPF and fIIP cases and 5,193 controls (Chicago and Colorado datasets). All signals passed quality control in both studies this replication analysis apart from rs3132684 (Colorado, imputation quality = 0.31, MAF =0.009, Chicago, imputation quality = 0.37, MAF =0.008) (table 3.2). None of the signals passed the Bonferroni corrected threshold for 12 variants for replication in this analysis (P<0.004) (table 3.4). The signals all have similar imputation quality and allele frequency across the UK, Colorado and Chicago datasets (Table 3.2 & 3.4). rs3132684 and rs9278834 have P<0.05 and similar effect sizes (to the UK dataset) in the Chicago dataset but was not significant in the Colorado dataset. Some of the variants in table 3.4 have different direction of effects in the Chicago and Colorado datasets, the most extreme case is rs190452998 which has an odds ratio of 0.65 (0.41-1.00) in the Colorado dataset and 2.40 (0.84-6.89) in the Chicago dataset. However, this signal has a low minor allele frequency and low imputation quality so this may account for the differences

Table 3.4: Replication results from a meta-analysis of the Colorado and Chicago datasets of the novel findings from the HLA-wide association analysis for IPF susceptibility in the UK IPF dataset.

Dataset	Rsid	Position – build 37 (HLA Class)	Nearest Gene	Coded/ non-coded allele	Imputation Quality	Coded allele frequency	OR (95% CI)	P-value	Meta-analysis OR (95% CI)	Meta-analysis P-Value
Colorado	rs112417354	31450567 (Class III)	*HCG26 (-10382bp) / MICB (+11487bp)*	G/A	1.00	0.10	0.88 (0.76-1.01)	0.07	0.91 (0.80-1.04)	0.17
Chicago					0.99	0.08	1.12 (0.80-1.56)	0.51		
Colorado	rs9267812	32128394 (Class III)	*PPT2*	C/T	0.99	0.14	0.94 (0.84-1.06)	0.33	0.96 (0.86-1.08)	0.51
Chicago					0.92	0.14	1.09 (0.82-1.44)	0.55		
Colorado	rs3132684	29990708 (Class I)	*ZNRD1-AS1*	A/G	1.00	0.33	1.02 (0.93-1.06)	0.73	1.06 (0.98-1.41)	0.15
Chicago					1.00	0.34	1.34 (1.09-1.63)	4.7x10-3		
Colorado	rs190452998	31606528 (Class III)	*BAG6*	G/A	0.47	0.02	0.65 (0.41-1.00)	0.05	0.79 (0.52-0.97)	0.25
Chicago					0.41	0.02	2.40 (0.84-6.89)	0.10		
Colorado	rs41268902	32071008 (Class III)	*TNXB*	G/A	0.88	0.11	0.89 (0.77-1.02)	0.10	0.92 (0.81-1.05)	0.23
Chicago					0.87	0.10	1.13 (0.82-1.55)	0.46		
Colorado	rs2442752	31351764 (Class III)	*HLA-S (-1500bp) / LOC101929072 (+10302bp)*	T/C	1.00	0.40	0.98 (0.90-1.06)	0.59	0.98 (0.91-1.06)	0.69
Chicago					0.99	0.39	1.02 (0.85-1.24)	0.81		
Colorado	rs9278834		*MUC22*	T/A	1.00	0.16	0.94	0.32	0.99	0.88

		30978884 (Class I)					(0.84-1.06)		(0.89-1.10)	
Chicago					0.94	0.15	1.31 (1.00-1.71)	0.05		
Colorado	rs9263726	31106499 (Class I)	*PSORS1C1/ PSORS1C2*	G/A	1.00	0.13	0.92 (0.81-1.04)	0.16	0.93 (0.83-1.04)	0.22
Chicago					0.94	0.13	1.02 (0.77-1.36)	0.89		
Colorado	rs116612059	28699440 (Class I)	*RPSAP2*	C/T	0.93	0.01	1.00 (0.66-1.50)	0.98	0.88 (0.60-1.28)	0.50
Chicago					0.90	0.01	0.45 (0.17-1.12)	0.10		
Colorado	rs709053	31324077 (Class I)	*HLA-B*	C/G	0.99	0.36	0.98 (0.90-1.07)	0.64	0.97 (0.90-1.05)	0.48
Chicago					0.99	0.37	0.94 (0.77-1.14)	0.51		
Colorado	rs17179108	29798642 (Class I)	*HLA-G*	T/C	0.96	0.10	1.04 (0.90-1.20)	0.57	1.02 (0.89-1.16)	0.78
Chicago					0.96	0.10	0.91 (0.66-1.25)	0.55		

3.4.3 Summary:

In this analysis I performed a discovery and replication association analysis in the HLA region in the UK, Chicago and Colorado datasets for IPF susceptibility. No signals passed the Bonferroni corrected threshold of $2.8x10^{-6}$ in the discovery dataset but 12 novel signals passed a suggestive significance threshold of $5x10^{-3}$. None of the novel signals identified in the IPF susceptibility discovery analysis in the UK dataset were replicated in a meta-analysis of the Chicago and Colorado datasets. Out of the 11 signals that passed quality control in this analyses, none passed $P<0.004$ (table 3.2). Rs3132684 and rs9278834 were nominally significant in the Chicago dataset, and this dataset has a stricter case control criteria (along with the UK dataset) and therefore these could be of interest. These results were all suggestive and must be treated with caution until they can be confirmed. An increased sample size for the discovery dataset would improve power and the ability to identify novel signals. The next section of this chapter, a HLA-wide association meta-analysis of IPF susceptibility was performed using the UK, Colorado and Chicago datasets.

3.5 Results: HLA-wide association meta-analyses of IPF susceptibility in UK, Colorado and Chicago datasets:

In this analysis the HLA-wide association analysis results from the UK, Chicago and Colorado datasets were meta-analysed to study the HLA region in the largest HLA-wide association analysis to date of the HLA region in IPF susceptibility. This will maximise the power to detect novel signals in the region.

2,769 IPF and fIIP cases and 8,591 controls with 35,043 variants from the UK dataset (figure 3.2), 33,323 variants from the Chicago dataset (supplementary figure 3.1) and 34,905 variants from the Colorado dataset (supplementary figure 3.2) were included in the meta-analysis of IPF susceptibility in the HLA region (figure 3.7). 33,979 variants (88%) were found in all three studies so when the datasets were merged, 37,212 SNPs 424 HLA alleles and 1,276 amino acid changes were included in this meta-analysis. 10,447 variants had an allele frequency of less than 5% (figure 3.3). The average imputation quality in this meta-analysis was 0.97 (0.97 for SNPs and 0.99 for HLA alleles and amino acids) and 87% of variants had an imputation quality of over 0.98 (figure 3.4). There was no inflation of the test statistic in any of the datasets (UK, $\lambda=1.02$ [figure 3.1], Colorado, $\lambda=1.08$ [figure 3.5], Chicago, $\lambda=1.04$ [figure 3.6]).

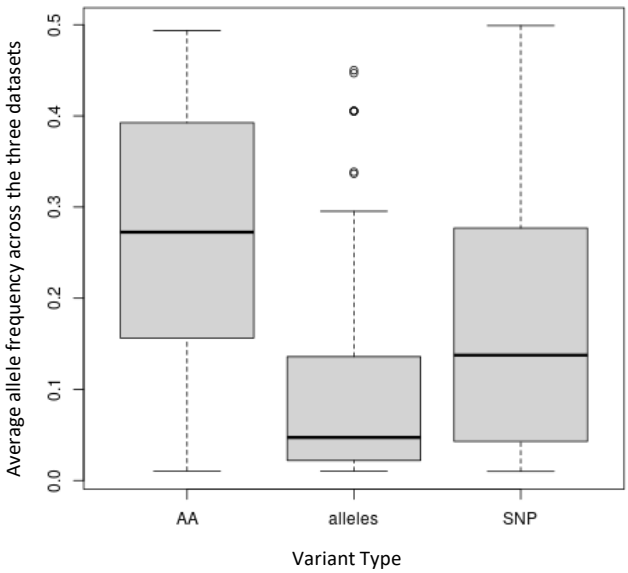


Figure 3.3: Histogram of average allele frequencies < 5% of variants in the meta-analysis of the HLA region in IPF susceptibility in the UK, Colorado and Chicago datasets.

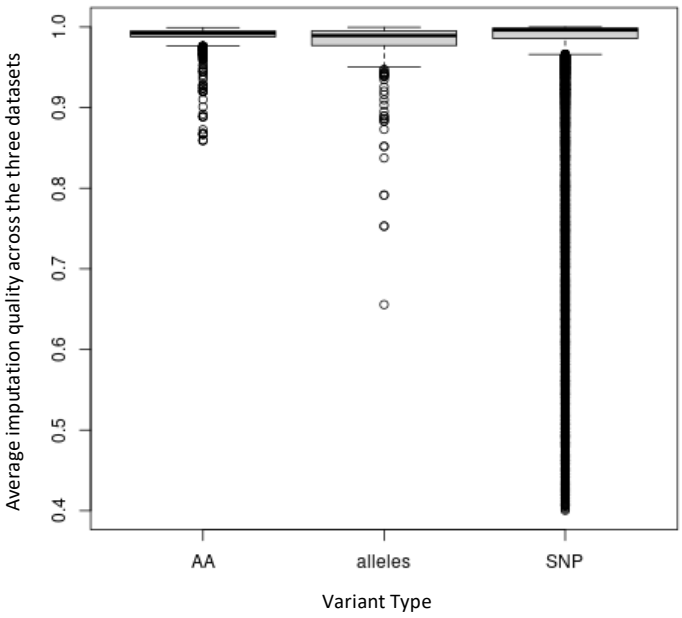


Figure 3.4: Histogram of average imputation qualities of variants in the meta-analysis of the HLA region in IPF susceptibility in the UK, Colorado and Chicago datasets.

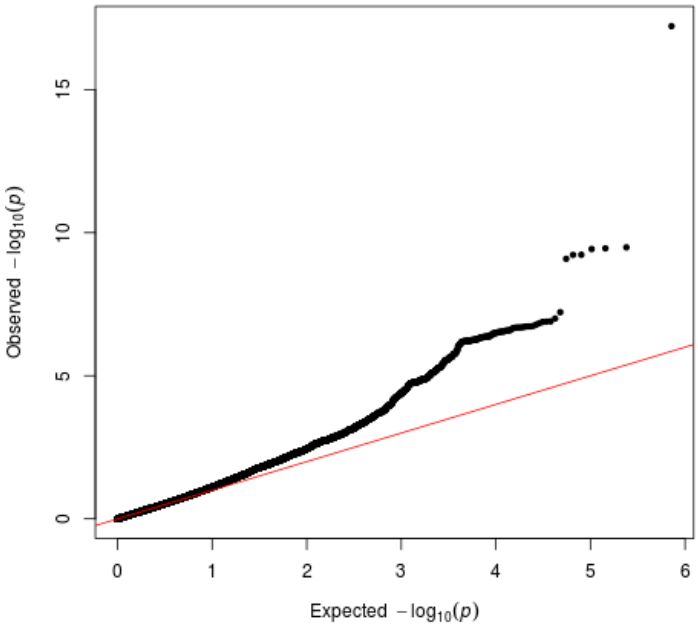


Figure 3.5: qq plot of the test statistics of an association analysis of IPF susceptibility across chromosome 6 in the Colorado dataset (λ=1.08).

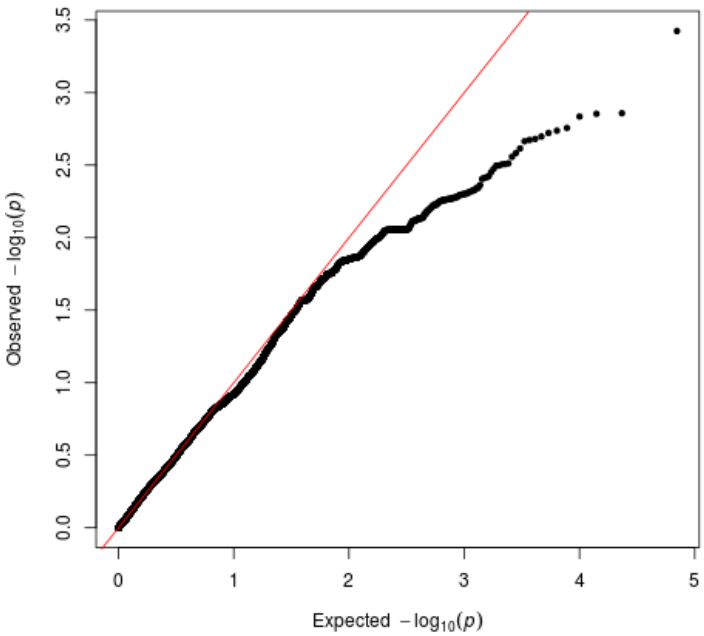


Figure 3.6: qq plot of the test statistics of an association analysis of IPF susceptibility across the HLA region (6:28,477,797 – 6:33,448,354) in the Chicago dataset (λ=1.04).

Of the variants that were present in all studies (figure 3.5), three independent signals passed the Bonferroni corrected threshold of $P<2.8x10^{-6}$ (table 3.5, supplementary figures 17, 18 and 19). In order for signals to be deemed as significant in this meta-analysis, the signals were also required to pass a threshold of $P<0.05$ in each dataset and be in the same direction of effect. Although these three variants had the same direction of effect in each dataset, they did not pass $P<0.05$ in each dataset independently (table 3.5). The three signals that passed $P<2.8x10^{-6}$ (rs7754402, rs3135350 and *HLA-DQB1*06:02*) were all completely driven by the Colorado dataset, which was also the largest of the three (table 3.5).

When considering all variants that were analysed across the HLA region, there were no variants that passed $P<0.05$ in all three studies (figure 3.8) but there were 18 independent variants that passed $P<0.05$ in two studies and had a meta-$P<5x10^{-3}$ (table 3.6). rs3132684 (identified in discovery and replication, section 3.5) was also identified here with a P-value of $4x10^{-3}$ (table 3.6, supplementary figure 38). Two signals identified in the UK and Chicago datasets were missense coding sequence variants (rs1042337 in *HLA-DMB* and rs17207895 in *TNXB,* supplementary figures 31 and 35 respectively*)* (table 3.6). rs115478552 and rs3132684 were nominally significant ($P<0.05$) in the UK and the Chicago datasets (supplementary figures 37 and 38 respectively).

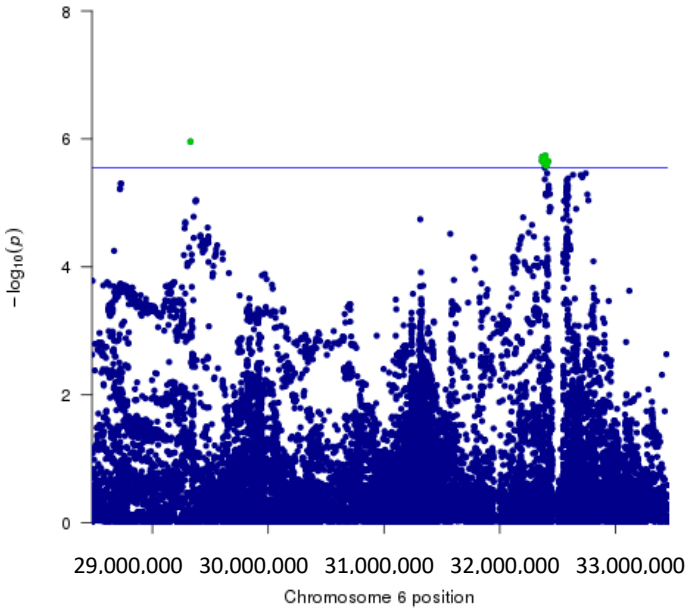


Figure 3.7: Manhattan plot of the meta-analysis in HLA region for IPF susceptibility in the UK, Colorado and Chicago IPF datasets (the green variants are all the variants that passed the Bonferroni corrected significance threshold). Blue line is Bonferroni corrected significance threshold of $P<2.8x10^{-6}$*).*

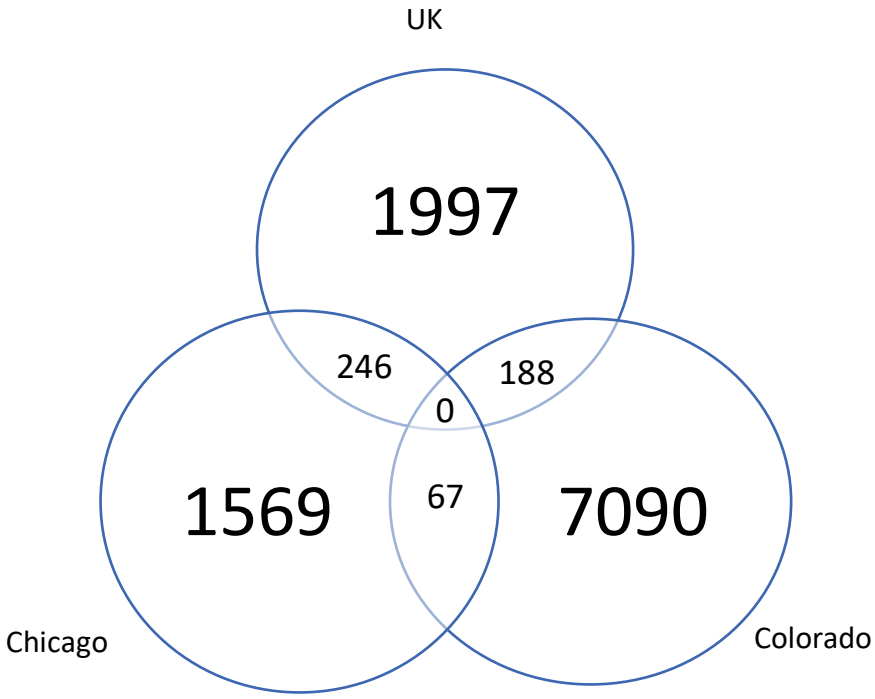


Figure 3.8: Number of variants with P<0.05 in each dataset in the meta-analysis of IPF susceptibility in the UK, Chicago and Colorado datasets.

Table 3.5: Independent signals ($P<2.8x10^{-6}$) in a HLA-wide meta-analyses of IPF susceptibility in the UK, Colorado and Chicago IPF datasets.

Dataset	Rsid	Position – build 37 (HLA Class)	Nearest Gene	Coded/ non-coded allele	Imputation quality	Coded allele frequency	OR (95% CI)	P-value	Meta-analysis OR (95% CI)	Meta-analysis P-value
UK	rs7754402	29328834 (Class I)	*OR5V1*	C/T	1.00	0.36	0.90 (0.78-1.05)	0.20	0.84 (0.78-0.90)	$1.10x10^{-6}$
Colorado					1.00	0.33	0.81 (0.75-0.89)	$5.34x10^{-6}$		
Chicago					1.00	0.33	0.85 (0.69-1.03)	0.10		
UK	rs3135350	32392981 (Class III)	Intergenic (*BTNL2* -18081, *HLA-DRA* +14638)	C/T	1.00	0.15	0.93 (0.76-1.15)	0.52	0.79 (0.71-0.84)	$1.83x10^{-6}$
Colorado					1.00	0.13	0.72 (0.64-0.81)	$1.89x10^{-7}$		
Chicago					1.00	0.14	0.89 (0.68-1.16)	0.39		
UK	HLA-DQB1*06:02	32634302 (Class II)	*HLA-DQB1*	P/A	0.99	0.15	0.90 (0.73-1.10)	0.29	0.79 (0.75-0.82)	$2.15x10^{-6}$
Colorado					0.98	0.13	0.72 (0.64-0.82)	$4.20x10^{-7}$		
Chicago					0.99	0.14	0.92 (0.70-1.22)	0.57		

Table 3.6: Independent signals ($P<2.8x10^{-6}$) in a HLA-wide meta-analyses of IPF susceptibility in the UK, Colorado and Chicago IPF datasets.

Dataset	Rsid	Position – build 37 (HLA Class)	Nearest Gene	Coded/ non-coded allele	Imputation quality	Coded allele frequency	OR (95% CI)	P-value	Meta-analysis OR (95% CI)	Meta-analysis P-value
UK	rs9265912	31313733 (Class I)	*WASF5P* (-56792), *HLA-B* (+7916)	T/C	1.00	0.39	0.98 (0.83-1.17)	0.86	0.85 (0.79-0.92)	$1.81x10^{-5}$
Colorado					0.98	0.35	0.83 (0.76-0.91)	$3.28x10^{-5}$		
Chicago					0.95	0.35	0.80 (0.67-0.99)	0.04		
UK	rs1233385	29559238 (Class I)	*OR2H2*	T/C	1.00	0.12	1.40 (1.03-1.89)	0.03	1.28 (1.14-1.44)	$4.59x10^{-5}$
Colorado					0.99	0.10	1.30 (1.13-1.50)	$3.65x10^{-4}$		
Chicago					0.99	0.11	1.11 (0.83-1.49)	0.49		
UK	rs2071627	32809223 (Class II)	*PSMB8*	C/A	1.00	0.42	0.96 (0.84-1.10)	0.60	0.87 (0.82-0.93)	$8.18x10^{-5}$
Colorado					1.00	0.42	0.86 (0.79-0.94)	$4.86x10^{-4}$		
Chicago					0.99	0.40	0.77 (0.64-0.93)	0.01		
UK	rs241436	32797876 (Class II)	*TAP2*	A/G	1.00	0.44	0.87 (0.76-0.99	0.04	0.89 (0.83-0.95)	$4.73x10^{-4}$
Colorado					1.00	0.46	0.90 (0.82-0.98)	0.01		
Chicago					0.96	0.49	0.88 (0.73-1.06)	0.19		
UK	rs6457256	30661007 (Class I)	*NRM*	T/C	1.00	0.02	0.61 (0.39-0.94)	0.02	0.69 (0.56-0.85)	$5.45x10^{-4}$
Colorado					0.99	0.03	0.75	0.03		

							(0.58-0.97)			
Chicago					0.99	0.03	0.61 (0.35-1.07)	0.08		
UK	rs2734945	29855945 (Class I)	*HLA-H*	A/G	1.00	0.32	1.21 (1.02-1.44)	0.03	1.14 (1.06-1.23)	$5.53x10^{-4}$
Colorado					0.94	0.31	1.12 (1.02-1.23)	0.01		
Chicago					0.92	0.33	1.14 (0.93-1.40)	0.20		
UK	rs9468541	29487535 (Class I)	*RPS17P1* (-30065), *LINC01015* (+9648)	G/A	1.00	0.07	0.72 (0.56-0.92)	$8.58x10^{-3}$	0.81 (0.72-0.92)	$8.48x10^{-4}$
Colorado					0.99	0.08	0.81 (0.69-0.95)	0.01		
Chicago					0.99	0.08	0.99 (0.72-1.38)	0.97		
UK	rs504653	31840766 (Class III)	*SLC44A4*	A/G	0.99	0.35	1.04 (0.90-1.20)	0.61	0.89 (0.83-0.95)	$8.76x10^{-4}$
Colorado					0.98	0.39	0.86 (0.79-0.93)	$4.95x10^{-4}$		
Chicago					0.96	0.39	0.81 (0.67-0.98)	0.03		
UK	rs2621331	32780470 (Class II)	*HLA-DOA*	C/T	1.00	0.40	0.90 (0.78-1.03)	0.12	0.89 (0.83-0.98)	$1.32x10^{-3}$
Colorado					1.00	0.37	0.91 (0.84-0.99)	0.03		
Chicago					1.00	0.35	0.80 (0.66-0.98)	0.03		
UK	rs1042337	32904980 (Class II)	*HLA-DMB*	C/T	0.98	0.26	0.92 (0.80-1.07)	0.29	0.88 (0.82-0.95)	$1.48x10^{-3}$
Colorado					0.98	0.24	0.90 (0.81-0.99)	0.03		
Chicago					0.92	0.21	0.71	$3.93x10^{-3}$		

							(0.56-0.90)			
UK	rs75561043	31414959 (Class III)	*LINC01149*	T/G	1.00	0.09	1.03 (0.82-1.30)	0.79	0.84 (0.76-0.94)	$1.61x10^{-3}$
Colorado					1.00	0.11	0.82 (0.72-0.94)	$3.29x10^{-3}$		
Chicago					1.00	0.11	0.70 (0.52-0.93)	0.01		
UK	rs28752951	31304674 (Class I)	*WASF5P* (-47733), *HLA-B* (+16975)	A/G	0.96	0.05	1.13 (0.76-1.69)	0.54	1.42 (1.14-1.77)	$1.97x10^{-3}$
Colorado					0.95	0.03	1.49 (1.12-1.98)	$6.40x10^{-3}$		
Chicago					0.62	0.03	2.11 (1.04-4.29)	0.04		
UK	AA_B_97_31432180_R	31324201 (Class I)	*HLA-B*	P/A	0.99	0.47	0.86 (0.74-1.00)	0.05	1.11 (1.04-1.19)	$2.79x10^{-3}$
Colorado					0.99	0.49	1.20 (1.12-1.31)	$1.51x10^{-5}$		
Chicago					0.98	0.48	1.10 (0.91-1.32)	0.31		
UK	rs17207895	32020512 (Class III)	*TNXB*	T/C	0.92	0.01	0.90 (0.51-1.59)	0.73	0.69 (0.54-0.88)	$3.32x10^{-3}$
Colorado					0.89	0.02	0.71 (0.53-0.96)	0.03		
Chicago					0.87	0.02	0.38 (0.18-0.76)	$6.85x10^{-3}$		
UK	rs2293751	31907837 (Class III)	*C2*	G/A	1.00	0.44	1.16 (1.00-1.35)	0.05	0.90 (0.84-0.97)	$3.91x10^{-3}$
Colorado					1.00	0.49	0.83 (0.77-0.91)	$2.03x10^{-5}$		
Chicago					0.98	0.47	0.92 (0.77-1.11)	0.39		
UK		30669762	*MDC1*	G/A	0.99	0.01	0.56	0.05	0.66	$3.94x10^{-3}$

	rs115478552	(Class I)					(0.31-1.00)		(0.50-0.88)	
Colorado					0.97	0.01	0.76 (0.53-1.01)	0.13		
Chicago					0.97	0.02	0.47 (0.22-0.99)	0.048		
UK	rs3132684	29990708 (Class I)	*ZNRD1ASP*	A/G	1.00	0.32	1.29 (1.11-1.50)	7.97×10^{-4}	1.11 (1.04-1.19)	4.00×10^{-3}
Colorado					1.00	0.33	1.02 (0.93-1.12)	0.73		
Chicago					1.00	0.34	1.34 (1.10-1.63)	4.71×10^{-3}		
UK	rs11244	32780724 (Class II)	*HLA-DOB*	G/A	1.00	0.26	0.92 (0.79-1.08)	0.30	0.90 (0.83-0.97)	4.25×10^{-3}
Colorado					1.00	0.27	0.91 (0.83-1.00)	0.05		
Chicago					1.00	0.28	0.80 (0.65-0.98)	0.04		

3.5.1 Replication analysis of the HLA-DQB1*06:02 signal in the UK and Chicago datasets

*HLA-DQB1*06:02* (table 3.7, supplementary figure 3.7) was previously found to be associated with idiopathic interstitial pneumonia (IIP) susceptibility in an analysis of the Colorado dataset in 2016 (5). In the replication analysis, the variant was in the same direction of effect in the UK and Chicago datasets (but the confidence intervals were wide and crossed one), however the signal did not pass the replication threshold of 0.05 (P=0.23). Power calculations were undertaken, and it was determined that there was sufficient power to replicate the HLA-DQB1*06:02 signal in this replication analysis (100% power for analysis in 1,112 cases, odds ratio of 0.72 and allele frequency of 0.15).

*Table 3.7: Replication results from a meta-analysis of the UK and Chicago datasets of the HLA-DQB1*06:02 signal*

Dataset	**Variant ID**	**Position – build 37 (HLA class)**	**Imputation quality**	**Allele frequency**	**Odds ratio (95% CI)**	**P-value**	**Meta Odds ratio (95% CI)**	**Meta P-value**
UK	*HLA-DQB1*06:02*	32631061 (Class II)	0.99	0.14	0.90 (0.73-1.10)	0.29	0.91 (0.77-1.10)	0.23
Chicago			0.99	0.15	0.92 (0.70-1.22)	0.57		

Summary

Although three signals passed the Bonferroni corrected threshold in this HLA-wide association meta-analysis of IPF susceptibility, none of these signals passed a nominal significance threshold in each separate data set (*P*<0.05). The *HLA-DQB1*06:02* signal previously identified in the Colorado dataset (5) did not replicate in the UK or the Chicago datasets, suggesting that this signal may be specific to the Colorado dataset (table 3.7). There were two signals (rs115478552 and rs3132684) which were nominally significant (P<0.05) in both the UK and the Chicago datasets, this could be of interest as these datasets had a stricter case inclusion criteria (only IPF) whereas the Colorado dataset included a wider range of interstitial pneumonias.

3.6 HLA-wide association meta-analysis of IPF susceptibility in the UK, Chicago and UUS datasets:

I have been identifying differences when comparing the cases of the Colorado dataset and the cases in the UK and Chicago datasets. Because of this I excluded the Colorado dataset from the meta-analysis and replace it with a new IPF dataset; the UUS dataset (see Chapter 2, section 2.1.4). Removing the Colorado dataset and replacing it with the UUS dataset will increase my power to detect signals associated with IPF as some of the heterogeneity has been removed.

1,905 IPF cases and 13,876 controls with 35,043 variants from the UK dataset (figure 3.2), 33,323 variants from the Chicago dataset (supplementary figure 3.1) and 36,965 variants from the UUS dataset (supplementary figure 3.8) were included in the meta-analysis of IPF susceptibility in the HLA region (figure 3.8). 88% of variants were shared between studies and so 39,570 variants were studied in this meta-analysis. 8,449 variants in this analysis had an allele frequency of less than 5% (figure 3.9). The average imputation quality in this analysis was 0.99 (0.99 For SNPs, 0.97 for HLA alleles and 0.99 for amino acid changes) and 86% of variants had an imputation quality of more than 0.98 (figure 3.10). There was slight inflation of the test statistic in the UUS dataset across chromosome 6 (UUS, λ=1.17 [figure 3.11]), this could be due to increased power with the increased sample size. There was no inflation of the test statistic in any the other datasets (UK, λ=1.02 [figure 3.1] and Chicago, λ=1.04 [figure 3.6]).

One signal passed the Bonferroni corrected significance threshold of 2.8×10^{-6} (Table 3.8, figure 3.8, supplementary figure 3.9). Rs3132684 was well imputed and had a high coded allele frequency (>30%) across all three datasets (Table 3.8). Rs3132684 is found in HLA Class I in an intron of the gene *ZNRD1ASP* (supplementary figure 3.9). This signal was in the same direction of effect and had P<0.05 in all three studies. In the analysis of the UK and Chicago datasets, the HLA-DQB1*06:02 signal did not replicate in this meta-analysis (P=0.57, Table 3.9).

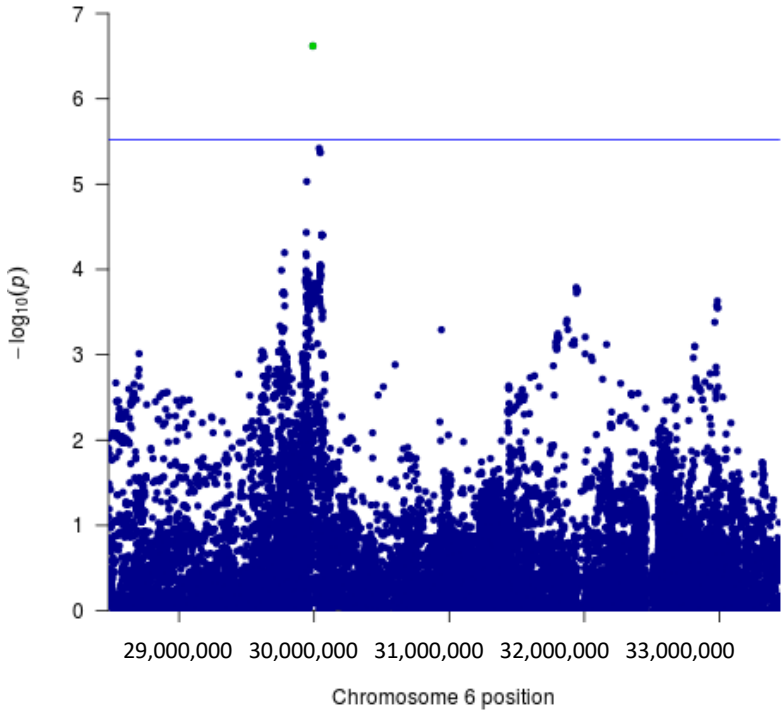


Figure 3.8: Manhattan plot of the meta-analysis in HLA region for IPF susceptibility in the UK, UUS and Chicago IPF datasets (the green variants are all the variants that passed the Bonferroni corrected significance threshold). Blue line is Bonferroni corrected significance threshold of $P<2.8\text{x}10^{-6}$).

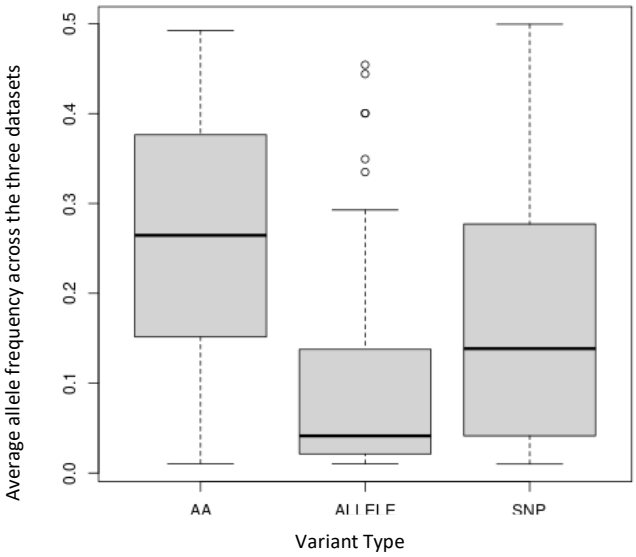


Figure 3.9: Histogram of average allele frequencies < 5% of variants in the meta-analysis of the HLA region in IPF susceptibility in the UK, UUS and Chicago datasets.

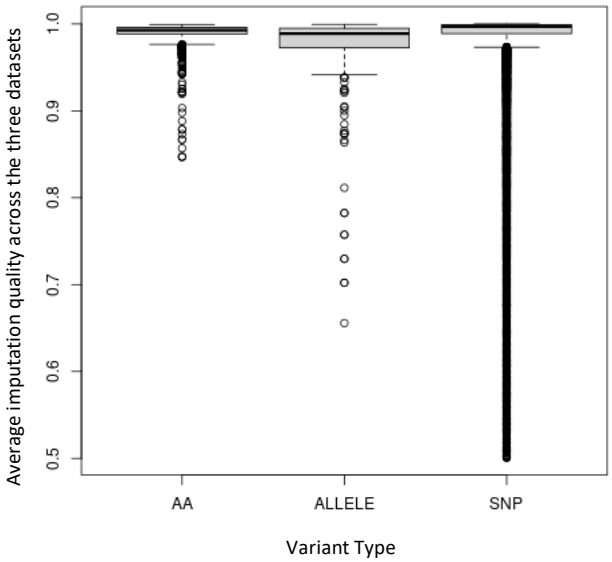


Figure 3.10: Histogram of average imputation qualities of variants in the meta-analysis of the HLA region in IPF susceptibility in the UK, UUS and Chicago datasets.

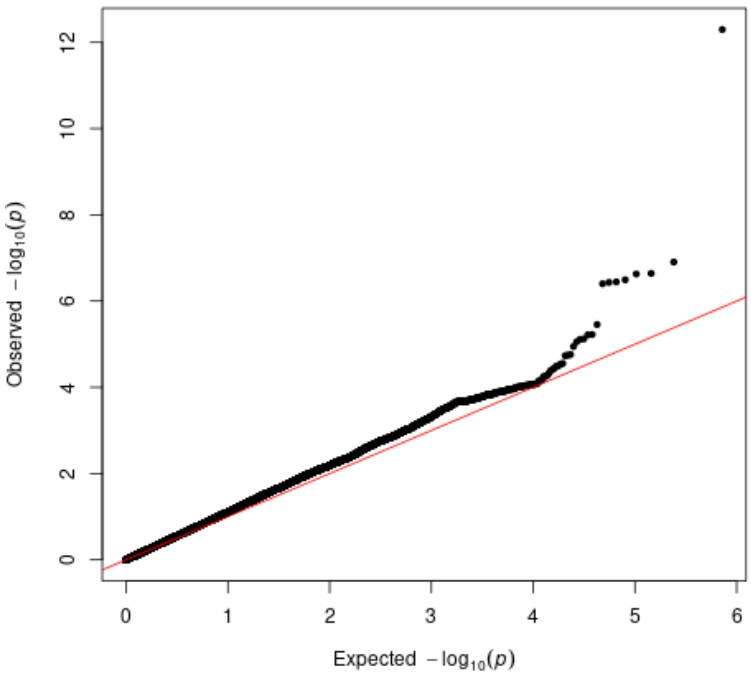


Figure 3.10: qq plot of the test statistics of an association analysis of IPF susceptibility across chromosome 6 in the UUS dataset (λ=1.17).

Table 3.8: Independent signals ($P<2.8x10^{-6}$) in a HLA-wide meta-analyses of IPF susceptibility in the UK, UUS and Chicago IPF datasets.

Dataset	rsid	Position – build 37 (HLA class)	Nearest gene	Coded/non-coded allele	Info score	Coded allele frequency	P-value	OR (95% CI)	Meta-analysis P-value	Meta-analysis OR (95% CI)
UK	rs3132684	29990708 (Class I)	*ZNRD1ASP*	A/G	1.00	0.32	$7.97x10^{-4}$	1.29 (1.11-1.50)	$2.41x10^{-7}$	1.24 (1.14-1.35)
UUS					1.00	0.32	$2.86x10^{-3}$	1.19 (1.02-1.32)		
Chicago					1.00	0.34	$4.71x10^{-3}$	1.34 (1.15-1.63)		

*Table 3.9: Results of the replication analysis of HLA-DQB1*06:02 in the UK, UUS and Chicago datasets.*

Dataset	rsid	Position – build 37 (HLA class)	Nearest gene	Coded/non-coded allele	Info score	Coded allele frequency	P-value	OR (95% CI)	Meta-analysis P-value	Meta-analysis OR (95% CI)
UK	HLA-DQB1*06:02	32631061 (Class II)	*HLA-DQB1*	P/A	0.99	0.15	0.29	0.90 (0.73-1.10)	0.57	0.97 (0.87-1.08)
UUS					0.99	0.12	0.75	1.02 (0.88-1.21)		
Chicago					0.99	0.14	0.57	0.92 (0.70-1.22)		

Rs3132684 is a common SNP (frequency of 32-34% in the datasets) and is in an intron of the gene *ZNRD1ASP* in HLA class II (supplementary figure 3.9). Because this signal did not map to a particular HLA gene allele or amino acid allele statistical fine mapping was undertaken to produce a set of SNPs in which we were 95% confident the causal SNP is located (under the assumption that the causal SNP was tested in the data), for rs3132684 the credible set contained 190 SNPs. Of those 190 SNPs, rs3132684 was the most likely causal variant (55%) and there were four coding sequence variants (two synonymous and two missense [rs2074479 in *RNF39* and rs2301753 in *PPP1R11*] variants). Rs3132684 and SNPs in the 95% credible set were tested for association with respiratory, inflammatory, immunity and autoimmune phenotypes using phenoscanner (REF), the SNPs were found to be associated with many of these phenotypes including peak expiratory flow, eosinophil counts, IgA deficiency, Rheumatoid Arthritis and white blood cell counts (supplementary table 2). There were eQTL SNPs (from the credible set) across 24 genes in GTEx (Table 3.10) including eight HLA genes. Two eQTL SNPs (rs2256919 [HLA-H] and rs2256919 [HLA-W]) were in low-moderate linkage disequilibrium with the lead association SNP which is suggestive of colocalization between the signals (the association signal is the same signal affecting the gene expression).

Table 3.10: Table of the linkage disequilibrium (LD) between the lead SNP from the HLA-wide association meta-analysis of IPF susceptibility and the eQTL signal in the lung (from GTEx). R^2>0.2=minimal LD, R^2=0.2-0.5=low LD, R^2=0.5-0.8= moderate LD, R^2>0.8= high LD.

Lead SNP	eQTL SNP	Gene	R^2
rs3132684	rs28698309	*HCG4*	0.08
	rs10947050	*HCG4B*	0.18
	rs1048412	*HCGP3*	0.24
	rs3823374	*HCGP5*	0.18
	rs2245952	*HCGP7*	0.06
	rs10947051	*HCG9*	0.21
	rs2245952	*HLA-A*	0.06
	rs28698309	*HLA-G*	0.08
	rs2256919	*HLA-H*	0.49
	rs10947050	*HLA-J*	0.18
	rs1048412	*HLA-K*	0.24
	rs2735076	*HLA-U*	0.37
	rs1048412	*HLA-V*	0.24
	rs2256919	*HLA-W*	0.49
	rs1048412	*IFITM4P*	0.24
	rs10947050	*MICD*	0.18
	rs28698309	*MICE*	0.08
	rs10947050	*RNF39*	0.18

	rs1048412	*RPL23AP1*	0.24
	rs10947050	*TRIM31*	0.18
	rs10947050	*XXBAC-BPG170G13.32*	0.18
	rs1048412	*ZFP57*	0.24
	rs1048412	*ZNRD1*	0.24
	rs10947051	*ZNRD1_AS1*	0.21

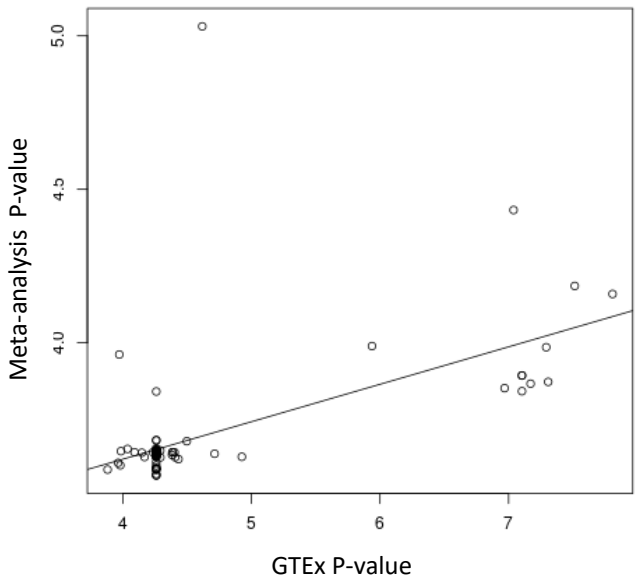


Figure 3.11: Comparison of -log10 p-values from the meta-analysis of IPF susceptibility and the -log10 p-values from lung tissue in GTEx.

3.7 Amino acid joint regression meta-analysis of IPF susceptibility in the UK, UUS and Chicago dataset

An amino acid joint regression meta-analysis was undertaken in the UK, UUS and Chicago datasets to identify further associations with IPF susceptibility. Both sets of amino acids were tested in order to identify the effect of collinearity on the results (set one was all remaining amino acids after quality control and set two was a subset of the quality controlled amino acids with the most frequent amino acid allele at each position removed (as described above)). The results of the joint regression analyses from the UK, UUS and Chicago datasets were merged using a fixed-effects meta-analysis.

In total, 604 amino acid alleles across 338 sites were tested for association with IPF susceptibility in set one of the 3-way meta-analysis in 1,905 cases and 13,876 controls and 441 amino acids across 238 sites were tested in set two (Table 3.11). Bonferroni corrected significance thresholds for these analyses were $1.5x10^{-4}$ and $2.1x10^{-4}$ for set one and set two respectively. The lambdas of the analysis of the two amino acid sets in the 3-way meta-analyses were under one indicating no inflation of the test statistics due to low power (figure 3.15). Amino acid tyrosine (Y) at position 9 in HLA-A (NM_002116:p.[Phe_9_Tyr]) was most significantly associated with IPF susceptibility in the 3-way meta-analysis of set one, however it was not nominally significant (P<0.05) in the UUS or Chicago datasets (Table 3.12, figure 3.14a). This amino acid was less significantly associated with IPF susceptibility in the 3-way meta-analysis of set two (frequency filtered) (P=0.05). NM_002116:p.[Phe_9_Tyr] had a frequency of around 15% and was well imputed across all three datasets (Chicago-17% [quality=0.98], UK-15% [quality=0.99], UUS-15% [quality=0.99]). NM_002116:p.[167_Trp]was most significantly associated with IPF susceptibility in the 3-way meta-analysis of set two (with the most common allele at each loci removed from the quality controlled dataset) (table 3.12, figure 3.14b). The p-value of this amino acid was less significant in the analysis of set one (P=0.2), this could be suggestive of collinearity at this locus in the analysis. NM_002116:p.[Phe_9_Tyr] was found in hundreds of *HLA-A* alleles including, including *HLA-A*01:43, A*02:05:01, A*02:06, A*03:12, A*11:01-A*11:198, A*25:01* and *A*06:01.* Tryptophan (W) at position 167 in *HLA-B* is the "wild type" amino acid and is found in most of the *HLA-B* alleles, however the amino acid was in high linkage disequilibrium (LD) with AA_B_167_S which was found across *HLA-B*44*.

Table 3.9: Number of amino acid alleles and sites with more than two or three alleles in set one and set two of the meta-analysis.

HLA Gene	**Set one**			**Set two**		
	Number of amino acid alleles	**Number of amino acid sites with > 2 alleles**	**Number of amino acid sites with > 3 alleles**	**Number of amino acid alleles**	**Number of amino acid sites with > 2 alleles**	**Number of amino acid sites with > 3 alleles**
A	85	32	14	61	24	9
B	97	33	18	88	31	15
C	98	39	15	71	28	7
DPA1	8	4	0	0	0	0
DPB1	36	12	12	24	12	0
DQA1	27	9	8	23	9	4

DQB1	131	39	38	92	38	19
DRB1	122	39	27	82	29	13

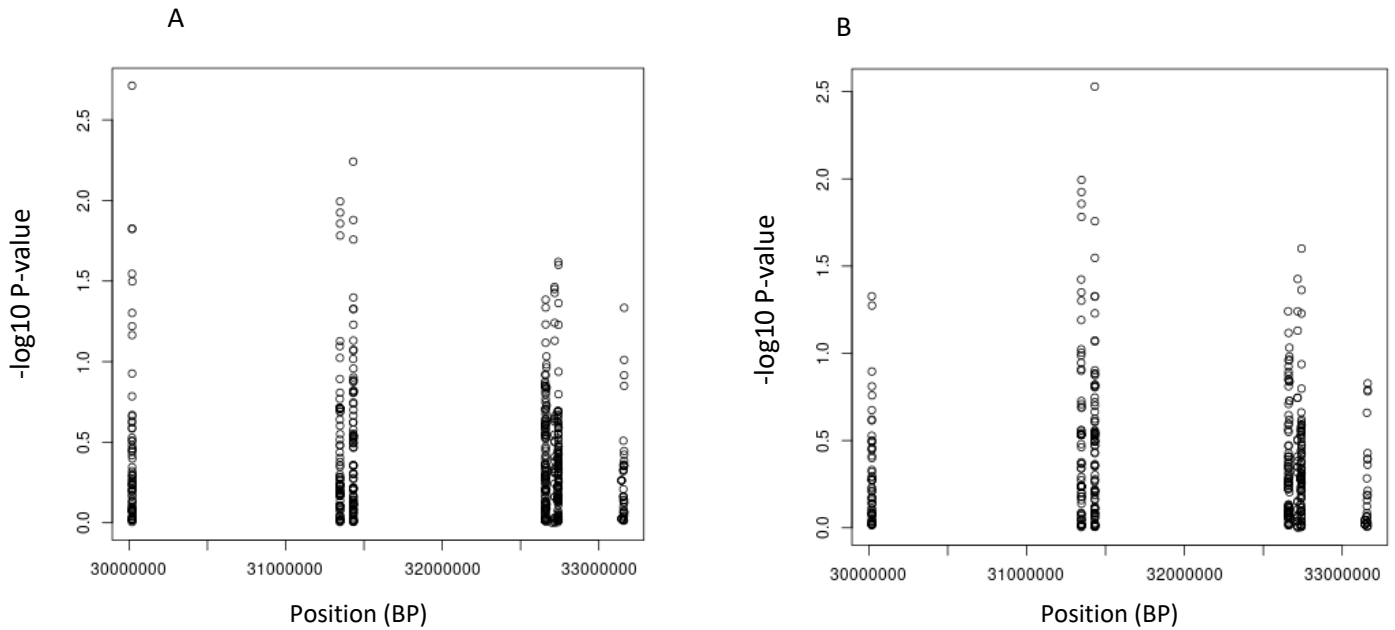


Figure 3.12: Manhattan plots of both joint regression meta-analyses of amino acids in the UK, UUS and Chicago datasets across the HLA region (set one [all amino acids with frequency > 1%] on the left and set two [with both rare amino acids and most common at each loci removed] on the right). Clustering of variZNRDants is due to only within gene variation being tested in this analysis.

Table 3.10: The top two amino acid results of the analysis of set one (full amino acid set) and set two (frequency filtered set) in the meta-analysis of the UK, UUS and Chicago datasets.

HLA Gene	**HLA amino acid variant**	**p-value**			**OR (95% CI)**			**Meta p-value**	**Meta OR (95% CI)**
		Chic	**UK**	**UUS**	**Chic**	**UK**	**UUS**		
Set one									
HLA-A	NM_002116:p.[Phe_9_Tyr]	0.20	0.0057	0.27	0.85 (0.65-1.09)	0.75 (0.61-0.92)	0.58 (0.23-1.51)	0.0019	0.78 (0.66-0.91)
Set two									
HLA-B	NM_002116:p.[167_Trp]	0.28	0.0040	NA	1.16 (0.88-1.53)	1.33 (1.10-1.62)	NA	0.0030	1.23 (1.09-1.49)

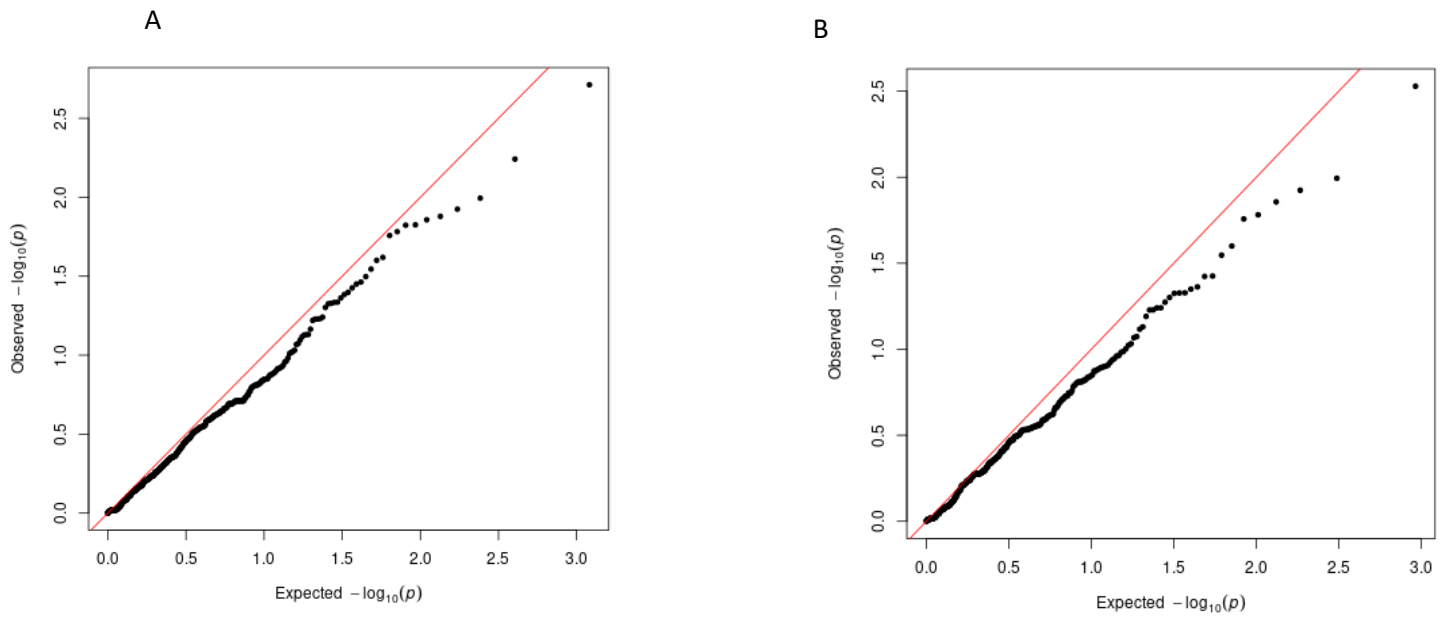


Figure 3.13: qq plots of the p-values from the joint regression meta-analysis of IPF susceptibility in the UK, UUS and Chicago datasets (A λ= 0.75, B λ= 0.86).

In the two joint regression meta-analyses of IPF susceptibility in the UK, UUS and Chicago datasets (set one and set two), the p-values (figure 3.16) and effect sizes (figure 3.17) were comparable. There was some spread around the least significant p-values (0-1 -log10 p-values in figure 3.16) and there was a suggestion of some attenuation to the null of variants in set one and in set two, this could be due to changes in power between the analyses.

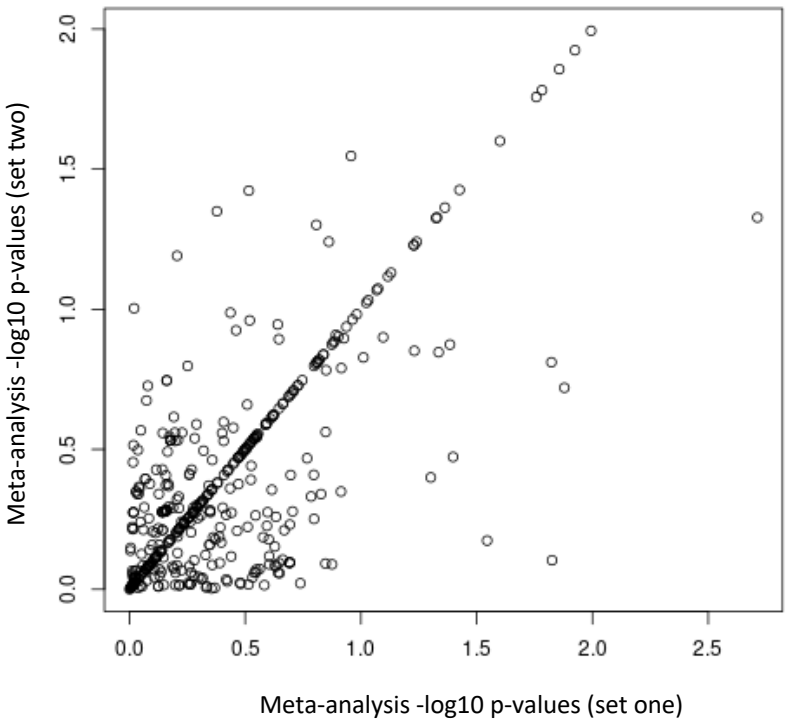


Figure 3.14: Comparison of the p-values from the HLA amino acid joint regression meta-analysis of IPF susceptibility in the UK, UUS and Chicago datasets in set one (X axis) and set two (Y axis).

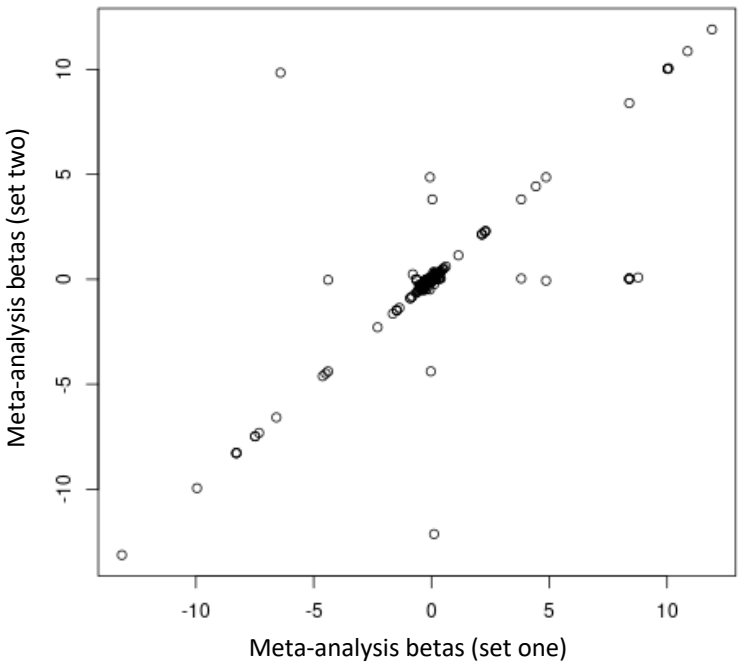


Figure 3.15: Comparison of the betas from the HLA amino acid joint regression meta-analysis of IPF susceptibility in the UK, UUS and Chicago datasets in set one (X axis) and set two (Y axis).

When comparing the -log10 p-values of the 3-way single-variant logistic regression meta-analysis (section 3.6) and the 3-way joint regression meta-analysis there was an attenuation to the null in the joint regression of both set one and set two (figure 3.18 A and B). There was a similar pattern with the effect sizes (figure 3.19 A and B). This could be due to a reduction in power in the joint regression analysis.

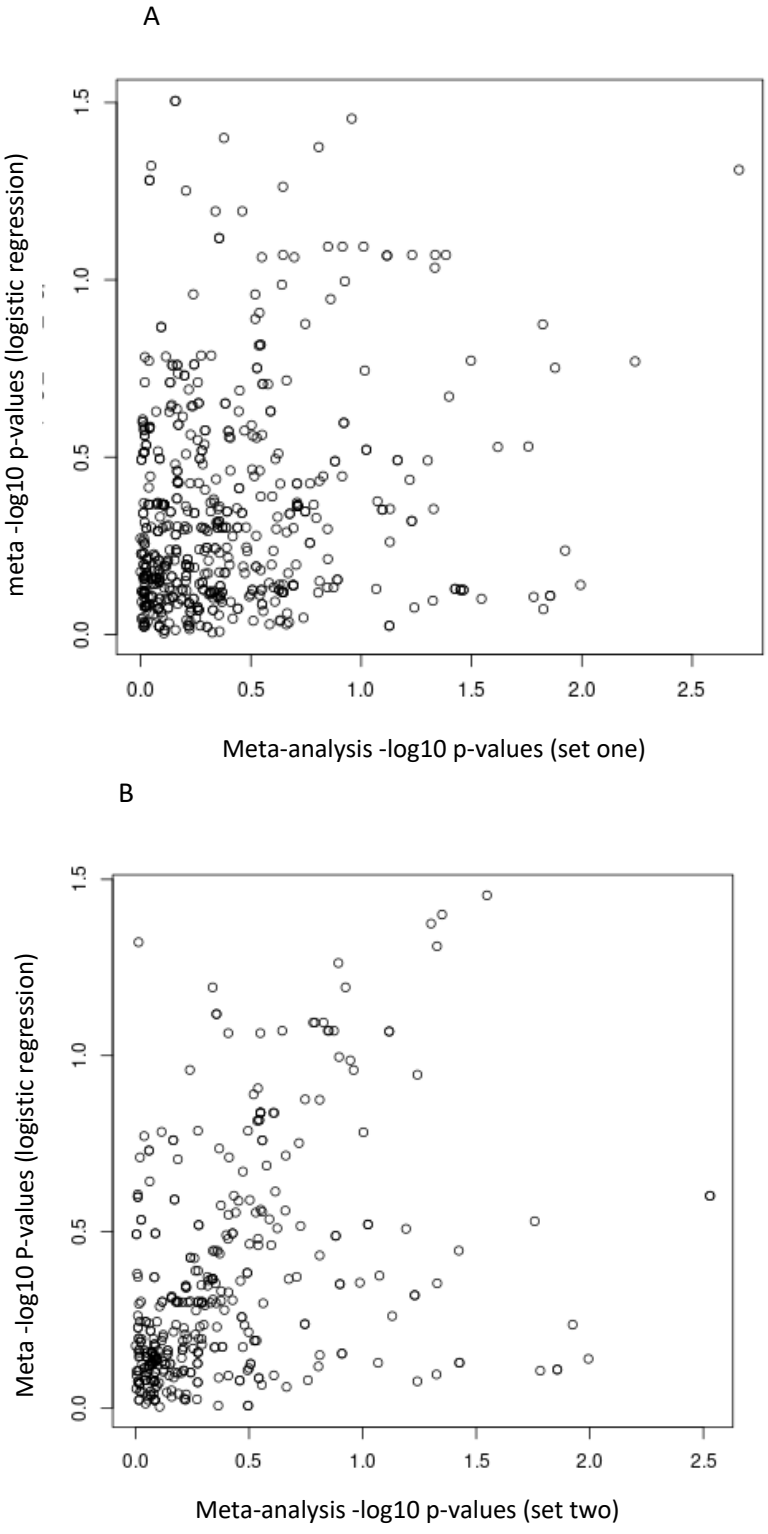
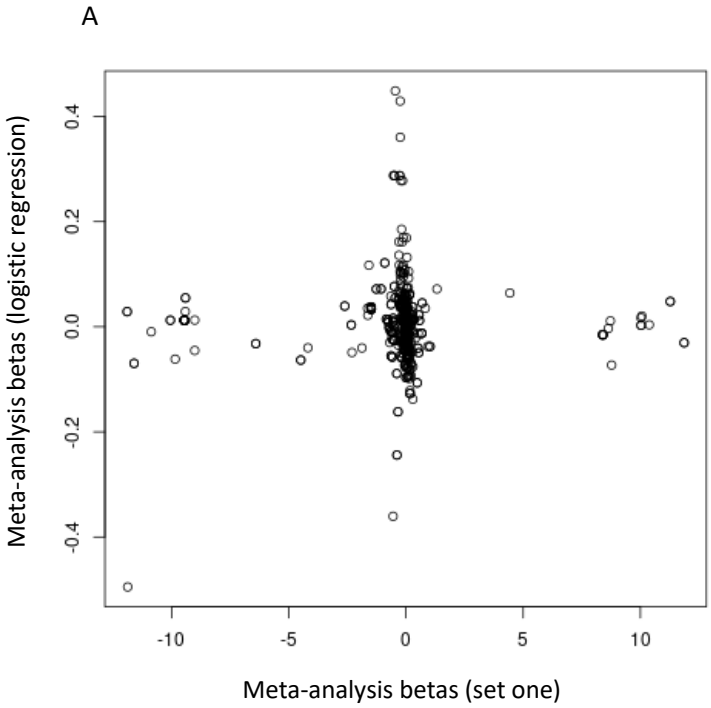


Figure 3.16: Comparison of p-values from the joint regression of amino acids from the three-way meta-analysis of IPF susceptibility in set one (full amino acid set) and set two (with both rare amino acids and most common at each loci removed) and the logistic regression in the UK, UUS and Chicago datasets.



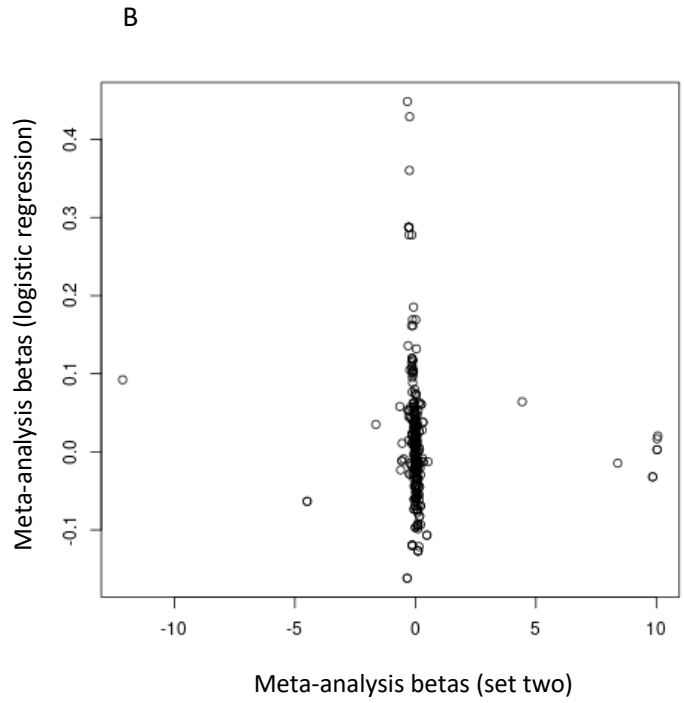


Figure 3.17: Comparison of betas from the joint regression of amino acids from the three-way meta-analysis of IPF susceptibility in set one (full amino acid set) and set two (with both rare amino acids and most common at each loci removed) and the logistic regression in the UK, UUS and Chicago datasets.

3.7 Discussion:

The aim of this chapter was to replicate the previously identified *HLA-DQB1*06:02* and to identify novel signals in the HLA region in IPF susceptibility. In this chapter the HLA region was analysed in the largest IPF study to date. Both HLA specific imputation and the most up to date SNP imputation was used to provide a high coverage of SNPs in the region and enabled fine mapping of signals to specific HLA gene alleles or amino acid alleles. The previously identified *HLA-DQB1*06:02* signal did not replicate in an analysis of the UK and Chicago datasets despite there being adequate power to detect the previously reported effect size in the UK dataset (98% power) and in the meta-analysis of the UK and Chicago datasets (100% power) (see section 3.5.1). The signal was replicated by Fingerlin *et al* 2016 in an independent set of cases and controls (5), however this was a similar group of fibrotic idiopathic interstitial pneumonias. IPF was diagnosed in all three studies using the American Thoracic Society and European Respiratory Society guidelines (195-197) however in the Colorado dataset there were also individuals diagnosed with non-specific pneumonia, cryptogenic organising pneumonia, respiratory bronchiolitis-associated interstitial lung disease and desquamative interstitial pneumonia and also those with unclassified interstitial pneumonia (4). In all, these alternative interstitial pneumonias make up 23% of the cases included in this dataset, it is possible that one of these groups is responsible for the *HLA-DQB1*06:02* signal because this is not seen when restricting to IPF. It is possible that the fIIP cases included some lung disease related to an auto-immune phenotype for example Rheumatoid Arthritis association interstitial lung disease (RA-ILD) in which *HLA-DRB1* and *HLA-DQB1* haplotypes and alleles (including *HLA-DQB1*06:02*) have previously been associated. (198, 199).

Given the differences in case definition in the Colorado dataset, compared to the other datasets, it was removed from the meta-analysis. In the IPF meta-analysis of the UK, Chicago and UUS, one signal in the HLA region passed the Bonferroni corrected threshold of $P<2.8x10^{-6}$. rs3132684 was a common variant in an intron of *ZNRD1ASP.* The variant (and variants in the 95% credible set) were associated with several respiratory, autoimmune, and inflammatory traits in Phenoscanner and it was associated with the differential expression of several HLA and non-HLA genes in tissues in GTEx. SNPs in the credible set were associated with the differential expression of several HLA and non-HLA genes in the lung, in most cases the eQTL SNP was only in weak linkage disequilibrium with the lead SNP from the association meta-analysis, there was suggestive evidence for colocalization with the eQTL SNPs in HLA-H and HLA-W however these are pseudogenes and identifying their role in disease processes would be challenging. Utilising the newly available UUS dataset instead of the Colorado dataset led to

identification of a new IPF susceptibility signal. More stringent quality control measures (higher allele frequency and imputation quality thresholds) could have been applied, however there would be limited power to detect rare signals in these analyses. All available datasets were including in the meta-analysis (to maximise power) however this meant that there was no data available for an independent replication stage. To partly address this, internal association P-value thresholds were applied to each dataset to ensure signals were not driven by one dataset.

To test the effect of all amino acids at each position simultaneously, a joint regression model was used. No amino acids passed the Bonferroni corrected significance threshold in the 3-way meta-analysis of set one or set two. Figure 3.18 A and B showed that the results from analysing these data using joint regression (section 3.7) and the logistic regression (section 3.6) were loosely similar however there appeared to be an attenuation to the null in the joint regression, suggesting there was power limitation. In the analysis of set two, the most frequent amino acids at each loci (with more than three alleles) were removed, however this significantly reduced the number of amino acid alleles that were analysed (441 in the meta-analysis) and HLA-DPA1 was completely removed. There was an underinflation of the test statistic in the 3-way meta-analyses of set one and set two, this suggested that the amino acids were less associated with IPF susceptibility than what would be expected, this could be due to power since joint regression analyses have lower power than logistic regression analyses.

The aims of this chapter were to replicate the previously identified *HLA-DQB1*06:02* signal and identify novel signals across the region. In summary, there were two main findings in this chapter, the *HLA-DQB1*06:02* signal did not replicate in any of the three available datasets, suggesting this signal may provide evidence for a role of HLA variation in fIIP but not IPF. Secondly, a common signal near *ZNRD1ASP* was consistently statistically significant across all three datasets in the meta-analysis, the credible set of this signal provided possible associations with respiratory, inflammatory and autoimmune disorders as well as acting as eQTLs in lung tissue. Suggesting that there may be a role for non-HLA genes in the HLA region in IPF susceptibility. *ZNRD1ASP* is a transcribed pseudogene, however little is known about the protein's specific function. Analysing amino acid alleles in the HLA region in IPF using a joint regression model provided no further insight into the role of HLA amino acids on IPF susceptibility and suggests that they are not associated with IPF susceptibility at an appreciable level in the UK, UUS and Chicago cohorts.

Chapter 4: SNP-SNP interaction analyses of variants in the HLA region and the *MUC5B* risk allele in IPF susceptibility

4.1 Introduction:

The most well characterised SNP association with IPF susceptibility is rs35705950 in the promoter region of the mucin gene *MUC5B* (2-4, 49, 200-202). *MUC5B* is the largest genetic risk factor for IPF with the *MUC5B* risk allele found in 30-35% of IPF cases and 11% of controls and each copy of the risk allele is associated with a 5-fold increase in odds of IPF (56). A recent study suggests 5.9-9.4% of IPF risk is explained by the *MUC5B* risk SNP (203). Despite the effect size of this SNP, not all individuals who carry the risk allele go on to develop IPF and not everyone with IPF has the *MUC5B* risk allele. Research in mouse models suggests that excess mucin will retain bacteria and cells in the lungs for an extended period of time, promoting the abnormal healing response associated with IPF (58). If carrying one or more *MUC5B* risk allele leads to increased mucin which in turn increases risk of IPF, we may hypothesise that those who don't carry *MUC5B* risk alleles might have alternative genetic risk factors that increase their risk of developing IPF. Alternatively, those who do carry one or more *MUC5B* risk alleles may have additional genetic risk factors contributing to their overall IPF risk. Standard association studies may not be able to detect these effects, therefore SNP**MUC5B* interaction analyses are needed.

The aim of this chapter was to identify whether the contribution of HLA variation to IPF risk is dependent on *MUC5B* risk allele status. This chapter describes a discovery HLA-wide variant**MUC5B* interaction analysis in 612 IPF cases and 3,366 controls with replication in 2,308 fibrotic idiopathic interstitial pneumonia (fIIP) cases and 14,683 controls (discovery and replication study design described in section 4.3.2). To maximise power, a 3-way interaction meta-analysis was undertaken in 2,920 IPF and fIIP cases and 18,049 controls (meta-analysis study design described in section 4.3.3).

4.1.1 Summary of Idiopathic Pulmonary Fibrosis Datasets:

The analyses in this chapter were performed using three IPF datasets; UK, Colorado and UUS. All three of these studies diagnosed IPF using the American Thoracic Society and European Respiratory Society guidelines (195, 197, 204) and were imputed using the HRC Reference Panel and HLA reference panel as described in Chapter 2. Briefly, the UK dataset is comprised of 612 IPF cases and 3,366 controls and has 36,743 well imputed variants (allele frequency

>1%, imputation quality >0.4) for analysis. The Colorado dataset is comprised of 1,515 fibrotic interstitial pneumonia cases and 4,683 controls with 34,905 well imputed variants (allele frequency >1%, imputation quality >0.4) for analysis. Finally, the UUS dataset has 793 IPF cases and 10,000 controls with 36,965 well imputed variants (allele frequency >1%, imputation quality >0.4) for analysis. The Chicago dataset was not included in this analysis because the *MUC5B* SNP was not well imputed (imputation quality 0.4). The *MUC5B* SNP was well imputed in all other datasets (imputation quality of 0.92 in UK, 0.77 in Colorado and directly genotyped in UUS). Despite being removed from the previous analysis (see chapter 3, section 3.6), Colorado was included in the analyses in this chapter to increase sample size because interaction analyses require larger sample sizes than marginal effects association analyses.

4.2 Methods:

4.2.2 Genome-wide SNP*_MUC5B_ risk allele interaction analysis:

Interactions between variants in the HLA region and the *MUC5B* risk SNP were tested using Plink V1.9 with sex and 10 principal components as covariates. The *MUC5B* risk allele was analysed using a dominant model; 1 = no *MUC5B* risk alleles, 2 = at least one *MUC5B* risk. The interaction model was as follows:

$$IPF \sim \beta_0 + \beta_1 G + \beta_2 M + \beta_3 G^* M$$

Where β_1 was the effect size for a HLA variant G (under an additive model) for those with no *MUC5B* risk allele, β_2 was the effect size for the *MUC5B* SNP M (under a dominant model) and β_3 was the effect for the interaction between G and M (G*M). Because all HLA gene alleles and amino acid alleles are modelled as biallelic variants (i.e. presence or absence), they can be modelled in the same way as SNPs.

4.2.3 Discovery and replication study design:

Variants in the discovery dataset were tested for an interaction with *MUC5B* using the model described above. Signals that passed a Bonferroni corrected significance threshold (described previously in section 3.2.3) in the discovery dataset were tested for association in a replication dataset. If no variants passed the Bonferroni corrected threshold, $P<5x10^{-3}$ was used to identify suggestive signals to follow up. Variants were required to pass a Bonferroni corrected significance threshold (corrected for the number of variants tested) in order to be described as statistically significant. Those that passed these thresholds were identified as independent by excluding those with an r^2 of more than 0.2 with the lead variant.

All variants were plotted on a Manhattan plot using qqman (189) and independent signals were visualised in region plots (created using Locuszoom (190) in python)

4.2.4 Meta-analysis study design:

Variants that passed quality control filters (allele frequency >1%, imputation quality >0.4) in the UK, UUS and Colorado dataset (see Chapter 2) were tested for an interaction with *MUC5B* risk allele status. A fixed effects weighted meta-analysis was performed on these data to provide a weighted P-value, beta and standard error for each variant. Variants were required to be present in at least two of the three studies to be included in the meta-analysis. Following meta-analysis, signals were also required to be in the same direction of effect and pass P<0.05 in at least two of the studies to be deemed as significant. Those that passed these significance thresholds were identified as independent from one another by excluding those with an r^2 of more than 0.2 with the lead variant.

All variants were plotted on a Manhattan plot using qqman (189) and independent signals were visualised in region plots (created using Locuszoom (190) in python).

4.2.5: Association effect sizes in individuals with and without MUC5B risk alleles for variants identified in interaction analyses

To estimate the effect of the interaction signals on *MUC5B* positive and *MUC5B* negative individuals independently, a stratified HLA-wide association meta-analysis was undertaken. Individuals were stratified based on if they had no *MUC5B* risk allele or at least one *MUC5B* risk allele and variants that had been significant in the interaction analyses were tested for association with IPF susceptibility in each *MUC5B* group separately. Each variant was tested assuming an additive model with 10 principal components (to adjust for fine-scale population structure) and sex included as covariates.

4.2.6: In silico characterisation of signals

Lead variants of signals identified in the analysis and variants in high LD (r^2>0.8) (SNPs or HLA alleles [amino acid alleles cannot be used as a search term]) were investigated using Phenoscanner (192) to identify associations with respiratory, autoimmune, inflammatory or immunity phenotype. Signals identified in the interaction analyses were said to be associated with a phenotype if it met a $P<5x10^{-8}$ threshold. GTEx consortium (n=15,253) was used to identify if the signals were associated with gene expression in one or more tissues or cell-types (193) (SNPs only).

4.3 Results: SNP-SNP interaction discovery and replication analysis of *MUC5B* risk allele status and the HLA region in IPF susceptibility:

4.3.1 Introduction:

A discovery interaction analysis between *MUC5B* risk allele (rs35705950) status and variants in the HLA region was completed using the 612 IPF cases and 3,366 controls from the UK IPF dataset. Any novel signals were tested for replication in two independent datasets (Colorado and UUS).

4.3.2 Results: Discovery Analysis

In the UK IPF dataset 36,743 well imputed variants were analysed for an interaction effect with *MUC5B* risk allele status on IPF susceptibility (figure 4.1).

No variants passed the Bonferroni significance threshold of $P<2.8x10^{-6}$, but there were seven independent variants that passed the suggestive significance threshold of $P<5x10^{-3}$ (figure 4.1, table 4.1). Three signals identified in this analysis were found in HLA class I and the remaining four were in class II. All seven signals were very well imputed (imputation quality between 0.93 and 1, table 4.1) apart from rs6299441 a coding variant in *MUC22* which has a relatively low imputation quality of r2=0.52. SNPs rs62399441 and rs62407874 had low coded allele frequencies (<5%). SNP rs62399441, a missense coding variant in *MUC22,* was in weak linkage disequilibrium with classical HLA allele *HLA-B*50:01* (R^2=0.25). HLA-B*50:01 itself was not tested in the interaction analysis because it was rare (0.7%).

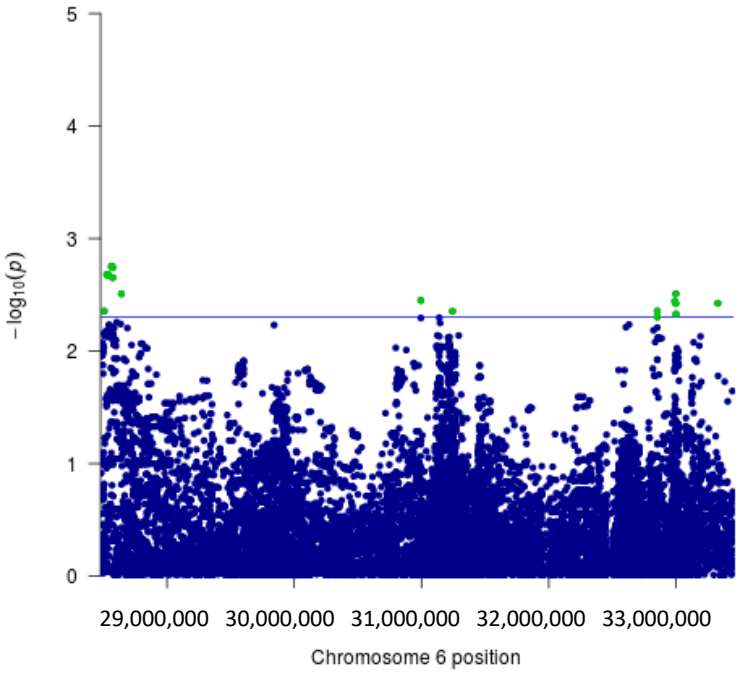


*Figure 4.1: Manhattan plot of the HLA-wide variant*MUC5B interaction analyses of IPF susceptibility in the UK IPF dataset (the green variants are all the variants that passed the suggestive significance threshold). Blue line is suggestive significance threshold of* $P<5x10^{-3}$*).*

*Table 4.1: Independent signals ($P<5x10^{-3}$) in the HLA-wide variant*MUC5B interaction analyses of IPF susceptibility in the UK IPF dataset using two imputation panels.*

RSID	**Nearest gene**	**Position – build 37 (HLA class)**	**Coded/ non-coded allele**	**Info**	**Coded allele freq**	**P-value**	**OR (95% CI)**
rs418204	ZBED9	28562957 (Class I)	A/G	1.00	0.33	$1.77x10^{-3}$	1.55 (1.18-2.04)
rs9277029	HLA DOA (-21417), HLA-DPA1 (+33540)	32998806 (Class II)	T/C	1.00	0.27	$3.10x10^{-3}$	0.65 (0.48-0.86)
rs62399441	MUC22	30995078 (Class I)	A/G	0.52	0.03	$3.55x10^{-3}$	0.23 (0.08-0.62)
rs62407874	MYL8P (-21399), LYPLA2P1 (+3830)	33328671 (Class II)	T/C	0.93	0.02	$3.76x10^{-3}$	4.79 (1.67-13.85)
rs7767277	PPP1R2P1 (-5191), LOC100294145 (+8911)	32853042 (Class II)	A/C	1.00	0.08	$4.41x10^{-3}$	0.49 (0.30-0.80)
rs56043329	GPX5 (-30457) 28,539,407 SCAND3 (+33222)	28506185 (Class I)	C/A	1.00	0.07	$4.41x10^{-3}$	2.23 (1.29-3.85)
rs13200569	USP8P1	31243615 (Class I)	A/G	1.00	0.28	$4.43x10^{-3}$	0.66 (0.49-0.88)

4.3.3 Results: Replication Analysis

The seven independent variants that reached a suggestive significance threshold of $P<5x10^{-3}$ (table 4.1) for an interaction with *MUC5B* risk allele status in the UK dataset were tested for replication in the Colorado and UUS datasets (2,308 cases and 14,683 controls). All the signals identified in the *MUC5B* interaction in the UK dataset passed quality control measures in the replication in the Colorado and UUS datasets (imputation quality>0.4, coded allele frequency>0.01) (table 4.2). However, none of these signals passed the Bonferroni corrected threshold for this replication analysis (seven variants, $P<0.007$) (table 4.2). Only one of the seven signals in the replication was in the same direction of effect (rs9277029), however all the signals had wide confidence intervals in this replication analysis (table 4.2). SNP rs62399441 was of low minor allele frequency and was weakly imputed in both the Discovery (quality=0.52) and Replication studies (quality=0.45 and 0.47).

*Table 4.2: Replication results from the Colorado and UUS datasets of the novel findings from the HLA-wide variant*MUC5B interaction analyses of IPF susceptibility in the UK IPF dataset.*

Dataset	Rsid	Position – build 37 (HLA class)	Nearest Gene	Coded/ Non-coded allele	Info Score	Coded allele frequency	P-Value	OR (95% CI)	Meta P-value	Meta OR (95% CI)
Colorado	rs418204	28562957 (Class I)	*ZBED9*	A/G	1.00	0.31	0.42	0.92 (0.76-1.12)	0.43	0.93 (0.77-1.12)
UUS					1.00	0.34	0.97	0.98 (0.46-2.1)		
Colorado	rs9277029	32998806 (Class II)	*HLA-DOA* (-21417), *HLA-DPA1* (+33540)	T/C	1.00	0.26	0.44	0.92 (0.75-1.13)	0.27	0.89 (0.73-1.09)
UUS					1.00	0.26	0.16	0.57 (0.26-1.26)		
Colorado	rs62399441	30995078 (Class I)	*MUC22*	A/G	0.45	0.02	0.50	1.40 (0.52-3.76)	0.55	1.33 (0.52-3.39)
UUS					0.47	0.01	0.88	0.80 (0.04-16.34)		
Colorado	rs62407874	33328671 (Class II)	*MYL8P* (-21399), *LYPLA2P1* (+3830)	T/C	0.85	0.03	0.25	0.73 (0.42-1.23)	0.33	0.77 (0.45-1.31)
UUS					0.92	0.02	0.62	1.64 (0.23-11.7)		
Colorado	rs7767277	32853042 (Class II)	*PPP1R2P1* (-5191) , *LOC100294145* (+8911)	A/C	1.00	0.09	0.77	1.05 (0.76-1.44)	0.77	1.05 (0.77-1.42)
UUS					1.00	0.08	0.96	1.03 (0.29-3.66)		
Colorado	rs56043329	28506185 (Class I)	*GPX5* (-30457) 28,539,407 *SCAND3* (+33222)	C/A	1.00	0.07	0.74	0.94 (0.66-1.35)	0.83	0.96 (0.68-1.36)
UUS					1.00	0.07	0.67	1.38 (0.31-6.09)		
Colorado	rs13200569	31243615 (Class I)	*USP8P1*	A/G	0.99	0.30	0.22	1.13 (0.93-1.38)	0.14	1.15 (0.95-1.40)
UUS					1.00	0.28	0.29	1.50 (0.71-3.20)		

4.4 Results: An interaction meta-analysis of *MUC5B* risk allele status and SNPs in the HLA region in IPF susceptibility in the UK, UUS and Colorado IPF datasets:

4.4.1 Introduction

For all three available studies (UK, UUS and Colorado), individual HLA-wide variant**MUC5B* interaction analyses for IPF susceptibility were undertaken and HLA-wide meta-analysis performed. The *MUC5B* SNP was poorly imputed in the Chicago dataset (imputation quality 0.4) and so Chicago was not included.

4.4.2 Results

A total of 2,920 IPF cases and 18,049 controls with 36,743 variants from the UK dataset (figure 4.3), 34,905 variants from the Colorado dataset (supplementary figure 4.1) and 36,965 variants from the UUS dataset (supplementary figure 4.2) were included in the *MUC5B* interaction meta-analysis of IPF susceptibility in the HLA region (figure 4.6). In total 40,376 variants, 178 HLA alleles and 985 amino acid changes were tested for association with IPF susceptibility in this meta-analysis. Twenty-four percent of variants had a coded allele frequency of less than 5% (figure 4.2). The average imputation quality in this data set was 0.99 (0.99 for SNPs, 0.98 for HLA alleles and 0.99 amino acids) and 85% of variants had an imputation quality of over 0.98 (figure 4.3).

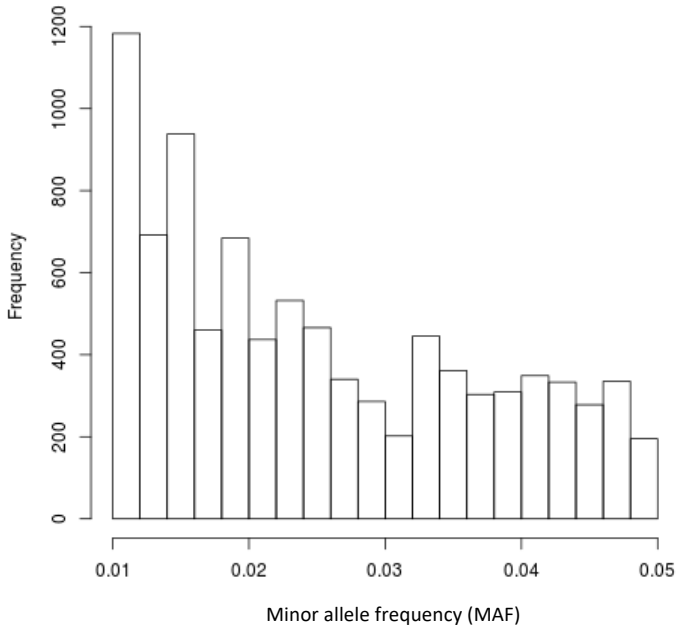


*Figure 4.2: Histogram of minor allele frequencies < 5% of variants in the meta-analysis of the HLA-wide variant*MUC5B interaction analyses of IPF susceptibility from the UK, UUS and Colorado datasets.*

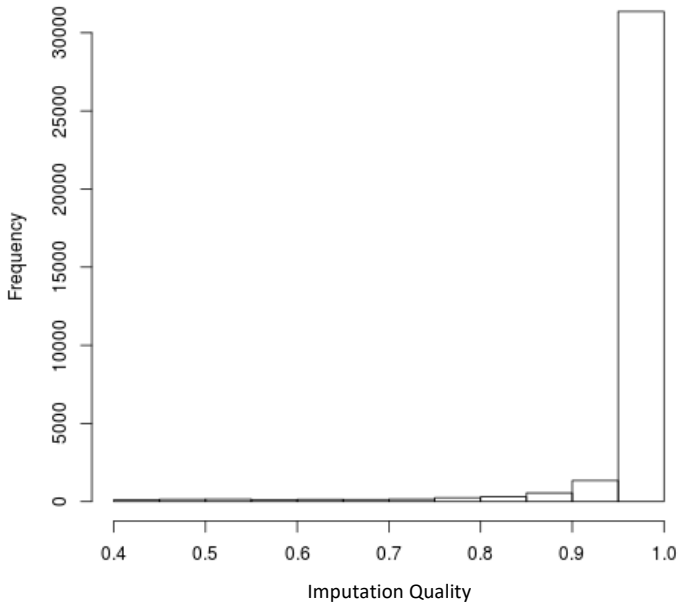


*Figure 4.3: Histogram of imputation qualities of variants in the meta-analysis of the HLA-wide variant*MUC5B interaction analyses of IPF susceptibility from the UK, UUS and Colorado datasets.*

No signals passed $P<3.4x10^{-6}$ in the meta-analysis of all 3 studies. Although 12 independent signals passed a suggestive significance threshold of P<5x10-3 (supplementary table 4.1), only five of these had P<0.05 and in the same direction of effect in at least two of the three studies (table 4.3). The five suggestively significant signals, which were all SNPs, were taken forward for characterisation. None of the five SNPs were in linkage disequilibrium with any HLA gene alleles or amino acid alleles.

To identify the effect of the suggestive variants (table 4.3) in *MUC5B* positive (one or more rs35705950 risk alleles, sample size: 5,323) and *MUC5B* negative (no copies of the rs35705950 risk allele, sample size: 15,757) individuals, the variants were tested for association with IPF susceptibility separately in those with and those without *MUC5B* risk alleles (supplementary table 4.3). When stratifying the *MUC5B* positive and negative groups, the interaction signal for the five suggestive variants appeared to be driven by opposing genetic effects in each group (supplementary table 4.3). The opposing genetic effects were statistically significant for variants rs145912914, rs9265961 and rs7774158. rs145912914 was significantly associated with higher odds of IPF in *MUC5B* negative individuals and significantly associated with lower odds of IPF in *MUC5B* positive individuals. rs9265961 was significantly associated with lower odds of IPF in the *MUC5B* negative group. rs7774158 was significantly associated with lower odds of IPF in the *MUC5B* positive individuals and was not associated with higher odds of IPF in *MUC5B* negative group.

The results of the phenoscanner search and eQTL search in GTEx are presented in supplementary table 4.2 and supplementary table 4.3. Three of the five signals identified in this analysis were found to be eQTLs for class I, class II or class III HLA genes in lung tissue and in tissues around the body. rs9265961 was found to be involved in the differential expression of several class I and class III HLA genes, for example *HLA-C* and *MICA* in the lung (supplementary table 4.3, supplementary figure 13).

Four of the five signals were also found to be associated with respiratory or immunity related phenotypes identified in a look-up using phenoscanner (supplementary table 4.3). For example, the top signal identified in this analysis (rs145912914, supplementary figure 4.3) was associated with self-reported ankylosing spondylitis and rs309115 and rs7774158 were associated with rheumatoid arthritis. rs9265961 (supplementary figure 4.6) was associated eosinophil counts, forced vital capacity (FVC) and forced expiratory volume in 1 second (FEV_1) and several autoimmune disorders or immunity related phenotypes including coeliac disease and psoriasis.

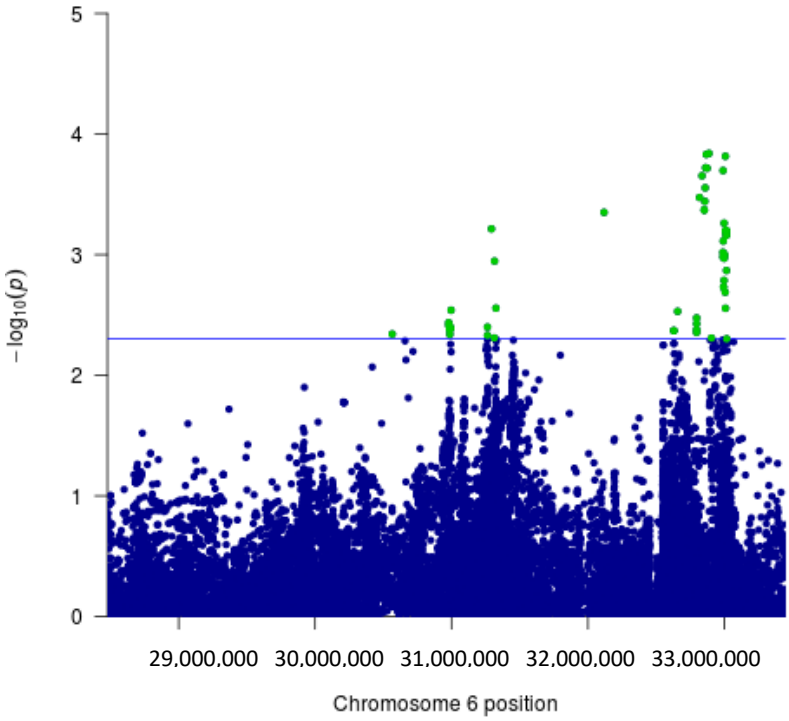


*Figure 4.4: Manhattan plot of the meta-analysis of the HLA-wide variant*MUC5B interaction analyses of IPF susceptibility in the UK, UUS and Colorado IPF datasets.*

*Table 4.3: The five independent signals ($P<5x10^{-3}$) in an interaction meta-analysis of the HLA-wide variant*MUC5B interaction analyses of IPF susceptibility in the UK, UUS and Colorado IPF datasets.*

Dataset	**rsid**	**Position – build 37 (HLA class)**	**Nearest Gene**	**Coded/ Non-coded allele**	**Info Score**	**Coded allele frequency**	**P-Value**	**OR (95% CI)**	**Meta-Analysis P-value**	**Meta-Analysis OR (95% CI)**
UK	rs145912914	32887150 (Class II)	*LOC100294145*	G/C	0.99	0.02	$7.70x10^{-3}$	0.30 (0.12-0.72)	$1.45x10^{-4}$	0.44 (0.29-0.67)
Colorado					0.99	0.02	0.036	0.53 (0.29-0.96)		
UUS					0.99	0.02	0.045	0.43 (0.19-0.98)		
UK	rs7774158	33007752 (Class II)	*HLA-DOA*	C/A	1.00	0.35	$9.42x10^{-3}$	0.70 (0.53-0.92)	$1.53x10^{-4}$	0.78 (0.69-0.89)
Colorado					1.00	0.35	0.065	0.84 (0.69-1.01)		
UUS					1.00	0.34	0.020	0.76 (0.60-0.96)		
UK	rs753712672	31292555 (Class I)	*LOC112267902*	A/C	0.56	0.22	$7.27x10^{-3}$	0.65 (0.48-0.89)	$6.10x10^{-4}$	0.70 (0.57-0.863)
Colorado					NA	NA	NA	NA		
UUS					0.57	0.21	0.028	0.74 (0.56-0.97)		
UK	rs9265961	31315501 (Class I)	*LOC112267902*	A/G	1.00	0.34	0.847	0.97 (0.74-1.28)	$1.13x10^{-3}$	0.80 (0.71-0.92)
Colorado					0.80	0.29	$6.16x10^{-3}$	0.76 (0.63-0.93)		
UUS					1.00	0.33	0.019	0.76 (0.60-0.96)		
UK	rs3909115	30993188 (Class I)	*MUC22*	A/C	0.99	0.25	0.578	1.09 (0.81-1.46)	$4.13x10^{-3}$	1.23 (1.07-1.41)
Colorado					0.99	0.27	0.038	1.24 (1.01-1.51)		

UUS					0.99	0.25	0.031	1.32 (1.03-1.71)		

4.4.3 Summary:

A meta-analysis of HLA-wide variant**MUC5B* interaction analyses for IPF susceptibility was performed that incorporated all available datasets to perform the most powerful *MUC5B* interaction meta-analysis in the HLA region for IPF susceptibility done to date. No signals passed $P<3.4x10^{-6}$ but five signals had a suggestive meta P-value ($P<5x10^{-3}$) and were nominally significant (P<0.05) in at least two of the three studies (table 4.3). Of these five, three of the signals had statistically significant opposing effects on IPF susceptibility in the *MUC5B* positive and *MUC5B* negative groups (supplementary table 4.3).

4.5 Discussion:

The aim of this chapter was to study the interaction between variants in the HLA region and the *MUC5B* risk allele in IPF susceptibility. A discovery analysis in the UK IPF dataset with replication in the Colorado and UUS datasets did not yield any signals that met a HLA-wide significance threshold in the discovery or Bonferroni corrected threshold in the replication study. No signals passed the Bonferroni corrected threshold of $P<2.8x10^{-6}$ in these analyses, however there were five signals that had support from at least two studies which suggests these may be true positives but they require larger sample sizes (and increased power) to confirm. Interestingly, two of these variants (rs145912914 and rs7774158) appeared to have an opposite direction of effect on IPF risk in *MUC5B* positive and *MUC5B* negative individuals.

Alleles associated with a reduced risk of IPF identified in these analyses were associated with reduced expression of *MICA, HLA-C, XXbac-BPG181B23.7 and LY6G5B. MICA* is a stress-induced self-antigen, associated with human cytomegalovirus infection and Spondylitis and Psoriatic Arthritis risk and is known to be involved in natural killer cell binding. This appears to correspond with a previous study of IPF lungs which identified significantly increased expression of *MICA* in IPF lungs (177). LY6G5B is involved in signal transduction and *XXbac-BPG181B23.7* is a long non-coding RNA. Interestingly, the rs9265961 IPF risk allele was associated with reduced expression of *HLA-C* (despite no *HLA-C* gene alleles in correlation with the signal). An increased expression of *HLA-C* could increase the production of immune-stimulatory cytokines and boost the immune response to infections, this could relate to what is known about the development of IPF.

Along with this, alleles associated with reduced risk of IPF were found to be associated with decreased eosinophil counts, increased forced vital capacity (FVC) and increased forced expiratory volume in one second (FEV_1) which is consistent with IPF pathogenicity (IPF causes a

reduction in lung function and eosinophil counts have been associated with exacerbations for example in chronic obstructive pulmonary disorder (COPD) patients (205)).

This study was limited by power, with only small sample sizes available and one dataset being removed due to the poor measurement of the *MUC5B* SNP. This study was also limited because the *MUC5B* SNP was only analysed at two levels (one copy vs at least one copy). To develop the analysis the SNP could be analysed assuming an additive model (no copies, one copy, two copies) or using dosages to determine if this strategy would increase the power to detect interactions or provide novel interaction signals. Since interaction analyses are lower powered than conventional association analyses, signals would not be expected to pass a stringent Bonferroni corrected threshold with the sample sizes available. This meta-analysis had only 17% power to detect interactions (at a coded allele frequency of 10%, interaction effect size of 1.1, alpha of 0.05). For 80% power to detect these interactions, a sample size of 23,124 cases would be required. This analysis could suggest that there is no interaction between the *MUC5B* SNP and variants in the HLA region on IPF susceptibility, however there were five suggestive signals ($P<3x10^{-3}$, $P<0.05$ in at least two datasets) that could be investigated further when larger sample sizes become available.

This chapter presented the largest HLA-wide variant*MUC5B interaction analysis for IPF risk undertaken to date which identified some suggestive signals that were reported in more than one independent study (but which did not reach HLA-wide significance).

Chapter 5: The imputation of variation within the Killer Immunoglobulin like Cell Receptor (KIR) region in four Idiopathic Pulmonary Fibrosis (IPF) datasets

Introduction:

Genes within the KIR region encode proteins found on the surface natural killer cells that detect foreign bodies presented on the surface of virally infected cells to initiate the immune response to bacterial and viral infection. They are also known to physically interact with HLA proteins (see Chapter 1 section 1.3.2). Since viral infection is thought to be a cause of micro-injury to the lung that triggers a fibrotic response in some individuals (7, 8, 60-63), the KIR region may provide further insight into the biological processes that underlie susceptibility to IPF. The KIR region harbours a considerable amount copy number variation which cannot be adequately measured using SNP genotypes (see Chapter 1: Introduction section 1.3.2). Analogous to the approach described in previous chapters for HLA region variation (Chapter 2), bespoke KIR imputation enables us to infer KIR gene copy number variation and KIR haplotypes which can be tested for association with IPF susceptibility.

The aim of this chapter was to impute KIR gene haplotypes and copy number variation using SNP data. Imputation usually uses directly genotyped variants to impute new variants, but this depends on the tag SNPs in the imputation panel being reliably genotyped in the input data set. Alternatively, the KIR imputation tag SNPs could be imputed to a high quality using standard SNP imputation and then the imputed KIR imputation tag SNPs can be used as input to the KIR imputation. This chapter describes an evaluation of the use of directly genotyped vs haplotype reference consortium (HRC) imputed SNPs to improve the accuracy of the KIR imputation in 612 IPF cases and 3,366 controls. The evaluated method was then applied to the final imputation approach for the other three IPF datasets to provide a set of KIR-imputed IPF datasets for association testing in chapter 6.

Methods:

IPF datasets:

The four IPF datasets are described in Chapter 2 section 2.2. Briefly, the UK IPF dataset was comprised of 612 IPF cases and 3,366 controls of European ancestry. The UUS dataset was comprised of 793 IPF cases and 10,000 controls. The Colorado dataset was comprised of 1,515 fibrotic idiopathic interstitial pneumonia (fIIP) cases and 4,683 controls. Finally, the Chicago dataset was comprised of 500 IPF cases and 510 controls.

SNP Imputation using the Haplotype Reference Consortium (HRC) panel:

Phasing and genome-wide SNP imputation using the Haplotype Reference Consortium (HRC) 1.1 panel (29) was previously completed on all IPF data sets for another study using the Michigan imputation server (206). This imputed data was then used in the KIR*IMP imputation.

KIR imputation panel:

KIR*IMP was the method previously developed for imputing KIR gene copy number variation and KIR haplotypes (152). This method used 301 SNPs from a UK reference panel comprising 698 individuals of European ancestry from the UK DNA banking network (DBN) which contained individuals who were selected for having atopic dermatitis or asthma (152). Copy number variation in genes *KIR2DL1, KIR2DL2, KIR2DL3, KIR2DL4, KIR2DL5, KIR2DP1, KIR2DS1, KIR2DS2, KIR2DS3, KIR2DS4DEL, KIR2DS4TOTAL, KIR2DS4WT, KIR2DS5, KIR3DL1ex4, KIR3DL1ex9, KIR3DP1* and *KIR3DS1* and KIR haplotypes 1-69 were imputed using the imputation panel. *KIR3DL1ex4* and *KIR3DL1ex9* were copy number variants in exons 4 or 9 in the gene *KIR3DL1*. *KIR2DS4* was split into *KIR2DS4DEL* and *KIR2DS4WT*, which corresponds to a 22-bp frameshift deletion and wild type allele. *KIR2DS4TOTAL* was the number of DS4 genes counting both deleted and wild-type alleles. A/B haplotype denotes stable copy number variation (A haplotype) vs extensive copy number variation (B haplotype).

KIR haplotypes are made up of varying numbers of the KIR genes named above, an example of how the gene copy number variations and haplotypes are related can be seen in table 5.1.

Table 5.1: Relationship between KIR haplotypes. KIR Haplotype was the haplotype classification defined by (148), A/B corresponds to the broad A/B haplotype classification, all the other gene columns show the copy number of each individuals KIR gene. Each KIR Haplotype was defined by the copy number values across each of the 17 KIR genes, table from (152), 0's are in red, 1's are in black and 2's are in blue.

KIR Haplotype	A/B	KIR3DL3	KIR2DS2	KIR2DL2	KIR2DL3	KIR2DP1	KIR2DL1	KIR3DP1	KIR2DL4	KIR3DL1ex4	KIR3DL1ex9	KIR3DS1	KIR2DL5	KIR2DS3	KIR2DS5	KIR2DS1	KIR2DS4TOTAL	KIR2DS4WT	KIR2DS4DEL	KIR3DL2
1	A	1	0	0	1	1	1	1	1	1	1	0	0	0	0	0	1	0	1	1
2	A	1	0	0	1	1	1	1	1	1	1	0	0	0	0	0	1	1	0	1
3	B	1	0	0	1	1	1	1	1	0	0	1	1	0	1	1	0	0	0	1
4	B	1	1	1	0	0	0	1	1	1	1	0	0	0	0	0	1	0	1	1
5	B	1	1	1	0	0	0	1	1	1	1	0	0	0	0	0	1	1	0	1
6	B	1	1	1	0	1	1	1	1	1	1	0	1	1	0	0	1	0	1	1
7	B	1	1	1	0	1	1	1	1	0	0	1	2	1	1	1	0	0	0	1
8	B	1	1	1	0	0	0	1	1	0	0	1	1	0	1	1	0	0	0	1
9	B	1	1	1	0	1	1	1	1	0	0	1	2	2	0	1	0	0	0	1
10	B	1	1	1	0	1	1	1	1	1	1	0	1	1	0	0	1	1	0	1
11	B	1	0	0	1	1	1	1	1	0	0	1	1	1	0	1	0	0	0	1
12	B	1	1	1	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	1
13	B	1	1	1	0	1	1	2	2	1	1	1	1	1	0	0	1	1	0	1
14	B	1	0	0	1	2	2	2	2	0	0	2	2	1	1	1	0	0	0	1
15	B	1	1	0	0	1	1	1	1	0	0	1	1	0	1	1	0	0	0	1
16	B	1	0	0	1	1	1	2	2	1	1	1	0	0	0	0	1	0	1	1
17	B	1	1	1	0	0	0	1	1	0	0	1	1	1	0	1	0	0	0	1
18	B	1	1	1	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	1
19	B	1	1	2	0	0	0	2	2	0	0	2	2	0	2	1	0	0	0	1
20	B	1	0	0	1	1	1	2	2	1	1	1	0	0	0	0	1	1	0	1
21	B	1	1	0	0	1	1	1	1	1	1	0	0	0	0	0	1	0	1	1
22	B	1	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
23	B	1	0	0	1	2	2	2	2	1	1	1	1	1	0	0	1	0	1	1

24	B	1	1	1	0	1	1	1	1	0	0	1	1	1	1	1	0	0	0	1
25	B	1	1	1	0	1	1	1	1	1	1	0	1	0	0	0	1	0	1	1
27	A	1	0	0	1	1	0	1	1	1	1	0	0	0	0	0	1	0	1	1
28	B	1	0	0	1	1	1	2	2	0	0	2	1	0	1	1	0	0	0	1
29	B	1	1	1	0	1	1	2	2	1	1	1	1	1	0	0	1	0	1	1
30	B	1	1	1	0	0	0	2	2	1	1	1	0	0	0	0	1	0	1	1
31	B	1	0	1	0	1	1	1	1	1	1	0	1	0	1	0	1	1	0	1
33	B	1	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1
34	A	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
36	B	1	1	1	0	2	2	2	2	0	0	2	1	2	0	1	0	0	0	1
38	B	1	0	0	1	2	2	2	2	0	0	2	1	0	1	1	0	0	0	1
40	B	1	1	1	0	2	2	1	1	0	0	1	2	1	1	0	0	0	0	1
41	B	1	1	1	0	1	2	1	1	0	0	1	2	2	0	1	0	0	0	1
42	B	1	1	2	0	0	0	2	2	1	1	1	1	0	1	0	1	0	1	1
44	B	1	1	1	0	1	1	1	1	0	0	1	2	0	1	1	0	0	0	1
42	B	1	1	1	0	1	1	1	1	0	0	1	2	1	1	0	0	0	0	1
46	B	1	0	1	0	1	1	1	1	0	0	1	1	1	1	1	0	0	0	1
48	A	1	0	0	1	1	1	1	1	1	2	0	0	0	0	0	2	2	0	1
50	B	1	0	0	0	1	1	1	1	1	1	0	1	1	0	0	1	1	0	1
52	B	1	0	0	0	1	1	1	1	1	1	0	1	0	1	0	1	0	1	1
53	B	1	1	0	1	1	1	1	1	1	1	0	0	0	0	0	1	0	1	1
55	A	1	0	0	0	1	2	1	1	1	1	0	0	0	0	0	1	1	0	1
56	B	1	1	1	0	0	0	0	1	0	0	1	1	0	1	1	0	0	0	1
57	A	1	0	0	1	1	1	0	1	1	1	0	0	0	0	0	1	0	1	1
58	A	1	0	0	1	1	1	1	0	1	1	0	0	0	0	0	1	1	0	1
59	A	1	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1
68	B	1	1	1	0	0	0	0	1	1	1	0	0	0	0	0	1	0	1	1
69	B	1	1	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1

Measures of imputation quality

There were two measures of imputation quality available in the data, average estimated imputation quality and posterior probability. Average estimated imputation accuracy is provided for each gene (covering all the CNVs in the gene) and for KIR haplotype (covering all the haplotypes) across all individuals. This measure provides a global view of the imputation quality across the dataset for each gene and haplotype, this enables the identification of trends in imputation quality.

Posterior probability is the likelihood that the assigned allele was the correct assignment. Therefore, the higher the posterior probability, the more confidence there is that the allele imputed is accurate. Because this measure is available for all assigned alleles, this measure can be used as a quality control measure to remove the low confidence alleles.

Imputation of the KIR region using KIR*IMP:

The alleles and minor allele frequencies of the 301 reference panel SNPs in KIR*IMP were compared to the alleles and minor allele frequencies of the same SNP that were directly genotyped or imputed in the imputation input dataset (HRC-imputed data). SNPs that were strand mismatches or SNPs from the imputation input dataset that had a minor allele frequency (MAF) that was +/- 10% from the MAF in the KIR*IMP panel were removed.

For each IPF dataset, SNPs that passed the pre-imputation quality control described above were uploaded to the KIR*IMP server (V1.2.0, available at http://imp.science.unimelb.edu.au/kir) for imputation.

An imputation strategy evaluation was undertaken in the UK IPF dataset in the first instance in which directly genotyped SNPs and HRC-imputed SNPs were evaluated as input to KIR*IMP in order to identify which would give the best estimated imputation accuracies across the KIR genes and haplotypes. Firstly, directly genotyped SNPs in the UK IPF dataset were checked against the 301 SNPs in the imputation panel, specifically the tag SNPs (SNPs significantly correlated with specific KIR genes) as these have been shown to significantly increase the imputation accuracy of the genes they tag (152). The imputation quality of the KIR genes and haplotypes imputed using directly genotyped SNPs was compared with the published expected imputation qualities when using the Axiom UK Biobank SNPs (152). To determine if there would be more SNPs available for imputation (out of the 301 in the imputation panel), SNPs imputed in the imputation input dataset (from the HRC panel) across several different imputation quality thresholds (info score of 0.3, 0.5, 0.7, 0.8, 0.9, 0.95 and no threshold) were also matched to the 301 SNPs in the imputation panel. SNPs in the imputation input dataset

that passed the pre-imputation quality control (described above) were then used to impute the KIR genes and haplotypes. The imputation accuracy of the KIR genes and haplotypes, and the posterior probability of the most likely allele were compared between the results from the different sets of SNPs used for the imputation input dataset.

KIR variants with an imputation posterior probability of less than 0.5 were removed from the datasets.

Results: Imputation of the KIR haplotypes and gene copy number variants using KIR*IMP

Imputation of the UK IPF dataset with KIR*IMP using directly genotyped SNPs:

A total of 3,172 directly genotyped variants across the KIR region (50,900,000-58,617,616 [19q13.4] on genome build 37) from the UK IPF dataset were phased using Shapeit (v2.837). Only 57/301 SNPs from the KIR imputation panel were directly genotyped in the UK IPF dataset and these were uploaded to the KIR*IMP imputation server. The allele frequency of the 57 SNPs that were used in the imputation were similar in the IPF dataset and the reference panel (figure 5.1). The imputed KIR genes and haplotypes had estimated imputation accuracies above 75% apart from “KIR haplotype” which was 67.6% (figure 5.2). An example of imputed haplotype and copy number variation (CNV) for a single IPF case can be seen in table 5.2, each individual has zero, one or two copies of each gene on each chromosome (maximum copy number of 4 across both chromosomes) and this defines the haplotype for each chromosome. The individual in table 5.2 for example has one copy of KIR haplotype 1 (which is an A haplotype) which is made up of a range of zero and one copies of the KIR genes and one copy of KIR haplotype 9 (which is a B haplotype) which is made up of zero, one and two copies of the KIR genes.

The imputed accuracies from the UK IPF dataset were comparable (and often better) to those calculated for the Axiom UK Biobank array (152) (figure 5.2). The Axiom UK Biobank array had 34/301 SNPs that overlapped with the SNPs used for imputation (152), the UK IPF dataset had 57 SNPs; this could explain the increased accuracy of the KIR imputation in the UK IPF dataset (figure 5.2).

Table 5.2: An example of results for each chromosome for a single IPF case. For the KIR haplotype locus, the "imputed type" column corresponds to the KIR haplotype number. For the A/B locus, the "imputed type" column corresponds to haplotype A or B. For the remaining KIR genes, the "imputed type" column corresponds to the copy number variation.

Locus	**Chromosome 1**		**Chromosome 2**	
	Imputed Type	**Posterior Probability**	**Imputed Type**	**Posterior Probability**
KIRhaplotype	1	0.633	9	0.446
A/B	A	0.626	B	0.983
KIR2DS2	0	0.613	0	0.656
KIR2DL2	0	0.662	0	0.654
KIR2DL3	1	0.626	1	0.636
KIR2DP1	1	0.616	1	0.954
KIR2DL1	1	0.594	1	0.96
KIR3DP1	1	0.975	1	0.985
KIR2DL4	1	0.977	1	0.982
KIR3DL1ex4	1	0.996	0	0.995
KIR3DL1ex9	1	0.999	0	0.993
KIR3DS1	0	1	1	0.941
KIR2DL5	0	0.939	1	0.556
KIR2DS3	0	0.933	2	0.477
KIR2DS5	0	0.997	0	0.661
KIR2DS1	0	0.992	1	0.999
KIR2DS4TOTAL	1	0.997	0	0.993
KIR2DS4WT	0	0.999	0	0.999
KIR2DS4DEL	1	0.994	0	0.99

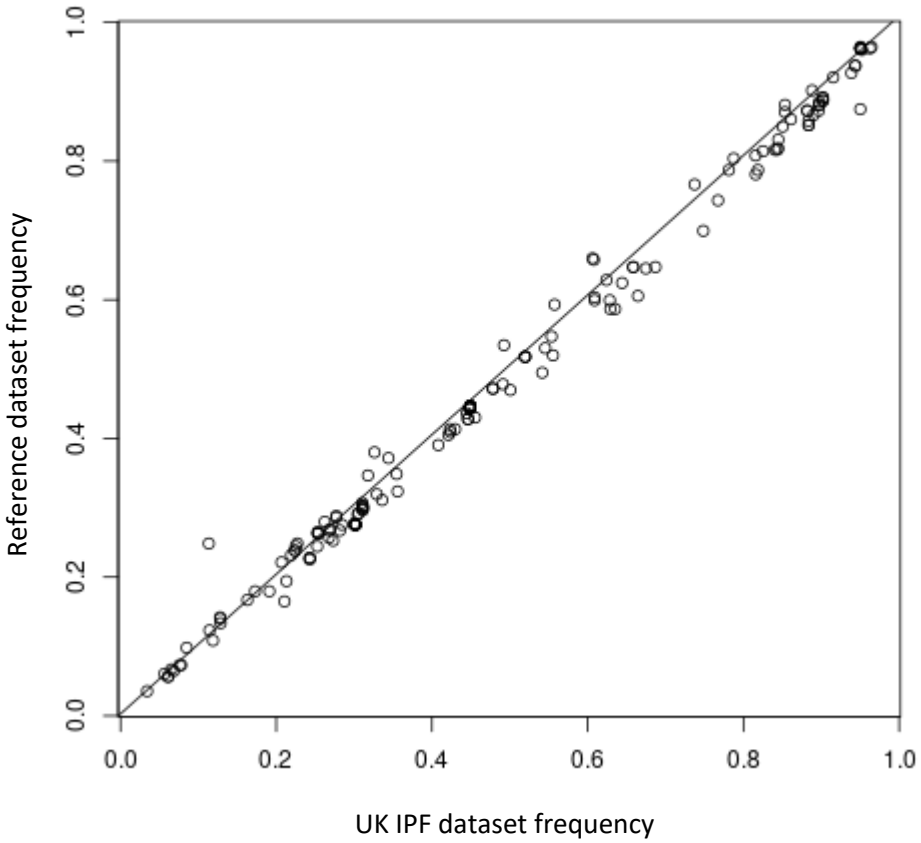


*Figure 5.1: Comparison of the frequencies of the SNPs used in the imputation in the input dataset (UK IPF dataset) and the KIR*IMP reference dataset.*

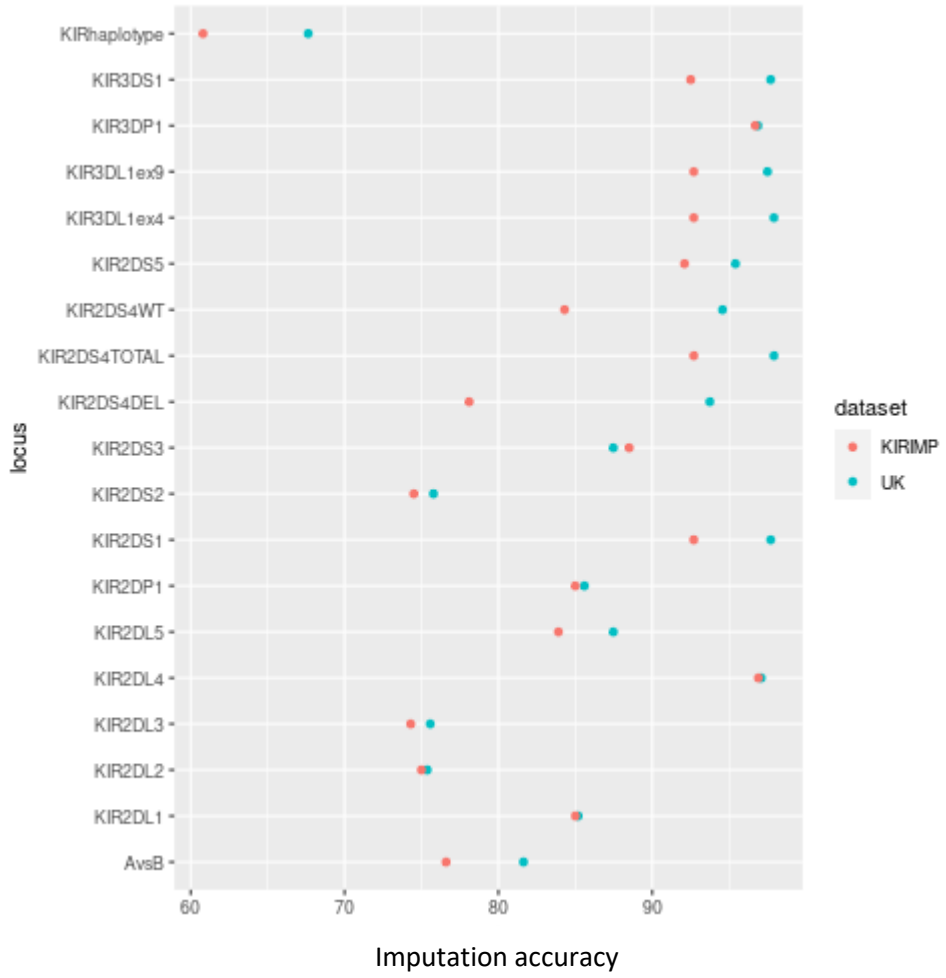


*Figure 5.2: Comparison of imputation accuracy of KIR variation (with KIR*IMP) using genotyped SNPs from UK IPF dataset and the Axiom UK biobank array estimates from KIR*IMP.*

Defining imputation quality thresholds for HRC-imputed SNPs as input to KIR imputation

Of the 301 SNPs used by KIR*IMP for the KIR imputation, only 57 were directly genotyped in the UK IPF dataset. To improve on the number of SNPs used for the imputation, HRC-imputed SNPs were additionally included as input to the imputation. A range of SNP HRC imputation quality thresholds were assessed (info scores threshold of greater than 0.3, 0.5, 0.7, 0.8, 0.9 and 0.95, and no threshold) to determine if including more HRC-imputed SNPs improved the KIR imputation output and to define the HRC imputation quality threshold that maximises the quality of the KIR imputation. The number of SNPs available for inclusion in the KIR imputation (out of the 301 SNPs used by KIR*IMP) increased as the SNP HRC imputation quality threshold was reduced (table 5.3). The minimum imputation quality of the 301 SNPs in the HRC imputation was 0.40.

Of the 301 SNPs used for KIR*IMP imputation, there were six tag SNPs which were known to significantly improve the accuracy of the imputation of the KIR genes/haplotypes they tag. The known tag SNPs (reported in (152)) included in KIR*IMP are presented in table 5.4 – five of the six are represented in the HRC-imputed data. All of these five tag SNPs are common in the UK IPF dataset (MAF >9%) and all have a high imputation quality (info>0.84) apart from rs587560 (which has an info score of 0.52) (table 5.4).

Table 5.3: Number of SNPs used for the imputation of the KIR genes and haplotypes with each imputation quality threshold of the SNPs in the imputation input dataset (number of SNPs out of 301).

HRC imputation threshold	**Number of KIR*IMP imputation panel SNPs represented amongst the directly genotyped and HRC-imputed SNPs in the UK IPF dataset (/301)**
Genotyped	57
No threshold	239
0.3	239
0.5	237
0.7	235
0.8	230
0.9	213
0.95	189

*Table 5.4: Minor allele frequency and imputation quality of the tag SNPs in the UK IPF dataset (*rs592645 was not imputed in the UK IPF dataset as it is not in the HRC imputation panel).*

RSID	**KIR gene/haplotype**	**Position**	**REF/ALT alleles**	**MAF (in UK dataset)**	**Imputation quality (in UK dataset, info score)**
rs587560	A/B, *KIR2DS2, KIR2DL2, KIR2DL3, KIR2DP1, KIR2DL1*	55245738	C/T	0.14	0.52
rs1010355	*KIR3DP1, KIR2DL4, KIR2DS3*	55102179	T/C	0.09	0.85
rs592645*	*KIR3DL1ex4/ex9, KIR2DS5, KIR2DS1, KIR2DS4TOTAL*	55320927	NA	NA	NA
seq-t1d-19-60034052-C-T	*KIR2DL5*	55342240	C/T	0.23	0.90
rs4806585	*KIR2DS4WT*	55346424	C/A	0.24	0.84
rs581623	*KIR2DS4DEL*	55326739	T/C	0.32	0.84

For each imputation quality threshold used for the input SNPs, the estimated KIR*IMP imputation accuracy was plotted for each KIR gene and haplotype (figure 5.3, supplementary table 5.1). In most cases, using only genotyped SNPs had the lowest accuracy and the accuracy improved as the input SNP imputation quality threshold decreased suggesting incorporating more variants in the imputation input dataset is improving the KIR imputation despite the lower quality of the input SNPs. For some KIR gene variants, the KIR*IMP imputation accuracy is high for all input SNP imputation thresholds (e.g. *KIR2DS1*, figure 5.3). The tag SNP rs587560 was HRC-imputed in the UK IPF dataset at a quality of 0.52 and was therefore excluded from input HRC imputation quality thresholds 0.7 and above. A corresponding decrease in accuracy (most significantly *KIR2DS2*, *KIR2DL2* and *KIR2DL3*) (figure 5.3) can clearly be seen between the HRC imputation quality thresholds of 0.5 and 0.7 (and above). There is also a notable difference in the "KIRhaplotype" imputation accuracy between the 0.5 and 0.7 thresholds. There was not a significant difference between the 0.5 and 0.3 thresholds for most of the KIR gene/haplotype accuracies however 0.3 slightly improved the accuracy of the KIR haplotypes (figure 5.3), however there were 2 additional SNPs included across the datasets when lowering the threshold from 0.5 to 0.3 (supplementary table 5.2).

Each KIR gene copy number variation and KIR haplotype had an associated posterior probability which correlates to the most likely copy number variation/haplotype for that individual. Imputation accuracy (average across the whole dataset for each haplotype/gene) and mean posterior probability (for all alleles for each gene/haplotype imputed in each individual) followed a similar pattern (figure 5.4), with the lower input SNP imputation quality thresholds providing the best (highest) posterior probabilities and accuracies (Figure 5.5). For KIRhaplotype, A/B, *KIR2DS2*, *KIR2DL1 KIR2DL2* and *KIR2DL3*, the 0.5 and 0.3 input SNP HRC imputation thresholds provided significantly higher mean posterior probabilities in the KIR imputation (figure 5.5). There were however, some instances where the imputation quality thresholds of 0.3 and 0.5 provided the lowest mean posterior probabilities (*KIR2DS4WT* and *KIR2DS4DEL*) (figure 5.5).

The HRC imputation quality threshold of info >0.3 was used for the remaining three datasets as, overall, the inclusion of the tag SNPs and additional panel SNPs improved the KIR imputation quality across the haplotypes and the genes, irrespective of the imputation quality of the input HRC-imputed SNPs. In the UK dataset, the lowest imputation quality for the input SNPs was 0.4 and therefore all available SNPs were included.

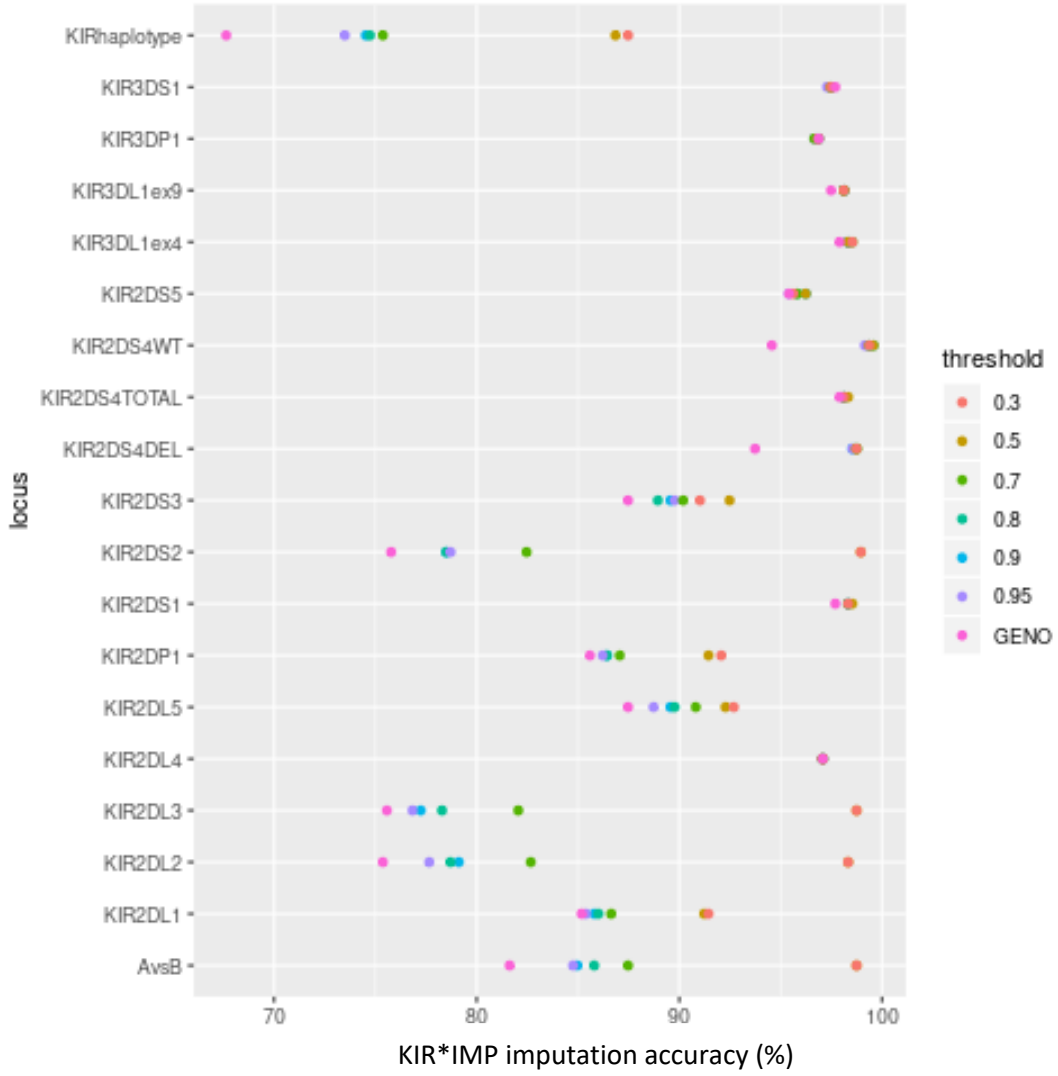


Figure 5.3: KIR gene/haplotype imputation accuracy comparison when using directly genotyped SNPs and a range of imputation qualities for HRC SNPs.

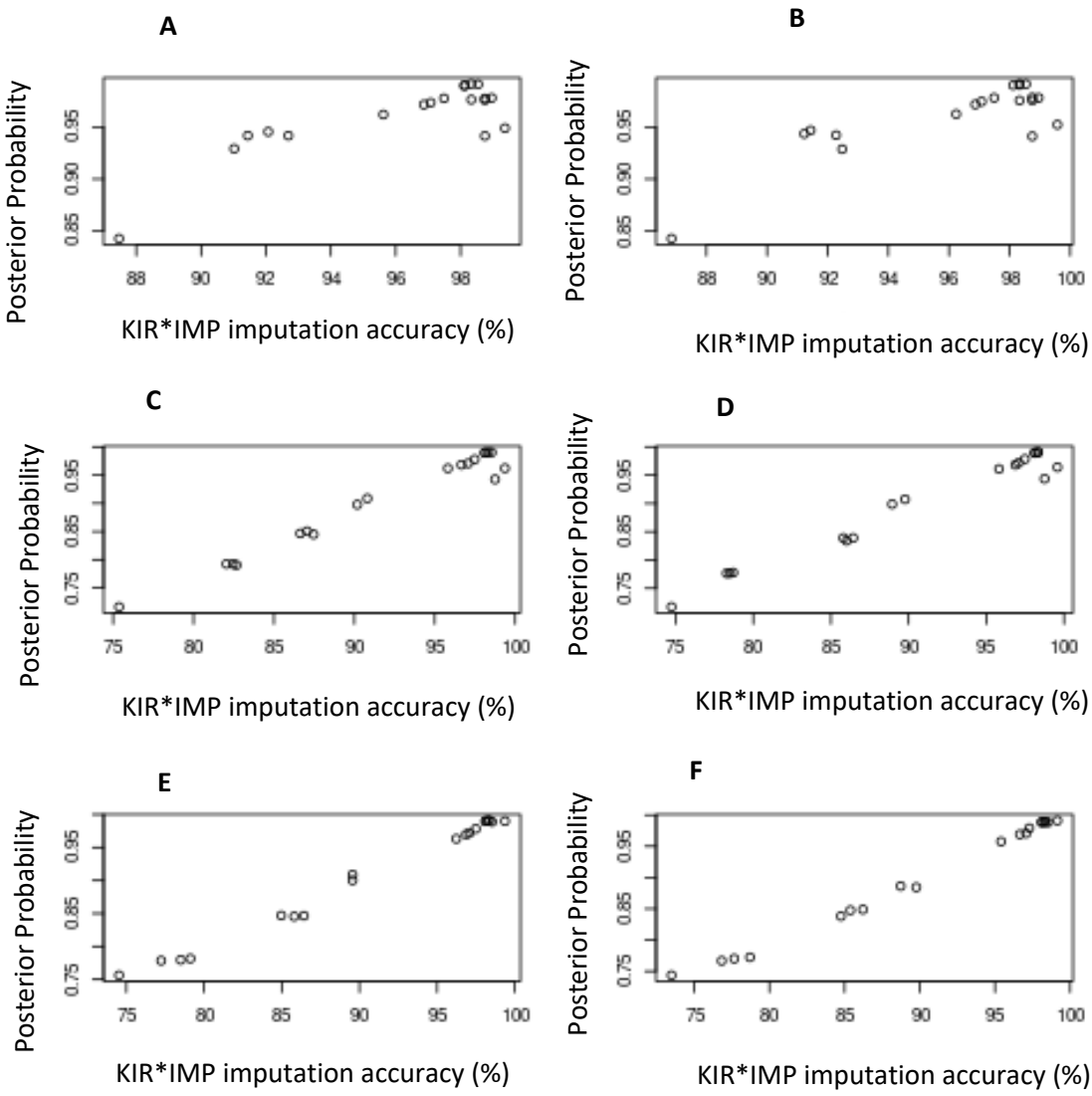


Figure 5.4: KIR gene/haplotype imputation accuracy and mean posterior probabilities (across all samples) for each HRC imputation quality threshold: A=0.3, B=0.5, C=0.7, D=0.8, E=0.9 and F=0.95), each point corresponds to a different locus.

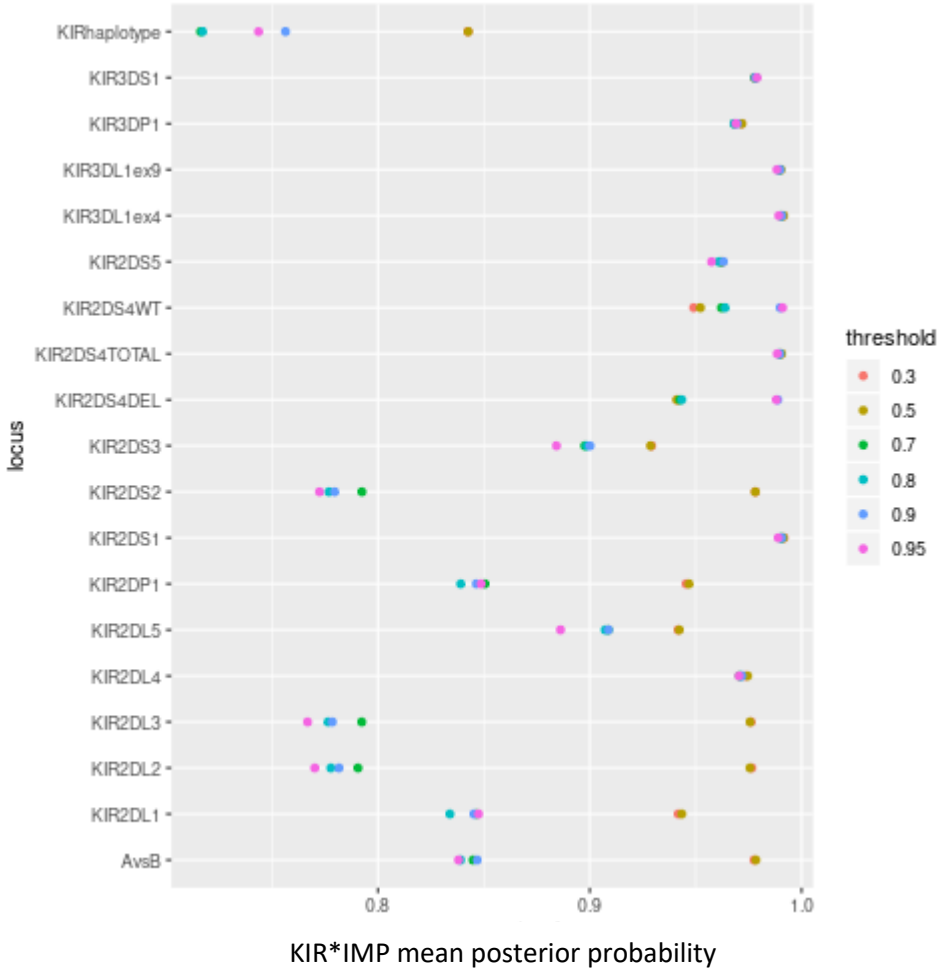


Figure 5.5: KIR gene/haplotype imputation mean posterior probabilities comparison of a range of imputation qualities for HRC SNPs

Imputation of the UK, UUS, Chicago and Colorado datasets IPF dataset using KIR*IMP:

Using HRC-imputed SNPs with a minimum imputation quality of 0.3 as the KIR imputation input dataset, there were three strand mismatches (rs8101660, rs10423751 and rs775900) and between 59 and 141 MAF mismatches (MAF in the IPF dataset was +/-10% compared to the reference dataset in KIR*IMP) in the four IPF datasets with between 157 and 239 SNPs remaining for analysis (table 5.5). Colorado had the highest number of strand and MAF mismatches and also had the lowest imputation qualities (table 5.5). Overall, the imputation accuracies of the haplotypes and the gene CNV imputation accuracy ranges were similar across the UK, UUS and Chicago datasets but lower in the Colorado dataset. (table 5.5). The similar can be said for the A vs B haplotypes (table 5.5). In the Colorado dataset, two of the six

tag SNPs were removed because they had a MAF more than 10% lower than the KIR*IMP reference panel (table 5.6). This mismatch could be partly explained by low HRC imputation quality scores in the Colorado dataset (supplementary table 5.2: score of 0.49 for seq-t1d-19-60034052 and 0.32 for rs587560).

*Table 5.5: Haplotype and gene CNV imputation accuracies across the four IPF dataset from imputation of the KIR region using KIR*IMP (152).*

Dataset	**MAF mismatches**	**SNPs remaining (/301)**	**Haplotype imputation accuracy (%)**	**A vs B haplotype imputation accuracy (%)**	**Gene CNV imputation accuracy range (%)**
UK	59	239	87.1	98.1	90.4-99.6
UUS	131	167	87.5	98.1	91.4-99.6
Chicago	127	171	87.9	97.5	91.7-98.9
Colorado	141	157	74.1	83.9	78.1.8-99.4

*Table 5.6: Minor allele frequencies of two tag SNPs in the Colorado dataset (imputed) and the KIR*IMP dataset.*

SNP ID	**position**	**allele0**	**allele1**	**Colorado MAF**	**KIR*IMP reference MAF**
seq-t1d-19-60034052-C-T	55342240	C	T	0.069	0.23
rs587560	55245738	C	T	0.026	0.25

The UK, UUS and Chicago datasets had similar estimated accuracies across all KIR genes and haplotypes (figure 5.6). In most cases, Colorado had significantly lower imputation accuracies than the rest of the datasets (KIR haplotype, *KIR2DS3, KIR2DS2, KIR2DP1, KIR2DL5, KIR2DL3, KIR2DL2, KIR2DL1* and A/B) (figure 5.6). This reduction in accuracy in the Colorado dataset was mirrored in the posterior probabilities, where a reduction was seen across almost all of the genes (*KIR2DS4WT/DEL and KIR2DL4* were the exceptions) (figure 5.7).

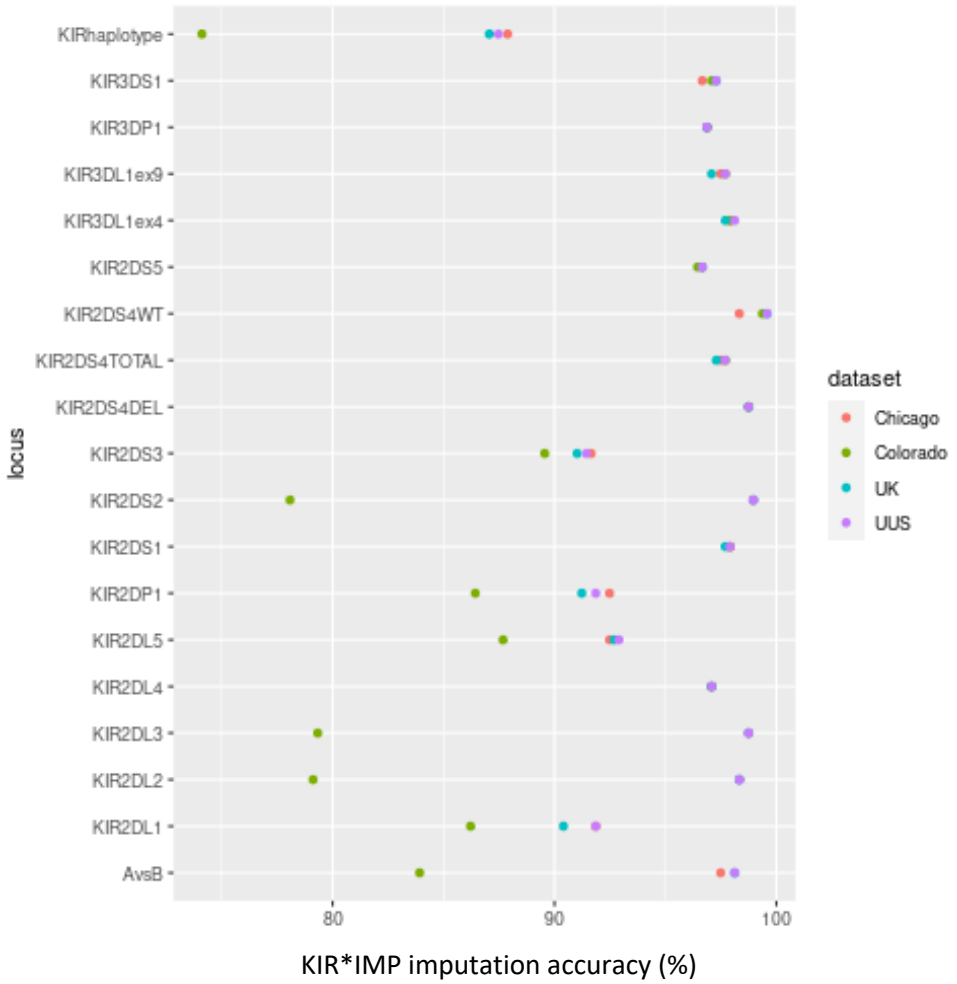


Figure 5.6: KIR gene/haplotype imputation accuracy comparison for the imputation of KIR genes and haplotypes across the four IPF datasets.

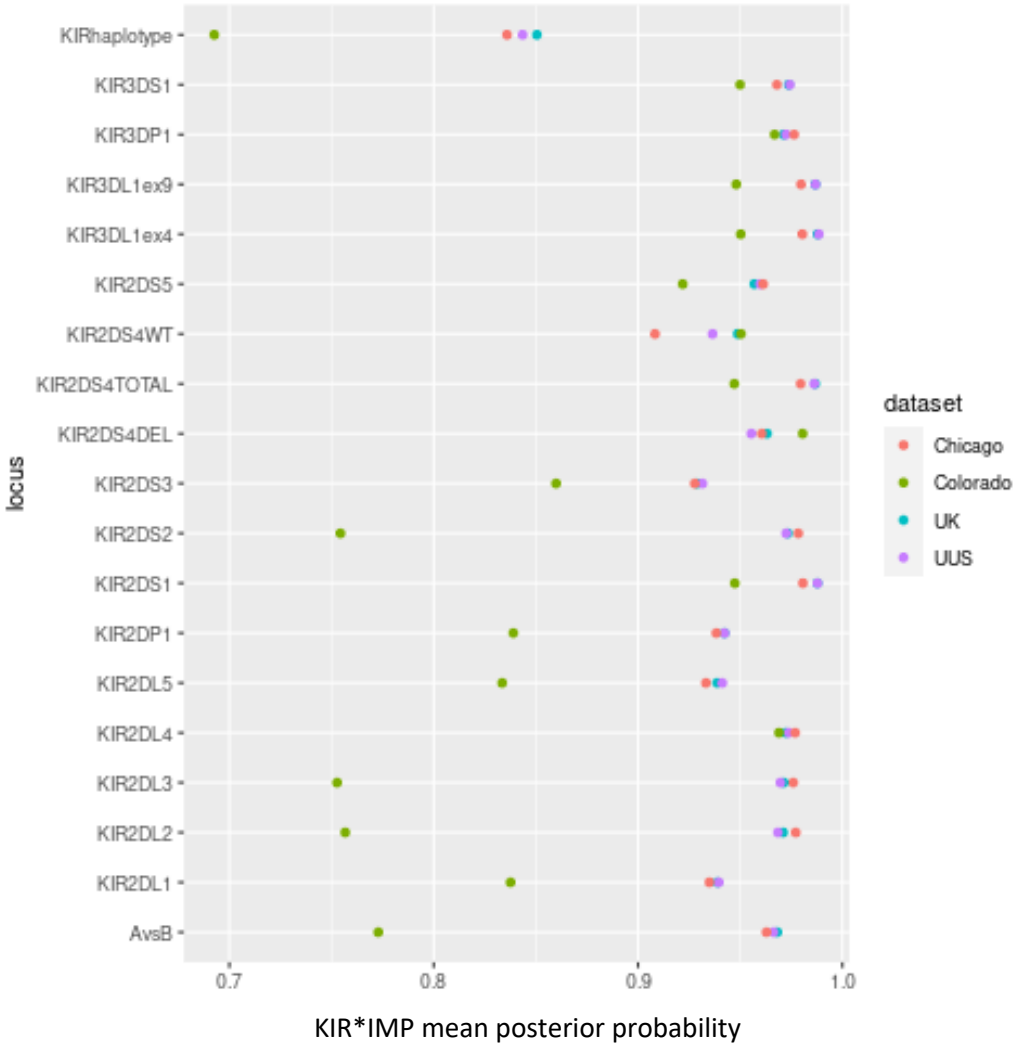


Figure 5.7: KIR gene/haplotype imputation mean posterior probabilities comparison across all four IPF datasets.

Each individual had the following imputed for each chromosome:

i) One of two possible haplotype groups (A or B),

ii) One of 56 possible haplotype numbers

iii) One of three possible copy numbers for each gene.

The number of these haplotypes and gene copy number variations for each IPF dataset can be seen in table 5.7. KIR haplotype 1 was the most common across all the datasets (between 48%-69% of the haplotypes in the datasets) and A haplotype was more common than B (table 5.7). KIR haplotypes 12, 13, 18, 19 and 56 were rare across all datasets (<0.5% in each dataset, table 5.7) and haplotypes 15, 16, 17, 20-55 and 57-69 were not imputed in these datasets

which was most likely due to rarity (the panel does provide the ability to impute these haplotypes if they are present), since only the first 11 haplotypes are thought to be found in >1% of European populations (148). In some cases, the KIR haplotypes were at a similar frequency across all datasets however there were some exceptions, for example haplotypes 1, 2 and 3 are very different across the four datasets (Haplotype 1 made up 65.4, 68.6%, 47.7% and 57.8% of the haplotypes in the UK, UUS, Chicago and Colorado datasets respectively). Rare KIR haplotypes in general (table 5.7) tended to have lower posterior probability than common haplotypes (figure 5.8). For example, in the UUS dataset, KIR haplotype 1 had a mean posterior probability of 0.89 and KIR haplotype 18 had a mean posterior probability of 0.11 (figure 5.8). Having two copies on either chromosome was rare across most genes in all datasets (*KIR2DL5* and *KIR2DS3* were the exceptions to this) (table 5.7). For *KIR2DL2* copy number variation 0 was most common (table 5.7). Although the datasets are not identical, they are all of European ancestry and so it would be expected that the frequencies of the haplotypes and CNVs would be similar. In many examples, the frequencies were vastly different, such as Haplotype 1, *KIR3DS1, KIR3DL1ex4 and KIR3DL1ex9*. In order to reduce these differences, a posterior probability threshold was applied to reduce the number of low confidence variants.

Variants with a posterior probability < 0.5 were removed from each dataset resulting in exclusion of between 0.5% and 1% of calls (haplotype or gene copy number (table 5.8). The number and respective percentages of some haplotypes still varied across datasets (tables 5.7 and 5.9). In several cases the percentages of the CNVs remained different in the Colorado dataset compared to the other three datasets, for example *KIR3DL1ex4, KIR3DL1ex9* and *KIR3DS1,* this could be because the haplotype posterior probabilities are much lower in the Colorado dataset – therefore more may be removed at PP < 0.5, see figure 5.6. Additionally, there is still variation in the proportions of the CNVs for some of the other genes in the Colorado dataset including KIR2DL5, KIR2DS1, KIR2DS4TOTAL, KIR3DL1ex4, KIR3DL1ex9, and KIR2DS4WT KIR3DS1 (table 5.8).

Table 5.7: Number of KIR haplotypes and copy number variation for each gene for each chromosome in each IPF dataset.

KIR gene/Haplotype	Haplotype /CNV number	Number of haplotype/CNV (% of total haplotypes/% of CNVs per gene)			
		UK	UUS	Chicago	Colorado
A vs B	A	5,958 (74.9)	17,231 (76.1)	1,489 (68.7)	10282 (81.6)
	B	1,998 (25.1)	5,399 (23.9)	679 (31.3)	2316 (18.4)
KIR Haplotype	1	5,206 (65.4)	15,516 (68.6)	1043 (47.7)	7280 (57.8)
	2	745 (9.4)	1,672 (7.4)	458 (21.0)	3322 (26.4)
	3	998 (12.5)	2,699 (11.9)	283 (13.0)	684 (5.4)
	4	106 (1.3)	322 (1.5)	38 (1.7)	50 (0.4)
	5	130 (1.6)	253 (1.1)	175 (8.0)	674 (5.4)
	6	37 (0.5)	93 (0.4)	6 (0.3)	76 (0.6)
	7	179 (2.3)	413 (1.8)	23 (1.1)	15 (0.1)
	8	132 (1.7)	375 (1.7)	45 (2.1)	61 (0.5)
	9	291 (3.7)	859 (3.8)	61 (2.8)	184 (1.5)
	10	13 (0.2)	16 (0.07)	8 (0.4)	109 (0.9)
	11	74 (0.9)	240 (1.1)	21 (1.0)	161 (1.0)
	12	2 (0.03)	7 (0.03)	0 (0)	0 (0)
	13	2 (0.03)	9 (0.04)	1 (0.05)	0 (0)
	14	31 (0.4)	114 (0.5)	4 (0.2)	10 (0.08)
	18	0 (0)	1 (0.004)	0 (0)	0 (0)
	19	10 (0.1)	17 (0.08)	0 (0)	2 (0.02)
	56	0 (0)	14 (0.06)	2 (0.09)	0 (0)
KIR2DL1	0	356 (4.5)	968 (4.3)	225 (10.4)	544 (4.3)
	1	7590 (95.4)	21628 (95.6)	1940 (89.5)	12053 (95.7)
	2	10 (0.1)	34 (0.2)	3 (0.1)	1 (0.01)
KIR2DL2	0	7050 (88.6)	20223(89.4)	1809 (83.4)	11238 (89.2)
	1	906 (11.4)	2402 (10.6)	359 (16.6)	1360 (10.8)
	2	0 (0)	5 (0.02)	0 (0)	0 (0)
KIR2DL3	0	906 (11.4)	2409 (10.6)	359 (16.6)	1449 (11.5)
	1	7050 (88.6)	20221 (89.4)	1809 (83.4)	11149 (88.5)
	2	0 (0)	0 (0)	0 (0)	0 (0)
KIR2DL4	0	0 (0)	0 (0)	0 (0)	0 (0)
	1	7944 (99.8)	22573 (99.7)	2164 (99.8)	12595 (100)
	2	12 (0.2)	57 (0.3)	4 (0.2)	3 (0.02)
KIR2DL5	0	6203 (78.0)	17827 (78.8)	1700 (78.4)	11163 (88.6)
	1	1389 (17.5)	3781 (16.7)	415 (19.1)	1274 (10.1)
	2	364 (4.5)	1022 (4.5)	53 (2.4)	161 (1.3)
KIR2DP1	0	356 (4.5)	949 (4.2)	225 (10.4)	555 (4.4)
	1	7590 (95.4)	21638 (95.6)	1941 (89.5)	12042 (95.6)
	2	10 (0.1)	43 (0.2)	2 (0.1)	1 (0.01)
KIR2DS1	0	6249 (78.5)	17937 (79.3)	1739 (80.2)	11554 (91.7)
	1	1707 (21.5)	4693 (20.7)	419 (19.3)	1044 (8.3)
	2	0 (0)	0 (0)	0 (0)	0 (0)
KIR2DS2	0	7048 (88.6)	20225 (89.4)	1809 (83.4)	11189 (88.8)
	1	908 (11.4)	2405 (10.6)	359 (16.6)	1409 (11.2)
	2	0 (0)	0 (0)	0 (0)	0 (0)
KIR2DS3	0	7491 (94.2)	21263 (94.0)	2067 (95.3)	12075 (95.8)
	1	306 (3.8)	882 (3.9)	59 (2.7)	339 (2.7)
	2	159 (2.0)	485 (2.1)	42 (1.9)	184 (1.5)
KIR2DS4DEL	0	2595 (32.6)	6670 (29.5)	1080 (49.8)	5193 (41.2)

	1	5361 (67.4)	15960 (70.5)	1088 (50.2)	7405 (58.8)
	2	0 (0)	0 (0)	0 (0)	0 (0)
KIR2DS4TOTAL	0	1709 (21.5)	4691 (20.7)	429 (19.8)	1042 (8.3)
	1	6247 (78.5)	17939 (79.3)	1739 (80.2)	11556 (91.7)
	2	0 (0)	0 (0)	0 (0)	0 (0)
KIR2DS4WT	0	6393 (80.4)	18067 79.8)	1539 (71.0)	8432 (66.9)
	1	1563 (19.6)	4563 (20.2)	629 (29.0)	4166 (33.1)
	2	0 (0)	0 (0)	0 (0)	0 (0)
KIR2DS5	0	6639 (83.4)	19055 (84.2)	1821 (84.0)	11012 (87.4)
	1	1316 (16.5)	3569 (15.8)	347 (16.0)	1583 (12.6)
	2	1 (0.01)	6 (0.03)	0 (0)	3 (0.02)
KIR3DL1ex4	0	1705 (21.4)	4697 (20.8)	429 (19.8)	1040 (8.3)
	1	6251 (78.6)	17933 (79.2)	1739 (80.2)	11558 (91.7)
	2	0 (0)	0 (0)	0 (0)	0 (0)
KIR3DL1ex9	0	1708 (21.5)	4691 (20.7)	430 (19.8)	1040 (8.3)
	1	6248 (78.5)	17939 (79.3)	1738 (80.2)	11558 (91.7)
	2	0 (0)	0 (0)	0 (0)	0 (0)
KIR3DP1	0	0 (0)	1 (0.004)	0 (0)	0 (0)
	1	7942 (99.8)	22577 (99.8)	2164 (99.8)	12596 (100)
	2	14 (0.2)	52 (0.2)	4 (0.2)	2 (0.02)
KIR3DS1	0	6265 (78.7)	17961 (79.4)	1735 (80.0)	11565 (91.8)
	1	1667 (21.0)	4587 (20.3)	428 (19.7)	1022 (8.1)
	2	24 (0.3)	82 (0.4)	5 (0.2)	11 (0.1)

Table 5.8: Number of variants before and after quality control (posterior probability <0.5).

Dataset	**Full number variants**	**QC number of variants**	**Difference**
Chicago	41,192	40,995	197 (0.5%)
Colorado	239,363	236,075	3,288 (1.3%)
UK	151,165	150,264	901 (0.6%)
UUS	429,971	427,495	2,476 (0.6%)

Table 5.9: Quality controlled (posterior probability > 0.5) set of haplotypes in each dataset (N.B haplotypes 13, 18 and 19 are now not present because posterior probability was < 0.5 in all individuals).

KIR gene/Haplotype	**Haplotype (A or B)/CNV number**	**Number of haplotype (%)**			
		UK	**UUS**	**Chicago**	**Colorado**
A vs B	A	5,958 (74.9)	17,231 (76.1)	1,489 (68.7)	10282 (81.6)
	B	1,998 (25.1)	5,399 (23.9)	679 (31.3)	2316 (18.4)
KIR Haplotype	1 (A)	5,203 (70.5)	15,448 (73.3)	1042 (51.2)	7041 (69.4)
	2 (A)	743 (10.0)	1,662 (7.9)	439 (21.7)	1909 (18.8)
	3 (B)	946 (12.8)	2,562 (12.2)	272 (13.4)	467 (4.6)
	4 (B)	76 (1.0)	237 (1.1)	32 (1.6)	28 (0.3)
	5 (B)	101 (1.4)	198 (0.9)	150 (7.4)	482 (4.7)

	6 (B)	4 (0.05)	29 (0.1)	4 (0.2)	51 (0.5)
	7 (B)	33 (0.4)	94 (0.4)	4 (0.2)	0 (0)
	8 (B)	44 (0.6)	133 (0.6)	22 (1.1)	7 (0.07)
	9 (B)	164 (2.2)	459 (2.2)	42 (2.1)	87 (0.9)
	10 (B)	12 (0.2)	9 (0.04)	1 (0.05)	26 (0.3)
	11 (B)	36 (0.5)	169 (0.8)	12 (0.6)	48 (0.5)
	12 (B)	0 (0)	1 (0.005)	0 (0)	0 (0)
	14 (B)	13 (0.2)	68 (0.3)	4 (0.2)	4 (0.04)
	56 (B)	0 (0)	1 (0.005)	0 (0)	0 (0)
KIR2DL1	0	356 (4.5)	967 (4.3)	225 (10.4)	540 (4.3)
	1	7590 (95.4)	21625 (95.6)	1940 (89.5)	12046 (95.7)
	2	10 (0.1)	34 (0.2)	3 (0.1)	1 (0.01)
KIR2DL2	0	7050 (88.6)	20223 (89.4)	1809 (83.4)	11213 (89.3)
	1	906 (11.4)	2400 (10.6)	359 (16.6)	1342 (10.7)
	2	0 (0)	3 (0.01)	0 (0)	0 (0)
KIR2DL3	0	906 (11.4)	2409 (10.6)	259 (12.5)	1449 (11.5)
	1	7050 (88.6)	20221 (89.4)	1809 (87.5)	11149 (88.5)
	2	0 (0)	0 (0)	0 (0)	0 (0)
KIR2DL4	0	0 (0)	0 (0)	0 (0)	0 (0)
	1	7943 (99.8)	22571 (99.8)	2164 (99.8)	12595 (100.0)
	2	12 (0.2)	53 (0.2)	4 (0.2)	2 (0.02)
KIR2DL5	0	6196 (78.3)	17807 (79.0)	1699 (78.7)	10923 (89.4)
	1	1362 (17.2)	3741 (16.6)	409 (18.9)	1160 (9.5)
	2	351 (4.4)	1002 (4.4)	52 (2.4)	141 (1.2)
KIR2DP1	0	356 (4.5)	948 (4.2)	225 (10.4)	554 (4.4)
	1	7590 (95.4)	21637 (95.6)	1941 (89.5)	12035 (95.6)
	2	10 (0.1)	43 (0.2)	2 (0.1)	1 (0.01)
KIR2DS1	0	6249 (78.5)	17937 (79.3)	1739 (80.2)	11554 (91.7)
	1	1707 (21.5)	4693 (20.7)	429 (19.8)	1044 (8.3)
	2	0 (0)	0 (0)	0 (0)	0 (0)
KIR2DS2	0	7048 (88.6)	20225 (89.4)	1809 (83.4)	11189 (88.8)
	1	908 (11.4)	2405 (10.6)	359 (16.6)	1409 (11.2)
	2	0 (0)	0 (0)	0 (0)	0 (0)
KIR2DS3	0	7402 (95.8)	21023 (95.9)	2051 (96.5)	11919 (97.4)
	1	214 (2.8)	578 (2.6)	50 (2.4)	207 (1.7)
	2	112 (1.4)	327 (1.5)	25 (1.2)	110 (0.9)
KIR2DS4DEL	0	2595 (32.6)	6670 (29.5)	1080 (49.8)	5193 (41.2)
	1	5361 (67.4)	15960 (70.5)	1088 (50.2)	7405 (58.8)
	2	0 (0)	0 (0)	0 (0)	0 (0)
KIR2DS4TOTAL	0	1709 (21.5)	4689 (20.4)	429 (19.8)	1042 (8.3)
	1	6247 (78.5)	17936 (79.3)	1739 (80.2)	11556 (91.7)
	2	0 (0)	0 (0)	0 (0)	0 (0)
KIR2DS4WT	0	6391 (80.4)	18067 (79.9)	1539 (71.0)	8429 (66.9)
	1	1562 (19.6)	4559 (20.1)	629 (29.0)	4166 (33.1)
	2	0 (0)	0 (0)	0 (0)	0 (0)
KIR2DS5	0	6623 (83.5)	19036 (84.3)	1820 (84.0)	11001 (87.5)
	1	1310 (16.5)	3549 (15.7)	347 (16.0)	1576 (12.5)
	2	0 (0)	1 (0.004)	0 (0)	0 (0)
KIR3DL1ex4	0	1705 (21.4)	4697 (20.7)	429 (19.8)	1040 (8.3)
	1	6251 (78.6)	17993 (79.3)	1739 (80.2)	11558 (91.7)
	2	0 (0)	0 (0)	0 (0)	0 (0)
KIR3DL1ex9	0	1708 (21.5)	4691 (20.7)	430 (19.8)	1040 (8.3)
	1	6247 (78.5)	17938 (79.3)	1738 (80.2)	11558 (91.7)
	2	0 (0)	0 (0)	0 (0)	0 (0)

KIR3DP1	0	0 (0)	1 (0.004)	0 (0)	0 (0)
	1	7941 (99.8)	22571 (99.8)	2164 (99.8)	12595 (100.0)
	2	14 (0.2)	49 (0.2)	4 (0.2)	1 (0.02)
KIR3DS1	0	6260 (78.8)	17947 (79.5)	1735 (80.1)	11561 (91.9)
	1	1661 (20.9)	4558 (20.2)	425 (19.6)	1015 (8.1)
	2	19 (0.2)	72 (0.3)	5 (0.2)	7 (0.1)

Having a total CNV count of 4 (i.e. 2 on each chromosome) was very rare across all the genes and was most commonly found in KIR2DL5 (table 5.10, figure 5.8. The distribution of the percentages of the CNVs in KIR2DL5 is considerably different in the Colorado dataset (table 5.10, figure 5.8. Additionally, the CNVs in several genes including KIR2DL1 and KIR2DL2 have remarkably different distributions in the Chicago dataset. There are visible copy number distribution differences for several genes (including KIR2DL5, KIR2DS1, KIR2DS4WT/TOTAL and KIR3DL1ex4/ex9) between the Colorado dataset and the other datasets (table 5.10).

Table 5.10: Percentages of total CNVs for each gene in each dataset. Percentage is calculated individually for each gene.

Gene	UK CNVs (%)					UUS CNVs (%)					Chicago CNVs (%)						Colorado CNVs (%)			
	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
KIR2DL1	0.2	8.5	91.0	0.3	0.0	0.1	8.3	91.3	0.3	0.0	1.2	18.3	80.4	0.2	0.0	0.2	8.4	91.4	0.0	0.0
KIR2DL2	78.4	20.4	1.2	0.0	0.0	79.8	19.2	1.0	0.0	0.0	70.1	26.7	3.2	0.0	0.0	80.0	18.8	1.3	0.0	0.0
KIR2DL3	1.2	20.4	78.4	0.0	0.0	1.0	19.3	79.7	0.0	0.0	3.2	26.7	70.1	0.0	0.0	1.4	20.0	78.7	0.0	0.0
KIR2DL4	0.0	0.0	99.7	0.3	0.0	0.0	0.1	99.5	0.5	0.0	0.0	0.0	99.6	0.4	0.0	0.0	0.0	100.0	0.0	0.0
KIR2DL5	61.8	26.5	9.9	1.7	0.2	62.6	25.8	10.1	1.3	0.2	63.0	27.8	8.2	0.9	0.1	80.3	16.7	2.8	0.1	0.0
KIR2DP1	0.2	8.6	91.0	0.3	0.0	0.1	8.1	91.4	0.4	0.0	1.2	18.3	80.4	0.1	0.0	0.2	8.6	91.2	0.0	0.0
KIR2DS1	61.9	33.3	4.8	0.0	0.0	62.9	32.7	4.4	0.0	0.0	65.0	30.4	4.6	0.0	0.0	84.1	15.2	0.7	0.0	0.0
KIR2DS2	78.4	20.4	1.2	0.0	0.0	79.8	19.2	1.0	0.0	0.0	70.1	26.7	3.2	0.0	0.0	79.1	19.5	1.4	0.0	0.0
KIR2DS3	91.9	5.3	2.7	0.1	0.0	92.1	4.9	2.9	0.0	0.0	93.2	4.5	2.2	0.1	0.0	95.0	3.2	1.7	0.1	0.0
KIR2DS4DEL	10.6	44.1	45.4	0.0	0.0	8.6	41.8	49.6	0.0	0.0	24.5	50.6	24.9	0.0	0.0	16.4	49.6	34.0	0.0	0.0
KIR2DS4TOTAL	4.8	33.4	61.8	0.0	0.0	4.4	32.7	62.9	0.0	0.0	4.6	30.4	65.0	0.0	0.0	0.7	15.2	84.1	0.0	0.0
KIR2DS4WT	64.9	31.0	4.1	0.0	0.0	64.0	31.8	4.3	0.0	0.0	50.5	41.1	8.5	0.0	0.0	44.5	44.8	10.7	0.0	0.0
KIR2DS5	70.0	27.1	2.9	0.0	0.0	71.4	25.8	2.8	0.0	0.0	70.8	26.5	2.8	0.0	0.0	76.7	21.6	1.7	0.0	0.0
KIR3DL1ex4	4.8	33.3	61.9	0.0	0.0	4.4	32.7	62.9	0.0	0.0	4.6	30.4	65.0	0.0	0.0	0.7	15.2	84.2	0.0	0.0
KIR3DL1ex9	4.8	33.4	61.8	0.0	0.0	4.4	32.7	62.9	0.0	0.0	4.6	30.4	64.9	0.0	0.0	0.7	15.2	84.2	0.0	0.0
KIR3DP1	0.0	0.0	99.6	0.4	0.0	0.0	0.1	99.5	0.4	0.0	0.0	0.0	99.6	0.4	0.0	0.0	0.0	100.0	0.0	0.0
KIR3DS1	62.4	32.5	5.0	0.1	0.0	63.4	31.8	4.7	0.1	0.0	64.9	30.1	4.9	0.1	0.0	84.5	14.8	0.8	0.0	0.0

Figure 5.8: Distributions of total CNV number for all KIR genes across the four datasets.

Discussion:

The KIR region harbours extensive copy number variation which cannot be accurately genotyped and are omitted by standard SNP-focused GWAS. The aim of this chapter was to use genotyped and imputed SNPs to impute KIR CNVs and haplotypes in four IPF datasets. KIR*IMP (152) enabled the imputation of copy number variation in 14 genes and 18 haplotypes. In order to produce the highest quality imputation, different numbers of SNPs (from different imputation quality thresholds of HRC-imputed SNPs) were used as the imputation input dataset (HRC-imputed data) to identify which number provides the highest accuracy and posterior probabilities. It was important that the selected threshold resulted in the highest possible accuracies and posterior probabilities across the genes and haplotypes and so an imputation quality threshold of 0.3 was chosen. An input SNP imputation quality threshold of 0.3 included all of the six tag SNPs (apart from rs592645 which was not included in the HRC imputation panel) in the KIR imputation input datasets across the UK, UUS and Chicago datasets (supplementary table 5.2). Rs592645 was not imputed in any of the four datasets. Although rs592645 was a tag SNP for *KIR3DL1ex4/ex9, KIR2DS5, KIR2DS1* and *KIR2DS4TOTAL,* these were imputed at an accuracy of 97.7%, 97.5%, 97.5%, 96.7% and 97.7% respectively using the other KIR imputation panel SNPs. This suggests this tag SNP may not have a strong positive effect on imputation of these genes. There were two tag SNP mismatches in the Colorado dataset (table 5.6), it is not clear why this occurred, they were imputed so the imputation may have been negatively affected by nearby poorly genotyped SNPs. These mismatches may have been the reason for lower overall accuracy of imputation for A/B, *KIR2DS2, KIR2DL2, KIR2DL3, KIR2DP1, KIR2DL1* and *KIR2DL5* in this dataset. The rs587560 T allele had a frequency of 0.026 in the Colorado dataset and 0.25 in the KIR*IMP reference dataset, the KIR*IMP reference dataset appeared to be closer to what is expected in European datasets, (dbSNP (207) presents a frequency of 0.30 in HapMap)).

KIR*IMP reference SNPs are from the Illumina Immunochip array which may explain the low SNP matches with the Axiom Affymetrix UK Biobank genotyping array (used in the UK IPF dataset) and although the reference dataset (for KIR*IMP) was a European dataset, there was a large quantity of MAF mismatches (between 20 and 48%, table 5.5) and three strand mismatches. It would be expected that HRC imputation would resolve the strand mismatches however, the three strand mismatches are multi-allelic which could have caused the problems.

There were significant differences between the frequencies and distributions of some haplotypes and CNVs between the datasets. The datasets are all of European ancestry and therefore it would be expected that the distributions would be similar. A posterior probability

filter of 0.5 was incorporated to remove the variant calls with low confidence imputation but there was still significant differences between the datasets, especially between the Chicago and Colorado datasets versus the UK and the UUS datasets. This could be due to the tag SNPs seq-t1d-19-60034052-C-T and rs587560 as these were missing in the Colorado dataset and had a significantly lower imputation quality in the Chicago dataset (compared to the UK and UUS datasets). Seq-t1d-19-60034052-C-T tags *KIR2DL5* which could explain the different distribution in the Colorado dataset. rs587560 tags A/B which is considerably different in the Chicago and Colorado datasets and KIR2DS2, KIR2DL2, KIR2DL3, KIR2DP1 and KIR2DL1 (the genes whose copy numbers determine the A type haplotype) are all different in the Chicago dataset.

This chapter described the methods for imputing KIR haplotypes and gene copy number across the four IPF datasets that will be used for association testing in Chapter 6. Using the HRC-imputed SNPs as input for the KIR imputation meant that these datasets had better KIR imputation accuracies than if only directly genotyped SNPs had been included. The gene copy number variations were imputed at an accuracy of over 90% and the haplotypes at an accuracy of over 87%, however there was considerable differences in the haplotype and CNV frequencies and distributions between the datasets which could cause spurious results. Some of the differences between the datasets will be accounted for in the subsequent analyses which include an internal validation step where the variant is required to be significantly associated with IPF (at nominal significance) in at least two of the datasets. This, together with evaluation of individual dataset level results, will ensure that spurious signals due to poor quality variant calls are excluded but without a loss of power due to exclusion of an entire dataset.

Chapter 6: KIR-wide association analysis of IPF susceptibility in four Idiopathic Pulmonary Fibrosis (IPF) datasets

6.1 Introduction

Natural killer (NK) cells play a vital role in the innate immune systems response to virus infection (139). NK cells are activated by a balance between activating and inhibitory receptors on their surface (including killer immunoglobulin-like receptors [KIRs]) (140-142). Activated NK cells detect virally infected cells by interacting with the HLA molecules to initiate cell death (139). KIR molecules are encoded by polygenic genes found on chromosome 19. The KIR region harbours significant copy number variation which requires bespoke imputation to adequately measure the variation (see chapter 1, section 1.3.2). Micro-injuries such as viral infection are believed to be triggers that lead to fibrosis in the lungs (7, 8, 60-63). It is therefore of interest to define the contribution of KIR haplotypes and copy number variation (CNV) of KIR genes to IPF susceptibility. KIR genes have not been analysed in IPF in this way previously and therefore this analysis could provide further insight into the biological processes underlying IPF development and pathogenesis.

This chapter tests the hypothesis that variation at the KIR gene locus contributes to genetic susceptibility to IPF. The methods and results of a KIR-wide association meta-analysis and a joint-regression meta-analysis of four IPF datasets, utilising the KIR imputation described in chapter 5, are described here.

6.2 Methods

KIR imputed IPF datasets:

The IPF datasets used in this chapter, and their phasing and KIR*IMP imputation, were described in Chapter 5.

The variation across the KIR region compromises copy number variation (CNV) of genes and haplotypes. Twelve KIR region haplotypes and copy number of 17 KIR genes were imputed across the four IPF datasets and individuals were assigned A and B haplotype classifications accordingly (see Chapter 5).

KIR-wide association analysis:

Across all four datasets, imputed KIR gene copy number variations (CNVs) (see chapter 5, table 10) and haplotypes (see chapter 5, table 9) were tested for association with IPF susceptibility

using R version 3.6.1 assuming an additive genetic model. Ten principal components (to adjust for fine scale population structure) and sex were included as covariates. A count threshold of equal to or less than three within each dataset was used to remove rare haplotypes and CNVs. There was no posterior probability filtering. The variants were analysed using the formula below, where β_0 is the intercept, G_i is the genotype (either presence vs absence [0,1] for the KIR haplotypes, as a genotype [AA, AB, BB] for AvsB and as a continuous variable for CNVs [0, 1, 2, 3 and 4]).

$$Phenotype\ (\log odds)\ \sim \beta_0 + \beta_1 G_i + Sex + \ PCs$$

KIR-wide association meta-analysis

The results of the KIR-wide association analyses from each IPF dataset were combined using a fixed-effects inverse weighted meta-analysis. A Bonferroni corrected significance threshold for the KIR-wide association analysis was calculated for each separate analysis based on the number of haplotypes or the number of genes tested across the region. Haplotypes and CNVs were required to pass the Bonferroni corrected significance threshold and be nominally significant (P<0.05) in at least two of the four studies.

6.3 Results

KIR-wide association meta-analysis of IPF susceptibility in the UK, UUS, Chicago and Colorado datasets

Four haplotypes in total were removed for low count (count <3), and these were haplotypes 12 and 56 in all four datasets, haplotype 10 in the Chicago dataset and haplotype 7 in the Colorado dataset. Twelve type A or B haplotypes, as well as haplotype types A and B themselves, were tested for association with IPF susceptibility in each of the UK (supplementary figure 6.1), UUS (supplementary figure 6.2), Chicago (supplementary figure 6.3) and Colorado datasets (supplementary figure 6.4). A Bonferroni corrected significance threshold of P<0.004 was applied based on the number of haplotypes tested (12, not including A and B). The mean posterior probabilities of the haplotypes across the four datasets was 0.85 and on average 51% of haplotypes had a posterior probability more than 0.9 (52% in the Chicago dataset, 18.7% in the Colorado dataset, 67% in the UK dataset, 66% in the UUS dataset) (figure 6.2). 12 haplotypes were tested across all four datasets and two haplotypes were tested across three datasets (there were no haplotypes tested in any less than three

datasets) (supplementary table 6.1). The P-values were mostly very similar across the four datasets (supplementary table 6.1, figure 6.2), but there was one outlier in the results which was haplotype 9 in the UK dataset ($P=6.11x10^{-60}$). No haplotypes passed the threshold of P<0.004 in any of the four datasets (supplementary table 6.1). The most significant haplotypes in each independent dataset were haplotype 2 (Chicago dataset, P=0.30 and UUS dataset, P=0.02), haplotype 8 (Colorado dataset, P=0.02) and haplotype 4 (UK dataset, P=0.13).

Following meta-analysis of results across four datasets, only one B-type haplotype passed the Bonferroni-corrected significance threshold (P<0.004); Haplotype 9 ($P=1.98x10^{-17}$) (table 6.1). However, the association with Haplotype 9 was only nominally significant (P<0.05) in the UK datasets and therefore did not pass the required significance thresholds (haplotype 9 was therefore excluded from figure 6.1). Haplotype 9 had an overall frequency of 2.2% and was well imputed in the UK dataset (average posterior probability of 0.7). The presence of this haplotype was associated with a significantly decreased risk of IPF in the UK dataset but with a non-significantly increased risk of IPF in the other three datasets. The mean posterior probabilities of haplotype 9 across all four datasets was 0.68, the values were similar across the four datasets (UK = 0.70, UUS = 0.69, Chicago = 0.68, Colorado = 0.65).

Haplotype 8 was the next most significantly associated haplotype in the meta-analysis (P=0.008). Haplotype 8 was tested in all four IPF datasets and although it did not reach nominal significance in at least two datasets, it had P=0.03 in the Colorado dataset and P=0.06 in the UUS dataset. The frequency of haplotype 8 is similar in the UK, UUS and Chicago datasets (0.6%, 0.6% and 1.1% respectively), however the frequency of haplotype 8 was much lower in the Colorado dataset (0.07%). The mean posterior probabilities of haplotype 8 across all four datasets was 0.605, and the values were consistent across the four datasets (UK=0.59, UUS=0.60, Chicago=0.60 and Colorado=0.63) (table 6.1).

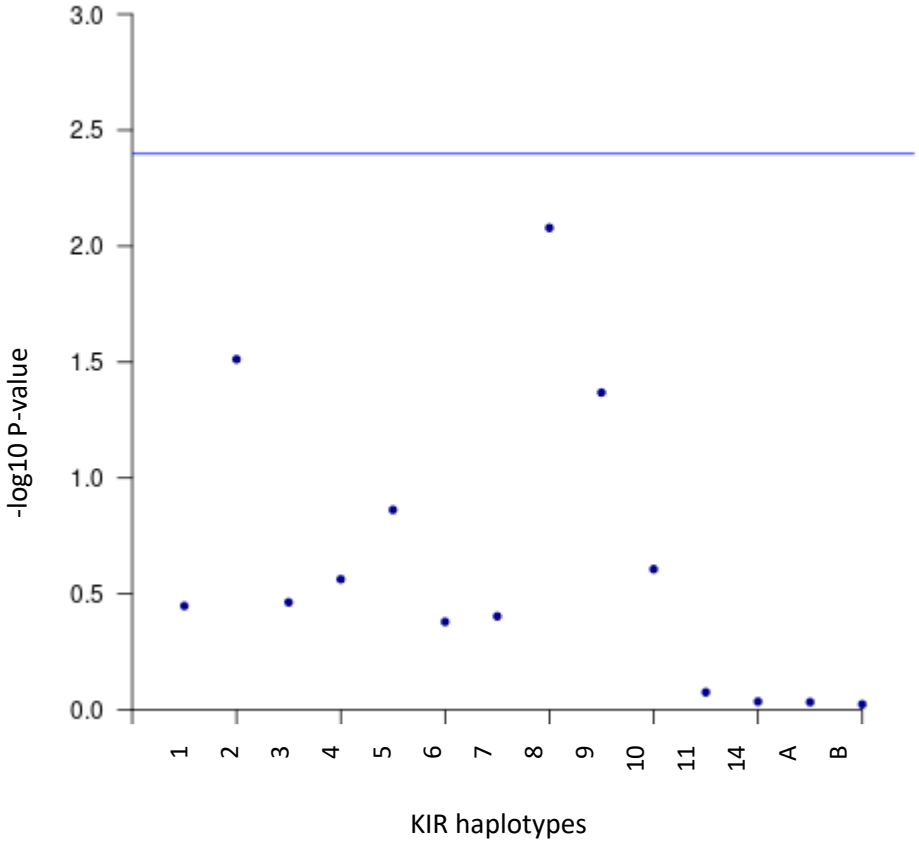


Figure 6.1: -log10 p-values of the association meta-analysis of association of KIR haplotypes with IPF susceptibility in the UK, UUS, Chicago and Colorado datasets. Blue line denotes the Bonferroni corrected significance threshold of 0.004.

The gene CNVs in the four datasets are described in chapter 5, table 10. 17 KIR genes were tested for association with IPF susceptibility. The Bonferroni corrected significance threshold (P<0.003) was calculated based on the 17 genes tested in this analysis. The mean posterior probability of the CNVs across the four datasets was 0.95 and an average of 84% of CNV had a posterior probability of more than 0.9 (89% in the Chicago dataset, 63% in the Colorado dataset, 91% in the UK dataset, 91% in the UUS dataset) (figure 6.2). No CNVs passed the Bonferroni corrected significance threshold in any of the four IPF datasets. The most significant CNVs across the four datasets were *KIR3DP1* (Chicago, P=0.69 and UK, P=0.58), *KIR2DS3* (Colorado, P=0.15) and *KIR2DS4WT* (UUS, P=0.03).

In the meta-analysis of the four IPF datasets, all 17 genes were tested for association with IPF susceptibility and genes passed the Bonferroni corrected threshold. The most significant CNV was in *KIR3DP1* (P=0.12) and the individual dataset P-values were not nominally significant

(P<0.05) (although it was close to the nominal significance threshold in the UUS dataset [P=0.06]) (table 6.1).

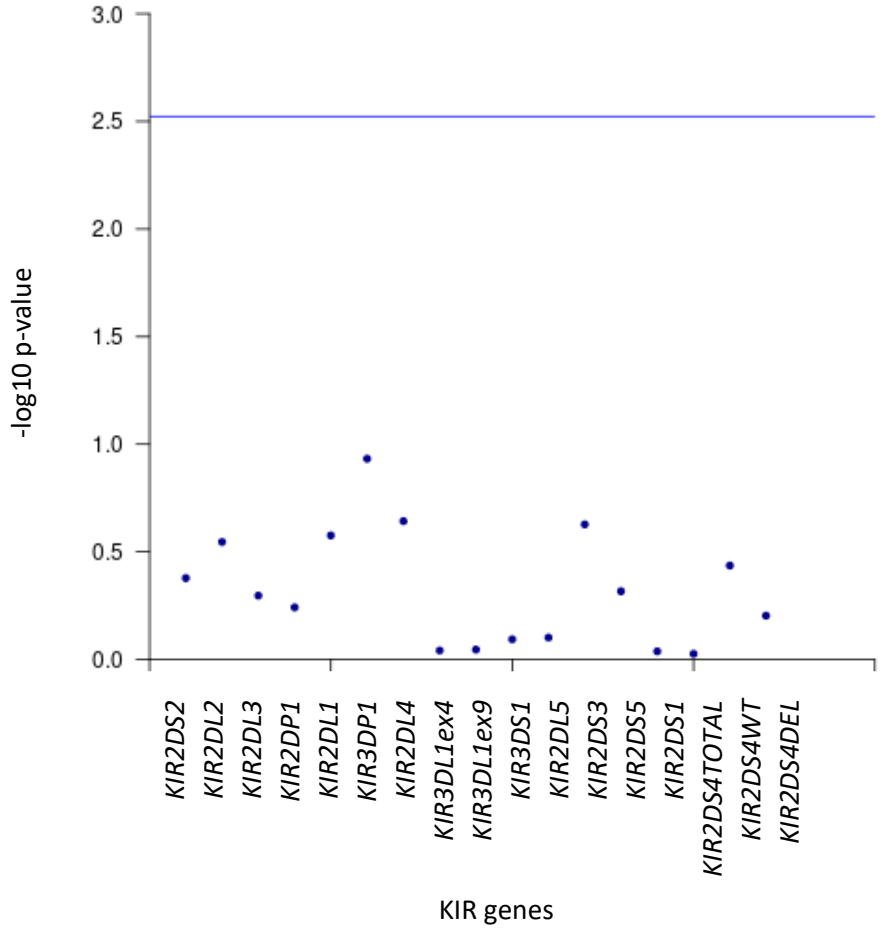


Figure 6.2: -log10 p-values of the association meta-analysis of KIR genes in IPF susceptibility in the UK, UUS, Chicago and Colorado datasets. Blue line denotes the Bonferroni corrected significance threshold of 0.003.

Table 6.1: Top signals from the association meta-analysis of CNVs and haplotypes in IPF susceptibility in the UK, UUS, Chicago and Colorado datasets.

Haplotype/ Gene CNV	**Posterior Probability**				**Mean frequency (%)**	**Odds ratio (95% CI)**				**P-value**				**Meta Odds ratio (95% CI)**	**Meta p-value**
	Chicago	Colorado	UK	UUS		Chicago	Colorado	UK	UUS	Chicago	Colorado	UK	UUS		
Haplotype 9	0.68	0.65	0.70	0.69	1.9	1.38 (0.55-3.47)	1.34 (0.75-2.40)	0.29 (0.25-0.34)	1.12 (0.99-1.26)	0.49	0.33	$6.11x10^{-60}$	0.073	0.67 (0.61-0.73)	$1.98x10^{-17}$
Haplotype 8	0.60	0.62	0.59	0.60	2.0	1.90 (0.51-7.05)	11.08 (1.23-100.00)	1.81 (0.72-4.61)	2.10 (0.97-4.38)	0.34	0.03	0.21	0.06	2.15 (1.22-3.79)	0.008
KIR3DP1	0.98	0.97	0.97	0.97	5.9	2.41 (0.20-29.23)	1.56 (0.10-23.94)	0.61 (0.11-3.46)	2.14 (0.96-4.73)	0.49	0.75	0.58	0.06	1.75 (0.71-2.87)	0.12

6.4 Discussion

The aim of this chapter was to test the hypothesis that KIR gene region variation contributes to IPF susceptibility. Although SNPs in the KIR region have been included in GWAS, no significant signals of association with risk were identified for this region in previous IPF studies Furthermore, the complex haplotype structure and gene copy number variation of the KIR region has not previously been explored. Utilising the bespoke KIR imputation of haplotypes this study, which included up to 3,420 IPF cases and 18,559 controls, did not identify any signals of association with IPF susceptibility.

Weak signals of association were observed for haplotype 9, haplotype 8 and copy number of *KIR3DP1*. However, none of these met the specified criteria for significance which required signals to meet an overall Bonferroni-corrected threshold in the meta-analysis and be supported across at least two of the tested datasets at nominal significance. Haplotype 8 was close to Bonferroni-corrected significance in the meta-analysis and close to nominal significance in two studies, but this signal would require additional studies to confirm it.

In previous analyses (HLA association analysis in chapter 3) there appeared to be heterogeneity between the cases in the Colorado dataset compared to the UK, UUS and Chicago datasets in terms of HLA association signals. This does not appear to be the case for associations in the KIR region; the Colorado dataset had similar odds ratios, frequencies, posterior probabilities, and P-values (figures 4 and 5) to the results from the other IPF datasets. In the haplotype meta-analysis, the Colorado dataset had a lower frequency of haplotype 8, this could be due to the loss of two of the six tag SNPs in the imputation for Colorado (see chapter 5). SNP seq-t1d-19-60034052-C-T was removed from the input SNPs for the imputation of the Colorado dataset because there was a frequency mismatch; this SNP was a tag SNP for *KIR2DL5* which was found in several B haplotypes including haplotype 8 (208).

Measurement error may have had an impact on the analyses undertaken, as discussed in Chapter 5. There was concern associated with the quality of the imputation because there were frequency differences (for the haplotypes and CNVs) between the four IPF datasets. There was also a possibility that individuals were being misclassified by copy number because poorly imputed CNVs were removed in the previous chapter (Chapter 5 – KIR Imputation), for example, if an individual has one well-imputed copy of *KIR2DL1* on one chromosome and two copies of *KIR2DL1* on the second chromosome, but this was poorly imputed and excluded

during the quality control steps, they would have been classified as having only one copy of *KIR2DL1* when this may not be the case.

This first study of the contribution of KIR region haplotype and gene copy number variation to IPF susceptibility did not identify any significant associations, however haplotype 8 was close to meeting the significance thresholds. This could suggest that the KIR region has no role in IPF susceptibility, that the study sample size was underpowered to detect any true small effects, or because the accuracy and quality of the imputation was not high. In order to further improve the analyses undertaken, better measurement of KIR variation should be sought.

Chapter 7: Discussion

Idiopathic pulmonary fibrosis (IPF) is a chronic interstitial lung disease characterised by scarring and inflammation of the alveolar wall, resulting in decreased lung function and poor quality of life. Risk factors for IPF include age, smoking, infection, and genetics. IPF has poor prognosis, very limited treatment options and little is known about the development and progression of the disease. The HLA region codes for molecules that play a vital role in the immune response against bacterial and viral infection along with the KIR region which codes for receptors which are associated with the activation and inhibition of natural killer cells. Micro-injuries such as viral infection are believed to be triggers that lead to fibrosis in the lungs (7, 8, 60-63), therefore studying these regions could provide an insight into the development of IPF and could help to identify drug targets or diagnostic markers. Drug targets with genetic support are two-times more likely to be successful in clinical development (209) which further motivates genetic studies such as those undertaken in this thesis.

This final chapter presents a summary of the previous IPF susceptibility studies, a description of the main findings of this thesis and the clinical impact of these, the strengths and weaknesses of the analyses undertaken and recommendations for potential future work.

7.1 Summary of previous IPF susceptibility studies and HLA-DQB1*06:02 findings

Between the years of 2008 and 2017, 17 genome-wide significant signals were identified in GWAS of IPF susceptibility (2-6) (see Chapter 1, section 1.2.3, table 1.1), this included one signal in the HLA region, *HLA-DQB1*06:02 (*odds ratio 1.34 (confidence interval = 1.18-1.52, allele frequency of 15% in cases and 12% in controls) (5). The effect size of this variant is comparable to the IPF risk genetic signals (see Chapter 1, table 1.1) but was typically smaller than HLA signals for other common diseases (for example, HLA-B27 in Ankylosing Spondylitis OR=1.71-2.01 (96, 210) and HLA-DRB1 in Type 1 Diabetes OR = 0.28-0.29 (133)). The HLA-DQB1*06:02 signal was identified in 1,616 fibrotic idiopathic interstitial pneumonia (fIIP) cases and 4,683 controls from Colorado (of European ancestry) and replicated in 878 cases and 2,017 controls (identified in the same way as the discovery dataset).

The strongest association with IPF susceptibility was found in the promoter region of a mucin gene MUC5B which is associated with a four-fold increased risk and studies in mouse models suggest that the excess mucin makes the lining sticky increasing the quantity of cells and bacteria, initiating an abnormal response (58). In 2020, the largest genome wide association meta-analysis for IPF susceptibility confirmed 11 of the previously reported signals (not including HLA-DQB1*06:02) and reported 3 novel genome-wide signals, the signals implicated

telomere length, cell-cell adhesion and lung defence in IPF susceptibility (1) (see Chapter 1, section 1.2.3).

The 15 reported IPF susceptibility genetic association signals only explain around 12.4% of the disease in the general population (53). The imputation of complex variation that is not well-measured using standard SNP arrays and imputation could therefore explain more disease risk. Prior to this thesis, only the Colorado study had analysed the specific role of complex genetic variation across HLA genes (5) and complex variation in the KIR region has never been studied in IPF. This thesis aimed to address that gap in knowledge and utilised computational imputation techniques to enable an in-depth analysis of complex genetic variation across two groups of immune system genes and their role in IPF susceptibility.

7.2 Summary of work undertaken in this thesis

Chapter two describes the imputation of HLA gene alleles, amino acid alleles and SNPs across the HLA region in four IPF case-control datasets using the most up to date SNP panel (available at the time – haplotype reference consortium [HRC] (29)) and a bespoke HLA panel (T1DGC). Of the quality-controlled variants, there was a high mean imputation quality of 0.98 in the UK, UUS and Colorado datasets and 0.97 in the Chicago dataset and there appeared to be good concordance between the datasets when looking at the imputation quality and allele frequency. In total, 34,905-36,905 SNPs, HLA gene alleles and amino acid alleles across HLA-A, HLA-B, HLA-C, HLA-DPA, HLA-DPB, HLA-DQA, HLA-DQB and HLA-DRB passed the quality control measures across the four IPF datasets. Chapter three describes the largest HLA-wide association analyses of IPF susceptibility across European IPF datasets using two study designs (discovery-replication and meta-analysis). A discovery HLA-wide association analysis was undertaken using 612 IPF cases and 3,366 controls in which no signals passed the Bonferroni-corrected significance threshold (corrected for number of variants tested) and 12 suggestively significant signals did not replicate in 2,015 IPF and fibrotic idiopathic interstitial pneumonia (fIIP) cases and 5,193 controls. In the meta-analysis of the three available datasets (UK, Chicago and Colorado datasets, 2,769 IPF and fIIP cases and 8,591 controls), three independent signals passed the Bonferroni corrected significance threshold (corrected for number of variants tested) however they were only statistically significant in the Colorado dataset, one of these being the previously identified HLA-DQB1*06:02 signal. The HLA-DQB1*06:02 signal did not replicate in the meta-analysis of the UK and Chicago datasets suggesting that this signal was specific to the Colorado case population which included broader fIIP phenotypes. Due to these differences in the Colorado dataset, it was omitted in the following meta-analysis and replaced by the UUS dataset (which was not available

previously). In the meta-analysis of the UK, UUS and Chicago datasets, one signal passed the Bonferroni corrected significance threshold (rs3132684). The signal was in an intron, the nearest gene was *ZNRD1ASP,* and it was a common variant (frequency = 32-34%). rs3132684 and SNPs in the credible set were associated with several lung and immunity phenotypes including peak expiratory flow, eosinophil counts and Rheumatoid Arthritis suggesting the signal could be involved in relevant disease pathways. In the joint regression analysis (for the HLA amino acid changes), no amino acids were associated with IPF susceptibility.

Chapter four describes the first HLA-wide variant*_MUC5B_ interaction analyses of IPF susceptibility. The aim of this analysis was to identify if there were any differences between the *MUC5B* positive (individuals with at least one *MUC5B* risk allele) and negative groups (individuals with no *MUC5B* risk alleles) which could be attributed to signals in the HLA region. Chicago was excluded from these analyses because the *MUC5B* SNP was poorly imputed (imputation quality = 0.6). In the discovery interaction analysis of 612 IPF cases and 3,366 controls, no signals passed the Bonferroni corrected significance threshold (corrected for number of interactions tested) and seven suggestive signals ($P<5x10^{-3}$) did not replicate in 2,308 cases and 14,683 controls. In the meta-analysis of all three datasets (UK, UUS and Colorado) no signals passed the Bonferroni corrected significance threshold but five independent signals passed the suggestive significance threshold. Four of the five signals were associated with respiratory, and immunity related phenotypes including forced vital capacity and psoriasis and were significantly associated with differing risk of IPF in the two *MUC5B* groups (positive and negative). One signal was associated with reduced expression of several genes including *HLA-C* and *MICA* and was significantly associated with reduced risk of IPF in the *MUC5B* negative group.

Chapter five describes the imputation of KIR haplotypes and copy number variations (CNVs) across four IPF datasets using KIR*IMP (152). The KIR variation imputed in this chapter was copy number variation of the KIR genes and haplotypes (which are made up of various gene CNVs). The KIR imputation was dependent on a key set of tag SNPs and not all of these were directly genotyped in the input datasets. To address this, the inclusion of imputed (using HRC) SNPs was evaluated to identify if these SNPs could act as tags for the KIR imputation. SNPs that had a HRC imputation quality of more than 0.3 were included in the input for the KIR*IMP imputation. The haplotype and CNV frequencies were considerably different between the datasets which caused concern, since all four datasets were of European ancestry, it would be expected that the frequencies would be similar. No further exclusions were made in the four datasets because the study design required signals be significant in three of the four datasets

which mitigated the likelihood of false positive associations. Chapter six describes the first KIR association meta-analysis of IPF susceptibility across four IPF datasets. A four-way meta-analysis was undertaken on KIR haplotypes and gene copy number variations (CNVs). No haplotypes passed a Bonferroni corrected significance threshold (corrected for number of haplotypes tested). Haplotype 8 was the most statistically significant which was a B haplotype and made up on average 2% of the haplotypes in the individuals of the four IPF datasets. Haplotype 8 did not pass suggestive significance thresholds in two of the four datasets ($P<0.05$), however it did pass the threshold in the Colorado dataset and was close in the UUS dataset ($P=0.06$) suggesting this haplotype could be of interest. No CNVs passed the Bonferroni corrected significance threshold (corrected for the number of CNVs tested) but the most significant signal was copy number variation in *KIR3DP1*.

7.3 Clinical implications of the work undertaken in this thesis

Overall, the work in this thesis was not able to detect any associations int these regions which suggests that there are no large-effect associations with HLA and KIR variation and IPF susceptibility, although the suggestive findings may represent variants which make small contributions to genetic risk. Most notably, a novel signal was identified in the HLA region for IPF susceptibility (rs3132684 near *ZNRD1ASP)*. There was high confidence in this signal for several reasons: the signal was extremely well imputed across the three datasets analysed (imputed at an info score of 1.00), it was not rare (coded allele frequency ranged between 0.32-0.34 across studies) and the p-values and effect sizes were similar across the three datasets. In order to gain further confidence, this signal should be replicated in additional independent IPF datasets. One concern with this variant is why the SNP had not been detected in the previous GWAS of IPF susceptibility (1-5). In the IPF meta-analysis in 2020 (1), the variant was tested across the four IPF datasets, the variant was in the same direction of affect across the four datasets, however the P-value was not statistically significant ($P=0.017$) which could be due to the inclusion of all four datasets (including the Colorado dataset which included different phenotypes) or the different covariates used (1). The variant was associated with respiratory, autoimmune, and inflammatory traits and with the expression of several HLA and non-HLA genes, therefore suggests this variant is of interest for the underlying biological processes behind IPF development. Further characterisation of this signal to identify the pathways the variant is involved in and associations of this variant with gene expression in relevant cells and tissues could help identify new processes and pathways involved in the development of IPF and suggest new drug targets. Specifically targeting HLA class I, II or III HLA genes could reduce the downstream immune and inflammatory response initiated by Class I or

II genes as a response to for example a viral infection in the lung. In order to improve our understanding of the novel signal further, the analysis should be updated as new samples become available, they could either be incorporated into the meta-analysis or they could be used as replication datasets. Additionally, the credible set for the novel signal was very large (190 SNPs), it would therefore be useful to improve the fine mapping in these analyses to reduce the size of the credible sets in SNP signals that don't tag a specific HLA gene allele or amino acid. Fine mapping can be improved through increased sample sizes, use of denser panels, inclusion of multiple ancestries, incorporation of functional annotation and better methodologies. Utilising the HLA imputation method described in Chapter 2 potentially acts as a fine-mapping technique as you can map a SNP to a specific HLA gene allele or amino acid change however the signal identified in the HLA-wide meta-analysis was not associated with any specific HLA gene allele or amino acid change.

The previously reported HLA signal (HLA*DQB1*06:02) did not replicate in the three independent IPF datasets in this study (UK, UUS, Chicago). Although the signal was reported and replicated in independent populations by Fingerlin et al (5), it is notable that those populations included other fibrotic idiopathic interstitial pneumonias (fIIPs) (such as non-specific interstitial pneumonia, cryptogenic organizing pneumonia (COP), respiratory bronchiolitis-associated interstitial lung disease (RB-ILD) or desquamative interstitial pneumonia (DIP)). COP (211), RB-ILD (212) and DIP (213) all have an inflammatory component but they only make up ~1% of the cases in the Colorado dataset (5). 21% of the Colorado dataset consisted of non-specific pneumonia cases and unclassified interstitial pneumonias (not including COP, DIP or RB-ILD), it is possible that some of these cases have an ILD with a significant immune component such as Rheumatoid Arthritis ILD (RA-ILD) in which *HLA-DRB1* and *HLA-DQB1* alleles (including *HLA-DRB1*15:01* and *HLA-DQB1*06-02*) have already been identified as significantly associated (HLA-DRB1*15 with an effect size of 1.75 and HLA-DQB1*06 with an effect size of 0.57) (198, 199). The findings in this thesis suggests that the *HLA-DQB1*06-02* signal may reflect inclusion of non-IPF ILD in the Colorado dataset for which HLA variation is important.

Although this thesis did not confirm the presence of *MUC5B* interaction signals across the HLA region in IPF susceptibility, the analysis provided some interesting findings which could help identify different biological pathways in *MUC5B* positive and *MUC5B* negative groups. Five suggestively significant signals were identified including rs9265961 (intronic variant located near *LOC112267902* an RNA gene in the ncRNA class*)* where the minor allele was associated with reduced risk of IPF and was also found to be associated with reduced expression of

several genes including *HLA-C* and *MICA.* This finding appears to fit with a previous study that saw increased levels of *MICA* in IPF lungs (177). Since this variant was associated with a suggestively significantly reduced risk of IPF only in people who carried no copies of the *MUC5B* IPF risk allele, this could provide important insight when considering precision medicine approaches to treatment.

Since the work began on this thesis a new study has since emerged to suggest that viral load in IPF lungs is minimal (185), it was found that there was no statistical difference of the expression of many viruses including EBV and herpesvirus between healthy control lungs and IPF case lungs (185). However, another recent study found reduced activity of natural killer cells in IPF lungs compared to healthy control lungs (214). Also, if there is no difference in viral load between people who had IPF and people who don't, this does not rule out viral infection as a potential triggering factor before someone develops IPF. Further work could help decipher if infection prior to IPF diagnosis plays a role and also if viral infection has a role in the progression of IPF. Additional work could be undertaken to study viral infection genetic risk signals in IPF susceptibility to identify if there is any genetic overlap (for example a study in 2017 utilised HLA imputation tool HIBAG (89) to fine-map signals from a GWAS of common infections (215)). If a signal in the HLA or KIR regions were identified to be associated with IPF in the analyses in this thesis, further wet laboratory analyses would be required such as quantifying the amount of protein in the lungs and identifying the affect of the variant in IPF lungs vs healthy lungs. These analyses would enable further characterisation of the variant and identify possible clinical application such as drug targeting or diagnostic methods.

7.4 Strengths and limitations

The biggest strength of this thesis is that it represents the largest HLA-wide association and KIR-wide association analyses for IPF susceptibility undertaken to date. The analyses utilised imputation that enabled analysis beyond simple SNP and indel variation across two complex loci (HLA and KIR). In the HLA association analysis this thesis, a total of 1,905 IPF cases and 13,876 controls were analysed, and the cases included only clinician diagnosed IPF. The HLA imputation strategy is also novel, merging the SNP imputation panel and the bespoke HLA panel to incorporate a large quantity of SNPs as well as the HLA gene alleles and amino acid alleles has not been done in IPF before. Although SNPs in the KIR region have been included in previous GWAS, the structural complexity of the region means that the variation in the region cannot be appropriately capture by the analysis of only SNPs. This was the first IPF susceptibility study which incorporated KIR structural variation at the haplotype and gene copy number level.

The research involved in this thesis has some limitations. Most significantly the analyses across chapters 3, 4 and 6 were underpowered to detect association signals with small effects, particularly for low frequency and rare variants. Because IPF is a relatively rare lung disease, sample sizes were modest across all the analyses, despite multiple individual datasets from the UK, US and Europe being combined. To maximise power, meta-analyses were undertaken with all the datasets available at the time. This resulted in the largest HLA-wide association analysis and HLA-MUC5B interaction analysis undertaken to date. However, as all datasets were then included in the discovery endeavour, there were no further independent datasets available to replicate the findings. This limitation was mitigated by requiring signals to reach nominal significance in more than one contributing study, thereby reducing the likelihood of reporting false positive signals arising from one dataset. Even with incorporating all the datasets available, power was limited and this was reflected in power calculations especially in the interaction meta-analyses (which are classically less powered than logistic regression analyses) which suggested there was only a 17% power to detect associations (at a coded allele frequency of 10%, interaction effect size of 1.1, alpha of 0.05).

A limitation of the HLA analyses in this thesis is that the HLA imputation imputed only classical HLA genes, however, since the analyses in this thesis were undertaken, a new HLA imputation method was released (216). MHC*IMP also enables the imputation of non-classical HLA genes (including *HLA-E*, *HLA-F* and *HLA-G*) which have been shown to be important in KIR recognition (HLA-E (217)) and autoimmune disorders (HLA-E and psoriasis (218) and HLA-G and Crohn's disease(219)) and polymorphic non-HLA genes including *TAP1/TAP2* (involved in the HLA's recognition of foreign bodies) and *MICA/MICB* (encode ligands for natural killer cell receptors) (216). Studies have shown that there is significantly increased expression of *MICA* alleles in IPF lungs (177), therefore incorporating the data from this imputation will allow an even more in-depth analysis of polygenic classical HLA genes, non-classical HLA genes and non-HLA genes across the region and their role in IPF.

Another limitation of the work undertaken in this thesis was the quality of the KIR imputation. There were frequency differences for many of the KIR haplotypes and CNVs between the IPF datasets which would not be expected since all cases and controls were of European descent. The poor imputation may have obscured true positive signals if they were not imputed correctly or efficiently throughout the datasets. KIR association analyses were undertaken with the caveat that a positive finding would need replication and validation through laboratory testing or typing by whole genome sequencing typing.

7.5 Future work

Sample size, power, and genomic coverage

Definitive studies about the role of HLA and KIR genetic variation in IPF susceptibility require larger sample sizes of well-phenotyped IPF cases and controls. to improve power. Improved SNP coverage in the HLA and KIR regions may also further improve the ability to impute gene alleles, amino acid changes, haplotypes and copy number variation. Since the analysis for this thesis was undertaken, an updated SNP imputation panel (TOPMed) has been released for use which contains 308,107,085 genetic variants across the autosomes and X chromosome (17). The analyses in this thesis only considered common variation (frequency > 1%) and therefore to further develop this work, rare variation could be tested to identify if rare variation in the HLA or KIR region are associated with IPF susceptibility. This could be done by utilising sequencing techniques including exome sequencing or long read sequencing to identify KIR CNVs, KIR haplotypes and HLA gene alleles. In addition to this, this thesis only studied the role of HLA and KIR variation but additional immune system genes such as the interleukin (IL) family of genes are also associated with immune response to viral infection that could contribute to IPF risk. Finally, whole genome sequencing would enable the analysis of rare variants and direct inference of HLA gene alleles, amino acid alleles and KIR haplotypes and CNVs more effectively. Long read sequencing is typically considered the gold standard approach for HLA and KIR typing however this approach is costly and is not currently possible in such large sample sizes in IPF.

The role of HLA and KIR in IPF progression and survival:

This thesis covered only the role of immune system genes in susceptibility to IPF however viral infection is well known to also be a major cause of exacerbation events, therefore these genes may have a role in IPF disease progression and survival time. Treatment options for IPF are currently limited and ineffective, therefore identifying genetic determinants of progression and survival is important because it may discover genes or proteins that could be targeted for personalised treatment options. This knowledge is currently limited because sample sizes are limited.

7.5 Conclusion

This thesis explored the role of complex genetic variation in the HLA and KIR regions in IPF susceptibility. Bespoke HLA and KIR imputation methods were utilised to test the association of variants across these regions. These analyses suggested that the previously reported HLA association (HLA-DQB1*06:02) was likely driven by other forms of fIIP that were included in

the original study and was likely not an important risk factor for IPF itself. Additionally, a novel signal in the HLA region was identified which was found to be associated with respiratory and autoimmune traits as well as differential expression of several HLA and non-HLA genes in lung tissue. However, this signal should be replicated in an additional independent dataset and validated to confirm its potential importance.

Only suggestively significant signals were identified in the *MUC5B* interaction analysis. However, this study was underpowered, and these signals may pass significance thresholds when there is more data available. The findings in the interaction analyses seem to suggest a protective role of *MICA* in *MUC5B* negative IPF cases which could help to stratify IPF cases and to identify different biological pathways in the two case groups (*MUC5B* positive and *MUC5B* negative). No KIR CNVs or haplotypes passed statistical significance thresholds however measurement error in the imputation may have reduced power to detect true positive associations. As new datasets become available and power increases, the novel signal in the HLA-wide meta-analysis may be confirmed and the suggestive signals in the interaction meta-analysis could pass significance thresholds. Although the analyses presented here have not provided definitive evidence for a role of HLA or KIR variation in IPF susceptibility, the findings suggest that larger studies with more accurate and comprehensive imputation may identify new genetic variants that contribute to genetic risk, albeit with small effect sizes.

References:

1. Allen RJ, Guillen-Guio B, Oldham JM, Ma SF, Dressen A, Paynton ML, et al. Genome-Wide Association Study of Susceptibility to Idiopathic Pulmonary Fibrosis. Am J Respir Crit Care Med. 2020;201(5):564-74.
2. Allen RJ, Porte J, Braybrooke R, Flores C, Fingerlin TE, Oldham JM, et al. Genetic variants associated with susceptibility to idiopathic pulmonary fibrosis in people of European ancestry: a genome-wide association study. Lancet Respir Med. 2017.
3. Noth I, Zhang Y, Ma SF, Flores C, Barber M, Huang Y, et al. Genetic variants associated with idiopathic pulmonary fibrosis susceptibility and mortality: a genome-wide association study. Lancet Respir Med. 2013;1(4):309-17.
4. Fingerlin TE, Murphy E, Zhang W, Peljto AL, Brown KK, Steele MP, et al. Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. Nat Genet. 2013;45(6):613-20.
5. Fingerlin TE, Zhang W, Yang IV, Ainsworth HC, Russell PH, Blumhagen RZ, 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 Genet. 2016;17(1):74.
6. Mushiroda T, Wattanapokayakit S, Takahashi A, Nukiwa T, Kudoh S, Ogura T, 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.
7. Moore BB, Moore TA. Viruses in Idiopathic Pulmonary Fibrosis. Etiology and Exacerbation. Ann Am Thorac Soc. 2015;12 Suppl 2:S186-92.
8. Magro CM, Allen J, Pope-Harman A, Waldman WJ, Moh P, Rothrauff S, et al. The role of microvascular injury in the evolution of idiopathic pulmonary fibrosis. Am J Clin Pathol. 2003;119(4):556-67.
9. Papi A, Bellettato CM, Braccioni F, Romagnoli M, Casolari P, Caramori G, et al. Infections and airway inflammation in chronic obstructive pulmonary disease severe exacerbations. Am J Respir Crit Care Med. 2006;173(10):1114-21.
10. Bafadhel M, McKenna S, Terry S, Mistry V, Reid C, Haldar P, et al. Acute exacerbations of chronic obstructive pulmonary disease: identification of biologic clusters and their biomarkers. Am J Respir Crit Care Med. 2011;184(6):662-71.
11. Guilbert TW, Denlinger LC. Role of infection in the development and exacerbation of asthma. Expert Rev Respir Med. 2010;4(1):71-83.
12. Shiina T, Hosomichi K, Inoko H, Kulski JK. The HLA genomic loci map: expression, interaction, diversity and disease. J Hum Genet. 2009;54(1):15-39.
13. Yawata M, Yawata N, Abi-Rached L, Parham P. Variation within the human killer cell immunoglobulin-like receptor (KIR) gene family. Crit Rev Immunol. 2002;22(5-6):463-82.
14. Moffatt MF, Gut IG, Demenais F, Strachan DP, Bouzigon E, Heath S, et al. A large-scale, consortium-based genomewide association study of asthma. N Engl J Med. 2010;363(13):1211-21.
15. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 2001;409(6822):928-33.
16. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061-73.
17. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes

from the NHLBI TOPMed Program. BioRxiv. 2019.
18. Slatkin M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. Nat Rev Genet. 2008;9(6):477-85.

19. Bush WS, Moore JH. Chapter 11: Genome-wide association studies. PLoS Comput Biol. 2012;8(12):e1002822.
20. Traherne J. Human MHC architecture and evolution: implications for disease association studies. International Journal of Immunogenetics. 2008;35(3):179-92.
21. Miretti MM, Walsh EC, Ke X, Delagado M, Griffiths M, Hunt S, et al. A High-Resolution Linkage-Disequilibrium Map of the Human Major Histocompatibility Complex and First Generation of Tag Single-Nucleotide Polymorphisms. American Journal of Human Genetics. 2005;76(4):634-46.
22. EMBL-EBI. GWAS Catalog 2018 [cited 2018 29th Nov]. Available from: https://www.ebi.ac.uk/gwas/.
23. Genebass. Genebass 2021 [Available from: https://genebass.org/.
24. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. Genome Med. 2020;12(1):44.
25. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet. 2018;50(9):1219-24.
26. Sakornsakolpat P, Prokopenko D, Lamontagne M, Reeve NF, Guyatt AL, Jackson VE, et al. Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations. Nat Genet. 2019;51(3):494-505.
27. McDonald J. Handbook of Biological Statistics. 3rd ed. Baltimore,

Maryland: Sparky

House Publishing; 2014. 254-60 p.
28. Pulit SL, de With SA, de Bakker PI. Resetting the bar: Statistical significance in whole-genome sequencing-based association studies of global populations. Genet Epidemiol. 2017;41(2):145-51.
29. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016;48(10):1279-83.
30. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annu Rev Genomics Hum Genet. 2009;10:387-406.
31. Halperin E, Stephan DA. SNP imputation in association studies. Nat Biotechnol. 2009;27(4):349-51.
32. The Genomes Project C, Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, et al. A global reference for human genetic variation. Nature. 2015;526:68.
33. Wilde S, Timpson A, Kirsanow K, Kaiser E, Kayser M, Unterlander M, et al. Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. Proc Natl Acad Sci U S A. 2014;111(13):4832-7.
34. Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet. 2018;50(9):1335-41.
35. Wolters PJ, Collard HR, Jones KD. Pathogenesis of idiopathic pulmonary fibrosis. Annu Rev Pathol. 2014;9:157-79.
36. Betensley A, Sharif R, Karamichos D. A Systematic Review of the Role of Dysfunctional Wound Healing in the Pathogenesis and Treatment of Idiopathic Pulmonary Fibrosis. J Clin Med. 2016;6(1).
37. Noble PW, Albera C, Bradford WZ, Costabel U, Glassberg MK, Kardatzke D, et al. Pirfenidone in patients with idiopathic pulmonary fibrosis (CAPACITY): two randomised trials. Lancet. 2011;377(9779):1760-9.
38. Richeldi L, Costabel U, Selman M, Kim DS, Hansell DM, Nicholson AG, et al. Efficacy of a tyrosine kinase inhibitor in idiopathic pulmonary fibrosis. N Engl J Med. 2011;365(12):1079-87.

39. Richeldi L, du Bois RM, Raghu G, Azuma A, Brown KK, Costabel U, et al. Efficacy and safety of nintedanib in idiopathic pulmonary fibrosis. N Engl J Med. 2014;370(22):2071-82.
40. NICE. Nintedanib for treating idiopathic pulmonary fibrosis 2016 [Available from: https://www.nice.org.uk/guidance/TA379/chapter/1-Recommendations.
41. NICE. Pirfenidone for treating idiopathic pulmonary fibrosis 2013 [Available from: https://www.nice.org.uk/guidance/ta282/chapter/1-guidance.
42. 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.
43. Bellou V, Belbasis L, Konstantinidis A, Evangelou E. Tobacco smoking and risk for idiopathic pulmonary fibrosis: a prospective cohort study in UK Biobank. European Respiratory Journal. 2017;50(suppl 61):PA4887.
44. Armanios MY, Chen JJ, Cogan JD, Alder JK, Ingersoll RG, Markin C, et al. Telomerase mutations in families with idiopathic pulmonary fibrosis. N Engl J Med. 2007;356(13):1317-26.
45. Tsakiri KD, Cronkhite JT, Kuan PJ, Xing C, Raghu G, Weissler JC, et al. Adult-onset pulmonary fibrosis caused by mutations in telomerase. Proc Natl Acad Sci U S A. 2007;104(18):7552-7.
46. Strek ME. Gender in idiopathic pulmonary fibrosis diagnosis: time to address unconscious bias. Thorax. 2020;75(5):365-6.
47. Peabody JW, Peabody JW, Jr., Hayes EW, Hayes EW, Jr. Idiopathic pulmonary fibrosis; its occurrence in identical twin sisters. Dis Chest. 1950;18(4):330-44.
48. Javaheri S, Lederer DH, Pella JA, Mark GJ, Levine BW. Idiopathic pulmonary fibrosis in monozygotic twins. The importance of genetic predisposition. Chest. 1980;78(4):591-4.
49. Seibold MA, Wise AL, Speer MC, Steele MP, Brown KK, Loyd JE, et al. A common MUC5B promoter polymorphism and pulmonary fibrosis. N Engl J Med. 2011;364(16):1503-12.
50. Zheng X, Shen J, Cox C, Wakefield JC, Ehm MG, Nelson MR, et al. HIBAG--HLA genotype imputation with attribute bagging. Pharmacogenomics J. 2014;14(2):192-200.
51. Young AI. Solving the missing heritability problem. PLoS Genet. 2019;15(6):e1008222.
52. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747-53.
53. Leavy OC, Ma SF, Molyneaux PL, Maher TM, Oldham JM, Flores C, et al. Proportion of Idiopathic Pulmonary Fibrosis Risk Explained by Known Common Genetic Loci in European Populations. Am J Respir Crit Care Med. 2021;203(6):775-8.
54. Dhindsa RS, Mattsson J, Nag A, Wang Q, Wain LV, Allen R, et al. Identification of a missense variant in SPDL1 associated with idiopathic pulmonary fibrosis. Commun Biol. 2021;4(1):392.
55. 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.
56. Zhu QQ, Zhang XL, Zhang SM, Tang SW, Min HY, Yi L, et al. Association Between the MUC5B Promoter Polymorphism rs35705950 and Idiopathic Pulmonary Fibrosis: A Meta-analysis and Trial Sequential Analysis in Caucasian and Asian Populations. Medicine (Baltimore). 2015;94(43):e1901.
57. OMIM. Mucin 5, Subtype B, Tracheobronchal, MUC5B 2014 [Available from: https://www.omim.org/entry/600770.
58. Hancock LA, Hennessy CE, Solomon GM, Dobrinskikh E, Estrella A, Hara N, et al. Muc5b overexpression causes mucociliary dysfunction and enhances lung fibrosis in mice. Nat Commun. 2018;9(1):5363.
59. Hunninghake GM, Hatabu H, Okajima Y, Gao W, Dupuis J, Latourelle JC, et al. MUC5B promoter polymorphism and interstitial lung abnormalities. N Engl J Med. 2013;368(23):2192-200.

60. Tang YW, Johnson JE, Browning PJ, Cruz-Gervis RA, Davis A, Graham BS, et al. Herpesvirus DNA is consistently detected in lungs of patients with idiopathic pulmonary fibrosis. J Clin Microbiol. 2003;41(6):2633-40.
61. Yonemaru M, Kasuga I, Kusumoto H, Kunisawa A, Kiyokawa H, Kuwabara S, et al. Elevation of antibodies to cytomegalovirus and other herpes viruses in pulmonary fibrosis. Eur Respir J. 1997;10(9):2040-5.
62. Vergnon JM, Vincent M, de The G, Mornex JF, Weynants P, Brune J. Cryptogenic fibrosing alveolitis and Epstein-Barr virus: an association? Lancet. 1984;2(8406):768-71.
63. Manika K, Alexiou-Daniel S, Papakosta D, Papa A, Kontakiotis T, Patakas D, et al. Epstein-Barr virus DNA in bronchoalveolar lavage fluid from patients with idiopathic pulmonary fibrosis. Sarcoidosis Vasc Diffuse Lung Dis. 2007;24(2):134-40.
64. Folcik VA, Garofalo M, Coleman J, Donegan JJ, Rabbani E, Suster S, et al. Idiopathic pulmonary fibrosis is strongly associated with productive infection by herpesvirus saimiri. Mod Pathol. 2014;27(6):851-62.
65. Naik PK, Moore BB. Viral infection and aging as cofactors for the development of pulmonary fibrosis. Expert Rev Respir Med. 2010;4(6):759-71.
66. Sheng G, Chen P, Wei Y, Yue H, Chu J, Zhao J, et al. Viral Infection Increases the Risk of Idiopathic Pulmonary Fibrosis: A Meta-Analysis. Chest. 2020;157(5):1175-87.
67. Trowsdale J, Knight JC. Major Histocompatibility complex genomics and human disease. Annual Review of Genomics and Human Genetics. 2013;14(1):301-23.
68. Consortium GR. Human Genome Region MHC 2017 [Available from: https://www.ncbi.nlm.nih.gov/grc/human/regions/MHC?asm=GRCh37.
69. Alleles H. HLA Antigens 2017 [Available from: http://hla.alleles.org/antigens/recognised_serology.html.
70. Mytilineos J, Scherer S, Dunckley H, Chapman J, Middleton D, Opelz G. Comparison of serological and DNA HLA-DR typing results for transplantation in Western Europe, Eastern Europe, North America and South America. Transplant International. 1994;7(Suppl 1):519-21.
71. Roberts RJ, Carneiro MO, Schatz MC. The advantages of SMRT sequencing. Genome Biol. 2013;14(7):405.
72. Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. 2016;17(1):239.
73. Mayor NP, Hayhurst JD, Turner TR, Szydlo RM, Shaw BE, Bultitude WP, et al. Recipients Receiving Better HLA-Matched Hematopoietic Cell Transplantation Grafts, Uncovered by a Novel HLA Typing Method, Have Superior Survival: A Retrospective Study. Biol Blood Marrow Transplant. 2019;25(3):443-50.
74. Dilthey AT, Moutsianas L, Leslie S, McVean G. HLA*IMP--an integrated framework for imputing classical HLA alleles from SNP genotypes. Bioinformatics. 2011;27(7):968-72.
75. Reference GH. Chromosome 6 2018 [Available from: https://ghr.nlm.nih.gov/chromosome/6#idiogram.
76. Alleles H. Nomenclature for Factors of the HLA System 2017 [Available from: http://hla.alleles.org/nomenclature/naming.html.
77. Alleles H. HLA Genes 2017 [Available from: http://hla.alleles.org/genes/index.html.
78. EMBL-EBI. IPD-KIR 2018 [Available from: https://www.ebi.ac.uk/ipd/kir/.
79. Wieczorek M, Abualrous E, Sticht J, Alvaro-Benito M, Stolzenberg S, Noe F, et al. Major Histocompatibility Complex (MHC) Class I and MHC Class II Proteins: Conformational Plasticity in Antigen Presentation. Frontiers in Immunology. 2017;8(292):Epub.
80. Blum JS, Wearsch PA, Cresswell P. Pathways of Antigen Processing. Annual Reviews of Immunology. 2013;31(1):443-73.
81. Alleles H. HLA Alleles Numbers 2017 [Available from: http://hla.alleles.org/nomenclature/stats.html.

82. Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, Khodiyar VK, et al. Gene map of the extended human MHC. Nat Rev Genet. 2004;5(12):889-99.
83. Cullen M, Perfetto SP, Klitz W, Nelson G, Carrington M. High-Resolution patterns of Meiotic Recombination across the Human Major Histocompatibility Complex. American Journal of Human Genetics. 2002;71(4):759-76.
84. Lam TH, Shen M, Chia JM, Chan SH, Ren EC. Population-specific recombination sites within the human MHC region. Heredity (Edinb). 2013;111(2):131-8.
85. Jia X, Han B, Onengut-Gumuscu S, Chen WM, Concannon PJ, Rich SS, et al. Imputing amino acid polymorphisms in human leukocyte antigens. PLoS One. 2013;8(6):e64683.
86. Robinson J, Barker DJ, Georgiou X, Cooper MA, Flicek P, Marsh SGE. IPD-IMGT/HLA Database. Nucleic Acids Res. 2020;48(D1):D948-D55.
87. International HapMap C, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;449(7164):851-61.
88. Stephens M, Scheet P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet. 2005;76(3):449-62.
89. Zheng X. HIBAG – an R Package for HLA Genotype Imputation with Attribute Bagging 2016 [Available from: https://www.bioconductor.org/packages/devel/bioc/vignettes/HIBAG/inst/doc/HIBAG_Tutorial.html.
90. Karnes JH, Shaffer CM, Bastarache L, Gaudieri S, Glazer AM, Steiner HE, et al. Comparison of HLA allelic imputation programs. PLoS One. 2017;12(2):e0172444.
91. Janeway C, Travers P, Walport M, Shlomchik M. The Major Histocompatibility Complex and it's functions. Immunology: The Immune System in Health and Disease. 5 ed. New York: Garland Science; 2001. p. 155-85.
92. Kuiper JJ, Van Setten J, Ripke S, Van TSR, Mulder F, Missotten T, et al. A genome-wide association study identifies a functional ERAP2 haplotype associated with birdshot chorioretinopathy. Hum Mol Genet. 2014;23(22):6081-7.
93. Bei JX, Li Y, Jia WH, Feng BJ, Zhou G, Chen LZ, et al. A genome-wide association study of nasopharyngeal carcinoma identifies three new susceptibility loci. Nat Genet. 2010;42(7):599-603.
94. Tse KP, Su WH, Chang KP, Tsang NM, Yu CJ, Tang P, et al. Genome-wide association study reveals multiple nasopharyngeal carcinoma-associated loci within the HLA region at chromosome 6p21.3. Am J Hum Genet. 2009;85(2):194-203.
95. Jin Y, Andersen G, Yorgov D, Ferrara TM, Ben S, Brownson KM, et al. Genome-wide association studies of autoimmune vitiligo identify 23 new risk loci and highlight key pathways and regulatory variants. Nat Genet. 2016;48(11):1418-24.
96. International Genetics of Ankylosing Spondylitis C, Cortes A, Hadler J, Pointon JP, Robinson PC, Karaderi T, et al. Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci. Nat Genet. 2013;45(7):730-8.
97. Autism Spectrum Disorders Working Group of The Psychiatric Genomics C. Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. Mol Autism. 2017;8:21.
98. Tsoi LC, Spain SL, Knight J, Ellinghaus E, Stuart PE, Capon F, et al. Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. Nat Genet. 2012;44(12):1341-8.
99. Stuart PE, Nair RP, Tsoi LC, Tejasvi T, Das S, Kang HM, et al. Genome-wide Association Analysis of Psoriatic Arthritis and Cutaneous Psoriasis Reveals Differences in Their Genetic Architecture. Am J Hum Genet. 2015;97(6):816-36.

100. Capon F, Bijlmakers MJ, Wolf N, Quaranta M, Huffmeier U, Allen M, et al. Identification of ZNF313/RNF114 as a novel psoriasis susceptibility gene. Hum Mol Genet. 2008;17(13):1938-45.
101. Evans DM, Spencer CC, Pointon JJ, Su Z, Harvey D, Kochan G, et al. Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. Nat Genet. 2011;43(8):761-7.
102. Stanescu HC, Arcos-Burgos M, Medlar A, Bockenhauer D, Kottgen A, Dragomirescu L, et al. Risk HLA-DQA1 and PLA(2)R1 alleles in idiopathic membranous nephropathy. N Engl J Med. 2011;364(7):616-26.
103. Chu X, Pan CM, Zhao SX, Liang J, Gao GQ, Zhang XM, et al. A genome-wide association study identifies two new risk loci for Graves' disease. Nat Genet. 2011;43(9):897-901.
104. Genetic Analysis of Psoriasis C, the Wellcome Trust Case Control C, Strange A, Capon F, Spencer CC, Knight J, et al. A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. Nat Genet. 2010;42(11):985-90.
105. Nair RP, Duffin KC, Helms C, Ding J, Stuart PE, Goldgar D, et al. Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. Nat Genet. 2009;41(2):199-204.
106. Liu Y, Helms C, Liao W, Zaba LC, Duan S, Gardner J, et al. A genome-wide association study of psoriasis and psoriatic arthritis identifies new disease loci. PLoS Genet. 2008;4(3):e1000041.
107. Ellinghaus E, Ellinghaus D, Stuart PE, Nair RP, Debrus S, Raelson JV, et al. Genome-wide association study identifies a psoriasis susceptibility locus at TRAF3IP2. Nat Genet. 2010;42(11):991-5.
108. de Lange KM, Moutsianas L, Lee JC, Lamb CA, Luo Y, Kennedy NA, et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. Nat Genet. 2017;49(2):256-61.
109. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012;491(7422):119-24.
110. Jung ES, Cheon JH, Lee JH, Park SJ, Jang HW, Chung SH, et al. HLA-C*01 is a Risk Factor for Crohn's Disease. Inflamm Bowel Dis. 2016;22(4):796-806.
111. Quan C, Ren YQ, Xiang LH, Sun LD, Xu AE, Gao XH, et al. Genome-wide association study for vitiligo identifies susceptibility loci at 6q27 and the MHC. Nat Genet. 2010;42(7):614-8.
112. Hirota T, Takahashi A, Kubo M, Tsunoda T, Tomita K, Sakashita M, et al. Genome-wide association study identifies eight new susceptibility loci for atopic dermatitis in the Japanese population. Nat Genet. 2012;44(11):1222-6.
113. Raychaudhuri S, Remmers EF, Lee AT, Hackett R, Guiducci C, Burtt NP, et al. Common variants at CD40 and other loci confer risk of rheumatoid arthritis. Nat Genet. 2008;40(10):1216-23.
114. Hinks A, Cobb J, Marion MC, Prahalad S, Sudman M, Bowes J, et al. Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis. Nat Genet. 2013;45(6):664-9.
115. Lessard CJ, Li H, Adrianto I, Ice JA, Rasmussen A, Grundahl KM, et al. Variants at multiple loci implicated in both innate and adaptive immune responses are associated with Sjogren's syndrome. Nat Genet. 2013;45(11):1284-92.
116. Bentham J, Morris DL, Graham DSC, Pinder CL, Tombleson P, Behrens TW, et al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. Nat Genet. 2015;47(12):1457-64.

117. Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, Thomson BP, et al. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat Genet. 2010;42(6):508-14.
118. Eyre S, Bowes J, Diogo D, Lee A, Barton A, Martin P, et al. High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. Nat Genet. 2012;44(12):1336-40.
119. Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature. 2014;506(7488):376-81.
120. Orozco G, Viatte S, Bowes J, Martin P, Wilson AG, Morgan AW, et al. Novel rheumatoid arthritis susceptibility locus at 22q12 identified in an extended UK genome-wide association study. Arthritis Rheumatol. 2014;66(1):24-30.
121. Plenge RM, Seielstad M, Padyukov L, Lee AT, Remmers EF, Ding B, et al. TRAF1-C5 as a risk locus for rheumatoid arthritis--a genomewide study. N Engl J Med. 2007;357(12):1199-209.
122. Kochi Y, Okada Y, Suzuki A, Ikari K, Terao C, Takahashi A, et al. A regulatory variant in CCR6 is associated with rheumatoid arthritis susceptibility. Nat Genet. 2010;42(6):515-9.
123. Wellcome Trust Case Control C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447(7145):661-78.
124. Freudenberg J, Lee HS, Han BG, Shin HD, Kang YM, Sung YK, et al. Genome-wide association study of rheumatoid arthritis in Koreans: population-specific loci as well as overlap with European susceptibility loci. Arthritis Rheum. 2011;63(4):884-93.
125. Govind N, Choudhury A, Hodkinson B, Ickinger C, Frost J, Lee A, et al. Immunochip identifies novel, and replicates known, genetic risk loci for rheumatoid arthritis in black South Africans. Mol Med. 2014;20:341-9.
126. Saxena R, Plenge RM, Bjonnes AC, Dashti HS, Okada Y, Gad El Haq W, et al. A Multinational Arab Genome-Wide Association Study Identifies New Genetic Associations for Rheumatoid Arthritis. Arthritis Rheumatol. 2017;69(5):976-85.
127. De Jager PL, Jia X, Wang J, de Bakker PI, Ottoboni L, Aggarwal NT, et al. Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. Nat Genet. 2009;41(7):776-82.
128. Patsopoulos NA, Bayer Pharma MSGWG, Steering Committees of Studies Evaluating I-b, a CCRA, Consortium AN, GeneMsa, et al. Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci. Ann Neurol. 2011;70(6):897-912.
129. Australia, New Zealand Multiple Sclerosis Genetics C. Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. Nat Genet. 2009;41(7):824-8.
130. Sanna S, Pitzalis M, Zoledziewska M, Zara I, Sidore C, Murru R, et al. Variants within the immunoregulatory CBLB gene are associated with multiple sclerosis. Nat Genet. 2010;42(6):495-7.
131. Nischwitz S, Cepok S, Kroner A, Wolf C, Knop M, Muller-Sarnowski F, et al. Evidence for VAV2 and ZNF433 as susceptibility genes for multiple sclerosis. J Neuroimmunol. 2010;227(1-2):162-6.
132. Comabella M, Craig DW, Camina-Tato M, Morcillo C, Lopez C, Navarro A, et al. Identification of a novel risk locus for multiple sclerosis at 13q31.3 by a pooled genome-wide scan of 500,000 single nucleotide polymorphisms. PLoS One. 2008;3(10):e3490.
133. Cooper JD, Smyth DJ, Smiles AM, Plagnol V, Walker NM, Allen JE, et al. Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. Nat Genet. 2008;40(12):1399-401.
134. Hakonarson H, Grant SF, Bradfield JP, Marchand L, Kim CE, Glessner JT, et al. A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. Nature. 2007;448(7153):591-4.

135. Liu JZ, van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet. 2015;47(9):979-86.
136. Anderson CA, Boucher G, Lees CW, Franke A, D'Amato M, Taylor KD, et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nat Genet. 2011;43(3):246-52.
137. Taylor KE, Wong Q, Levine DM, McHugh C, Laurie C, Doheny K, et al. Genome-Wide Association Analysis Reveals Genetic Heterogeneity of Sjogren's Syndrome According to Ancestry. Arthritis Rheumatol. 2017;69(6):1294-305.
138. Li Y, Zhang K, Chen H, Sun F, Xu J, Wu Z, et al. A genome-wide association study in Han Chinese identifies a susceptibility locus for primary Sjogren's syndrome at 7q11.23. Nat Genet. 2013;45(11):1361-5.
139. Horowitz A, Behrens RH, Okell L, Fooks AR, Riley EM. NK cells as effectors of acquired immune responses: effector CD4+ T cell-dependent activation of NK cells following vaccination. J Immunol. 2010;185(5):2808-18.
140. Moretta A, Tambussi G, Bottino C, Tripodi G, Merli A, Ciccone E, et al. A novel surface antigen expressed by a subset of human CD3- CD16+ natural killer cells. Role in cell activation and regulation of cytolytic function. J Exp Med. 1990;171(3):695-714.
141. Ljunggren HG, Karre K. In search of the 'missing self': MHC molecules and NK cell recognition. Immunol Today. 1990;11(7):237-44.
142. Parham P. MHC class I molecules and KIRs in human history, health and survival. Nat Rev Immunol. 2005;5(3):201-14.
143. Robinson J, Mistry K, McWilliam H, Lopez R, Marsh SG. IPD--the Immuno Polymorphism Database. Nucleic Acids Res. 2010;38(Database issue):D863-9.
144. Bondzio I, Arendt A, Schmitz J, Huppert V. KIR Positive Subsets of Human CD56+CD3+ NKT Cells Are Different in Frequency from Equivalant KIR Positive CD56+CD3- NK Cell Subsets. Blood. 2007;110(11):3878-.
145. Anfossi N, Doisne JM, Peyrat MA, Ugolini S, Bonnaud O, Bossy D, et al. Coordinated expression of Ig-like inhibitory MHC class I receptors and acquisition of cytotoxic function in human CD8+ T cells. J Immunol. 2004;173(12):7223-9.
146. Snyder MR, Nakajima T, Leibson PJ, Weyand CM, Goronzy JJ. Stimulatory killer Ig-like receptors modulate T cell activation through DAP12-dependent and DAP12-independent mechanisms. J Immunol. 2004;173(6):3725-31.
147. Marsh SG, Parham P, Dupont B, Geraghty DE, Trowsdale J, Middleton D, et al. Killer-cell immunoglobulin-like receptor (KIR) nomenclature report, 2002. Hum Immunol. 2003;64(6):648-54.
148. Jiang W, Johnson C, Jayaraman J, Simecek N, Noble J, Moffatt MF, et al. Copy number variation leads to considerable diversity for B but not A haplotypes of the human KIR genes encoding NK cell receptors. Genome Res. 2012;22(10):1845-54.
149. Hsu KC, Liu XR, Selvakumar A, Mickelson E, O'Reilly RJ, Dupont B. Killer Ig-like receptor haplotype analysis by gene content: evidence for genomic diversity with a minimum of six basic framework haplotypes, each with multiple subsets. J Immunol. 2002;169(9):5118-29.
150. Ivarsson MA, Michaelsson J, Fauriat C. Activating killer cell Ig-like receptors in health and disease. Front Immunol. 2014;5:184.
151. Pende D, Falco M, Vitale M, Cantoni C, Vitale C, Munari E, et al. Killer Ig-Like Receptors (KIRs): Their Role in NK Cell Modulation and Developments Leading to Their Clinical Exploitation. Front Immunol. 2019;10:1179.
152. Vukcevic D, Traherne JA, Naess S, Ellinghaus E, Kamatani Y, Dilthey A, et al. Imputation of KIR Types from SNP Variation Data. Am J Hum Genet. 2015;97(4):593-607.

153. De Re V, Caggiari L, De Zorzi M, Repetto O, Zignego AL, Izzo F, et al. Genetic diversity of the KIR/HLA system and susceptibility to hepatitis C virus-related diseases. PLoS One. 2015;10(2):e0117420.
154. Stern M, Elsasser H, Honger G, Steiger J, Schaub S, Hess C. The number of activating KIR genes inversely correlates with the rate of CMV infection/reactivation in kidney transplant recipients. Am J Transplant. 2008;8(6):1312-7.
155. Martin MP, Gao X, Lee JH, Nelson GW, Detels R, Goedert JJ, et al. Epistatic interaction between KIR3DS1 and HLA-B delays the progression to AIDS. Nat Genet. 2002;31(4):429-34.
156. Pelak K, Need AC, Fellay J, Shianna KV, Feng S, Urban TJ, et al. Copy number variation of KIR genes influences HIV-1 control. PLoS Biol. 2011;9(11):e1001208.
157. Ozturk OG, Gun FD, Polat G. Killer cell immunoglobulin-like receptor genes in patients with breast cancer. Med Oncol. 2012;29(2):511-5.
158. Marin D, Gabriel IH, Ahmad S, Foroni L, de Lavallade H, Clark R, et al. KIR2DS1 genotype predicts for complete cytogenetic response and survival in newly diagnosed chronic myeloid leukemia patients treated with imatinib. Leukemia. 2012;26(2):296-302.
159. Gabriel IH, Sergeant R, Szydlo R, Apperley JF, DeLavallade H, Alsuliman A, et al. Interaction between KIR3DS1 and HLA-Bw4 predicts for progression-free survival after autologous stem cell transplantation in patients with multiple myeloma. Blood. 2010;116(12):2033-9.
160. Hou YF, Zhang YC, Jiao YL, Wang LC, Li JF, Pan ZL, et al. Disparate distribution of activating and inhibitory killer cell immunoglobulin-like receptor genes in patients with systemic lupus erythematosus. Lupus. 2010;19(1):20-6.
161. Luszczek W, Manczak M, Cislo M, Nockowski P, Wisniewski A, Jasek M, et al. Gene for the activating natural killer cell receptor, KIR2DS1, is associated with susceptibility to psoriasis vulgaris. Hum Immunol. 2004;65(7):758-66.
162. Garcia-Leon JA, Pinto-Medel MJ, Garcia-Trujillo L, Lopez-Gomez C, Oliver-Martos B, Prat-Arrojo I, et al. Killer cell immunoglobulin-like receptor genes in Spanish multiple sclerosis patients. Mol Immunol. 2011;48(15-16):1896-902.
163. Culley FJ. Natural killer cells in infection and inflammation of the lung. Immunology. 2009;128(2):151-63.
164. Takeuchi M, Nagai S, Nakajima A, Shinya M, Tsukano C, Asada H, et al. Inhibition of lung natural killer cell activity by smoking: the role of alveolar macrophages. Respiration. 2001;68(3):262-7.
165. Ferson M, Edwards A, Lind A, Milton GW, Hersey P. Low natural killer-cell activity and immunoglobulin levels associated with smoking in human subjects. Int J Cancer. 1979;23(5):603-9.
166. Zeidel A, Beilin B, Yardeni I, Mayburd E, Smirnov G, Bessler H. Immune response in asymptomatic smokers. Acta Anaesthesiol Scand. 2002;46(8):959-64.
167. Lu LM, Zavitz CC, Chen B, Kianpour S, Wan Y, Stampfli MR. Cigarette smoke impairs NK cell-dependent tumor immune surveillance. J Immunol. 2007;178(2):936-43.
168. Lin SJ, Chang LY, Yan DC, Huang YJ, Lin TJ, Lin TY. Decreased intercellular adhesion molecule-1 (CD54) and L-selectin (CD62L) expression on peripheral blood natural killer cells in asthmatic children with acute exacerbation. Allergy. 2003;58(1):67-71.
169. Jira M, Antosova E, Vondra V, Strejcek J, Mazakova H, Prazakova J. Natural killer and interleukin-2 induced cytotoxicity in asthmatics. I. Effect of acute antigen-specific challenge. Allergy. 1988;43(4):294-8.
170. Di Lorenzo G, Esposito Pellitteri M, Drago A, Di Blasi P, Candore G, Balistreri C, et al. Effects of in vitro treatment with fluticasone propionate on natural killer and lymphokine-induced killer activity in asthmatic and healthy individuals. Allergy. 2001;56(4):323-7.

171. Burke SM, Issekutz TB, Mohan K, Lee PW, Shmulevitz M, Marshall JS. Human mast cell activation with virus-associated stimuli leads to the selective chemotaxis of natural killer cells by a CXCL8-dependent mechanism. Blood. 2008;111(12):5467-76.
172. Prieto A, Reyes E, Bernstein ED, Martinez B, Monserrat J, Izquierdo JL, et al. Defective natural killer and phagocytic activities in chronic obstructive pulmonary disease are restored by glycophosphopeptical (inmunoferon). Am J Respir Crit Care Med. 2001;163(7):1578-83.
173. Sykes A, Johnston SL. Etiology of asthma exacerbations. J Allergy Clin Immunol. 2008;122(4):685-8.
174. Sethi S, Murphy TF. Infection in the pathogenesis and course of chronic obstructive pulmonary disease. N Engl J Med. 2008;359(22):2355-65.
175. Khalil N, Parekh TV, O'Connor R, Antman N, Kepron W, Yehaulaeshet T, et al. Regulation of the effects of TGF-beta 1 by activation of latent TGF-beta 1 and differential expression of TGF-beta receptors (T beta R-I and T beta R-II) in idiopathic pulmonary fibrosis. Thorax. 2001;56(12):907-15.
176. Lee JC, Lee KM, Kim DW, Heo DS. Elevated TGF-beta1 secretion and down-modulation of NKG2D underlies impaired NK cytotoxicity in cancer patients. J Immunol. 2004;172(12):7335-40.
177. Aquino-Galvez A, Perez-Rodriguez M, Camarena A, Falfan-Valencia R, Ruiz V, Montano M, et al. MICA polymorphisms and decreased expression of the MICA receptor NKG2D contribute to idiopathic pulmonary fibrosis susceptibility. Hum Genet. 2009;125(5-6):639-48.
178. Tsao CC, Tsao PN, Chen YG, Chuang YH. Repeated Activation of Lung Invariant NKT Cells Results in Chronic Obstructive Pulmonary Disease-Like Symptoms. PLoS One. 2016;11(1):e0147710.
179. Pichavant M, Remy G, Bekaert S, Le Rouzic O, Kervoaze G, Vilain E, et al. Oxidative stress-mediated iNKT-cell activation is involved in COPD pathogenesis. Mucosal Immunol. 2014;7(3):568-78.
180. Finkelstein R, Fraser RS, Ghezzo H, Cosio MG. Alveolar inflammation and its relation to emphysema in smokers. Am J Respir Crit Care Med. 1995;152(5 Pt 1):1666-72.
181. Saetta M, Di Stefano A, Turato G, Facchini FM, Corbino L, Mapp CE, et al. CD8+ T-lymphocytes in peripheral airways of smokers with chronic obstructive pulmonary disease. Am J Respir Crit Care Med. 1998;157(3 Pt 1):822-6.
182. Asimit JL, Zeggini E. Imputation of rare variants in next-generation association studies. Hum Hered. 2012;74(3-4):196-204.
183. Shrine N, Guyatt AL, Erzurumluoglu AM, Jackson VE, Hobbs BD, Melbourne CA, et al. New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. Nat Genet. 2019;51(3):481-93.
184. Irving WL, Day S, Johnston ID. Idiopathic pulmonary fibrosis and hepatitis C virus infection. Am Rev Respir Dis. 1993;148(6 Pt 1):1683-4.
185. Yin Q, Strong MJ, Zhuang Y, Flemington EK, Kaminski N, de Andrade JA, et al. Assessment of viral RNA in idiopathic pulmonary fibrosis using RNA-seq. BMC Pulm Med. 2020;20(1):81.
186. Rubicz R, Yolken R, Drigalenko E, Carless MA, Dyer TD, Bauman L, et al. A genome-wide integrative genomic study localizes genetic factors influencing antibodies against Epstein-Barr virus nuclear antigen 1 (EBNA-1). PLoS Genet. 2013;9(1):e1003147.
187. Crosslin DR, Carrell DS, Burt A, Kim DS, Underwood JG, Hanna DS, et al. Genetic variation in the HLA region is associated with susceptibility to herpes zoster. Genes Immun. 2015;16(1):1-7.
188. Wain LV, Shrine N, Artigas MS, Erzurumluoglu AM, Noyvert B, Bossini-Castillo L, et al. Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets. Nat Genet. 2017;49(3):416-25.

189. Turner SD. qqman: an R package for visualising GWAS results using Q-Q and manhattan plots. bioRxiv. 2014;DOI 10.1101/005165.
190. Locuszoom. Locuszoom 2020 [cited 2020 02/09]. Available from: http://locuszoom.org/.
191. Wakefield J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. Am J Hum Genet. 2007;81(2):208-27.
192. Staley JR, Blackshaw J, Kamat MA, Ellis S, Surendran P, Sun BB, et al. PhenoScanner: a database of human genotype-phenotype associations. Bioinformatics. 2016;32(20):3207-9.
193. Consortium GT. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45(6):580-5.
194. Midi H, Sarkar SK, Rana S. Collinearity diagnostics of binary logistic regression model. Journal of Interdisciplinary Mathematics. 2010;13(3):253-67.
195. American Thoracic S, European Respiratory S. American Thoracic Society/European Respiratory Society International Multidisciplinary Consensus Classification of the Idiopathic Interstitial Pneumonias. This joint statement of the American Thoracic Society (ATS), and the European Respiratory Society (ERS) was adopted by the ATS board of directors, June 2001 and by the ERS Executive Committee, June 2001. Am J Respir Crit Care Med. 2002;165(2):277-304.
196. Raghu G, Collard HR, Egan JJ, Martinez FJ, Behr J, Brown KK, et al. An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. Am J Respir Crit Care Med. 2011;183(6):788-824.
197. Raghu G, Rochwerg B, Zhang Y, Garcia CA, Azuma A, Behr J, et al. An Official ATS/ERS/JRS/ALAT Clinical Practice Guideline: Treatment of Idiopathic Pulmonary Fibrosis. An Update of the 2011 Clinical Practice Guideline. Am J Respir Crit Care Med. 2015;192(2):e3-19.
198. Furukawa H, Oka S, Shimada K, Sugii S, Ohashi J, Matsui T, et al. Association of human leukocyte antigen with interstitial lung disease in rheumatoid arthritis: a protective role for shared epitope. PLoS One. 2012;7(5):e33133.
199. Oka S, Furukawa H, Shimada K, Sugii S, Hashimoto A, Komiya A, et al. Association of human leukocyte antigen alleles with chronic lung diseases in rheumatoid arthritis. Rheumatology (Oxford). 2016;55(7):1301-7.
200. Peljto AL, Steele MP, Fingerlin TE, Hinchcliff ME, Murphy E, Podlusky S, et al. The pulmonary fibrosis-associated MUC5B promoter polymorphism does not influence the development of interstitial pneumonia in systemic sclerosis. Chest. 2012;142(6):1584-8.
201. Borie R, Crestani B, Dieude P, Nunes H, Allanore Y, Kannengiesser C, et al. The MUC5B variant is associated with idiopathic pulmonary fibrosis but not with systemic sclerosis interstitial lung disease in the European Caucasian population. PLoS One. 2013;8(8):e70621.
202. Peljto AL, Zhang Y, Fingerlin TE, Ma SF, Garcia JG, Richards TJ, et al. Association between the MUC5B promoter polymorphism and survival in patients with idiopathic pulmonary fibrosis. Jama. 2013;309(21):2232-9.
203. Leavy OC, Shwu-Fan M, Molyneaux PL, Maher TM, Oldham JM, Flores C, et al. Proportion of idiopathic pulmonary fibrosis risk explained by known genetic loci. medRxiv. 2020;preprint.
204. Raghu G, Collard HR, Egan JJ, Martinez FJ, Behr J, Brown KK, et al. An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. Am J Respir Crit Care Med. 2011;183(6):788-824.
205. Kerkhof M, Sonnappa S, Postma DS, Brusselle G, Agusti A, Anzueto A, et al. Blood eosinophil count and exacerbation risk in patients with COPD. Eur Respir J. 2017;50(1).
206. Consortium HR. Michigen Imputation Server 2020 [cited 14th January 2020]. Available from: https://imputationserver.sph.umich.edu/index.html#!
207. NCBI. dbSNP 2020 [cited 2020 8th October]. Available from: https://www.ncbi.nlm.nih.gov/snp/.

208. Cisneros E, Moraru M, Gomez-Lozano N, Muntasell A, Lopez-Botet M, Vilches C. Haplotype-Based Analysis of KIR-Gene Profiles in a South European Population-Distribution of Standard and Variant Haplotypes, and Identification of Novel Recombinant Structures. Front Immunol. 2020;11:440.
209. 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 Genet. 2019;15(12):e1008489.
210. Brown MA, Pile KD, Kennedy LG, Calin A, Darke C, Bell J, et al. HLA class I associations of ankylosing spondylitis in the white population in the United Kingdom. Ann Rheum Dis. 1996;55(4):268-70.
211. Cordier JF. Cryptogenic organising pneumonia. Eur Respir J. 2006;28(2):422-46.
212. Sieminska A, Kuziemski K. Respiratory bronchiolitis-interstitial lung disease. Orphanet J Rare Dis. 2014;9:106.
213. Hellemons ME, Moor CC, von der Thusen J, Rossius M, Odink A, Thorgersen LH, et al. Desquamative interstitial pneumonia: a systematic review of its features and outcomes. Eur Respir Rev. 2020;29(156).
214. Cruz T, Jia M, Sembrat J, Tabib T, Agostino N, Bruno TC, et al. Reduce Proportion and Activity of NK Cells in the Lung of Idiopathic Pulmonary Fibrosis Patients. Am J Respir Crit Care Med. 2021.
215. Tian C, Hromatka BS, Kiefer AK, Eriksson N, Noble SM, Tung JY, et al. Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. Nat Commun. 2017;8(1):599.
216. Squire DM, Motyer A, Ahn R, Nititham J, Huang Z-M, Oksenberg JR, et al. MHC*IMP – Imputation of Alleles for Genes in the Major

Histocompatibility Complex BioRxiv. 2020.
217. Braud VM, Allan DS, O'Callaghan CA, Soderstrom K, D'Andrea A, Ogg GS, et al. HLA-E binds to natural killer cell receptors CD94/NKG2A, B and C. Nature. 1998;391(6669):795-9.
218. Zeng X, Chen H, Gupta R, Paz-Altschul O, Bowcock AM, Liao W. Deletion of the activating NKG2C receptor and a functional polymorphism in its ligand HLA-E in psoriasis susceptibility. Exp Dermatol. 2013;22(10):679-81.
219. Rizzo R, Melchiorri L, Simone L, Stignani M, Marzola A, Gullini S, et al. Different production of soluble HLA-G antigens by peripheral blood mononuclear cells in ulcerative colitis and Crohn's disease: a noninvasive diagnostic tool? Inflamm Bowel Dis. 2008;14(1):100-5.
220. Ferreira MA, Vonk JM, Baurecht H, Marenholz I, Tian C, Hoffman JD, et al. Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. Nat Genet. 2017;49(12):1752-7.
221. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell. 2016;167(5):1415-29 e19.
222. Bronson PG, Chang D, Bhangale T, Seldin MF, Ortmann W, Ferreira RC, et al. Common variants at PVT1, ATG13-AMBRA1, AHI1 and CLEC16A are associated with selective IgA deficiency. Nat Genet. 2016;48(11):1425-9.
223. Ji SG, Juran BD, Mucha S, Folseraas T, Jostins L, Melum E, et al. Genome-wide association study of primary sclerosing cholangitis identifies new risk loci and quantifies the genetic relationship with inflammatory bowel disease. Nat Genet. 2017;49(2):269-73.
224. Shen X, Klaric L, Sharapov S, Mangino M, Ning Z, Wu D, et al. Multivariate discovery and replication of five novel loci associated with Immunoglobulin G N-glycosylation. Nat Commun. 2017;8(1):447.

Supplementary Data

Chapter one supplementary data:

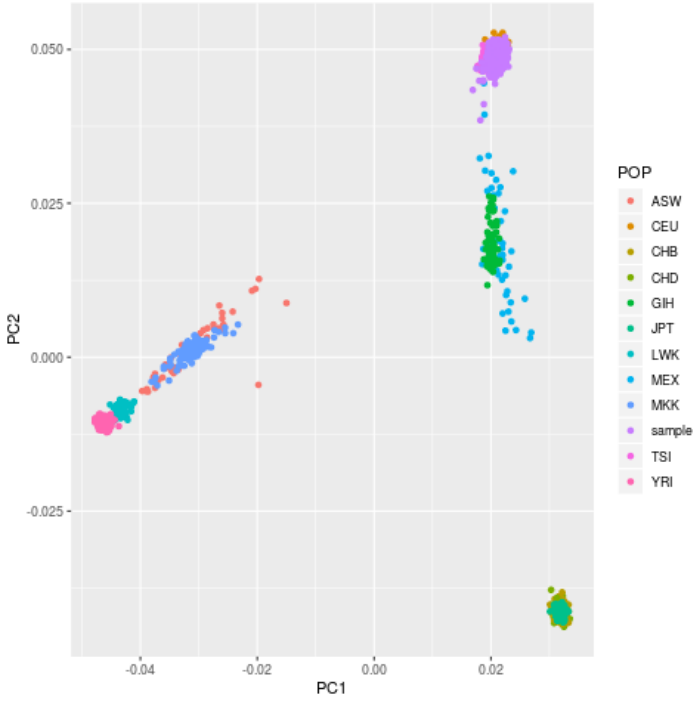
Supplementary table 1.1: Relationship between KIR types. KIR Haplotype was the haplotype classification defined by (148), AvsB corresponds to the broad A/B haplotype classification, all the other gene columns show the copy number of each individuals KIR gene. Each KIR Haplotype was defined by the copy number values across each of the 17 KIR genes, table from (152).

KIR Haplotype	A/B	KIR3DL3	KIR2DS2	KIR2DL2	KIR2DL3	KIR2DP1	KIR2DL1	KIR3DP1	KIR2DL4	KIR3DL1ex4	KIR3DL1ex9	KIR3DS1	KIR2DL5	KIR2DS3	KIR2DS5	KIR2DS1	KIR2DS4TOTAL	KIR2DS4WT	KIR2DS4DEL	KIR3DL2
1	A	1	0	0	1	1	1	1	1	1	1	0	0	0	0	0	1	0	1	1
2	A	1	0	0	1	1	1	1	1	1	1	0	0	0	0	0	1	1	0	1
3	B	1	0	0	1	1	1	1	1	0	0	1	1	0	1	1	0	0	0	1
4	B	1	1	1	0	0	0	1	1	1	1	0	0	0	0	0	1	0	1	1
5	B	1	1	1	0	0	0	1	1	1	1	0	0	0	0	0	1	1	0	1
6	B	1	1	1	0	1	1	1	1	1	1	0	1	1	0	0	1	0	1	1
7	B	1	1	1	0	1	1	1	1	0	0	1	2	1	1	1	0	0	0	1
8	B	1	1	1	0	0	0	1	1	0	0	1	1	0	1	1	0	0	0	1
9	B	1	1	1	0	1	1	1	1	0	0	1	2	2	0	1	0	0	0	1
10	B	1	1	1	0	1	1	1	1	1	1	0	1	1	0	0	1	1	0	1
11	B	1	0	0	1	1	1	1	1	0	0	1	1	1	0	1	0	0	0	1
12	B	1	1	1	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	1
13	B	1	1	1	0	1	1	2	2	1	1	1	1	1	0	0	1	1	0	1
14	B	1	0	0	1	2	2	2	2	0	0	2	2	1	1	1	0	0	0	1
15	B	1	1	0	0	1	1	1	1	0	0	1	1	0	1	1	0	0	0	1
16	B	1	0	0	1	1	1	2	2	1	1	1	0	0	0	0	1	0	1	1
17	B	1	1	1	0	0	0	1	1	0	0	1	1	1	0	1	0	0	0	1

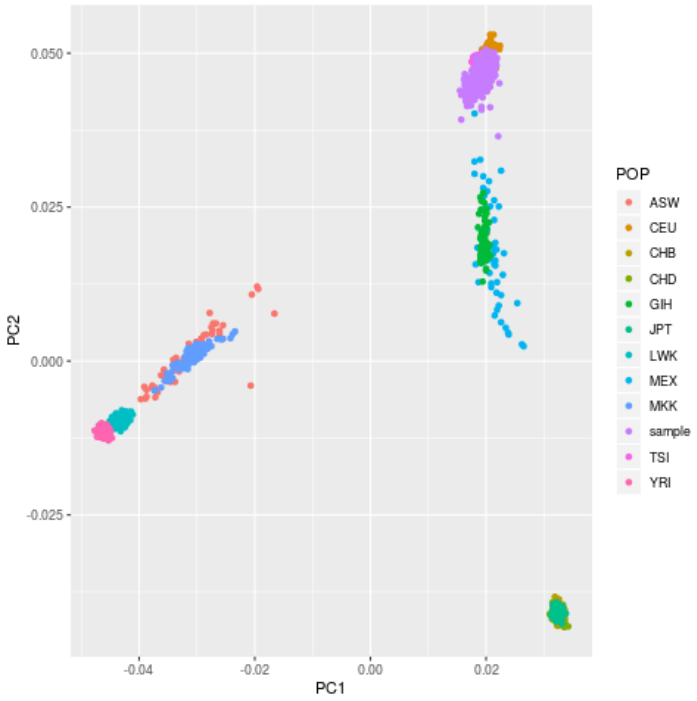
18	B	1	1	1	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	1
19	B	1	1	2	0	0	0	2	2	0	0	2	2	0	2	1	0	0	0	1
20	B	1	0	0	1	1	1	2	2	1	1	1	0	0	0	0	1	1	0	1
21	B	1	1	0	0	1	1	1	1	1	1	0	0	0	0	0	1	0	1	1
22	B	1	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
23	B	1	0	0	1	2	2	2	2	1	1	1	1	1	0	0	1	0	1	1
24	B	1	1	1	0	1	1	1	1	0	0	1	1	1	1	1	0	0	0	1
25	B	1	1	1	0	1	1	1	1	1	1	0	1	0	0	0	1	0	1	1
27	A	1	0	0	1	1	0	1	1	1	1	0	0	0	0	0	1	0	1	1
28	B	1	0	0	1	1	1	2	2	0	0	2	1	0	1	1	0	0	0	1
29	B	1	1	1	0	1	1	2	2	1	1	1	1	1	0	0	1	0	1	1
30	B	1	1	1	0	0	0	2	2	1	1	1	0	0	0	0	1	0	1	1
31	B	1	0	1	0	1	1	1	1	1	1	0	1	0	1	0	1	1	0	1
33	B	1	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1
34	A	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
36	B	1	1	1	0	2	2	2	2	0	0	2	1	2	0	1	0	0	0	1
38	B	1	0	0	1	2	2	2	2	0	0	2	1	0	1	1	0	0	0	1
40	B	1	1	1	0	2	2	1	1	0	0	1	2	1	1	0	0	0	0	1
41	B	1	1	1	0	1	2	1	1	0	0	1	2	2	0	1	0	0	0	1
42	B	1	1	2	0	0	0	2	2	1	1	1	1	0	1	0	1	0	1	1
44	B	1	1	1	0	1	1	1	1	0	0	1	2	0	1	1	0	0	0	1
42	B	1	1	1	0	1	1	1	1	0	0	1	2	1	1	0	0	0	0	1
46	B	1	0	1	0	1	1	1	1	0	0	1	1	1	1	1	0	0	0	1
48	A	1	0	0	1	1	1	1	1	1	2	0	0	0	0	0	2	2	0	1
50	B	1	0	0	0	1	1	1	1	1	1	0	1	1	0	0	1	1	0	1
52	B	1	0	0	0	1	1	1	1	1	1	0	1	0	1	0	1	0	1	1
53	B	1	1	0	1	1	1	1	1	1	1	0	0	0	0	0	1	0	1	1
55	A	1	0	0	0	1	2	1	1	1	1	0	0	0	0	0	1	1	0	1
56	B	1	1	1	0	0	0	0	1	0	0	1	1	0	1	1	0	0	0	1
57	A	1	0	0	1	1	1	0	1	1	1	0	0	0	0	0	1	0	1	1
58	A	1	0	0	1	1	1	1	0	1	1	0	0	0	0	0	1	1	0	1
59	A	1	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1
68	B	1	1	1	0	0	0	0	1	1	1	0	0	0	0	0	1	0	1	1
69	B	1	1	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1

Chapter two supplementary data:

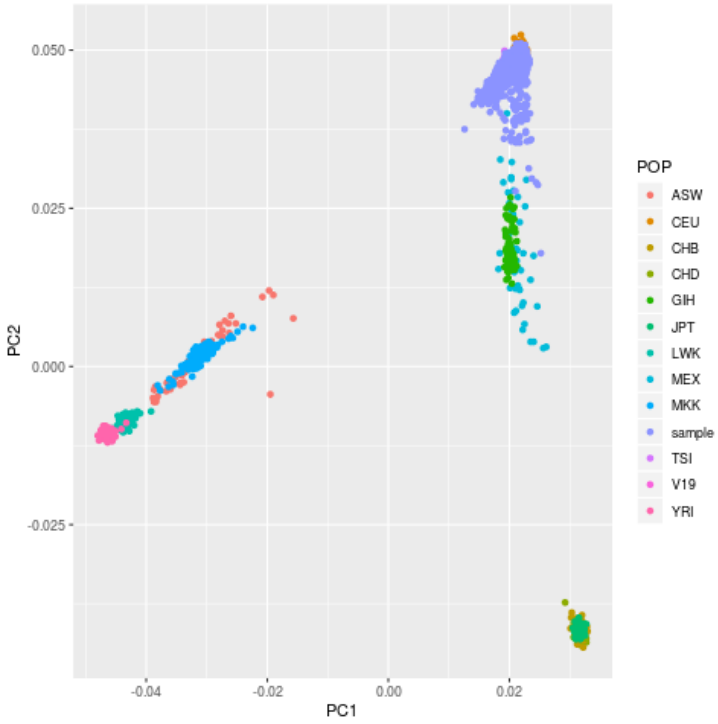
Principal Component Analysis:



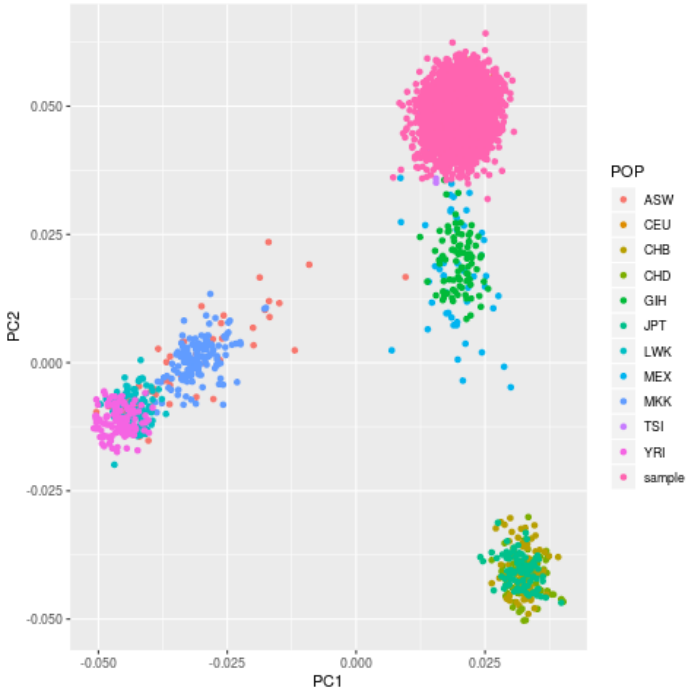
Supplementary figure 2.1: Plot of principal component one and principal component two in the UK IPF dataset. Cases and controls from the UK dataset (named 'sample' [pink]) overlay European ancestry samples from HapMap.



Supplementary figure 2.2: Plot of principal component one and principal component two in the Colorado IIP dataset. Cases and controls from the Colorado dataset (named 'sample' [pink]) overlay European ancestry samples from HapMap.



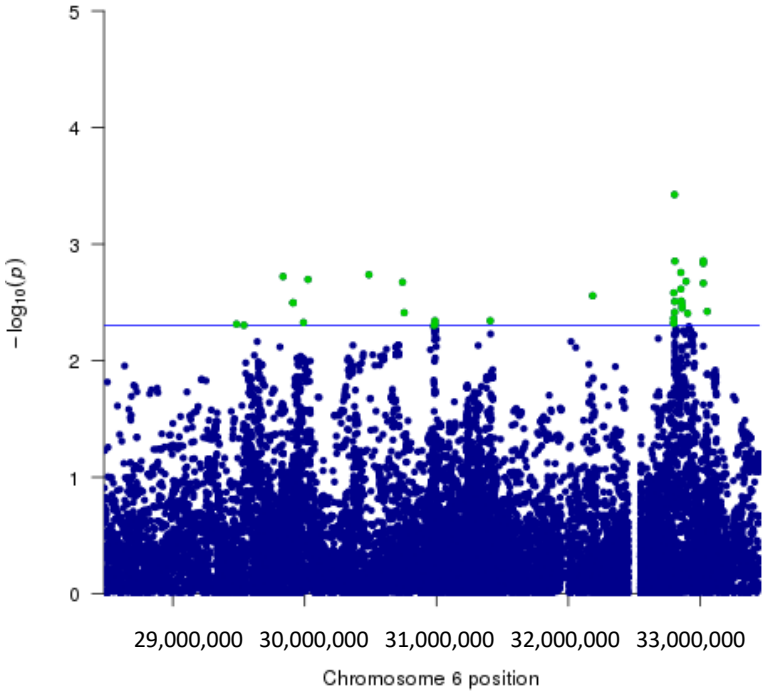
Supplementary figure 2.3: Plot of principal component one and principal component two in the Chicago IPF dataset. Cases and controls from the Chicago dataset (named 'sample' [purple]) overlay European ancestry samples from HapMap.



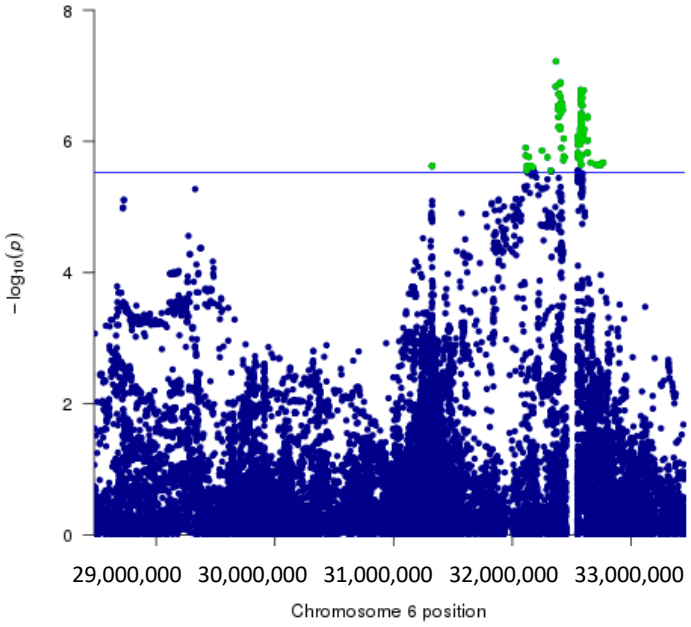
Supplementary figure 2.4: Plot of principal component one and principal component two in the UUS IPF dataset. Cases and controls from the UUS dataset (named 'sample' [pink]) overlay European ancestry samples from HapMap.

Chapter three supplementary data:

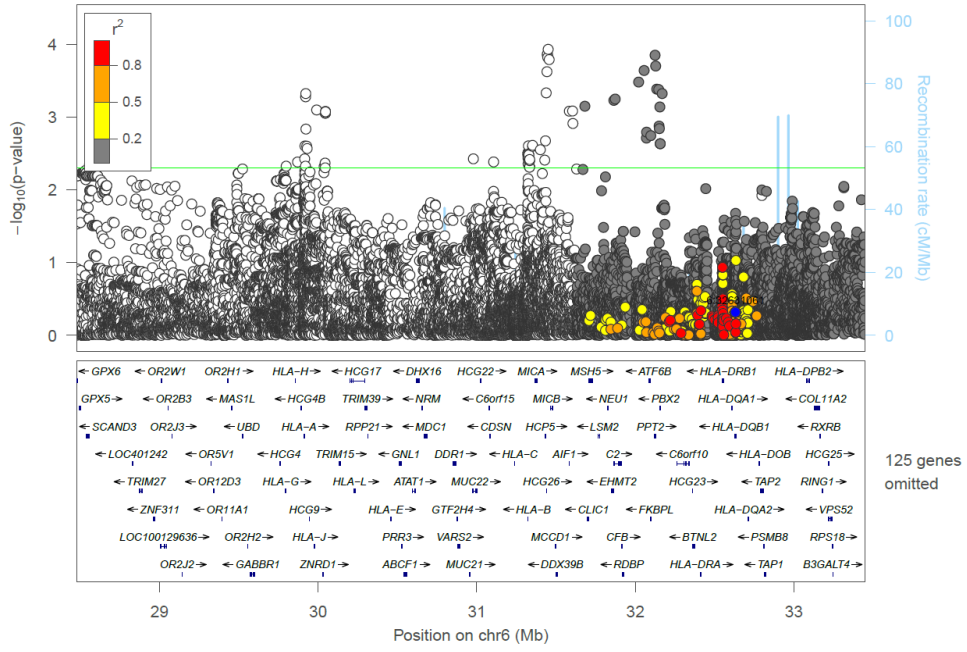
Meta-analyses of the HLA region of IPF susceptibility in UK, Colorado and Chicago datasets:



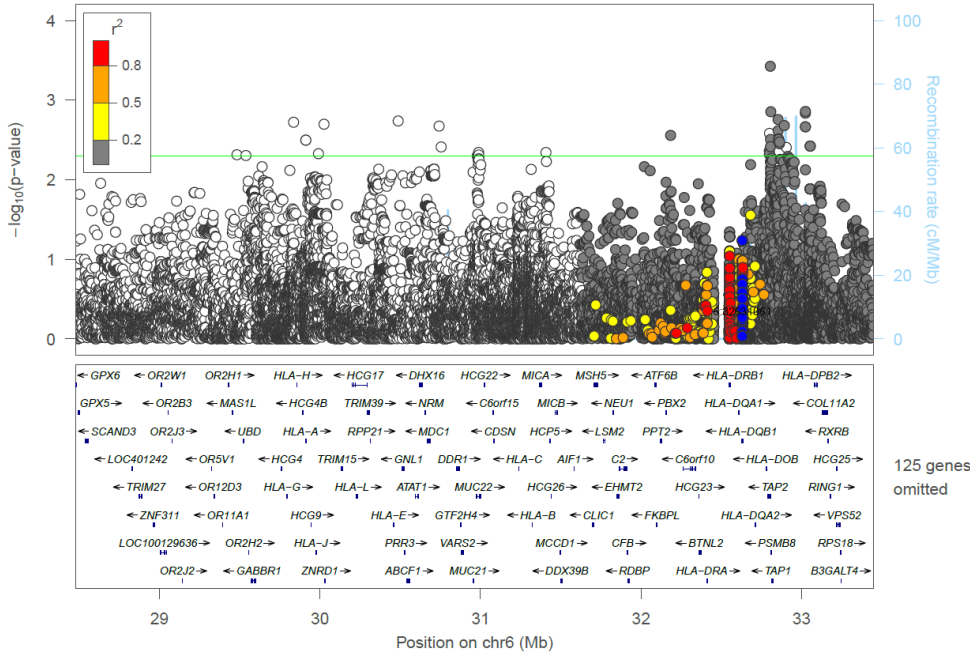
Supplementary figure 3.1: Manhattan plot of the analysis of the HLA region for IPF susceptibility in the Chicago IPF cohort (the green variants are all the variants that passed the suggestive significance threshold). Blue line is suggestive significance threshold of $P<5x10^{-3}$*).*



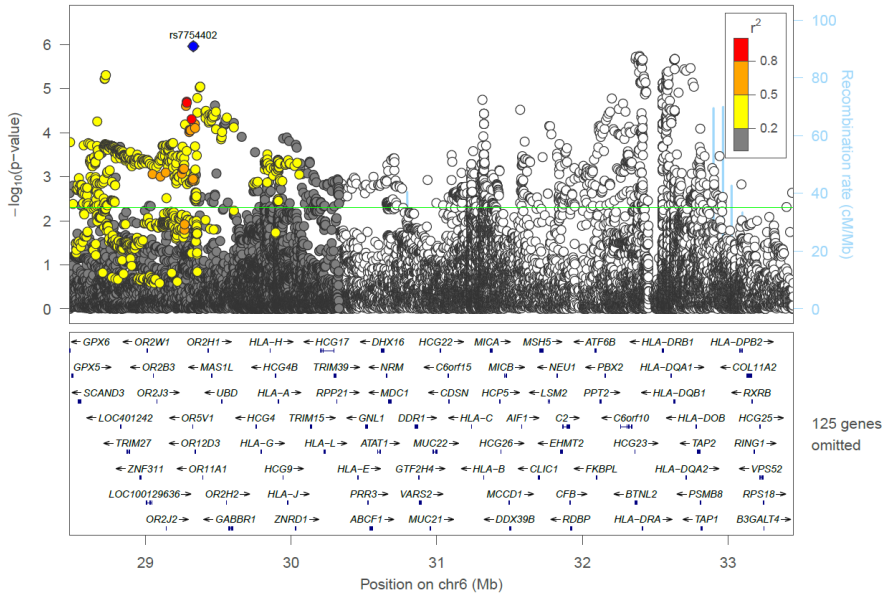
Supplementary figure 3.2: Manhattan plot of the analysis of the HLA region for IPF susceptibility in the Colorado IPF cohort (the green variants are all the variants that passed the suggestive significance threshold). Blue line is Bonferroni significance threshold of P<2.8x10-^{6}).



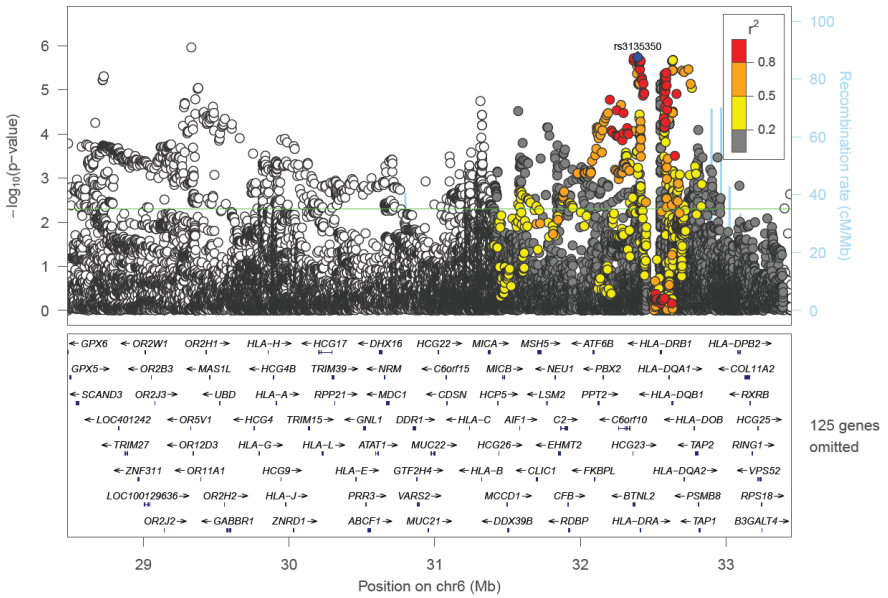
*Supplementary figure 3.3; Region plot of HLA-DQB1*06:02 over the whole HLA region (28477797-33448354 bp) for the replication of this signal in the UK IPF dataset. Green line is suggestive significance threshold of P<5x10-^{3}).*



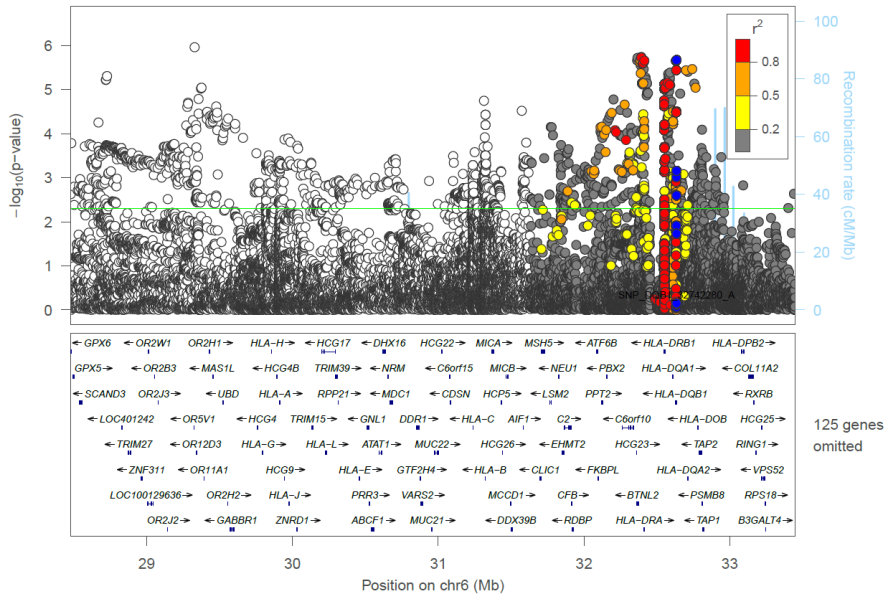
*Supplementary figure 3.4; Region plot of HLA-DQB1*06:02 over the whole HLA region (28477797-33448354 bp) for the replication of this signal in the Chicago IPF dataset. Green line is suggestive significance threshold of P<5x10-^{3}).*



Supplementary figure 3.5: Region plot of rs7754402 over the whole HLA region (28477797-33448354 bp) in a meta-analysis for IPF susceptibility using the UK, Colorado and Chicago IPF datasets. Green line is suggestive significance threshold of P<5x10-[3]).

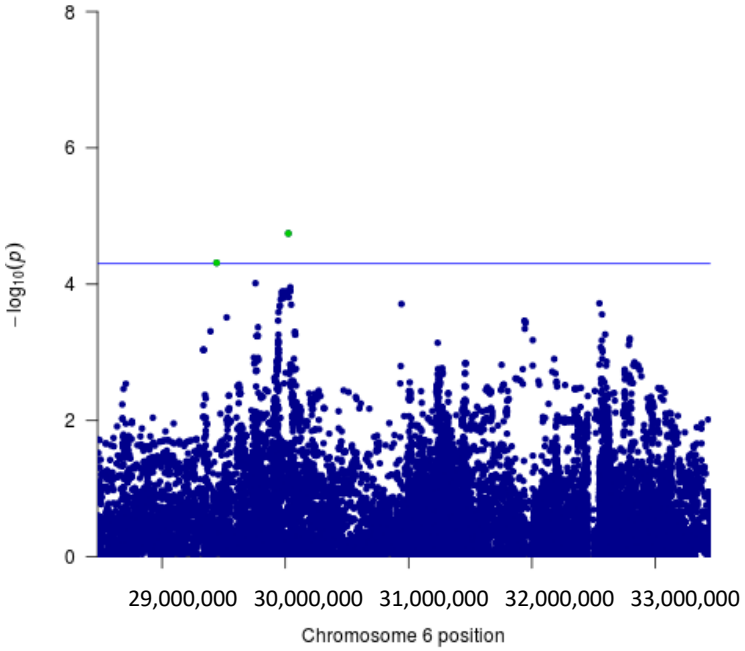


Supplementary figure 3.6: Region plot of rs3135350 over the whole HLA region (28477797-33448354 bp) in a meta-analysis for IPF susceptibility using the UK, Colorado and Chicago IPF datasets. Green line is suggestive significance threshold of P<5x10-[3]).

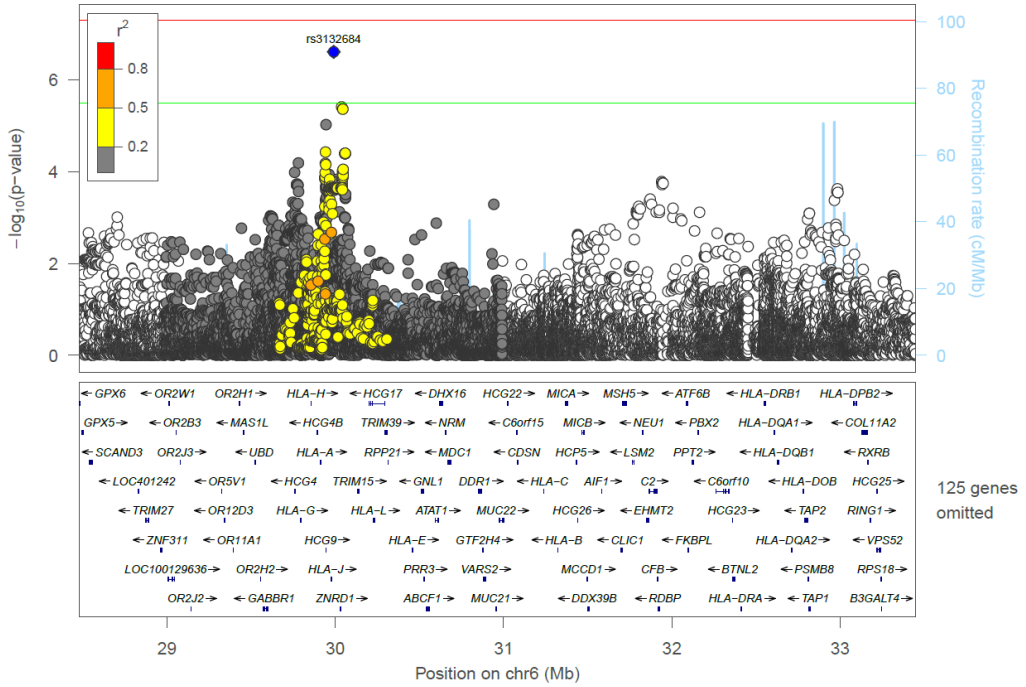


*Supplementary figure 3.7: Region plot of HLA-DQB1*06:02 over the whole HLA region (28477797-33448354 bp) in a meta-analysis for IPF susceptibility using the UK, Colorado and Chicago IPF datasets. Green line is suggestive significance threshold of P<5x10-^{3}).*

HLA-wide association meta-analysis of IPF susceptibility in the UK, Chicago and UUS datasets:



Supplementary figure 3.8: Manhattan plot of the analysis of the HLA region for IPF susceptibility in the UUS IPF cohort (the green variants are all the variants that passed the suggestive significance threshold). Blue line is suggestive significance threshold of P<5x10-^{5}).



Supplementary figure 3.9: Region plot of rs3132684 over the whole HLA region (28477797-33448354 bp) in a meta-analysis for IPF susceptibility using the UK, UUS and Chicago IPF datasets. Green line is Bonferroni significance threshold of P<2.8x10^{-6} and red line is genome-wide significance of P<5x10^{-8}).

Supplementary table 3.1: Table of Phenoscanner results for the lead SNP and SNPs in the credible set. Showing results for the most significant SNP for each trait.

SNP	Position	Alleles	Trait	Reference	P-value
rs2735076	29943490	G/A	Allergic disease	(220)	2.35x10^{-9}
rs4713279	29944896	A/T	Eosinophil count	(221)	4.91x10^{-42}

rs4713279	29944896	A/T	Eosinophil percentage of granulocytes	(221)	$2.11x10^{-25}$
rs4713279	29944896	A/T	Eosinophil percentage of white cells	(221)	$1.45x10^{-29}$
rs2517917	29781020	T/C	Granulocyte count	(221)	$4.62x10^{-14}$
rs2735076	29943490	G/A	Hayfever, allergic rhinitis or eczema	UKBB	$5.56x10^{-15}$
rs3823377	29944253	A/C	IgA deficiency	(222)	$3.49x10^{-12}$
rs28698309	29758393	A/C	Lymphocyte count	(221)	$2.06x10^{-20}$
rs28698309	29758393	A/C	Monocyte count	(221)	$1.31x10^{-15}$
rs2517917	29781020	T/C	Myeloid white cell count	(221)	$4.62x10^{-14}$
rs2517917	29781020	T/C	Neutrophil count	(221)	$4.27x10^{-15}$
rs4713279	29944896	A/T	Neutrophil percentage of granulocytes	(221)	$1.83x10^{-21}$
rs3823377	29944253	A/C	Peak expiratory flow	UKBB	$3.34x10^{-11}$
rs2517917	29781020	T/C	Primary sclerosing cholangitis	(223)	$1.80x10^{-74}$
rs4713279	29944896	A/T	Rheumatoid arthritis	(119)	$5.90x10^{-35}$
rs3823377	29944253	A/C	Self-reported ankylosing spondylitis	UKBB	$1.20x10^{-8}$
rs9295829	30028800	A/G	Self-reported malabsorption or coeliac disease	UKBB	$2.76x10^{-28}$
rs9295829	30028800	A/G	Self-reported multiple sclerosis	UKBB	$2.63x10^{-14}$
rs380924	29939885	G/A	Self-reported psoriasis	UKBB	$9.18x10^{-23}$
rs2523957	29940260	G/A	Self-reported psoriatic arthropathy	UKBB	$2.89x10^{-8}$
rs4713279	29944896	A/T	Sum eosinophil basophil counts	(221)	$3.41x10^{-38}$
rs2517917	29781020	T/C	Sum neutrophil eosinophil counts	(221)	$3.18x10^{-14}$
rs3823375	29944158	C/T	Ulcerative colitis	(109)	$2.26x10^{-14}$
rs9261306	30045731	C/T	White blood cell count	(221)	$8.18x10^{-17}$

Supplementary table 3.2: Results from a GTEx search for the rs3132684 SNP that pass Bonferroni corrected threshold of 0.001 (corrected for number of tissues).

Lead SNP	**Gene Symbol**	**P-Value**	**Tissue**
rs3132684	*HCP5B*	$6.60X10^{-52}$	Adipose - Subcutaneous
	HCG4P3	$1.50X10^{-27}$	Adipose - Subcutaneous
	HLA-H	$7.70X10^{-10}$	Adipose - Subcutaneous
	HCG4P7	$3.50X10^{-6}$	Adipose - Subcutaneous
	HLA-J	$1.20X10^{-5}$	Adipose - Subcutaneous
	ZFP57	$1.40X10^{-5}$	Adipose - Subcutaneous
	MICE	$2.00X10^{-5}$	Adipose - Subcutaneous
	RPL23AP1	$1.80X10^{-4}$	Adipose - Subcutaneous
	HCP5B	$7.90X10^{-30}$	Adipose - Visceral (Omentum)
	HCG4P3	$2.10X10^{-22}$	Adipose - Visceral (Omentum)
	HLA-J	$7.70X10^{-13}$	Adipose - Visceral (Omentum)
	HCG4P7	$2.30X10^{-7}$	Adipose - Visceral (Omentum)
	HLA-H	$3.40X10^{-7}$	Adipose - Visceral (Omentum)
	ZFP57	$2.60X10^{-6}$	Adipose - Visceral (Omentum)
	RPL23AP1	$1.60X10^{-4}$	Adipose - Visceral (Omentum)
	HCP5B	$1.80X10^{-27}$	Adrenal Gland
	HCG4P3	$1.20X10^{-14}$	Adrenal Gland
	HLA-K	$1.20X10^{-6}$	Adrenal Gland
	HCG4P7	$2.00X10^{-5}$	Adrenal Gland
	ZNRD1ASP	$1.20X10^{-4}$	Adrenal Gland
	HCP5B	$3.40X10^{-24}$	Artery - Aorta
	HCG4P3	$3.80X10^{-6}$	Artery - Aorta
	HLA-K	$1.10X10^{-9}$	Artery - Aorta
	HLA-J	$2.10X10^{-8}$	Artery - Aorta
	ZNRD1ASP	$4.60X10^{-6}$	Artery - Aorta
	ZFP57	$6.30X10^{-5}$	Artery - Aorta
	RNF39	$1.10X10^{-4}$	Artery - Aorta
	HCG4	$1.30X10^{-4}$	Artery - Aorta

	HCG4B	$3.10X10^{-4}$	Artery - Aorta
	HCP5B	$7.80X10^{-12}$	Artery - Coronary
	HCG4P3	$8.90X10^{-9}$	Artery - Coronary
	HCP5B	$6.20X10^{-39}$	Artery - Tibial
	HCG4P3	$2.20X10^{-24}$	Artery - Tibial
	HLA-J	$1.30X10^{-11}$	Artery - Tibial
	ZNRD1ASP	$2.00X10^{-8}$	Artery - Tibial
	ZFP57	$1.60X10^{-7}$	Artery - Tibial
	RNF39	$1.60X10^{-7}$	Artery - Tibial
	HLA-H	$3.60X10^{-5}$	Artery - Tibial
	HCP5B	$3.10X10^{-9}$	Brain - Amygdala
	HLA-K	$6.30X10^{-8}$	Brain - Amygdala
	HCG4P3	$1.20X10^{-5}$	Brain - Amygdala
	HCP5B	$3.70X10^{-16}$	Brain - Anterior cingulate cortex (BA24)
	HCG4P3	$2.20X10^{-5}$	Brain - Anterior cingulate cortex (BA24)
	HLA-K	$4.00X10^{-5}$	Brain - Anterior cingulate cortex (BA24)
	HCP5B	$2.50X10^{-20}$	Brain - Caudate (basal ganglia)
	HCG4P3	$1.20X10^{-11}$	Brain - Caudate (basal ganglia)
	HLA-K	$5.60X10^{-11}$	Brain - Caudate (basal ganglia)
	RNF39	$4.10X10^{-8}$	Brain - Caudate (basal ganglia)
	HLA-H	$4.30X10^{-6}$	Brain - Caudate (basal ganglia)
	HCP5B	$1.10X10^{-18}$	Brain - Cerebellar Hemisphere
	HLA-H	$2.80X10^{-12}$	Brain - Cerebellar Hemisphere
	HCG4P3	$5.60X10^{-11}$	Brain - Cerebellar Hemisphere
	HCG4	$6.10X10^{-7}$	Brain - Cerebellar Hemisphere
	RNF39	$8.40X10^{-7}$	Brain - Cerebellar Hemisphere
	HLA-F	$9.70X10^{-7}$	Brain - Cerebellar Hemisphere
	HLA-V	$1.60X10^{-6}$	Brain - Cerebellar Hemisphere
	HLA-W	$3.30X10^{-5}$	Brain - Cerebellar Hemisphere
	ZNRD1ASP	$5.60X10^{-5}$	Brain - Cerebellar Hemisphere
	HCP5B	5.00X10-18	Brain - Cerebellum
	HCG4P3	$2.10X10^{-15}$	Brain - Cerebellum
	HLA-H	$1.60X10^{-12}$	Brain - Cerebellum

	HLA-F	$2.30X10^{-8}$	Brain - Cerebellum
	HCG4	$5.20X10^{-8}$	Brain - Cerebellum
	RNF39	$5.30X10^{-8}$	Brain - Cerebellum
	HLA-W	$3.10X10^{-7}$	Brain - Cerebellum
	HLA-V	$3.80X10^{-7}$	Brain - Cerebellum
	ZNRD1ASP	$2.90X10^{-6}$	Brain - Cerebellum
	HCG4P7	$4.80X10^{-5}$	Brain - Cerebellum
	HCP5B	$3.00X10^{-19}$	Brain - Cortex
	HLA-K	$3.20X10^{-13}$	Brain - Cortex
	HCG4P3	$2.50X10^{-10}$	Brain - Cortex
	RNF39	$5.10X10^{-6}$	Brain - Cortex
	HLA-H	$2.60X10^{-5}$	Brain - Cortex
	HLA-J	$8.80X10^{-5}$	Brain - Cortex
	HCP5B	$1.90X10^{-11}$	Brain - Frontal Cortex (BA9)
	HCG4P3	$4.40X10^{-9}$	Brain - Frontal Cortex (BA9)
	HLA-K	$6.10X10^{-8}$	Brain - Frontal Cortex (BA9)
	HCP5B	$2.40X10^{-11}$	Brain - Hippocampus
	HCG4P3	$2.30X10^{-6}$	Brain - Hippocampus
	HLA-H	$9.10X10^{-6}$	Brain - Hippocampus
	HLA-K	$7.50X10^{-5}$	Brain - Hippocampus
	HCP5B	$3.10X10^{-23}$	Brain - Hypothalamus
	HLA-K	$3.80X10^{-10}$	Brain - Hypothalamus
	HCG4P3	$1.00X10^{-9}$	Brain - Hypothalamus
	HCG4	$6.30X10^{-6}$	Brain - Hypothalamus
	HLA-H	$6.20X10^{-5}$	Brain - Hypothalamus
	RNF39	$7.60X10^{-5}$	Brain - Hypothalamus
	HCP5B	$2.90X10^{-20}$	Brain - Nucleus accumbens (basal ganglia)
	HLA-K	$2.70X10^{-11}$	Brain - Nucleus accumbens (basal ganglia)
	HCG4P3	$4.60X10^{-10}$	Brain - Nucleus accumbens (basal ganglia)

	RNF39	$1.70X10^{-6}$	Brain - Nucleus accumbens (basal ganglia)
	HCP5B	$3.30X10^{-17}$	Brain - Putamen (basal ganglia)
	RNF39	$2.30X10^{-10}$	Brain - Putamen (basal ganglia)
	HLA-K	$4.10X10^{-10}$	Brain - Putamen (basal ganglia)
	HCG4P3	$1.20X10^{-7}$	Brain - Putamen (basal ganglia)
	HCP5B	$5.20X10^{-8}$	Brain - Spinal cord (cervical c-1)
	HLA-K	$6.50X10^{-8}$	Brain - Spinal cord (cervical c-1)
	HCG4P3	$5.80X10^{-5}$	Brain - Spinal cord (cervical c-1)
	RNF39	$9.60X10^{-6}$	Brain - Substantia nigra
	HLA-K	$1.60X10^{-5}$	Brain - Substantia nigra
	HCP5B	$4.20X10^{-5}$	Brain - Substantia nigra
	HCP5B	$2.30X10^{-30}$	Breast - Mammary Tissue
	HCG4P3	$2.10X10^{-12}$	Breast - Mammary Tissue
	HLA-H	$3.40X10^{-6}$	Breast - Mammary Tissue
	HLA-K	$4.20X10^{-6}$	Breast - Mammary Tissue
	ZFP57	$5.50X10^{-5}$	Breast - Mammary Tissue
	HCG4P5	$1.60X10^{-4}$	Breast - Mammary Tissue
	MICE	$2.30X10^{-4}$	Breast - Mammary Tissue
	HCG4P3	$1.60X10^{-14}$	Cells - Cultured fibroblasts
	HLA-J	$5.70X10^{-10}$	Cells - Cultured fibroblasts
	HCG4	$1.90X10^{-8}$	Cells - Cultured fibroblasts
	HLA-H	$3.50X10^{-8}$	Cells - Cultured fibroblasts
	HLA-A	$1.50X10^{-5}$	Cells - Cultured fibroblasts
	ZFP57	$1.50X10^{-5}$	Cells - Cultured fibroblasts
	HCG18	$9.80X10^{-5}$	Cells - Cultured fibroblasts
	HLA-F	$1.80X10^{-4}$	Cells - Cultured fibroblasts
	HLA-U	$2.40X10^{-4}$	Cells - Cultured fibroblasts
	HCG4P7	$4.30X10^{-4}$	Cells - Cultured fibroblasts
	HCP5B	$1.70X10^{-9}$	Cells - EBV-transformed lymphocytes
	HLA-F	$6.90X10^{-5}$	Cells - EBV-transformed lymphocytes
	HCP5B	$1.10X10^{-20}$	Colon - Sigmoid
	HCG4P3	$6.30X10^{-20}$	Colon - Sigmoid

	HLA-K	$1.10X10^{-11}$	Colon - Sigmoid
	ZFP57	$2.80X10^{-5}$	Colon - Sigmoid
	HLA-J	$8.10X10^{-5}$	Colon - Sigmoid
	HCP5B	$3.80X10^{-31}$	Colon - Transverse
	HCG4P3	$4.40X10^{-22}$	Colon - Transverse
	HLA-K	$2.60X10^{-8}$	Colon - Transverse
	HCG4P5	$1.20X10^{-6}$	Colon - Transverse
	ZFP57	$2.50X10^{-6}$	Colon - Transverse
	HLA-H	$8.20X10^{-5}$	Colon - Transverse
	HLA-J	$1.10X10^{-4}$	Colon - Transverse
	HCP5B	$5.00X10^{-23}$	Esophagus - Gastroesophageal Junction
	HCG4P3	$7.60X10^{-15}$	Esophagus - Gastroesophageal Junction
	HLA-J	$4.10X10^{-9}$	Esophagus - Gastroesophageal Junction
	HLA-K	$2.30X10^{-7}$	Esophagus - Gastroesophageal Junction
	HLA-A	$6.10X10^{-6}$	Esophagus - Gastroesophageal Junction
	HCP5B	$1.80X10^{-26}$	Esophagus - Mucosa
	HLA-K	$4.00X10^{-22}$	Esophagus - Mucosa
	HCG4P3	$9.90X10^{-18}$	Esophagus - Mucosa
	HLA-J	$2.10X10^{-9}$	Esophagus - Mucosa
	HLA-H	$7.00X10^{-8}$	Esophagus - Mucosa
	ZFP57	$1.30X10^{-6}$	Esophagus - Mucosa
	HCG4B	$2.90X10^{-6}$	Esophagus - Mucosa
	SFTA2	$2.80X10^{-5}$	Esophagus - Mucosa
	ZNRD1	$9.00X10^{-5}$	Esophagus - Mucosa
	HCG4	$1.70X10^{-4}$	Esophagus - Mucosa
	HCP5B	$7.50X10^{-44}$	Esophagus - Muscularis
	HCG4P3	$2.80X10^{-32}$	Esophagus - Muscularis
	HLA-J	$1.40X10^{-11}$	Esophagus - Muscularis
	HLA-K	$2.60X10^{-9}$	Esophagus - Muscularis
	HLA-A	$4.70X10^{-7}$	Esophagus - Muscularis
	HCG9	$4.50X10^{-6}$	Esophagus - Muscularis
	ZFP57	$2.80X10^{-4}$	Esophagus - Muscularis
	RPL23AP1	$3.70X10^{-4}$	Esophagus - Muscularis

	HCP5B	$6.60X10^{-18}$	Heart - Atrial Appendage
	HCG4P3	$1.40X10^{-17}$	Heart - Atrial Appendage
	HCG9	$3.00X10^{-8}$	Heart - Atrial Appendage
	HLA-K	$1.60X10^{-6}$	Heart - Atrial Appendage
	ZFP57	$5.40X10^{-5}$	Heart - Atrial Appendage
	HLA-A	$1.30X10^{-4}$	Heart - Atrial Appendage
	HCG4P3	$1.10X10^{-23}$	Heart - Left Ventricle
	HCP5B	$7.70X10^{-20}$	Heart - Left Ventricle
	HCG9	$2.20X10^{-7}$	Heart - Left Ventricle
	HLA-H	$2.50X10^{-6}$	Heart - Left Ventricle
	HLA-V	$7.60X10^{-6}$	Heart - Left Ventricle
	HLA-K	$1.70X10^{-5}$	Heart - Left Ventricle
	ZNRD1	$1.20X10^{-4}$	Heart - Left Ventricle
	HCP5B	$4.80X10^{-7}$	Kidney - Cortex
	HCP5B	$1.30X10^{-10}$	Liver
	HLA-K	$3.60X10^{-8}$	Liver
	HLA-F	$5.40X10^{-5}$	Liver
	HCP5B	$1.80X10^{-39}$	Lung
	HCG4P3	$1.00X10^{-25}$	Lung
	HLA-H	$3.00X10^{-9}$	Lung
	HLA-V	$2.60X10^{-7}$	Lung
	HCG4P7	$1.60X10^{-6}$	Lung
	HCP5B	$1.10X10^{-7}$	Minor Salivary Gland
	HLA-K	$3.70X10^{-6}$	Minor Salivary Gland
	HCG4P3	$1.60X10^{-5}$	Minor Salivary Gland
	HCP5B	$3.20X10^{-30}$	Muscle - Skeletal
	HLA-H	$1.20X10^{-14}$	Muscle - Skeletal
	HCG4P3	$1.40X10^{-10}$	Muscle - Skeletal
	HLA-J	$9.30X10^{-9}$	Muscle - Skeletal
	ZNRD1ASP	$1.80X10^{-5}$	Muscle - Skeletal
	TRIM26	$2.80X10^{-4}$	Muscle - Skeletal
	HCP5B	$8.10X10^{-45}$	Nerve - Tibial
	HCG4P3	$3.20X10^{-27}$	Nerve - Tibial

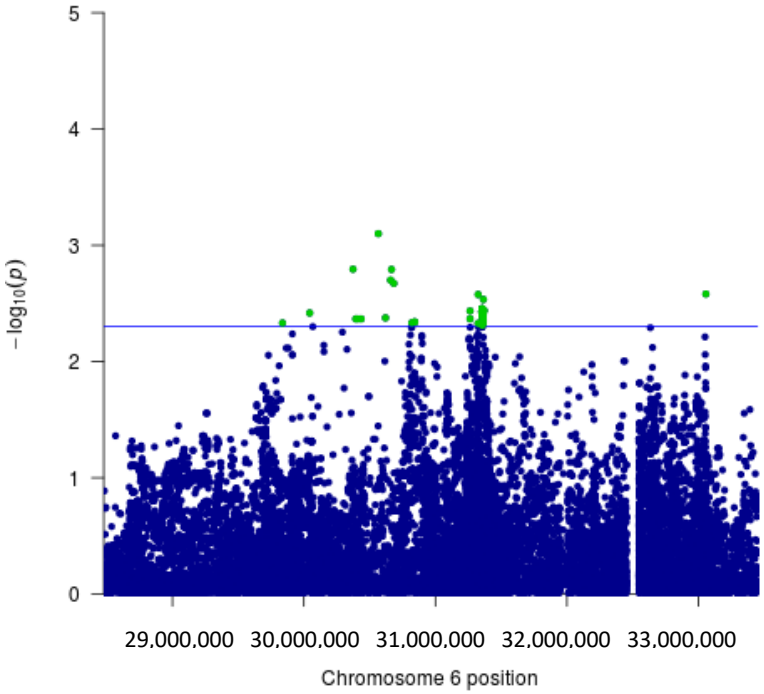
	HLA-J	$2.20X10^{-15}$	Nerve - Tibial
	HLA-H	$1.80X10^{-10}$	Nerve - Tibial
	RNF39	$5.00X10^{-6}$	Nerve - Tibial
	ZNRD1	$2.20X10^{-5}$	Nerve - Tibial
	MICD	$3.60X10^{-5}$	Nerve - Tibial
	ZFP57	$5.20X10^{-5}$	Nerve - Tibial
	MICE	$1.80X10^{-4}$	Nerve - Tibial
	HCP5B	$9.30X10^{-16}$	Ovary
	HCG4P3	$2.40X10^{-13}$	Ovary
	HLA-J	$3.50X10^{-9}$	Ovary
	HCP5B	$7.40X10^{-23}$	Pancreas
	HLA-J	$2.80X10^{-5}$	Pancreas
	HCG4P3	$9.20X10^{-5}$	Pancreas
	HCP5B	$1.80X10^{-27}$	Pituitary
	HCG4P3	$4.20X10^{-12}$	Pituitary
	HLA-V	$6.30X10^{-5}$	Pituitary
	HCP5B	$5.20X10^{-17}$	Prostate
	HCG4P3	$6.20X10^{-11}$	Prostate
	HCP5B	$1.20X10^{-28}$	Skin - Not Sun Exposed (Suprapubic)
	HCG4P3	$3.00X10^{-15}$	Skin - Not Sun Exposed (Suprapubic)
	HLA-K	$4.60X10^{-13}$	Skin - Not Sun Exposed (Suprapubic)
	HLA-V	$1.40X10^{-11}$	Skin - Not Sun Exposed (Suprapubic)
	HLA-A	$1.10X10^{-4}$	Skin - Not Sun Exposed (Suprapubic)
	HLA-H	$2.40X10^{-4}$	Skin - Not Sun Exposed (Suprapubic)
	HCP5B	$1.90X10^{-46}$	Skin - Sun Exposed (Lower leg)
	HCG4P3	$1.50X10^{-20}$	Skin - Sun Exposed (Lower leg)
	HLA-K	$1.50X10^{-13}$	Skin - Sun Exposed (Lower leg)
	ZFP57	$6.10X10^{-9}$	Skin - Sun Exposed (Lower leg)
	HLA-V	$1.50X10^{-7}$	Skin - Sun Exposed (Lower leg)
	HLA-F	$2.40X10^{-7}$	Skin - Sun Exposed (Lower leg)
	HLA-H	$1.20X10^{-5}$	Skin - Sun Exposed (Lower leg)
	MICE	$4.40X10^{-5}$	Skin - Sun Exposed (Lower leg)
	ZNRD1ASP	$2.30X10^{-4}$	Skin - Sun Exposed (Lower leg)

	HCP5B	$1.30X10^{-14}$	Small Intestine - Terminal Ileum
	HCG4P3	$8.20X10^{-12}$	Small Intestine - Terminal Ileum
	HLA-K	$5.10X10^{-7}$	Small Intestine - Terminal Ileum
	HLA-J	$5.80X10^{-5}$	Small Intestine - Terminal Ileum
	HCG4P3	$2.20X10^{-19}$	Spleen
	HCP5B	$7.60X10^{-17}$	Spleen
	HLA-J	$2.40X10^{-7}$	Spleen
	HLA-H	$2.70X10^{-6}$	Spleen
	HCP5B	$4.70X10^{-29}$	Stomach
	HCG4P3	$7.10X10^{-19}$	Stomach
	HLA-K	$1.10X10^{-10}$	Stomach
	HLA-J	$6.40X10^{-8}$	Stomach
	MICE	$7.50X10^{-5}$	Stomach
	HCG4P3	$7.30X10^{-22}$	Testis
	HCP5B	$7.50X10^{-20}$	Testis
	HLA-J	$2.00X10^{-14}$	Testis
	HLA-K	$2.40X10^{-9}$	Testis
	HLA-H	$1.00X10^{-6}$	Testis
	ZDHHC20P1	$3.80X10^{-6}$	Testis
	HLA-A	$3.80X10^{-6}$	Testis
	HCG4B	$4.70X10^{-5}$	Testis
	HLA-G	$1.20X10^{-4}$	Testis
	HCP5B	$5.30X10^{-60}$	Thyroid
	HCG4P3	$2.40X10^{-39}$	Thyroid
	HLA-J	$1.20X10^{-18}$	Thyroid
	HLA-A	$2.40X10^{-14}$	Thyroid
	ZFP57	$1.00X10^{-9}$	Thyroid
	RNF39	$8.10X10^{-8}$	Thyroid
	ZNRD1ASP	$9.70X10^{-6}$	Thyroid
	FLOT1	$1.40X10^{-5}$	Thyroid
	HLA-V	$2.00X10^{-5}$	Thyroid
	HLA-H	$2.10X10^{-5}$	Thyroid

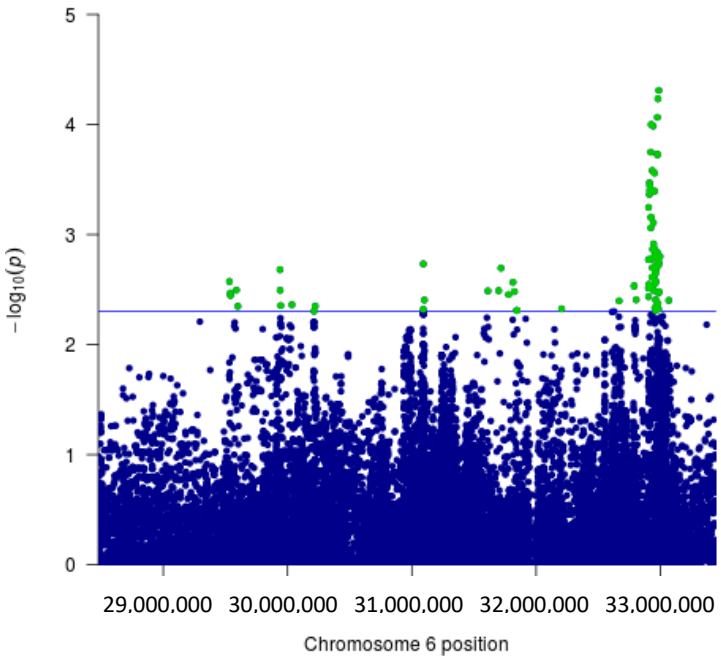
	MICE	$9.20X10^{-5}$	Thyroid
	HCG4	$9.30X10^{-5}$	Thyroid
	HLA-F-AS1	$4.70X10^{-4}$	Thyroid
	HCG4P3	$8.30X10^{-10}$	Uterus
	HCP5B	$1.20X10^{-9}$	Uterus
	HCP5B	$6.10X10^{-13}$	Vagina
	HCG4P3	$4.70X10^{-7}$	Vagina
	HLA-K	$4.60X10^{-5}$	Vagina
	HCP5B	$6.30X10^{-25}$	Whole Blood
	HCG4P3	$2.10X10^{-22}$	Whole Blood
	HLA-H	$2.40X10^{-10}$	Whole Blood
	HCG9	$1.70X10^{-8}$	Whole Blood
	HLA-F-AS1	$2.00X10^{-8}$	Whole Blood
	DDX39BP2	$2.20X10^{-8}$	Whole Blood
	ZFP57	$4.30X10^{-7}$	Whole Blood
	ZNRD1	$1.90X10^{-6}$	Whole Blood
	HCG4P7	$5.30X10^{-6}$	Whole Blood
	IFITM4P	$1.60X10^{-5}$	Whole Blood
	HLA-J	$4.20X10^{-5}$	Whole Blood
	MICD	$3.20X10^{-4}$	Whole Blood

Chapter four supplementary data:

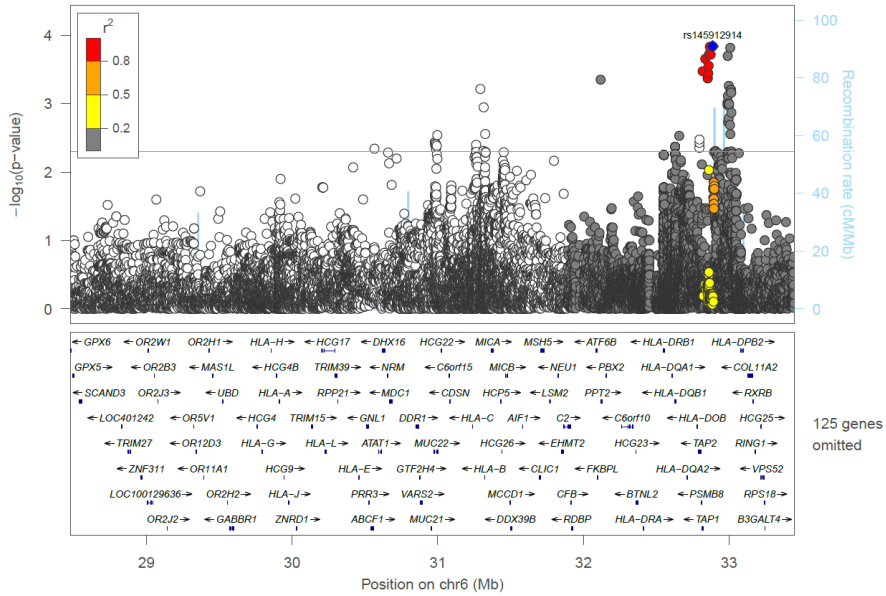
A meta-analysis of results from interaction analyses of MUC5B risk allele status and SNPs in the HLA region in IPF susceptibility in the UK and Colorado IPF datasets



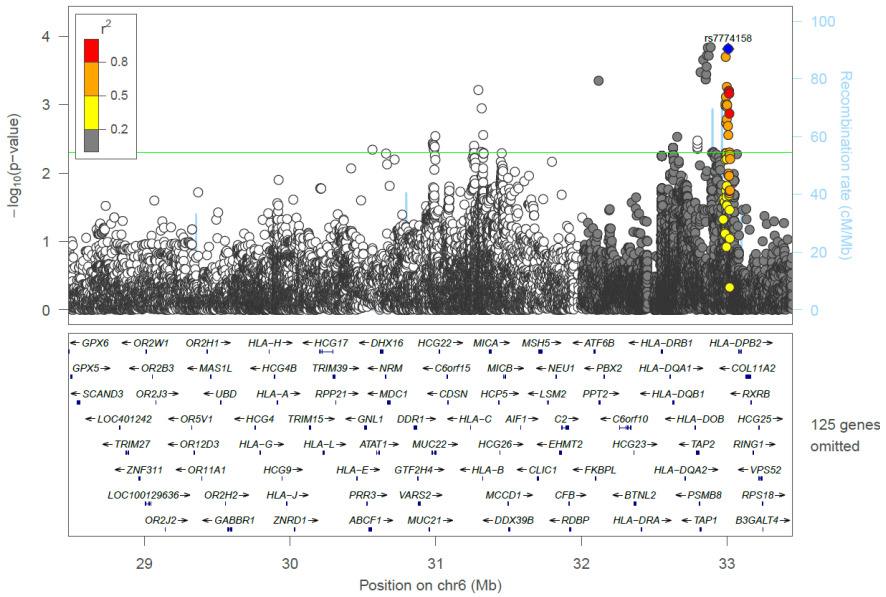
Supplementary figure 4.1: Manhattan plot of the HLA region for the MUC5B interaction analysis of IPF susceptibility in the Colorado IPF dataset (the green variants are all the variants that passed the suggestive significance threshold). Blue line is suggestive significance threshold of $P<5x10^{-3}$).



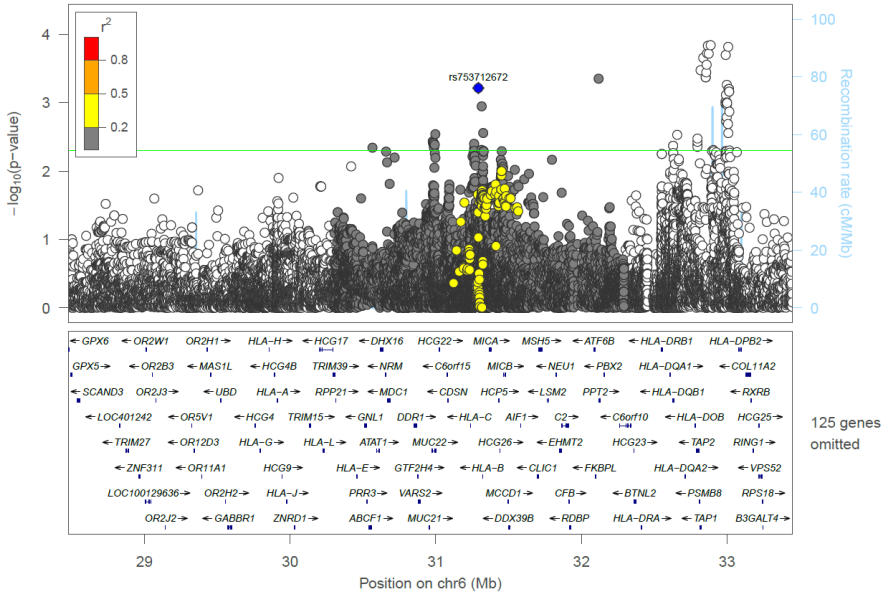
Supplementary figure 4.2: Manhattan plot of the HLA region for the MUC5B interaction analysis of IPF susceptibility in the UUS IPF dataset (the green variants are all the variants that passed the suggestive significance threshold). Blue line is suggestive significance threshold of $P<5x10^{-3}$).



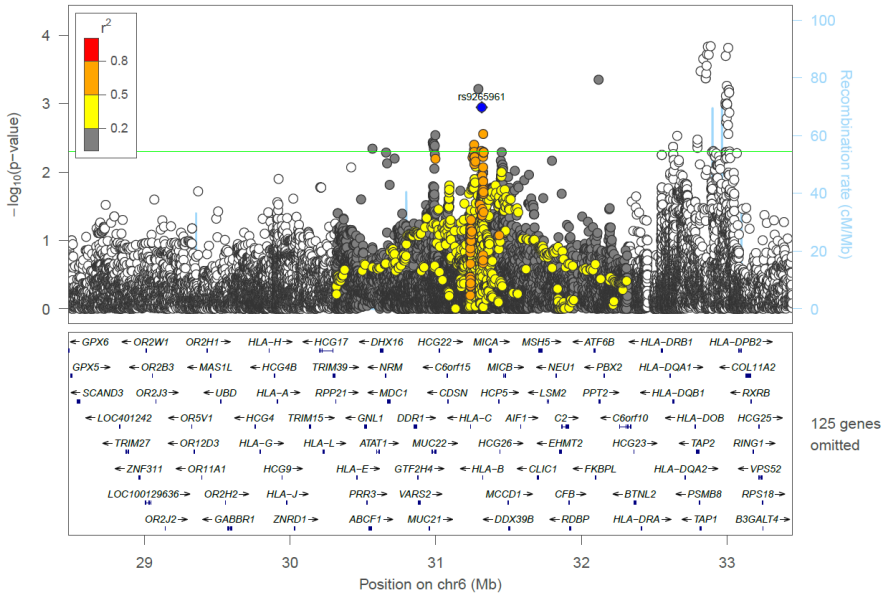
Supplementary figure 4.3: Region plot of rs145912914 over the whole HLA region (28477797-33448354 bp) for a SNP-SNP interaction meta-analysis with MUC5B in IPF susceptibility. Green line is suggestive significance threshold of P<5x10^{-3}).



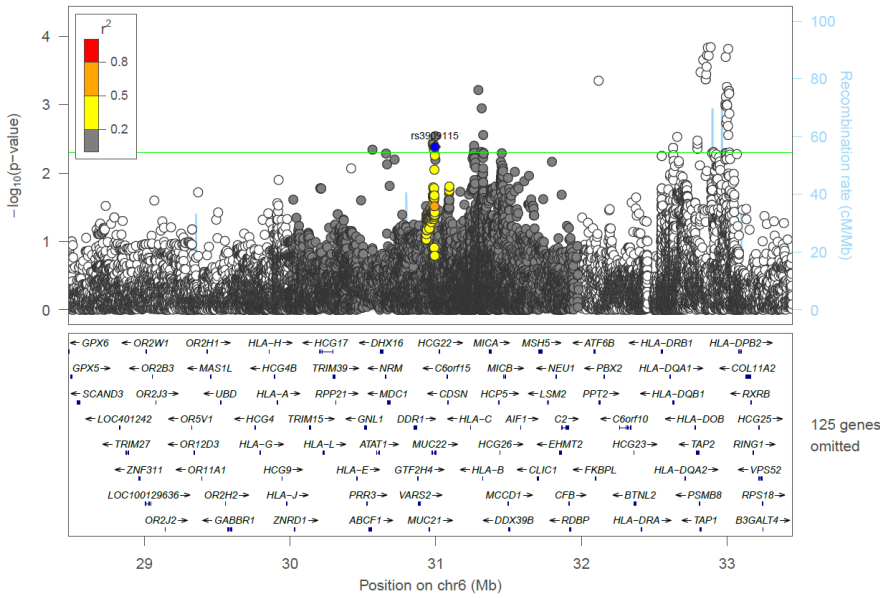
Supplementary figure 4.4: Region plot of rs7774158 over the whole HLA region (28477797-33448354 bp) for a SNP-SNP interaction meta-analysis with MUC5B in IPF susceptibility. Green line is suggestive significance threshold of P<5x10^{-3}).



Supplementary figure 4.5: Region plot of rs753712672 over the whole HLA region (28477797-33448354 bp) for a SNP-SNP interaction meta-analysis with MUC5B in IPF susceptibility. Green line is suggestive significance threshold of P<5x10^{-3}).



Supplementary figure 4.6: Region plot of rs9265961 over the whole HLA region (28477797-33448354 bp) for a SNP-SNP interaction meta-analysis with MUC5B in IPF susceptibility. Green line is suggestive significance threshold of P<5x10^{-3}).



Supplementary figure 4.7: Region plot of rs3909115 over the whole HLA region (28477797-33448354 bp) for a SNP-SNP interaction meta-analysis with MUC5B in IPF susceptibility. Green line is suggestive significance threshold of $P<5x10^{-3}$).

Supplementary table 4.1: Independent signals ($P<5x10^{-3}$) in an interaction meta-analyses of MUC5B risk allele status and the HLA region analyses of IPF susceptibility in the UK, UUS and Colorado IPF datasets. Variants with an asterisk are in the same direction of effect in all datasets and has p-values of < 0.05 in at least two datasets.

Dataset	**rsid**	**BP Position**	**Nearest Gene**	**Coded/ Non-coded allele**	**Info Score**	**Coded allele frequency**	**P-Value**	**OR (95% CI)**	**Meta-Analysis P-value**	**Meta-Analysis OR (95% CI)**
UK	rs145912914*	32887150	*LOC100294145*	G/C	0.99	0.02	$7.70x10^{-3}$	0.30 (0.12-0.72)	$1.45x10^{-4}$	0.44 (0.29-0.67)
Colorado					0.99	0.02	0.036	0.53 (0.29-0.96)		
UUS					0.99	0.02	0.045	0.43 (0.19-0.98)		
UK	rs7774158*	33007752	*HLA-DOA*	C/A	1.00	0.35	$9.42x10^{-3}$	0.70 (0.53-0.92)	$1.53x10^{-4}$	0.78 (0.69-0.89)
Colorado					1.00	0.35	0.065	0.84 (0.69-1.01)		
UUS					1.00	0.34	0.020	0.76 (0.60-0.96)		
UK	rs114334832	32118085	*PRRT-1*	A/G	0.65	0.01	0.056	0.35 (0.12-1.03)	$4.46x10^{-4}$	0.38 (0.22-0.65)
Colorado					0.65	0.01	0.051	0.42 (0.18-1.00)		
UUS					0.63	0.01	0.026	0.36 (0.15-0.89)		
UK	rs753712672*	31292555	*LOC112267902*	A/G	0.56	0.22	$7.27x10^{-3}$	0.65 (0.48-0.89)	$6.10x10^{-4}$	0.70 (0.57-0.863)
Colorado					NA	NA	NA	NA		
UUS					0.57	0.21	0.028	0.74 (0.56-0.97)		
UK	rs9265961*	31315501	*LOC112267902*	A/G	1.00	0.34	0.847	0.97 (0.74-1.28)	$1.13x10^{-3}$	0.80 (0.71-0.92)
Colorado					0.8	0.29	$6.16x10^{-3}$	0.76 (0.63-0.93)		

UUS					1.300	0.33	0.019	0.76 (0.60-0.96)		
UK	rs200777418	30996959	*MUC22* (missense coding)	G/A	0.62	0.04	0.062	0.54 (0.28-1.03)	$2.88x10^{-3}$	0.62 (0.45-0.85)
Colorado					0.60	0.05	0.097	0.66 (0.40-1.08)		
UUS					0.64	0.05	0.089	0.63 (0.37-1.07)		
UK	rs9275206	32657565	*HLA-DQB1*	G/A	1.00	0.10	0.150	0.73 (0.47-1.12)	$2.95x10^{-3}$	0.72 (0.57-0.89)
Colorado					1.00	0.10	0.148	0.80 (0.58-1.08)		
UUS					1.00	0.10	0.013	0.56 (0.36-0.88)		
UK	rs241447	32796750	*TAP2* (missense coding)	G/A	1.00	0.25	0.480	1.12 (0.82-1.51)	$3.35x10^{-3}$	1.24 (1.07-1.44)
Colorado					1.00	0.26	0.104	1.19 (0.97-1.46)		
UUS					1.00	0.24	$6.01x10^{-3}$	1.46 (1.12-1.92)		
UK	rs3869094	30979558	*MUC22*	T/C	1.00	0.41	0.566	1.08 (0.83-1.40)	$3.66x10^{-3}$	1.20 (1.06-1.36)
Colorado					1.00	0.43	0.093	1.17 (0.97-1.40)		
UUS					1.00	0.41	$7.35x10^{-3}$	1.36 (1.09-1.70)		
UK	rs3909115*	30993188	*MUC22*	A/C	0.99	0.25	0.578	1.09 (0.81-1.46)	$4.13x10^{-3}$	1.23 (1.07-1.41)
Colorado					0.99	0.27	0.038	1.24 (1.01-1.51)		
UUS					0.99	0.25	0.031	1.32 (1.03-1.71)		
UK	rs139105056	30565628	*ABCF1*	G/A	0.91	0.01	0.859	0.90 (0.27-3.00)	$4.55x10^{-3}$	2.07 (1.24-3.41)
Colorado					0.89	0.02	$7.94x10^{-4}$	3.52 (1.01-7.34)		

UUS					0.85	0.02	0.301	1.55 (0.68-3.56)		
UK	rs714289	32905811	*HLA-DMB*	A/G	1.00	0.06	0.062	0.58 (0.33-1.03)	$4.91x10^{-3}$	0.69 (0.54-0.89)
Colorado					1.00	0.06	0.303	0.82 (0.56-1.20)		
UUS					1.00	0.06	0.030	0.62 (0.40-0.96)		

*Supplementary table 4.2: Effects of the signals from the HLA-wide variant*MUC5B interaction analyses in MUC5B positive and MUC5B negative individuals.*

rsid	***MUC5B* positive results**		***MUC5B* negative results**	
	Meta-Analysis P-value	**Meta-Analysis OR (95% CI)**	**Meta-Analysis P-value**	**Meta-Analysis OR (95% CI)**
rs145912914	0.0028	0.63 (0.46-0.85)	0.018	1.41 (1.06-1.89)
rs7774158	0.0088	0.88 (0.80-0.97)	0.13	1.07 (0.98-1.17)
rs753712672	0.026	0.80 (0.66-0.97)	0.79	1.03 (0.82-1.29)
rs9265961	0.76	1.02 (0.92-1.13)	$5.16x10^{-4}$	0.84 (0.76-0.93)
rs3909115	0.091	1.09 (0.99-1.21)	0.24	0.94 (0.86-1.04)

Supplementary table 4.3: Summary of results from a GTEx search of the five lead SNPs identified in the MUC5B interaction meta-analysis of IPF susceptibility.

SNP	SNP position	Gene	P-value	Expression effect size	Tissue
rs9265961	31315501	*PSORS1C1*	$1.46x10^{-10}$	0.44	Lung
rs9265961	31315501	*PSORS1C2*	$3.30x10^{-9}$	0.45	Lung
rs9265961	31315501	*HCG27*	$2.24x10^{-11}$	0.29	Lung
rs9265961	31315501	*HLA-C*	$7.77x10^{-15}$	-0.41	Lung
rs9265961	31315501	*XXbac-BPG248L24.12*	$5.50x10^{-8}$	0.39	Lung
rs9265961	31315501	*HLA-S*	$3.00x10^{-13}$	0.53	Lung
rs9265961	31315501	*XXbac-BPG181B23.7*	$5.26x10^{-11}$	-0.48	Lung
rs9265961	31315501	*MICA*	$2.97x10^{-8}$	-0.26	Lung
rs9265961	31315501	*ATP6V1G2*	$7.13x10^{-5}$	0.20	Lung
rs9265961	31315501	*LY6G5B*	$7.69x10^{-6}$	-0.11	Lung
rs3909115	30993188	*PSORS1C1*	$4.68x10^{-5}$	0.29	Lung

Supplementary table 4.4: Summary of results from a phenoscanner search of the five suggestively significant lead SNPs (and SNPs in high LD [r^2>0.8]) from the MUC5B interaction meta-analysis of IPF susceptibility.

SNP	Alleles	Trait	P-value	Beta	Number in study	Reference
rs145912914	C/G	Self-reported ankylosing spondylitis	$3.59x10^{-8}$	0.002359	337159	UKBB
rs145912914		Self-reported type 1 diabetes	$3.67x10^{-9}$	0.001375	337159	UKBB
rs3909115	C/A	Primary sclerosing cholangitis	$1.38x10^{-13}$	-0.2882	14890	(223)
rs3909115		Rheumatoid arthritis	$2.60x10^{-18}$	-0.157	58284	(119)
rs7774158	A/C	Intestinal malabsoption	$8.05x10^{-13}$	-0.00079	337199	UKBB
rs7774158		Self-reported malabsorption or coeliac disease	$3.65x10^{-29}$	-0.00188	337159	UKBB
rs7774158		Rheumatoid arthritis	$6.90x10^{-11}$	0.1044	58284	(119)
rs9265961	G/A	Basophil count	$2.65x10^{-16}$	-0.03046	173480	(221)
rs9265961		Eosinophil count	$3.07x10^{-33}$	-0.04555	173480	
rs9265961		Eosinophil percentage of granulocytes	$5.28x10^{-9}$	-0.02221	173480	
rs9265961		Eosinophil percentage of wbc	$3.00x10^{-9}$	-0.02247	173480	
rs9265961		Granulocyte count	$3.60x10^{-44}$	-0.05322	173480	
rs9265961		Granulocyte percentage of myeloid white cells	$1.27x10^{-9}$	-0.02306	173480	
rs9265961		Lymphocyte count	$4.22x10^{-74}$	-0.06971	173480	
rs9265961		Monocyte count	$5.40x10^{-11}$	-0.02491	173480	

rs9265961		Monocyte percentage of white cells	$1.14x10^{-15}$	0.03034	173480	
rs9265961		Myeloid white cell count	$1.47x10^{-44}$	-0.05359	173480	
rs9265961		Neutrophil count	$1.93x10^{-38}$	-0.04938	173480	
rs9265961		Neutrophil percentage of granulocytes	$1.45x10^{-8}$	0.02156	173480	
rs9265961		Sum basophil neutrophil counts	$1.01x10^{-38}$	-0.04964	173480	
rs9265961		Sum eosinophil basophil counts	$1.60x10^{-38}$	-0.04927	173480	
rs9265961		Sum neutrophil eosinophil counts	$1.30x10^{-43}$	-0.05279	173480	
rs9265961		White blood cell count	$2.54x10^{-78}$	-0.07141	173480	
rs9265961		IgA deficiency	$2.37x10^{-12}$	0.3433	6487	(222)
rs9265961		Primary sclerosing cholangitis	$4.83x10^{-93}$	0.6946	14890	(223)
rs9265961		Doctor diagnosed sarcoidosis	$1.24x10^{-11}$	0.002266	83529	UKBB
rs9265961		FEV1	$5.87x10^{-9}$	0.0168	110423	
rs9265961		FVC	$1.68x10^{-12}$	0.01477	307638	
rs9265961		FVC best measure	$7.73x10^{-11}$	0.01492	255492	
rs9265961		Height	$1.02x10^{-59}$	0.02985	336474	
rs9265961		Intestinal malabsorption	$2.64x10^{-46}$	0.001588	337199	
rs9265961		Self-reported adrenocortisol insufficient or Addison's disease	$1.38x10^{-8}$	0.0003	337159	
rs9265961		Self-reported ankylosing spondylitis	$4.33x10^{-13}$	-0.001	337159	
rs9265961		Self-reported malabsorption or coeliac disease	$4.72x10^{-124}$	0.004011	337159	

rs9265961		Self-reported multiple sclerosis	$6.42x10^{-11}$	0.001018	337159	
rs9265961		Self-reported psoriasis	$3.59x10^{-48}$	-0.00402	337159	
rs9265961		Self-reported sarcoidosis	$2.46x10^{-14}$	0.000865	337159	
rs9265961		Sitting height	$9.90x10^{-64}$	0.03358	336172	
rs9265961		IgG galactosylation	$1.11x10^{-8}$	NA	1960	(224)

Chapter five supplementary data:

Supplementary table 5.1: Table of KIR haplotype/gene imputation accuracies from KIR*IMP for each input SNP imputation threshold.

KIR gene/haplotype	Input SNPs imputation threshold	Imputation accuracy (%)
KIRhaplotype	Genotyped	67.64
	0.3	87.47
	0.5	86.85
	0.7	75.37
	0.8	74.74
	0.9	74.53
	0.95	73.49
A/B	Genotyped	81.63
	0.3	98.75
	0.5	98.75
	0.7	87.47
	0.8	85.8
	0.9	84.97
	0.95	84.76
KIR2DS2	Genotyped	75.78
	0.3	98.96
	0.5	98.96
	0.7	82.46
	0.8	78.5
	0.9	78.5
	0.95	78.71
KIR2DL2	Genotyped	75.37
	0.3	98.33
	0.5	98.33
	0.7	82.67
	0.8	78.71
	0.9	79.12
	0.95	77.66
KIR2DL3	Genotyped	75.57
	0.3	98.75
	0.5	98.75
	0.7	82.05
	0.8	78.29
	0.9	77.24
	0.95	76.83
KIR2DP1	Genotyped	85.59
	0.3	92.07
	0.5	91.44
	0.7	87.06
	0.8	86.43
	0.9	86.43
	0.95	86.22

KIR2DL1	Genotyped	85.18
	0.3	91.44
	0.5	91.23
	0.7	86.64
	0.8	86.01
	0.9	85.8
	0.95	85.39
KIR3DP1	Genotyped	96.87
	0.3	96.87
	0.5	96.87
	0.7	96.66
	0.8	96.87
	0.9	96.87
	0.95	96.66
KIR2DL4	Genotyped	97.08
	0.3	97.08
	0.5	97.08
	0.7	97.08
	0.8	97.08
	0.9	97.08
	0.95	97.08
KIR3DL1ex4	Genotyped	97.91
	0.3	98.54
	0.5	98.33
	0.7	98.54
	0.8	98.33
	0.9	98.33
	0.95	98.33
KIR3DL1ex9	Genotyped	97.49
	0.3	98.12
	0.5	98.12
	0.7	98.12
	0.8	98.12
	0.9	98.12
	0.95	98.12
KIR3DS1	Genotyped	97.70
	0.3	97.49
	0.5	97.49
	0.7	97.49
	0.8	97.49
	0.9	97.49
	0.95	97.29
KIR2DL5	Genotyped	87.47
	0.3	92.69
	0.5	92.28
	0.7	90.81
	0.8	89.77
	0.9	89.56
	0.95	88.73
KIR2DS3	Genotyped	87.47

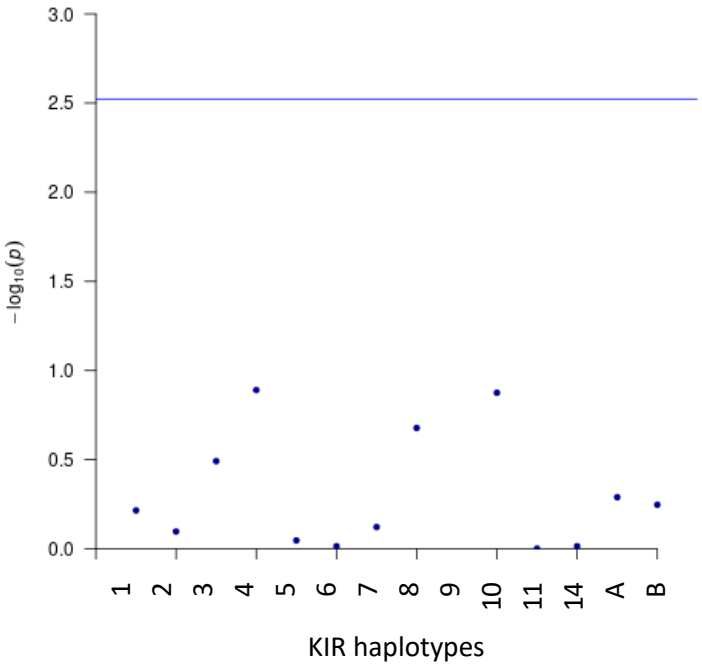
	0.3	91.02
	0.5	92.48
	0.7	90.19
	0.8	88.94
	0.9	89.56
	0.95	89.77
KIR2DS5	Genotyped	95.41
	0.3	95.62
	0.5	96.24
	0.7	95.82
	0.8	95.82
	0.9	96.24
	0.95	95.41
KIR2DS1	Genotyped	97.70
	0.3	98.33
	0.5	98.54
	0.7	98.33
	0.8	98.33
	0.9	98.33
	0.95	98.33
KIR2DS4TOTAL	Genotyped	97.91
	0.3	98.12
	0.5	98.33
	0.7	98.12
	0.8	98.12
	0.9	98.12
	0.95	98.12
KIR2DS4WT	Genotyped	94.57
	0.3	99.37
	0.5	99.58
	0.7	99.37
	0.8	99.58
	0.9	99.37
	0.95	99.16
KIR2DS4DEL	Genotyped	93.74
	0.3	98.75
	0.5	98.75
	0.7	98.75
	0.8	98.75
	0.9	98.54
	0.95	98.54

Supplementary table 5.2: Imputation quality of each tag SNP across all four IPF datasets

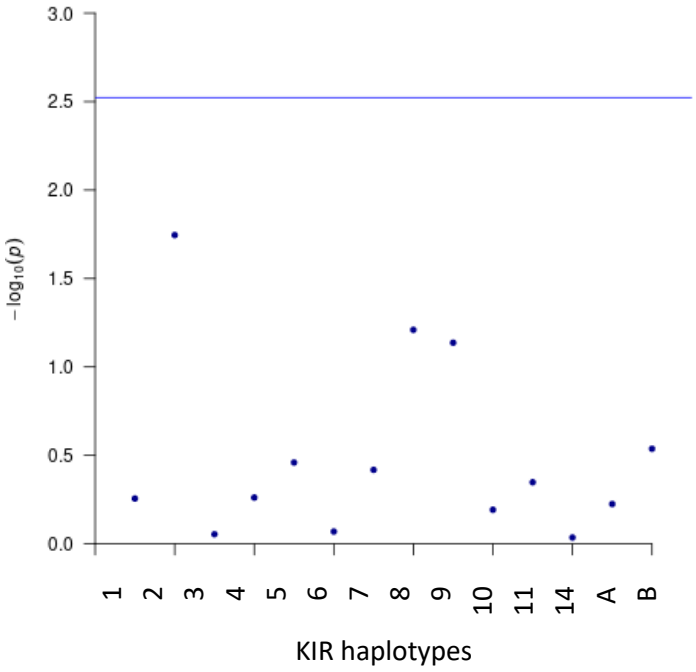
Tag SNP	Imputation quality			
	UK	UUS	Colorado	Chicago
rs587560	0.52	0.53	0.32	0.42
rs1010355	0.85	0.82	0.81	0.60
rs592645	NA	NA	NA	NA
seq-t1d-19-60034052-C-T	0.90	0.91	0.49	0.53
rs4806585	0.84	0.84	0.72	0.32
rs581623	0.84	0.82	Genotyped	0.59

Chapter six supplementary data:

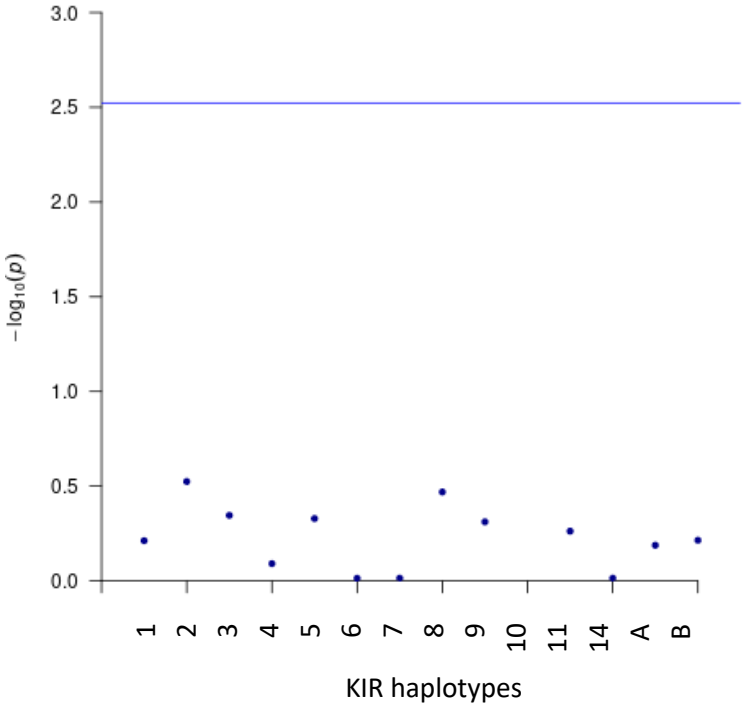
KIR-wide association meta-analysis of IPF susceptibility in the UK, UUS, Chicago and Colorado datasets



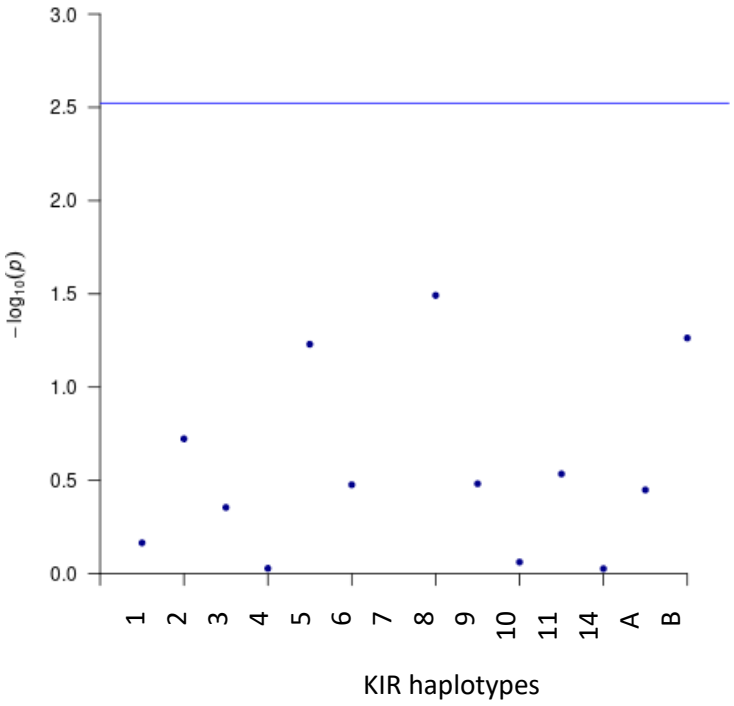
Supplementary figure 6.1: -log10 p-values of the association analysis of KIR haplotypes in IPF susceptibility in the UK dataset (haplotype 9 has been removed). Blue line denotes the Bonferroni corrected significance threshold of 0.004.



Supplementary figure 6.2: -log10 p-values of the association analysis of KIR haplotypes in IPF susceptibility in the UUS dataset. Blue line denotes the Bonferroni corrected significance threshold of 0.004.



Supplementary figure 6.3: -log10 p-values of the association analysis of KIR haplotypes in IPF susceptibility in the Chicago dataset (haplotype 10 was removed for low count). Blue line denotes the Bonferroni corrected significance threshold of 0.004.

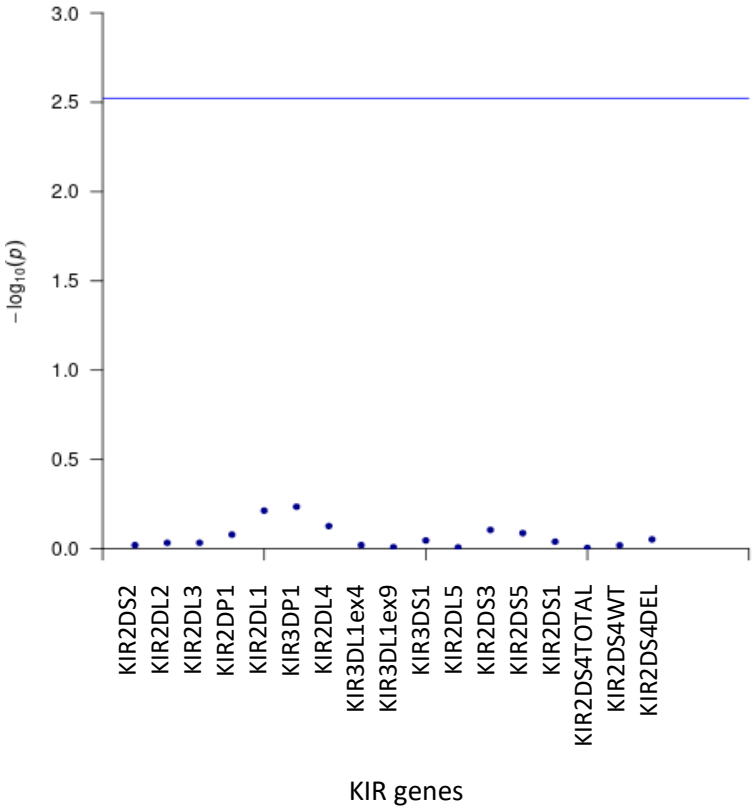


Supplementary figure 6.4: -log10 p-values of the association analysis of KIR haplotypes in IPF susceptibility in the Colorado dataset (haplotype 7 was removed for low count). Blue line denotes the Bonferroni corrected significance threshold of 0.004.

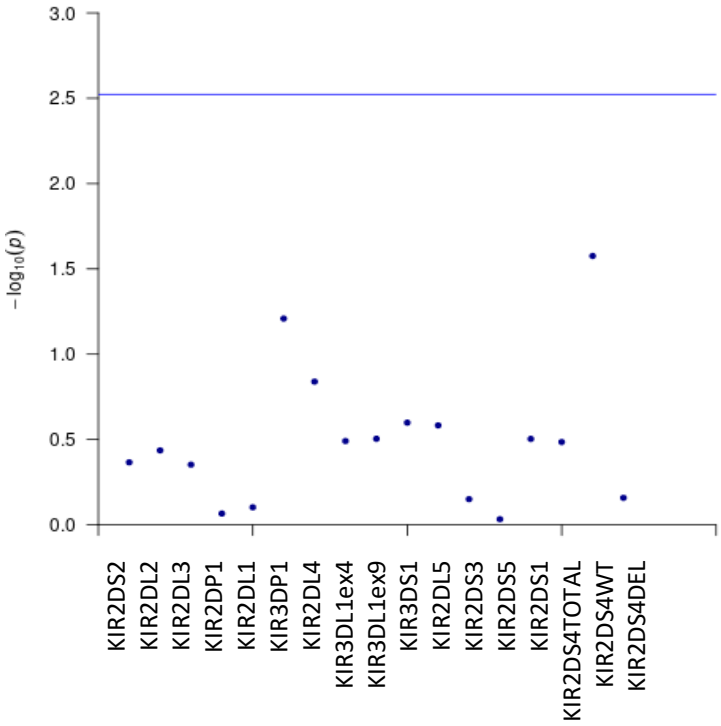
Supplementary table 6.1: Results for the KIR haplotype association analysis in IPF susceptibility for each of the four IPF datasets.

Locus	Chicago P-Value	Chicago odds ratio (95% CI)	Colorado P-Value	Colorado odds ratio (95% CI)	UK P-Value	UK odds ratio (95% CI)	UUS P-Value	UUS Odds ratio (95% CI)
1	0.61	1.06 (0.85-1.32)	0.68	1.02 (0.93-1.12)	0.61	1.04 (0.91-1.18)	0.55	1.03 (0.93-1.16)
2	0.30	0.85 (0.63-1.15)	0.19	0.90 (0.76-1.05)	0.80	0.96 (0.72-1.29)	0.02	0.68 (0.49-0.94)
3	0.45	0.87 (0.60-1.26)	0.44	0.87 (0.62-1.23)	0.32	0.87 (0.66-1.15)	0.88	1.02 (0.81-1.27)
4	0.81	0.87 (0.29-2.64)	0.94	1.05 (0.33-3.34)	0.13	1.77 (0.85-3.71)	0.55	1.23 (0.63-2.40)
5	0.47	1.20 (0.73-1.99)	0.06	1.32 (0.99-1.76)	0.90	1.05 (0.48-2.32)	0.35	0.62 (0.22-1.69)
6	0.97	829632.67 ($1.50x10^{-295}$-$4.59x10^{+306}$)	0.33	1.56 (0.63-3.86)	0.97	5.37E-05 ($2.66x10^{-200}$-$1.09x10^{+191}$)	0.85	0.82 (0.10-6.52)
7	0.97	1.50E-06 ($1.19x10^{-290}$-$1.88x10^{+278}$)	NA	NA	0.75	0.78 (0.17-3.60)	0.38	0.52 (0.12-2.23)
8	0.34	1.90 (0.51-7.05)	0.03	11.08 (1.23-100.00)	0.21	1.81 (0.72-4.61)	0.06	2.06 (0.97-4.38)
9	0.49	1.38 (0.55-3.47)	0.33	1.34 (0.75-2.40)	$2.17x10^{-60}$	0.29 (0.25-0.34)	0.07	1.12 (0.99-1.26)
10	NA	NA	0.87	1.10 (0.35-3.48)	0.13	3.05 (0.71-13.05)	0.64	1.85 (0.14-24.40)
11	0.55	1.73 (0.29-10.16)	0.29	0.56 (0.19-1.66)	1.00	1.00 (0.21-4.81)	0.45	1.33 (0.63-2.80)
14	0.97	441404.81 ($1.65x10^{-271}$-$1.18x10^{+282}$)	0.94	2.93E-05 ($2.26x10^{-123}$-$3.78x10^{+113}$)	0.97	$5.52x10^{-05}$ ($2.16x10^{-200}$-$1.41x10^{+191}$)	0.92	0.93 (0.22-3.95)
A	0.65	0.96 (0.78-1.16)	0.36	1.04 (0.95-1.14)	0.51	0.96 (0.84-1.09)	0.60	0.97 (0.87-1.08)

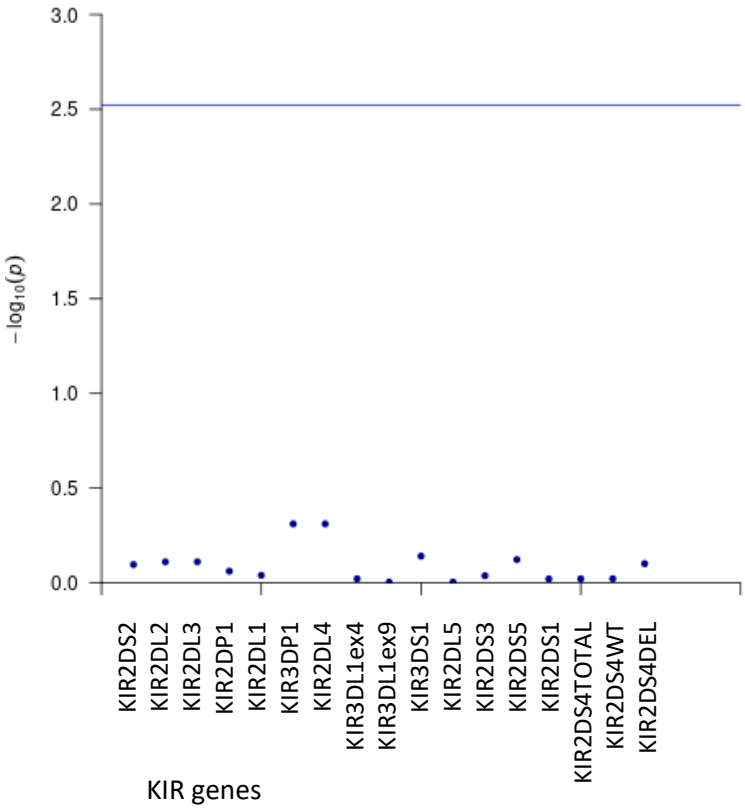
B	0.61	1.07 (0.83-1.38)	0.05	0.86 (0.74-1.00)	0.57	1.06 (0.88-1.28)	0.29	1.09 (0.93-1.27)



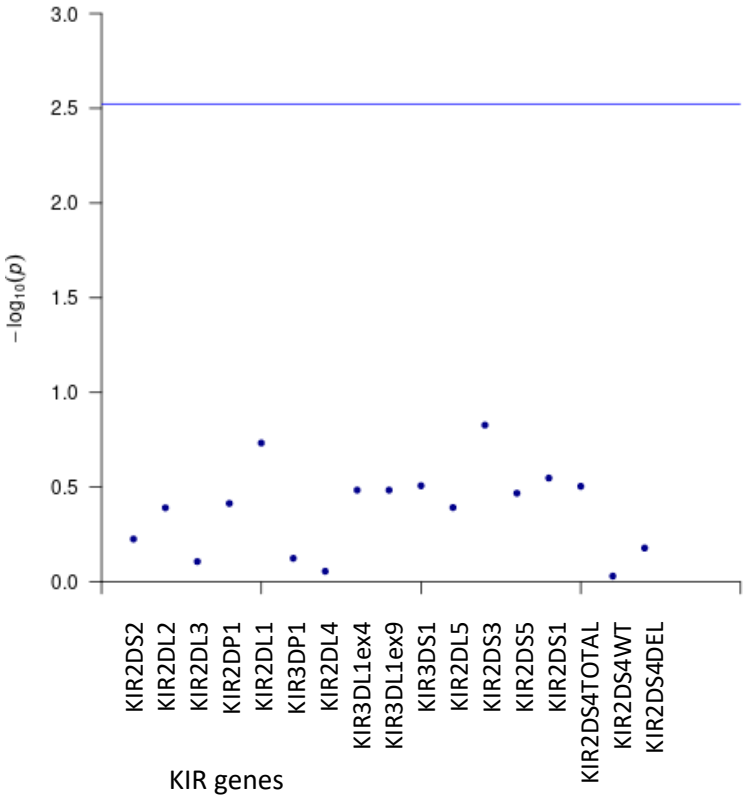
Supplementary figure 6.5: -log10 p-values of the association analysis of KIR CNVs in IPF susceptibility in the UK dataset. Blue line denotes the Bonferroni corrected significance threshold of 0.003.



Supplementary figure 6.6: -log10 p-values of the association analysis of KIR CNVs in IPF susceptibility in the UUS dataset. Blue line denotes the Bonferroni corrected significance threshold of 0.003.



Supplementary figure 6.7: -log10 p-values of the association analysis of KIR CNVs in IPF susceptibility in the Chicago dataset. Blue line denotes the Bonferroni corrected significance threshold of 0.003.



Supplementary figure 6.8: -log10 p-values of the association analysis of KIR CNVs in IPF susceptibility in the Colorado dataset. Blue line denotes the Bonferroni corrected significance threshold of 0.003.

Supplementary table 6.2: Results for the KIR CNV association analysis in IPF susceptibility for each of the four IPF datasets.

Locus	Chicago P-Value	Chicago odds ratio (95% CI)	Colorado P-Value	Colorado odds ratio (95% CI)	UK P-Value	UK odds ratio (95% CI)	UUS P-Value	UUS Odds ratio (95% CI)
KIR2DS2	0.80	1.03 (0.81-1.31)	0.59	1.04 (0.91-1.18)	0.95	1.01 (0.83-1.23)	0.43	1.07 (0.90-1.27)
KIR2DL2	0.77	1.04 (0.81-1.32)	0.41	1.06 (0.93-1.21)	0.93	1.01 (0.83-1.23)	0.37	1.08 (0.91-1.28)
KIR2DL3	0.77	0.97 (0.76- 1.23)	0.78	0.98 (0.86-1.12)	0.93	0.99 (0.81-1.21)	0.44	0.94 (0.79-1.11)
KIR2DP1	0.87	0.97 (0.72-1.31)	0.39	0.92 (0.75-1.12)	0.83	0.97 (0.72-1.30)	0.86	1.02 (0.79-1.32)
KIR2DL1	0.91	0.98 (0.73-1.32)	0.18	0.87 (0.72-1.07)	0.61	0.93 (0.69-1.24)	0.79	0.97 (0.75-1.24)
KIR3DP1	0.49	2.41 (0.20-29.23)	0.75	1.56 (0.10-23.94)	0.58	0.61 (0.11-3.46)	0.06	2.14 (0.96-4.73)
KIR2DL4	0.49	2.41 (0.20-29.23)	0.88	0.81 (0.06-11.97)	0.75	0.75 (0.13-4.41)	0.14	1.84 (0.81-4.15)
KIR3DL1ex4	0.95	1.01 (0.80-1.26)	0.33	1.08 (0.93-1.26)	0.95	1.00 (0.86-1.16)	0.32	0.94 (0.83-1.06)
KIR3DL1ex9	0.99	1.00 (0.80-1.26)	0.33	1.08 (0.93-1.26)	0.98	1.00 (0.86-1.16)	0.31	0.94 (0.83-1.06)
KIR3DS1	0.72	1.04 (0.83-1.30)	0.31	0.92 (0.79-1.08)	0.90	0.99 (0.85-1.15)	0.25	1.07 (0.95-1.22)
KIR2DL5	0.99	1.00 (0.83-1.21)	0.41	0.95 (0.84-1.07)	0.98	1.00 (0.89-1.12)	0.26	1.06 (0.96-1.17)
KIR2DS3	0.92	1.02 (0.70-1.48)	0.15	1.14 (0.95-1.36)	0.78	1.03 (0.83-1.28)	0.71	1.04 (0.86-1.24)
KIR2DS5	0.75	0.96 (0.75-1.23)	0.34	0.94 (0.83-1.07)	0.82	0.98 (0.83-1.16)	0.93	1.01 (0.87-1.16)
KIR2DS1	0.95	0.99 (0.79-1.25)	0.28	0.92 (0.79-1.07)	0.91	1.01 (0.87-1.17)	0.31	1.07 (0.94-1.21)

KIR2DS4TOTAL	0.95	1.01 (0.80-1.26)	0.31	1.08 (0.93-1.26)	0.99	1.00 (0.86-1.16)	0.33	0.94 (0.83-1.07)
KIR2DS4WT	0.95	1.01 (0.82-1.23)	0.93	1.00 (0.92-1.10)	0.96	1.00 (0.87-1.17)	0.03	0.86 (0.75-0.98)
KIR2DS4DEL	0.79	1.03 (0.85-1.23)	0.66	1.02 (0.94-1.11)	0.89	0.99 (0.87-1.13)	0.69	1.02 (0.91-1.15)