Investigation into the role of rare genetic variation in lung function and chronic obstructive pulmonary disease.

Thesis submitted for the degree of

Doctor of Philosophy

at the University of Leicester

by

Victoria E Jackson BA MSc

Department of Health Sciences

University of Leicester

September 2016

Abstract

Investigation into the role of rare genetic variation in lung function and chronic obstructive pulmonary disease.

Victoria E Jackson

Lung Function is a physiological measurement used for monitoring respiratory health and in the diagnosis of chronic obstructive pulmonary disease (COPD), a leading cause of morbidity and mortality worldwide. Lung function and COPD are influenced by a combination of environmental and genetic factors. This thesis aims to investigate the genetic basis of these traits, with a particular focus on the effect of low frequency and rare genetic variants, so far largely overlooked in genome-wide association studies (GWAS).

An analysis of exome array data and COPD identifies novel associations between COPD risk and low frequency single nucleotide polymorphisms (SNPs) in *MOCS3* and *IFIT3* and between a rare SNP in *SERPINA12* and percent predicted forced expiratory volume in one second (FEV_1) in COPD cases. Recently developed methods for the meta-analysis of gene-based tests are empirically evaluated and shown to be approximately equivalent to a mega-analysis using individual level data for a quantitative trait. These methods are then applied in a meta-analysis of exome array data and quantitative lung function measures. This meta-analysis identifies no gene-based associations; however genome-wide significant ($P<5x10^{-8}$) single variant associations are identified in two novel regions: a SNP near *LY86* associated with the ratio of FEV_1 to forced vital capacity (FVC) and a SNP near *FGF10* associated with FVC in ever smokers. Finally the largest GWAS to date of two lung function flow measures (peak expiratory flow [PEF] and forced expiratory flow between 25% and 75% of FVC [FEF_{25-75}]) is described. The overlap in variants associated with PEF and FEF_{25-75} and volumetric measures of lung function (FEV_1, FVC and FEV_1/FVC) is examined, and 10 SNPs are identified as showing association with PEF ($P<5x10^{-8}$), but no other lung function trait with $P<5x10^{-5}$. These findings have the potential to provide insight into the biological mechanisms underlying lung health and disease.

Acknowledgements

Firstly I would like to extend my gratitude to my supervisors Dr Louise Wain and Professor Martin Tobin for the guidance, support and patience they have provided throughout my time in the department. I am also grateful to all past and current members of the Genetic Epidemiology group for the help they have given along the way and to the wider department of Health Sciences for creating an enjoyable place to work.

I further wish to acknowledge all of the investigators who have provided data and contributed to the work in this thesis, in particular the collaborators from the UK COPD Exome Chip and SpiroMeta Consortia, along with the UK Medical Research Council who funded my studentship.

Finally, I would like to thank my family, Mum and Dad, James and Jennifer, my esteemed friend Liz and my enduring partner Paul for the encouragement and reassurance they have bestowed upon me throughout my years of academic endeavour.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

This chapter provides an introduction to genetic epidemiology, including the basic genetic concepts underlying genetic epidemiology, a description of genome-wide association studies and an introduction to the study of rare genetic variation. Key lung function measures and chronic obstructive pulmonary disease, the traits which are investigated within this thesis, are then described. Finally the aims of the thesis are set out and the structure of the remainder of the thesis is described.

Table 1-1 provides a list of fundamental terms and abbreviations, which are used throughout the thesis. These terms are highlighted in bold throughout this chapter.

Table 1-1: Key terms and abbreviations used throughout the thesis.

A. Genetics

Term	**Definition**
allele	The variant forms of a SNP or other genetic variant.
amino acid	Building blocks of proteins.
base pair	Pairs of nucleotides connected by hydrogen bond. Genetic distance may be measured in base pairs.
chromosome	Structure into which DNA is organised.
CNV	Copy Number Variation - Type of structural variation where sections of DNA are either inserted or deleted.
codon	A sequence of three nucleotides that codes a specific amino acid or stop signal during translation.
DNA	Deoxyribonucleic acid - molecule which carries genetic information
exons / exonic	Coding sections of a gene which are translated into a protein.
gene	Section of DNA, which codes for a protein or functional RNA molecule.
genotype	The two alleles (from each of the chromosomal pair) collectively.
haplotype	Groups of SNPs which tend to be inherited together
heterozygous	An individual is heterozygous at a particular locus if they have two different alleles.
homozygous	An individual is homozygous at a particular locus if they have two copies of the same allele.
indel	A variant where one, or a small number of bases are inserted or deleted.
intergenic	Located in a region of the genome with no genes.
intron / intronic	Noncoding sections of a gene which are spliced out before RNA is translated into a protein.
Inversion	Section of DNA where the order of bases is reversed.
LD	Linkage Disequillibrium - the non-random association (correlation) of nearby genetic variants.
locus	A genetic location.
missense	A SNP which results in a codon that codes for a different amino acid.
mRNA	Messenger RNA.
nonsense	A SNP which results in a premature stop codon.

Term	Definition
nonsynonymous	A SNP which results in a change to the amino acid.
nucleotide	Building block of DNA. Contain one of four types of chemical base: A, T, C and G.
proteins	Molecules consisting of several amino acids that perform many functions within living organisms.
recombination	A process occurring during meiosis in which sections of DNA are broken and recombined to produce new chromosomes.
regulatory region	Region of DNA which regulates the transcription of a gene.
RNA	Ribonucleic acid - complementary molecule to DNA, made during transcription.
SNP	Single Nucleotide Polymorphism: type of genetic variation where a single nucleotide base is substituted.
splice site	A SNP which changes the sequence at a site at which splicing takes place.
spliced / splicing	Process in which introns are removed from RNA and exons are joined together, to form mRNA.
synonymous	A SNP which does not affect the amino acid sequence.
transcription	Process in which RNA is made from a DNA molecule.
translation	Process in which proteins are produced from mRNA.

B. Genetic epidemiology and statistical genetics

Term	Definition
additive	Occurs where each allele contributes one unit of an effect to a trait for each copy of that allele, in a linear fashion.
co-dominant	Occurs where the effect of both alleles on a trait can be observed.
complex trait	A trait influenced by a combination of several genetic and environmental factors.
dominant	Interaction of alleles at a locus, where the effect of the dominant allele masks the effect of the other (recessive) allele.
dosage	Imputed genotypes value take the form of allele dosages, which relate to the expected count of the minor allele on a continuous scale from 0 to 2.
epistatic / epistasis	Interactions between alleles at different loci.
familial aggregation	The clustering of a trait or disease within families.
GWAS	Genome-wide association study - An investigation of statistical associations between alleles and a trait, for variants across the genome.
heritability	The variance of a trait which can be attributed to genetic effects
h^2	Narrow sense heritability - heritability due to additive genetic effects.
H^2	Broad sense heritability - heritability due to all genetic effects (additive, dominant and epistatic effects).
HWE	Hardy-Weinberg equillibrium - principle which assumes that allele and genotype frequencies in a population will remain constant through the generations of a population, assuming random mating and the absence of other evolutionary influences.
IBS	Identity (Identical) by state - Where two individuals share the same alleles at a locus, they are said to be IBS.
IBD	Identity (identical) by decent - Where two individuals share the same alleles at a locus, inherited from a recent common ancestor, they are said to be IBD.
imputation	Method of inferring genotypes that are not directly measured through genotyping.

Term	Definition
kinship coefficient	Measure of relatedness between individuals - the probability that two individuals are IBD at a given locus.
λ	Genomic inflation factor - the ratio of the median of the observed test statistics of a GWAS to the median of the expected test statistics under the null. In GWAS, test statistics may be scaled using λ to correct for population stratification; this is known as genomic control.
MAF	Minor allele frequency - frequency of the less common allele in a population.
monogenic	A trait that is a result of a single gene.
PCA / PCs	Principal components analysis / principal components - statistical method used to cluster sample by ancestry, which can be used to account for population structure.
polygenic	A trait that is a result of the effects of multiple genes.
recessive	An allele at a locus is recessive, where it's effect is masked by the effect of the other (dominant) allele.

C. Respiratory function and disease

Term	Definition
AAT	Alpha1-antitrypsin deficiency - a rare disorder caused by mutations in the *SERPINA1* gene which leads to early onset COPD.
airflow limitation	A reduction of FEV_1, and FEV_1/FVC (airflow limitation in COPD defined as $FEV_1/FVC<0.7$)
alveoli	Small air sacs in the lungs where gas exchange takes place.
bronchioles	Small airways of the lung, which branch off from the bronchi.
bronchi	Airways which branch off from the trachea into the lungs, subsequently branching into bronchioles.
bronchodilation	Expansion of the bronchi and bronchioles in response to a pharmacologically active substance.
COPD	Chronic Obstructive Pulmonary Disease - lung disease characterised by fixed airflow limitation.
FEF_{25-75}	The forced expiratory flow between 25% and 75% of vital capacity.
FEV_1	Forced expiratory volume in 1 second - the amount of air that can be forcibly exhaled in the first second of an FVC manoeuvre.
fibrosis	Thickening and scarring of (lung) fibrous tissue in response to injury or damage.
FEV_1/FVC	The ratio of FEV_1 to FVC.
FVC	Forced vital capacity - the total amount of air that can be forcibly exhaled.
GOLD	Global Initiative for Chronic Obstructive Lung Disease - organisation which has provided a strategy document for the diagnosis, management and prevention of COPD, which includes a grading system for severity of airflow limitation in COPD (GOLD 1-4).
lung parenchyma	The part of the lung involved in gas transfer, including the alveoli, alveolar ducts and bronchioles
PEF	Peak expiratory flow - a measure of maximum instantaneous expiratory flow
%pred (eg FEV_1)	Percent predicted (eg FEV_1) - a measure which compares an individual's measured spirometric values with that which would be expected, given their age, sex, height and ethnicity.
pneumocytes	Cells lining the alveoli.
spirometry	A physiological test, measuring lung function.

Term	Definition
trachea	The largest airway in the lower respiratory tract that connects the pharynx/larynx with the lungs.

1.1 Introduction to genetic epidemiology

1.1.1 Genetic concepts

The human genome is made up of deoxyribonucleic acid (**DNA**), which in turn consists of **nucleotides**, each containing one of four types of chemical base, namely adenine (denoted A), cytosine (C), guanine (G) or thymine (T), joined with a sugar (deoxyribose) and phosphate group. Covalent bonds join the nucleotide bases together to form strands of DNA. Strands of DNA have directionality, with the two different ends denoted 5' and 3'. Two strands running in opposite directions and connected by hydrogen bonds form the double helix structure of a DNA molecule. In this double stranded structure, the A nucleotides consistently pair with T nucleotides, with the C and G nucleotides pairing equivalently, to form base pairs. The human genome consists of approximately 3.3 billion **base pairs** and is arranged into 46 **chromosomes**, consisting of 22 homologous pairs of autosomes (numbered 1-22) and two sex chromosomes (XX in females and XY in males). For each chromosomal pair, one is inherited from the individual's mother, and one from the father (1, 2).

DNA contains biological information which instructs the synthesis of **proteins**. In a process known as **transcription**, the two strands of DNA break apart, with one of the strands forming a template for a complementary second molecule called ribonucleic acid (**RNA**). RNA is similar to DNA, but it contains the sugar ribose instead of deoxyribose and an alternative base, Uracil (U) is in place of T. Similarly to DNA, an RNA strand contains a 5' and 3' end, and is aligned in the opposite direction to the DNA strand in transcription. Certain sections of the DNA sequence are known as **genes**; most genes contain **regulatory regions**, non-coding regions known as **introns** and coding regions known as **exons**. After transcription, the sections of RNA from the exons are **spliced** together to form messenger RNA (**mRNA**). **Translation** of the mRNA molecule then occurs in which groups of three bases, known as **codons**, are read from

the 5’ to the 3’ end of the mRNA to form a chain of **amino acids** and in turn produce proteins (1, 2).

The majority of the genome is identical across all humans; however there are several ways in which DNA can vary between individuals. The variation present in the human genome can broadly be categorised into two classes: structural variants and single nucleotide polymorphisms (**SNPs**). SNPs are the most common type of variation, and are where a single nucleotide is substituted for a different nucleotide at a particular position, or **locus** (3). SNPs may be located in regions of the genome with no genes (**intergenic**), or they may be located within genes. SNPs which are located within the coding regions of genes (exonic) may be either **nonsynonymous** or **synonymous**. A nonsynonymous SNP results in a codon which codes for a different amino acid within the translated protein (**missense**). Sometimes this alternative codon is a stop codon, which results in truncation of the protein product (**nonsense**). If an exonic SNP does not affect the amino acid sequence, it is known as synonymous. SNPs may also be located at the site where splicing occurs during the formation of mRNA (**splice site**), or in the non-coding regions of genes (**intronic**).

Structural variants include insertion-deletions (**indels**), **inversions** and copy number variants (**CNVs**). Indels and CNVs are where sections of bases are either inserted or deleted. Indels tend to consist of a small number of bases, whereas CNVs are longer sections of DNA, which may also contain repeats of bases. Inversions occur where the order of a section of bases in a chromosome is reversed (3).These types of structural variation and SNPs are shown schematically in Figure 1-1.

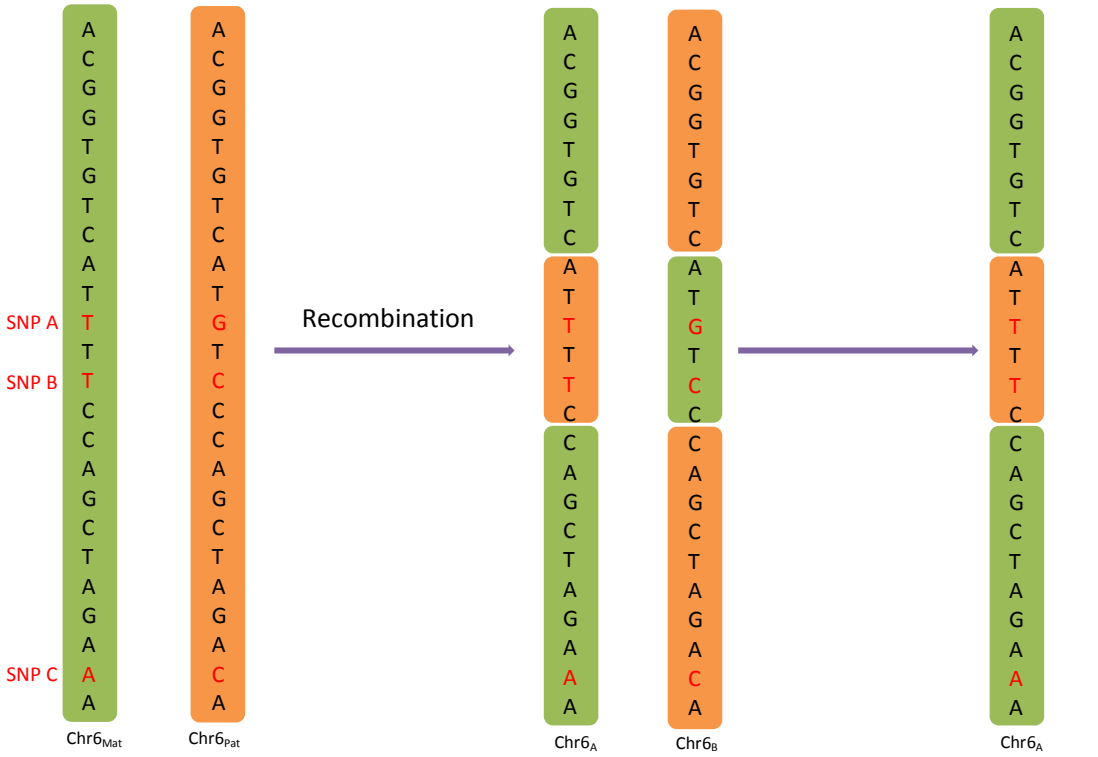
Figure 1-1: Classes of Genetic Variation.
For each type of variation, the bases of one strand of DNA are shown from a particular locus on two copies of a chromosome.

Where there are these differences in a particular position in the genome, the variants at that locus are known as **alleles**, and the two alleles an individual has (one from each of the chromosomal pair) are collectively known as the **genotype**. Using the SNP illustrated in Figure 1-1 as an example, the alleles are A and G, and an individual may have one of three genotypes: AA, AG or GG. If an individual has the same allele on both chromosomes (AA or GG), their genotype is termed **homozygous**, whereas if the two alleles are different (AG), their genotype is **heterozygous**. Given that the two strands of a DNA molecule are complementary, the same information may be obtained from either strand; therefore only the alleles and genotypes from one strand are ever reported. If the alleles from the preceding example were taken from the other strand, the resulting genotypes would be TT, TC and CC.

Chromosomes are passed on to offspring from their parents via gametes (sperm from the father and ovum from the mother). The gametes are formed by a cell division process known as meiosis in which cells are created with only one member of each chromosomal pair: 22 autosomes and one sex chromosome (X in an ovum cell and either an X or Y in a sperm cell). During meiosis, a process called **recombination** takes place where the two parental chromosomes overlap and sections of DNA are

exchanged. Consequently chromosomes are not transmitted to the gametes as a whole, rather a mixture of the two homologous chromosomes is passed on, as illustrated in Figure 1-2 (1, 2).

Figure 1-2: Illustration of recombination, during meiosis.



Genes and genetic variants that are located close to each other on a chromosome are less likely to have undergone recombination than those located farther away. As a consequence, nearby genes and variants are more likely to be inherited together and are therefore found to be correlated within populations. For example, in Figure 1-2, SNP A and SNP B are more likely to be inherited together than are SNP A and SNP C. This non-random association is known as linkage disequilibrium (**LD**) and the stretch of genes or variants which are inherited together are known as a **haplotype** (1, 4).

The majority of the work in this thesis focusses on SNPs; the terms SNP and (genetic) variant have therefore been used interchangeably for the remainder of this thesis.

1.1.2 Overview of Genetic Epidemiology

Genetic epidemiology is the study of how genetic factors affect health and disease in populations. The uncovering of the genetics of a disease or a trait, can lead to a greater understanding of the mechanisms of disease and may highlight molecular targets for novel therapeutics.

The initial steps in determining whether a trait or disease is influenced by genetic determinants do not require measurement of any genetic information. Firstly, we might want to show there is **familial aggregation** of a disease or quantitative trait. For a binary disease trait, this is where there is on average, a greater prevalence of disease amongst the relatives of individuals with the disease, compared to amongst relatives of individuals who are disease free. For a quantitative trait, familial aggregation may be assessed using measures such as the intra-family correlation coefficient, which measures the proportion of the total trait variance that is due to variation between families. Familial aggregation is usually a result of a combination of genetic and shared environmental factors (1, 5).

The genetic contribution to the variability of a trait or disease is known as the **heritability** and is defined as a ratio of the variance of a trait that can be attributed to genetic effects, to the total trait variance. There are two types of heritability: firstly narrow sense heritability ($\mathbf{h^2}$), which considers only additive genetic effects; secondly broad sense heritability ($\mathbf{H^2}$), which comprises all genetic effects, including interactions within loci (**dominant** effects), and between loci (**epistatic** effects). Estimates of heritability are made by partitioning known variation into components of unmeasured genetic and environmental effects. This estimation is straightforward for quantitative traits; for binary disease traits, heritability is usually estimated using a hypothetically assumed underlying normally distributed liability trait, which determines the probability of an individual developing the disease (1, 5-7). Heritability is an important measure to assess the level of genetic contribution to a trait or disease, however it has a number of limitations: heritability estimates are population specific, can change over time, and are not informative about the actions and interactions of specific genes (7).

Once a genetic contribution to a trait has been established, segregation analysis has historically been used to determine the mode of inheritance. A genetic trait may be consistent with a dominant, **recessive**, **co-dominant** or **additive** genetic model and be a result of a single locus (**monogenic**), a combination of a small number of genes (oligogenic), or a result of a large number of genes, each with small effects (**polygenic**) (8). There are a number of limitations to segregation analysis methods however, and with the developments in technologies for measuring genetic variation, they are no longer widely utilised (1).

The first studies which attempted to implicate specific regions of the genome with a trait were based on genetic linkage. These family-based genetic linkage studies are reliant on the tendency for short regions of the genome to be transmitted from parent to offspring as a whole, without being subject to recombination. If a genetic marker is passed down through families of affected individuals, it follows that there might be a disease gene close to that marker (1, 8).These studies involve genome-wide scans of sparsely distributed markers spaced several Centimorgans (cM, a measure of genetic distance based on recombination frequency) apart. These analyses have been most successful for identifying genetic causes of monogenic Mendelian disease, such as cystic fibrosis and Huntington disease (9). For most polygenic **complex traits**, influenced by several genetic and environmental factors, linkage studies have had limited success.

A previously popular approach for identifying genes associated with complex traits was through candidate gene studies. These studies are usually population-based and examine variants in specific gene regions chosen due to a priori hypotheses about their role in the trait of interest. Results from candidate gene studies have seldom been replicated in follow-up studies however and are reliant on our ability to predict biologically plausible candidate genes (10). Due to these limitations, hypothesis-free genome-wide association studies (**GWAS**) have been most widely used in recent years and have proved a powerful method for identifying genes associated with complex traits. These GWAS are fully described in the following section.

1.1.3 Genome-wide Association Studies (GWAS)

Over the past decade, GWAS have proved a powerful method for identifying associations between common SNPs and a number of complex traits. GWAS are usually large, population based studies which involve testing for associations between a trait and multiple individual SNPs in turn, from across the genome.

The development of GWAS was made possible due to efforts such as the International HapMap Project (11), which provided knowledge of LD structures, alongside the development of genotyping technologies, allowing for the measurement of several hundreds of thousands of SNPs, genome-wide, in large numbers of individuals. Due to the correlation (LD) patterns existing across the genome, the SNPs measured by these arrays are able to capture a large proportion of common variation genome-wide (12). Any association between genotyped SNPs and a trait may then be a result of a direct association (where the associated SNP is the causal variant), or an indirect association (where the associated SNP is in LD with [or tagging] the causal variant) (1, 12). The remainder of this section outlines how GWAS are carried out, including quality control (QC) of genotype data, methods for association testing and interpretation of results.

1.1.3.1 Quality control of GWAS genotype data

SNP data from genotyping arrays are in the form of allele probe intensities from which genotypes may be estimated by genotype calling algorithms, usually implemented in software accompanying the genotyping platform (e.g. Gencall by Illumina (13)). Once genotypes have been estimated, there are several quality control (QC) metrics which are routinely implemented, prior to any analysis (14, 15). These QC metrics are undertaken on both a per-sample and per-SNP basis and are detailed below.

Per-sample QC

i. **High level of missing data:** Samples with a high number of missing genotypes are likely a result of poor DNA quality and should be excluded.

ii. **Outlying heterozygosity rate:** The mean level of autosomal heterozygosity across all individuals should be determined and any sample with an outlying mean heterozygosity rate should be identified for exclusion. Samples with an excessive heterozygosity rate may be subject to DNA sample contamination, whilst samples with a lower than expected heterozygosity rate may be due to inbreeding.

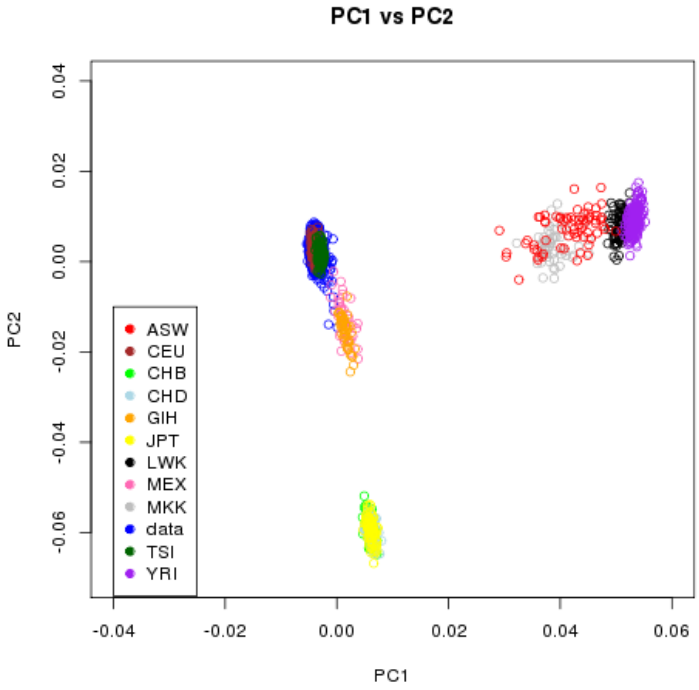
iii. **Discordant sex information:** Samples whose genetically inferred sex is inconsistent with that supplied within the phenotype data may be subject to sample mix-ups, or DNA sample contamination.

iv. **Duplicate and related samples:** Metrics known as identity by state (**IBS**) and identity by decent (**IBD**) can be calculated for each pair of samples. IBS is an estimation of the average proportion of alleles shared by each sample pair, across genotyped SNPs. IBD is a measure of recent shared ancestry, and can be estimated with IBS data. Sample pairs with IBD=1 are likely duplicates (or monozygotic twins). IBD values of 0.5 and 0.25 correspond with first and second degree relatives. In studies with related individuals, one sample of each pair of related individuals may be removed, to leave only unrelated samples; removal of related samples is usually carried out where there are a small number of samples who are related, so there is only small decrease in overall sample size and therefore power. The benefit of restricting to only unrelated individuals is that association analyses are computationally more straightforward. Alternatively, related samples may be retained and their relatedness taken into account during analyses (Section 1.1.3.2).

v. **Ancestral outliers:** Principal components analysis (**PCA**) is a statistical method which calculates a number of uncorrelated variables (principal components, **PCs**) each accounting for variability in the data. These PCs can be calculated jointly with samples of known ancestry (e.g. from HapMap (11) or 1000 Genomes projects (16)); the first two resulting PCs are sufficient to cluster individuals from different ancestral populations (Figure 1-3) and any sample with an ancestry different to

the population being studied may be identified for exclusion. These calculated PCs may also be used in the analysis, to account for smaller scale population structure (Section 1.1.3.2).

Figure 1-3: Example of PCA plot.
First two principal components for samples from the 1958 British Birth Cohort (samples labelled data), and HapMap samples of different ancestries. HapMap ancestry codes: African populations: ASW; LWK; MKK; YRI. East Asian Ancestry: CHB; CHD; JPT. South Asian Ancestry: GIH. Admixed American Ancestry: MEX. European Ancestry: CEU; TSI.



Per-SNP QC

i. **High level of missing data:** SNPs with a low call rate (non-missing genotypes) are excluded.

ii. **SNPs deviating from Hardy Weinberg equilibrium:** Hardy-Weinberg equilibrium (**HWE**) assumes that under random mating and no evolutionary influences, the relationship between genotype and allele frequencies in a population should remain stable. SNPs which show significant deviation from HWE may be a result of genotyping, or genotype calling errors; these SNPs may be removed or flagged as potentially problematic.

iii. **SNPs with a low minor allele frequency:** In GWAS of common SNPs, it has been common practice to exclude variants under a given minor allele frequency (**MAF**) threshold, as these SNPs are more difficult to call using genotype-calling

algorithms. More recently, there has been a greater focus on low frequency and rare SNPs (Section 1.2) and as such this filter is not always applied.

1.1.3.2 Analysis of GWAS data

GWAS analyses generally consist of testing for a statistical association between a trait of interest and all genotyped SNPs. GWAS may be undertaken assuming an additive, dominant or recessive genetic model; however the additive model is most commonly assumed in the analyses of polygenic complex traits, and is described in this section.

For a SNP with two alleles C and T, individuals may have one of three genotypes: CC; CT; or TT. If we assign T as the effect allele, then an individual with a CC genotype would have no copies of the effect allele, an individual with a CT genotype would have one copy of the effect allele and an individual with a TT genotype would have two copies of the effect allele. Associations may then be tested between the number of copies of the effect allele an individual has (0, 1 or 2) and the trait. For a quantitative trait, this genetic association may be tested using a linear model, and for a binary trait a logistic model may be utilised, as follows:

Quantitative Trait:

$$\boldsymbol{Y} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 G} + \boldsymbol{\beta_2 X} + \varepsilon \quad \textbf{(1-1)}$$
$$where\ \varepsilon \sim N(0, \sigma_\varepsilon^2)$$

Binary Trait:

$$logit(\boldsymbol{Y}) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 G} + \boldsymbol{\beta_2 X} \quad \textbf{(1-2)}$$

where **Y** is an vector of trait values and **G** is an matrix of the genotypes {0,1,2}. Additional covariates (denoted by the matrix **X**) may additionally be included in these models, including non-genetic risk factors (e.g. sex or smoking), and PCs (Section 1.1.3.1) to adjust for fine-scale genetic differences due to population structure (population stratification (17)). ε is a random error term, accounting for the residual variation of **Y.** In both models, $\boldsymbol{\beta_0}$ is a vector of intercept terms and $\boldsymbol{\beta_2}$ is a vector of covariate effects. For a quantitative trait, $\boldsymbol{\beta_1}$ provides a vector of effect estimates for the effect of each copy of the effect allele on the trait value, for each SNP. For a binary

trait, $\boldsymbol{\beta_1}$ provides a vector of log odds ratio, again for each copy of the effect allele. In each case, the null hypothesis being tested is that the genotype has no effect on the trait, that is each element of $\boldsymbol{\beta_1}$ is equal to 0 (H_0: β_1=0).

In the analyses undertaken in this thesis, I have utilised both linear models (equation (1-1)) and logistic models (equation (1-2)) described above, as implemented in the software packages PLINK (18) and SNPTEST (19).

GWAS may also be undertaken in samples which contain related individuals; for these analyses, linear mixed models are the most widely used approach. This method firstly involves the calculation of **kinship coefficients** between all pairs of individuals, using genome-wide genotype data. Secondly, a linear mixed model is fitted, which includes the SNP as a fixed effect, along with a random effect which incorporates the kinship estimates and models the genetic correlation between individuals, as in equation (1-3) (20).

$$\boldsymbol{y} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 G} + \boldsymbol{\beta_2 X} + \boldsymbol{U} + \varepsilon \qquad \textbf{(1-3)}$$
$$\boldsymbol{where\ U} \sim MVN\left(0, \sigma_g^2 \boldsymbol{K}\right)$$
$$\varepsilon \sim N(0, \sigma_\varepsilon^2)$$

where **U** is a vector of random effects where σ_g^2 is the component of the overall variance of **Y,** which is due genetic factors and **K** is a matrix of kinship coefficients.

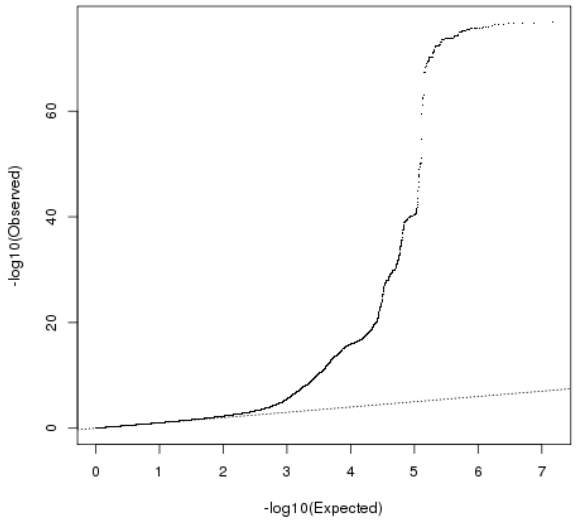
Once genetic associations have been tested for all SNPs, plots and summaries may be produced to examine the results. Firstly, the observed $-\log_{10}$ P-values of the GWAS may be plotted against those expected under the null hypothesis, in a quantile-quantile (**QQ**) plot (Figure 1-4). This plot shows whether there were more significant associations than would be expected by chance. If the distribution of P-values broadly follows the expected distribution under the null, but with some deviation at the highly significant end of the distribution, this is suggestive of there being significant trait-associated SNPs. Alternatively, if there is inflation of the P-values across the distribution, this may be indicative of underlying population stratification, that has not been properly accounted for during the analysis. Another related indicator of population stratification is the genomic inflation factor (**λ**) (21). λ is defined in equation

(1-4) as the ratio of the median of the observed test statistics $\left(X_1^2, \ldots, X_j^2\right)$ to the median of the expected test statistics under the null.

$$\lambda = median\left(X_1^2, \ldots, X_j^2\right)/0.456 \qquad \textbf{(1-4)}$$

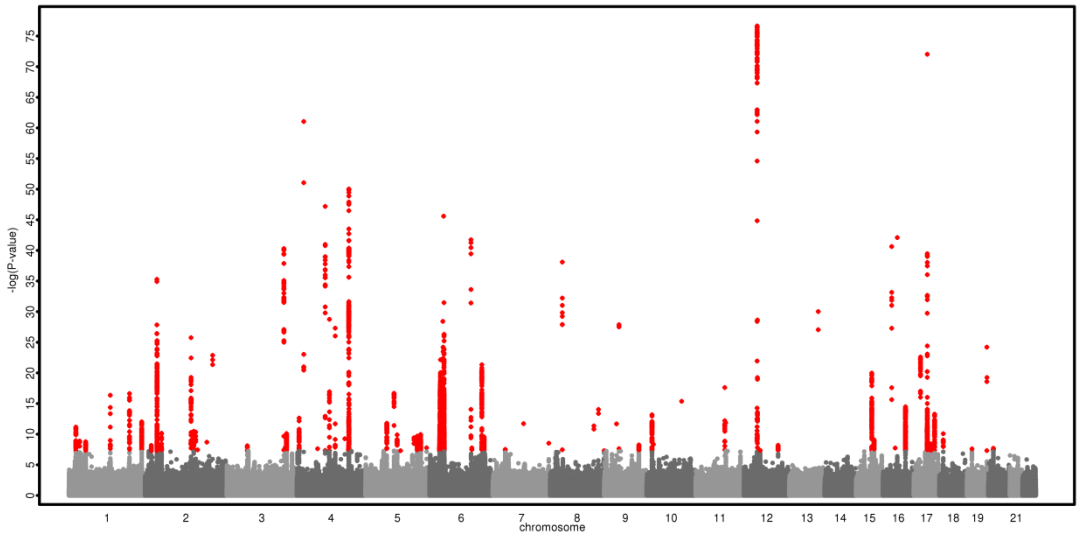
where 0.456 is the median of a chi-squared distribution with one degree of freedom. A value of λ that is well in excess of 1 may be an indication over-inflation of test statistics due to population stratification. Where λ>1, test statistics may be scaled using the value of λ to give corrected P-values; this method is known as genomic control.

Figure 1-4: Example of quantile-quantile (QQ Plot).



Manhattan plots are usually produced to summarise GWAS results (Figure 1-5). These plots show the $-\log_{10}$ P-values (y-axis) for each SNP, ordered by chromosome and position (x-axis). The higher a SNP's position on the y-axis, the more highly significant the association for that SNP. Due to the LD structure of SNPs usually included in GWAS, there is a tendency of "peaks" of association to form.

Figure 1-5: Example of Manhattan plot.



1.1.3.3 Meta-analysis of GWAS

In order to increase sample size, and therefore statistical power to detect SNP associations, meta-analyses combining GWAS data from several studies are commonly undertaken. In these meta-analyses, it is not necessary to share individual level data; rather each contributing study performs association analyses according to a centrally agreed analysis plan. The study-level results are then combined by a central meta-analyst to give overall genome-wide association results. Combining data in this way has been shown to be as powerful as conducting analyses using individual level data (22).

1.1.3.4 Statistical significance and replication

In a GWAS there are a large number of hypotheses being tested; the significance threshold at which we reject the null hypothesis must reflect this. A Bonferroni correction, in which the desired type 1 error rate (α=0.05) is divided by the number of tests undertaken, may be used to correct for multiple testing. However, the Bonferroni correction assumes that all tests are independent and this is not the case in GWAS due to the correlation of the SNPs, and therefore the Bonferroni correction is likely to be overly conservative. It is possible to estimate the number of effective independent tests, for example using the SNPSpD package (23); however it has been common practice to utilise a significance level of $P<5x10^{-8}$. This widely used significance threshold is equivalent to a Bonferroni correction for 1 million independent tests and

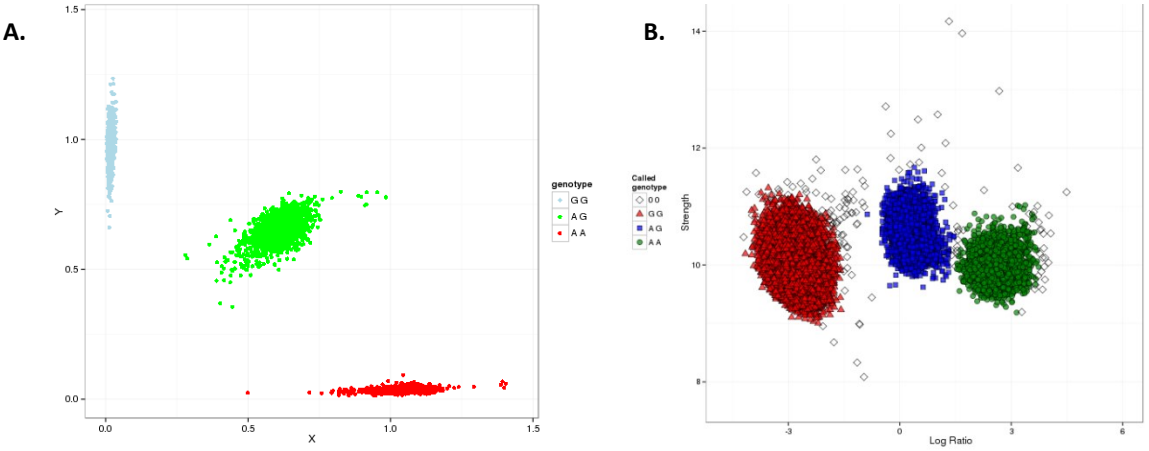
SNPs meeting this significance level are deemed genome-wide significant (24). Other methods for determining significance levels have also been proposed, such as permutation-based (25, 26), or Bayesian approaches (27), however these have been less widely adopted.

In order to verify that SNPs identified in a GWAS reflect true associations, replication should be sought. A common approach is to select SNPs from a GWAS under a given P-value threshold (usually less stringent than the genome-wide significance level), which are then taken forward to a follow-up replication stage of analyses undertaken in an independent set of samples. The results from the first discovery-stage GWAS and the follow-up stage may then be combined to give an overall result; SNPs meeting $P<5x10^{-8}$ overall are then declared as genome-wide significant results (28).

A quality control check that should be undertaken to eliminate false-positive associations is the examination of cluster plots of probe intensities from the genotyping experiment. Either the normalised probe intensities for each allele can be plotted (Figure 1-6 A), or the log ratio may be plotted against the strength (Figure 1-6 B), where log ratio is defined as $\log_2\left(\frac{int_X}{int_Y}\right)$ and the strength as $\frac{\log_2(int_X)+\log_2(int_Y)}{2}$, where int_X and int_Y are the two probe intensities. Plotted on either scale, three distinct clusters should form, corresponding to the three possible genotypes (homozygous for the major allele, heterozygous, and homozygous for the minor allele). If these three clusters are not well defined, then there may be inaccuracies with the called genotypes, which could lead to spurious associations.

Figure 1-6: Examples of cluster plots of probe intensity data.

A. Normalised probe intensities. B. log ratio versus strength of probe intensities.



1.2 "Missing-heritability" and rare variants

1.2.1 Missing Heritability

GWAS to date have generally focussed on investigating the effects of common SNPs (MAF ≥ 5%), and whilst they have had some success in discovering genetic variants that influence complex diseases and traits, most genome-wide significant associated variants have only shown modest effects, and collectively only explain small amounts of the expected heritability. Some recent examples include body mass index (BMI), for which 97 genome-wide significant loci were found to explain approximately 2.7% of phenotypic variance (29) (h^2 estimates between 47-90% (30)) and schizophrenia for which 108 genome-wide significant loci explained 3.4% of variation on the liability scale (31) (h^2 estimates 44-87% (32)). However, these studies and others (29, 31, 33-35) have estimated that collectively, common SNPs do explain a substantial proportion of the heritability estimated from family studies.

Nevertheless, there still remains a proportion of the heritability that is unexplained and several suggestions have been offered for where this so-called "missing heritability" might be found. These include structural variations; gene-environment interactions; epigenetics; epistasis; transgenerational effects and rare variants with large effects (36). The work in this thesis investigates the last of these hypotheses. There are a number of arguments in support of the rare variant hypothesis: firstly, evolutionary theory predicts that disease causing variants are likely to be rare, since

variants which are deleterious should be selected against (37). Furthermore, most nonsynonymous mutations, which lead to changes in the protein structure and therefore are more likely to be deleterious, are strongly skewed to the lower end of the allele frequency spectrum (38). There have also been a number of highly penetrant, rare CNVs identified that confer a substantial risk in common neuropsychiatric disorders, including epilepsy, schizophrenia and autism (37, 39).

This remainder of this section describes the types of studies used to investigate the effect of rare SNPs on a phenotype, and outlines some of the methods for association testing.

1.2.2 Sequencing Studies

Sequencing is the process of determining the full sequence of nucleotides in DNA and is the most useful tool for identifying very rare genetic variation. Sequencing may be carried out in specific regions (targeted sequencing), in exonic regions (whole exome sequencing), or across the entire genome (whole genome sequencing). Although the cost of sequencing has fallen dramatically in recent years, it remains prohibitively expensive for studying genome-wide genetic variation in large numbers of individuals (39).

1.2.3 Imputation

Genotype **imputation** is the prediction of genotypes that are not directly measured in a sample of individuals. Given the tendency for stretches of DNA, or haplotypes, to be shared amongst individuals, by measuring genotypes at a selection of SNPs, it is possible to infer the unobserved genotypes at other SNPs within that region. Several reference panels containing haplotypes from a number of individuals are available, which genotyped samples may be imputed against. More recent reference panels utilising haplotypes from an increasing number of samples (recent efforts include the 1000 Genomes (16) and UK10K (40) projects and the Haplotype Reference Consortium (41)), are allowing for imputation of SNPs at the lower end of the allele frequency spectrum, and analyses using data imputed to these panels are beginning to identify

associations with low frequency SNPs and a number of common complex traits (42-45).

Several software packages have been developed to facilitate imputation, with the most widely used being IMPUTE2 (46), MaCH/minimac (47, 48) and Beagle (49). Imputed genotypes are in the form of allele **dosages**, on a continuous scale from 0 to 2. These dosages relate to the expected count of the minor allele, with the non-integer nature of the value reflecting the genotype uncertainty. Imputation packages also provide metrics for each SNP regarding the quality of the imputation; SNPs with low imputation qualities are usually excluded from any analyses.

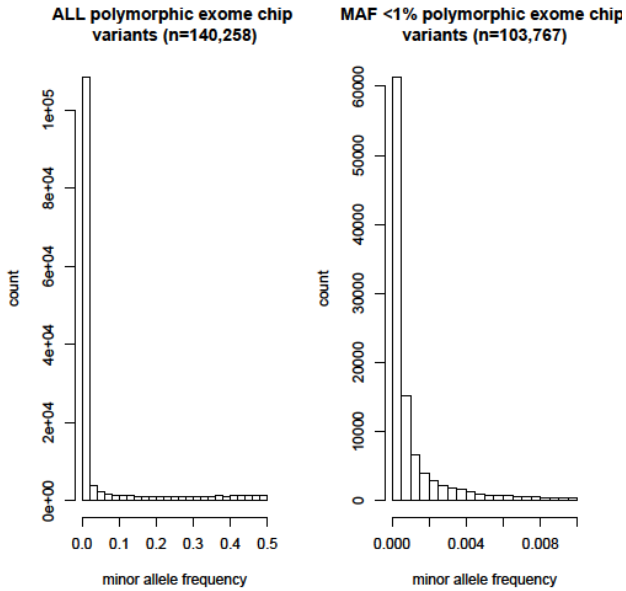
1.2.4 The exome chip

Whilst genotyping arrays (or chips) have widely been used as a cost-effective method for examining common genetic variation, low frequency and rare SNPs have been largely underexplored in GWAS using array data. To date, the influence of rare genetic variation in disease has mainly been examined through sequencing studies; however as mentioned in Section 1.2.2, the cost of this technology remains high. The exome chip is an array that was designed to act as an intermediary between existing genotyping array and sequencing technologies, allowing for the investigation of rare genetic variation in large sample sizes.

The variants included on the chip were selected as they were observed numerous times in the sequenced exomes or genomes of a set 12,000 individuals taken from 16 sample collections, enriched for a range of traits. The chip focusses on SNPs within the exons, the part of the gene which when transcribed to RNA, codes for proteins. Most variants on the chip are either missense, nonsense or splice site variants and are likely to affect protein structure and function. Nonsynonymous variants were included if they were observed at least three times, in two or more sample collections, with less stringent inclusion criteria for splice and nonsense variants. Additional content on the chip includes previously described GWAS hits, ancestry informative markers, IBD markers, fingerprint SNPs, variants from the mitochondria, the major histocompatibility complex genes and chromosome Y, and a random sample of synonymous SNPs. The exome chip is estimated to include 97-98% of missense

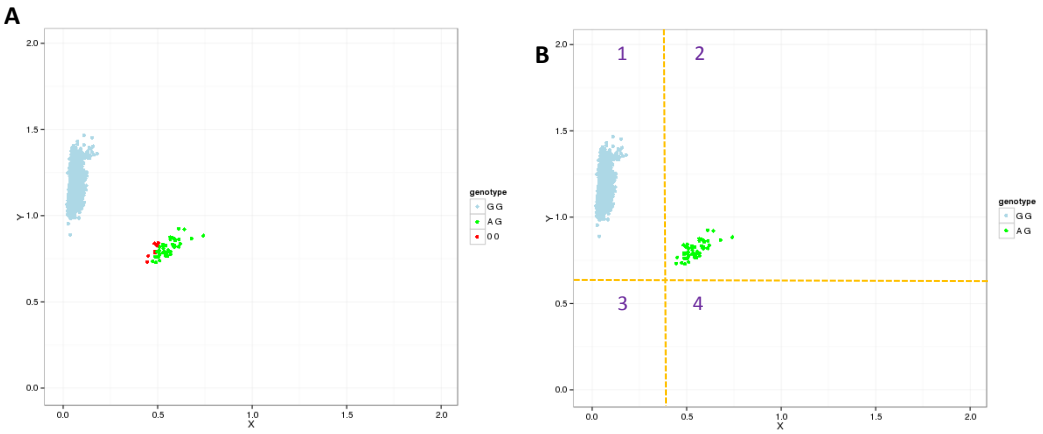
variants and 94-95% splice and nonsense variants detected in an average sequenced genome. The array includes variants with all ranges of MAFs, but the majority are low frequency (1-5% MAF) and rare (>1% MAF) (Figure 1-7) (50).

Figure 1-7: Distribution of allele frequencies of SNPs genotyped by the exome chip and polymorphic in the 1958 British Birth Cohort.



One of the challenges for the exome array was that the genotype calling algorithms previously used for calling array genotype data were less accurate for SNPs with a low MAF. Consequentially, new algorithms were developed which were intended to either refine genotype calls from existing algorithms, (zCall (51)), or to more accurately call genotypes for SNPs from across the allele frequency spectrum (iCall (52)). The CHARGE Consortium have also developed a "Best Practices and Joint Calling Protocol" (53) and made available a cluster file to be used for calling Illumina exome array data in Gencall. In the analyses described in Chapters 2 and 4 of this thesis, I have utilised zCall for refining genotype calls; Figure 1-8 shows a schematic of how this algorithm assigns genotypes to individuals called as missing by the Gencall calling algorithm, for a particular SNP.

Figure 1-8: Comparison of genotype calling by A. Gencall and B. zCall.
A. Shows a SNP with genotypes based on the Gencall calling algorithm. B. Shows the same SNP with genotypes refined by the zCall algorithm: zCall separates the plot of normalised intensities into four quadrants based on the means and standard errors of the X and Y intensities across common SNPs on the array. Samples whose genotypes were called as missing by Gencall (coloured red in A) were then reassigned as follows: samples in quadrant 1 as homozygote GG; samples in quadrant 2 as heterozygote AG, samples in quadrant 3 as missing and samples in quadrant 4 as homozygote AA. Samples which were assigned a genotyped by Gencall were not reassigned by zCall.



The exome array has been used to investigate the effect of low frequency and rare variation and a number of traits, with some success. Exome array analyses have identified low frequency SNPs associated with asthma (54), diabetes and related traits (55-57), lipids (58) and haematological traits (59). Other exome array studies have identified only common variants associated with disease, for example with glaucoma (60) and psoriasis (61).

More recently, other related arrays have been developed which include the original exome array content, alongside a grid of SNPs which allow for genome-wide imputation, to provide even greater coverage. Two such arrays are the UK BiLEVE array and UK Biobank array, which are currently being utilised to generate genome-wide genotype data from half a million individuals from UK Biobank.

1.2.5 Methods for testing associations with rare variants

Single variant association analyses (methods described in Section 1.1.3.2), where the effect of each SNP is tested in turn, with the trait of interest, have proved successful in identifying common SNPs which influence complex disease; for rare variants however, they are somewhat underpowered (62). A number of collapsing methods have been developed, which combine information from several variants within a specified genomic region, for example a gene, into a single quantity which is then used for

association testing with the trait (63). The first category of methods is the burden test, which attempts to assess the overall "genetic burden", attributable to rare variants. The Cohort Allelic Sum Test (CAST) tests for association between genes and binary traits by collapsing the counts of alternate alleles for a set of variants below a given MAF threshold, into single quantities in cases and controls, and then comparing these quantities (64). The Combined Multivariate & Collapsing Test (CMC) utilises a similar collapsing method to sets of variants, and performs a multivariate test to determine whether any of the sets of variants show association with a binary disease trait (62). These approaches involve selecting variants below a pre-specified allele frequency threshold; a variable threshold test (VTT) applies the collapsing methods under a series of MAF thresholds, and selects the threshold which has the greatest statistical power (65). The VTT can be applied to a binary or quantitative trait, and additionally allows the incorporation weights, for example based of predictions of functional effect. The Weighted Sum Test (WST) is a further collapsing method which weights variants, such that the effects of rare variants are accentuated, and allows for testing association with both binary and quantitative traits (66).

The main disadvantage of burden tests is that they fail to account for the magnitude and direction of effect of each variant so will be low powered when variants within the region act in opposing directions. A more flexible approach, which does not make assumptions about the direction of effect of variants, is through testing genetic similarities using nonparametric kernel functions and variance component models. The Kernel Based Association Test (KBAT) uses a range of kernel functions in an analysis of variance (ANOVA) formulation, for use with binary disease traits (67). The Sequence Kernel Association test (SKAT) is a more flexible regression approach which uses a weighted linear kernel, and allows for covariate adjustment and both binary and quantitative traits (68). These tests are powerful where a region has a combination of protective, deleterious and neutral variants, but where the majority of rare variants in a region influence the trait in the same direction, they tend to be outperformed by burden tests. A further method, SKAT-O attempts to unify the WST method and SKAT, by providing a weighted average of the two methods, optimised for each region (69).

The methods mentioned above are summarised in Table 1-2; this is not an exhaustive list of all methods developed for gene-based association analyses however.

Table 1-2: Summary of statistical methods for testing gene-based associations.

Test	Category	Traits	Description
CAST	Burden test	Binary	Collapsing method, using univariate tests of association.
CMC	Burden test	Binary	Collapsing method, using multivariate tests.
VTT	Burden test	Binary & Quantitative	Performs collapsing methods under various MAF thresholds and selects MAF threshold to give optimal P-value. Also allows for weighting of variants.
WST	Burden test	Binary and Quantitative	Collapsing method which incorporates weights, such that rarer variants have greater weights.
KBAT	Test of genetic similarity	Binary	Compares variation in cases and controls using kernel functions.
SKAT	Test of genetic similarity	Binary and Quantitative	Compares variation using a weighted kernel function, and allows for covariate adjustment.
SKAT-O	Optimising approach	Binary and Quantitative	Weighted combination of WST and SKAT tests.

In practice, the underlying genetic architecture of a trait is generally unknown, so it is beneficial to utilise both burden tests and methods which compare genetic similarity, or alternatively to use an optimising approach (SKAT-O). In the remainder of this thesis, the WST, VTT, SKAT and SKAT-O methods are considered, and are more formally defined in equations (1-5) to (1-10).

The WST, VTT, SKAT and SKAT-O tests, may all be constructed using the score statistic. The score statistic (U) for the jth variant in a study with n individuals is:

$$U_j = \sum_{i=1}^{n} \frac{g_{ij}(y_i - \hat{\mu}_i)}{\phi} \qquad \textbf{(1-5)}$$

where g_{ij} is the genotype (g_{ij}={0,1,2}) of j^{th} variant of the i^{th} individual, y_i is the phenotype of the i^{th} individual and $\hat{\mu}_i$is the predicted y_i under the null. For a continuous trait, a linear model is utilised with $\hat{\mu}_i = \beta_0 + \boldsymbol{\beta_1 X_i}$, where $\boldsymbol{X_i}$is a vector of covariates with effects $\boldsymbol{\beta_1}$, and intercept β_0, and $\phi = \sigma_\varepsilon^2$. For a binary trait, a logistic model is used, where $logit(\hat{\mu}_i) = \beta_0 + \boldsymbol{\beta_1 X_i}$ and $\phi = 1$.

For a gene with m variants, the variance-covariance matrix $\boldsymbol{V}$ of the score statistics $\mathbf{U} = [U_1, U_2, \dots, U_m]^{\mathrm{T}}$ may be estimated as follows:

$$\mathbf{V} = \boldsymbol{G}(\boldsymbol{\Omega}^{-1} - \boldsymbol{\Omega}^{-1}\boldsymbol{X}(\boldsymbol{X}^{T}\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{T}\boldsymbol{\Omega}^{-1})\boldsymbol{G}^{T} \quad \textbf{(1-6)}$$

Where $\boldsymbol{G}$ is a matrix of genotpyes and $\boldsymbol{X} = [\mathbf{X_1}, \mathbf{X_2}, \dots, \mathbf{X_n}]$ is a matrix of covariates. For a continuous trait, $\boldsymbol{\Omega} = \sigma_{\varepsilon}^{2}\boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix, and for a binary trait $\boldsymbol{\Omega} = diag[\hat{\mu}_1(1-\hat{\mu}_1), \dots, \hat{\mu}_n(1-\hat{\mu}_n)]$, where $\hat{\mu}_i$ is the probability that the i^{th} individual is a case, under the null. The jj^{th} element of $\boldsymbol{V}$ is the variance of the score statistic for variant *j*:

$$V_{jj} = Var(U_j)$$

For a gene with m variants, each gene-based test may be constructed as follows.

Weighted Sum Test (WST):

$$Q_{WST} = \sum_{j=1}^{m} w_j U_j \sim N\left(0, \sum_{j=1}^{m} w_j^2 V_{jj}\right) \quad \textbf{(1-7)}$$

where w_j is a weight for the j^{th} variant. The WST test statistic Q_{WST} asymptotically follows a normal distribution.

Variable Threshold Test (VTT):

$$Q_T(F) = \sum_{j=1}^{m} T_F U_j \quad \textbf{(1-8)}$$

$$Q_{VTT} = \max\big(Q_T(F)\big) \sim N\left(0, \sum_{j=1}^{m} T_F V_{jj}\right)$$

Where T_F is an indicator variable that equals 1 if the MAF of j is less than F, or 0 otherwise. Q_T is calculated using a series of MAF thresholds (F), and the Q_{VTT} test statistic is then defined as the maximum Q_T over values of F.

SKAT:

$$Q_{SKAT} = \sum_{j=1}^{m} w_j^2 U_j^2 \sim \sum_{j=1}^{m} \lambda_j \chi_{1j}^2 \quad \textbf{(1-9)}$$

The Q_{SKAT} statistic follows a mixture of chi-squared distributions where $\lambda_{1, \ldots,} \lambda_M$ are eigenvalues of $\boldsymbol{V^{1/2} W \left(V^{1/2}\right)^T}$, where $\boldsymbol{V}$ is the variance-covariance matrix (equation (1-6)), $\boldsymbol{W} = diag[w_1^2, \ldots, w_m^2]$ is a diagonal matrix of weights and χ_{11}^2, …, χ_{1m}^2 are independent χ_1^2 variables.

SKAT-O:

$$Q_p(\rho) = (1-\rho) Q_{SKAT} + \rho Q_{WST}$$
$$Q_{SKAT-O} = \max\left(Q_p(\rho)\right) \sim \sum_{j=1}^{m} \lambda_j \chi_{1j}^2 \quad \textbf{(1-10)}$$

Q_p is calculated using a series of values (ρ), with the $Q_{SKAT\text{-}O}$ test statistic defined using the value ρ which gives the maximum Q_p.

There have been several challenges regarding the testing of joint effects of several variants on a trait. One question surrounds how to define the genomic region for association testing; a region might be defined as a gene or an exon, or alternatively a "moving window" approach could be adopted (70). Another often cited issue is that of population structure. It is thought that rare variants are likely to exhibit stronger patterns of population stratification compared to common variants and that existing methods for correcting for population structure may be insufficient (63, 70-72). A further consideration is whether functional information regarding variants could be utilised. Variants thought to have a greater functional impact might be given more weight in pooled variant analyses (63), or pathway-based approaches might be used (70).

1.2.6 Meta-analysis of rare variant tests

The importance of sample size is even greater for studies attempting to identify low frequency SNPs that are associated with a trait. As mentioned in Section 1.1.3.3, meta-

analysis is a popular way of achieving large sample sizes in GWAS studies, and is commonly used for identifying single variant associations. Recently, there have been a number of software packages developed which additionally facilitate the meta-analysis of some of the gene-based tests mentioned in Section 1.2.5 (73-78). Similarly to the single variant meta-analysis, the meta-analysis of gene-based tests involves each contributing study undertaking analyses according to an agreed analysis plan, with the gene-based tests then being carried out centrally. These methods are more extensively reviewed in Chapter 3.

1.3 Introduction to pulmonary function measures & COPD

1.3.1 Spirometry & key pulmonary function measures

Spirometry is physiological test which measures the volume and flow of air, as an individual forcibly and completely, expels air from the lungs after maximal inspiration. This procedure provides several measures of pulmonary function that may be used for the diagnosis of a number of diseases and for monitoring general respiratory health (79). Some key volumetric measures include: forced vital capacity (**FVC**), the total amount of air that can be forcibly exhaled; forced expiratory volume in the first second (**FEV_1**), the amount of air that can be exhaled in the first second of the blow; the ratio of FEV_1 to FVC (**FEV_1/FVC**). Spirometry also provides flow measurements including the peak expiratory flow (**PEF**), a measure of maximum instantaneous expiratory flow, and the forced expiratory flow between 25% and 75% of vital capacity (**FEF_{25-75}**) which is the average forced expiratory flow rate over the middle 50 percent of the FVC.

During a spirometric manoeuvre, flow of air reaches a peak early in the blow, with airflow then declining as the volume of air within the lung decreases. Expiratory flow is determined by a combination of the power of the respiratory muscles, elasticity of the lung and airway resistance. Up to peak flow, airflow is largely determined by the expiratory muscles; following peak flow the force largely comes from elastic recoil. At peak flow, muscle force, elastic roil and airway calibre are all at their maximum (80, 81).

Around peak flow, airflow may be limited due to the concept that the flow of air cannot exceed the speed at which a pressure wave can travel along the airway (wave speed theory). Under this theory, air flow velocity will be determined by a combination of airway cross-sectional area and stiffness, along with gas density (82). Following peak flow, proposed mechanisms for airflow limitation include the equal pressure point theory (83) and the Starling resistor ("waterfall") effect (84), which describe the relationship between the elastic recoil of the lung and expiratory flow (80, 81).

Spirometry measures are influenced by the age, sex, ethnicity and height of an individual. Several initiatives, including the National Health and Nutrition Examination Survey (NHANES III) (85), the European Community for Steel and Coal (ECSC) (86) and more recently the Global Lung initiative (GLI) (87) have developed reference equations, which provide predicted spirometric measures for an individual given these characteristics. These equations can also be used to derive percent predicted measures (eg percent predicted FEV_1 [**%pred FEV_1**]) by comparing an individual's measured spirometric values with that which would be expected and allow the estimation of lower limits of normal and Z-scores. Elsewhere, efforts have been made to standardise procedure and quality control of spirometry, by both the the European Respiratory Society (ERS) and American Thoracic Society (ATS), who in 2005 appointed a joint task force to combine guidelines for pulmonary function testing (88).

1.3.2 Chronic Obstructive Pulmonary Disease - Symptoms and Diagnosis

Chronic obstructive pulmonary disease (**COPD**) is a major public health concern, being a leading cause of morbidity and mortality worldwide (89). According to World Health Organisation estimates, 65 million people globally have moderate to severe COPD (90) and the disease accounted for approximately 6% of deaths worldwide in 2012 (91).

The typical symptoms of COPD are chronic cough, shortness of breath (dyspnoea) and sputum production and the disease is characterised by progressive and irreversible **airflow limitation**. The airflow limitation in COPD is a result of a combination of disease of the small airways, leading to increased airway resistance (92), and destruction of the **lung parenchyma** (emphysema), causing a decrease in lung elastic recoil (93, 94). Airflow limitation in COPD is progressive and associated with an increased inflammatory response to long-term exposure to noxious particles or gases (94).

Airflow limitation is defined using spirometry as $FEV_1/FVC<0.70$, after **bronchodilation**. Severity of airflow limitation is further classified using %pred FEV_1, as per the Global Initiative for Chronic Obstructive Lung Disease (**GOLD**) (94) (Table 1-3). There is a weak correlation between severity of airflow limitation and quality of life however; as such, GOLD recommends that the impact of COPD on an individual patient should assessed by considering breathlessness, symptoms and risk of exacerbation, in combination with the level of airflow limitation (94).

Table 1-3: GOLD COPD Classification.

GOLD Classification	Spirometric Definition
GOLD I: mild	$FEV_1/FVC<0.70$ and %pred $FEV_1 \geq 80\%$
GOLD II: moderate	$FEV_1/FVC<0.70$ and $80\% >$ %pred $FEV_1 \geq 50\%$
GOLD III: severe	$FEV_1/FVC<0.70$ and $50\% >$%pred $FEV_1 \geq 30\%$
GOLD IV: very severe	$FEV_1/FVC<0.70$ and %pred $FEV_1 < 30\%$

1.3.3 Processes and risk factors for airflow limitation and COPD

There are a number of processes that may lead to impaired lung function and COPD that can occur throughout an individual's lifetime. In the womb, the lung bud initially develops at 6 weeks gestation, with the completion of lung airway branching (formation of the **trachea**, **bronchi** and **bronchioles**) occurring at 17 weeks. Subsequently, the airways increase in size and **alveoli** continue to develop until an individual reaches early adulthood. During this period of growth, lung volume and airflow increase, until a maximum lung volume is attained, at 20-25 years. Lung function then tends to plateau, for a period of approximately 5-10 years in healthy individuals, before a gradual decline with age (95, 96).

A recent epidemiological study found there were broadly two trajectories to impaired lung function (97). Firstly, a reduced maximally attained lung function, resulting from impaired lung development and growth either in utero or during childhood, can increase the risk of airflow limitation and COPD in later life, even where the rate of lung function decline is normal (97). Lung development may be compromised by a number of factors including maternal smoking and nutrition during pregnancy, prematurity, low birthweight, respiratory infections during childhood and exposure to cigarette smoke through either passive or active smoking (95, 96).

Secondly, the development of airflow limitation and COPD might also be a result of a shortened plateau phase at maximal lung function, or by an accelerated decline (97). Cigarette smoking is recognised as the most significant risk factor for this decline in lung function, however there are other environmental factors thought to play a role (95). In low and middle income countries, exposure to biomass fuels used for heating and cooking in a domestic setting has a great effect on lung health and COPD risk (93). Other risk factors include air pollution and workplace exposure to dust or fumes (98-100). The exposure to noxious particles from cigarette smoking or occupational or environmental exposures causes an abnormal inflammatory response in the airways and in the lung parenchyma. (94, 96, 101). Such inflammatory responses can lead to the accumulation of exudates in the bronchial lumen and an increased infiltration of

inflammatory cells, which have been associated with a thickening of the walls of the small airways, by means of repair or remodelling processes (102).

Whilst it is helpful for understanding the mechanisms underlying lung function decline and COPD to consider these two separate trajectories, for any individual, airflow limitation might be a result of a combination of both a low maximal lung function and accelerated decline in lung function (97). Aside from environmental factors affecting both lung development and the rate of lung function decline, there is additionally a genetic component, with COPD and lung function tending to aggregate within families (103, 104).

1.3.4 Known genetics of lung function and COPD

Estimates of the narrow sense heritability of lung function vary widely, with the proportion of variance of FEV_1 attributable to additive genetic effects estimated at between 11% and 50% (105-110). Heritability estimates for FVC range from 37% to 54% (107, 108, 111), whilst for FEV_1/FVC estimates range from 14% to 66% (106-108, 110). For flow measures, estimates of heritability have similar ranges: 14% to 43% for PEF (112, 113) and 35% to 45% for FEF_{25-75} (109, 114, 115).

To date, a number of GWAS have had success in identifying SNPs which show association with three volumetric measures of lung function: FEV_1, FVC and FEV_1/FVC. In 2009, Wilk and colleagues in the Framingham Heart Study (116) identified a region on 4q31 associated with both FEV_1 and FEV_1/FVC, close to the hedgehog-interacting protein (*HHIP*) gene. Two large scale meta-analyses, carried out by the SpiroMeta (117) and CHARGE (118) consortia in 2010, verified this association and identified a further nine regions showing association with either FEV_1 or FEV_1/FVC. Two joint SpiroMeta-CHARGE efforts identified an additional 16 regions associated with FEV_1 or FEV_1/FVC in 2012 (119), and 6 novel regions associated with FVC in 2014.

Running concurrently to the analyses described in this thesis were two studies, whose results were published at the end of 2015. Firstly, a further SpiroMeta analysis, using genotype data imputed to the 1000 Genomes (16) reference panel identified 16 novel lung function loci, including two low frequency variants (rs113473882, in *LTBP4* and

rs148274477 near *GPR126*) (43). Secondly, the UK Biobank Lung Exome Variant Evaluation (UK BiLEVE) effort examined the genotypes from 48,943 individuals, who were sampled from the extremes and the middle of the distribution of FEV_1, separately in never smokers and ever smokers. Case-control analyses of samples with high FEV_1 compared to samples with low FEV_1 identified an additional 6 loci, including a rare intergenic SNP between *RBM19* and *TBX5* (120).

Less is known about the genetic determinants of flow lung function measures and to date, very few GWAS of PEF or FEF_{25-75} have been carried out. A genome-wide interaction study of FEF_{25-75} and small particulate matter (PM10) identified SNPs in *CDH13* associated with FEF_{25-75} decline; this association met genome-wide significance and showed evidence of replication (121). A study undertaken as part of the Framingham Heart Study identified several other, less statistically significant associations with FEF_{25-75}, or FEF_{25-75} decline at *NIP2*, *IL6R*, *CCBL2*, *LIPF*, *SYT10* and *ETAA16*, though replication of these signals was not undertaken (122). A GWAS of allergic disease in a Russian population identified associations between PEF and SNPs in *RIT2* and *ADAD2*, however these did not reach genome-wide significant, and no replication was undertaken (123).

Through the study of genetic influences of lung function measures, in particular of FEV_1 and FEV_1/FVC, it is hoped that a greater understanding of the genetics of COPD may also be gained, given the diagnosis of the disease is based on these measures and so they are likely to share many of the same genetic determinants. The advantage of studying quantitative lung function traits in addition to COPD is that there is greater statistical power to detect associations with continuous outcomes, rather than binary traits (124). So far, nineteen of the 54 loci associated with lung function have also been shown to be associated with COPD risk or airflow obstruction through GWAS (120, 125-129). In addition to these lung function regions, a locus on 15q25 has been implicated in COPD susceptibility (129, 130), and has since shown to be associated with smoking behaviour (131-134). Finally, approximately 1-2% of COPD cases can be attributed to alpha1-antitrypsin (AAT) deficiency, a rare inherited disorder, caused by mutations within the *SERPINA1* gene (129, 135). Low levels of AAT leave pulmonary tissue more susceptible to degradation and often results in early onset COPD (98, 135).

All loci identified through GWAS as showing genome-wide significant associations to date with one or more of FEV_1, FVC, FEV_1/FVC and COPD are listed in Table 1-4. The genes cited in this table are generally the nearest gene(s) to the GWAS signal, and are not necessarily the causal gene in each case.

Table 1-4: Genetic loci showing genome-wide significant association with at least one lung function / COPD trait to date.

		Quantitative lung function		COPD
Chr	**Gene(s)**	**Phenotype**	**Reference**	**Reference**
1	*MFAP2*	FEV_1/FVC	Soler Artigas et al. 2011 (119)	-
1	*MCL1,ENSA*	FEV_1	Soler Artigas et al. 2015 (43)	-
1	*TGFB2*	FEV_1/FVC	Soler Artigas et al. 2011 (119)	Cho et al. 2014 (128)
1	*LYPLAL1,RNU5F-1*	FEV_1/FVC	Soler Artigas et al. 2015 (43)	-
2	*KCNS3,NT5C1B-RDH14*	FEV_1/FVC	Soler Artigas et al. 2015 (43)	-
2	*EFEMP1*	FVC	Loth et al. 2014 (136)	-
2	*TNS1*	FEV_1	Repapi et al. 2010 (117)	Soler Artigas et al 2011 (127)
2	*HDAC4*	FEV_1/FVC	Soler Artigas et al. 2011 (119)	-
3	*RARB*	FEV_1/FVC	Soler Artigas et al. 2011 (119)	Wilk et al. 2012 (130)
3	*RP11-538P18.2*	FVC	Soler Artigas et al. 2015 (43)	-
3	*MECOM*	FEV_1	Soler Artigas et al. 2011 (119)	-
4	*FAM13A*	FEV_1/FVC	Hancock et al. 2010 (118)	Cho et al. 2010 (126)
4	*TET2*	FEV_1	Wain et al. 2015 (120)	Wain et al. 2015 (120)
4	*GSTCD*	FEV_1	Repapi et al. 2010 (117), Hancock et al. 2010 (118)	Soler Artigas et al 2011 (127)
4	*NPNT*	FEV_1	Wain et al. 2015 (120)	Wain et al. 2015 (120)
4	*NPNT*	FEV_1/FVC	Soler Artigas et al. 2015 (43)	-
4	*HHIP*	FEV_1/FVC	Wilk et al. 2009 (116)	Pillai et al. 2009 (129)
5	*SPATA9*	FEV_1/FVC	Soler Artigas et al. 2011 (119)	-
5	*HTR4*	FEV_1	Repapi et al. 2010 (117), Hancock et al. 2010 (118)	Soler Artigas et al 2011 (127)
5	*ADAM19*	FEV_1/FVC	Hancock et al. 2010 (118)	Castaldi et al. 2011 (125)
6	*BMP6*	FVC	Loth et al. 2014 (136)	-
6	*ZKSCAN3*	FEV_1	Soler Artigas et al. 2011 (119)	-
6	*NCR3*	FEV_1/FVC	Soler Artigas et al. 2011 (119)	-
6	*AGER*	FEV_1/FVC	Repapi et al. 2010 (117), Hancock et al. 2010 (118)	Castaldi et al. 2011 (125)
6	*HLA-DQB1/ HLA-DQA2*	FEV_1	Wain et al. 2015 (120)	Wain et al. 2015 (120)
6	*ARMC2*	FEV_1/FVC	Soler Artigas et al. 2011 (119)	-
6	*GPR126, RP11-440G9.1*	FEV_1/FVC	Soler Artigas et al. 2015 (43)	-
6	*GPR126*	FEV_1/FVC	Hancock et al. 2010 (118)	Wilk et al. 2012 (130)
9	*PTCH1*	FEV_1/FVC	Hancock et al. 2010 (118)	-
9	*ASTN2*	FEV_1/FVC	Soler Artigas et al. 2015 (43)	-

		Quantitative lung function		COPD
Chr	**Gene(s)**	**Phenotype**	**Reference**	**Reference**
9	*LHX3*	FVC	Soler Artigas et al. 2015 (43)	-
10	*CDC123*	FEV_1/FVC	Soler Artigas et al. 2011 (119)	-
10	*C10orf11*	FEV_1	Soler Artigas et al. 2011 (119)	Wilk et al. 2012 (130)
11	*HSD17B12*	FVC	Loth et al. 2014 (136)	-
11	*PRDM11*	FVC	Loth et al. 2014 (136)	-
11	*MMP12*	-	-	Hunninghake et al. 2009 (137)
12	*PTHLH,CCDC91*	FVC	Soler Artigas et al. 2015 (43)	-
12	*LRP1*	FEV_1/FVC	Soler Artigas et al. 2011 (119)	-
12	*CCDC38*	FEV_1/FVC	Soler Artigas et al. 2011 (119)	-
12	*RBM19/TBX5*	FEV_1	Wain et al. 2015 (120)	Wain et al. 2015 (120)
12	*TBX3,MED13L*	FEV_1	Soler Artigas et al. 2015 (43)	-
14	*TRIP11*	FEV_1	Soler Artigas et al. 2015 (43)	-
14	*RIN3*	FEV_1	Soler Artigas et al. 2015 (43)	Cho et al. 2014 (128)
15	*CHRNA3, CHRNA5, IREB2*	-	-	Pillai et al. 2009 (129)
15	*THSD4*	FEV_1/FVC	Repapi et al. 2010 (117)	Wilk et al. 2012 (130)
16	*EMP2,TEKT5*	FEV_1/FVC	Soler Artigas et al. 2015 (43)	-
16	*MMP15*	FEV_1/FVC	Soler Artigas et al. 2011 (119)	-
16	*CFDP1*	FEV_1/FVC	Soler Artigas et al. 2011 (119)	-
16	*WWOX*	FVC	Loth et al. 2014 (136)	-
17	*17q21.31*	FEV_1	Wain et al. 2015 (120)	Wain et al. 2015 (120)
17	*KCNJ2*	FVC	Loth et al. 2014 (136)	-
17	*TSEN54*	FEV_1	Wain et al. 2015 (120)	Wain et al. 2015 (120)
19	*LTBP4*	FEV_1/FVC	Soler Artigas et al. 2015 (43)	-
21	*KCNE2*	FEV_1/FVC	Soler Artigas et al. 2011 (119)	-
22	*MIAT,MN1*	FEV_1	Soler Artigas et al. 2015 (43)	-
X	*AP1S2,GRPR*	FEV_1/FVC	Soler Artigas et al. 2015 (43)	-

Whilst GWAS have been successful in identifying many regions of the genome that are associated with lung function and COPD, it is not clear how these genetic regions might be influencing lung health and disease. There are several genes which have been implicated in these traits which are involved in processes such as growth, inflammation and tissue remodelling and repair, which might suggest biological mechanisms for disease. A number of Metalloproteinase genes, have been associated with COPD and/or lung function: matrix metalloproteinases (MMPs) *MMP12* and *MMP15* and a disintegrin and metalloprotease domain gene *ADAM19.* Metalloproteinase play a role in tissue remodelling and repair are involved in inflammatory processes (138-140). Genes in the Hedgehog (Hh) signalling pathway, including *HHIP* and *PTCH1* have also

been implicated in lung function traits; this pathway is crucial to the branching morphogenesis of the lung and other embryonic processes and disruption of the pathway can lead to severe foetal lung malformations (141, 142).

Amongst other genes implicated in COPD and lung function are *GSTCD,* a member of the glutathione S-transferase, family of genes involved in cellular detoxification and upregulated in response to oxidative stress (143), and *AGER* (advanced glycosylation end product-specific receptor), a multiligand receptor of the immunoglobulin superfamily which is highly expressed in the lung, particularly in the type I **pneumocytes**, and has been implicated in lung cancer and **fibrosis** (144). One of the genes identified recently through an association with a low frequency variant was *LTBP4* (latent transforming growth factor beta [TGFβ] binding protein 4); this gene belongs to a family of extracellular matrix (ECM) proteins which binds to TGFβ, and targets TGFβ to the ECM (145). *LTBP4* has been found to play a role in the regulation of fibulin-5 dependent elastic fiber assembly and mice deficient in the protein display defects in lung septation and elastogenesis (146).

A number of the regions associated with lung function and COPD have also been implicated in other traits, notably height (*CCDC91, TRIP11, TET2, HHIP, GPR126, MFAP2)*, smoking (*CHRNA3/5, IREB2*) and lung cancer (*CHRNA3/5, NCR3, ZSCAN3*). Several associations with lung function have also been identified in the major histocompatibility complex (MHC) region (*NCR3, ZSCAN3, AGER, HLA-DQB1, ARMC2*). This region is characterized by extended blocks of LD, making localisation of signals in this region challenging, and has been implicated in a large number of diseases, particularly autoimmune and inflammatory diseases such as rheumatoid arthritis, type 1 diabetes and ankylosing spondylitis (147). Whilst some of these associations may represent pleiotropic effects (for example in the case of height), for the loci which have been found to be associated with both smoking and COPD risk, it is likely that the observed effect on the risk of COPD is driven by the genetic effect on smoking.

1.4 Aims and Outline of Thesis

The primary aim of the analyses described in this thesis was to identify low frequency and rare SNPs which influence lung function and COPD susceptibility, and to explain some of the heritability not accounted for by the common variants already identified as being associated with these traits.

In Chapter 2, I describe an analysis of COPD cases and controls with exome array data. This study aimed to investigate the role of low frequency SNPs in both the risk of COPD, and the severity of airflow limitation within COPD cases. In this chapter, I fully describe the quality control of the data undertaken and the results of both single variant and gene-based analyses.

In Chapter 3, I describe the methods for the meta-analysis of gene-based tests and review the software packages which implement these tests. I then evaluate the performance of one of the software packages (RAREMETAL (74)), using real data from 48,943 individuals from the UK BiLEVE study (120). Through these analyses I compare the concordance of the meta-analysis methods with analyses carried out using individual level data, and compare the new meta-analysis methods to the meta-analysis of gene-based tests using Fisher's and Z-score methods of combining P-values.

Following this evaluation of the RAREMETAL package, I then utilise the software in the analyses I describe in Chapter 4, consisting of a meta-analysis of exome array data and three lung function measures: FEV_1, FVC and FEV_1/FVC. In this study, I designed an analysis plan, which was used by 11 contributing studies to undertake study level analyses. I then undertook quality control of all study level analyses, and combined the results using the RAREMETAL software package, undertaking both single variant and gene-based tests.

In Chapter 5, I describe a GWAS of two flow measures: PEF and FEF_{25-75}. This analysis includes 102,929 individuals from UK Biobank, with 14.6 million genotypes imputed to the combined 1000 Genomes (16) and UK10K (40) reference panel. I describe firstly how the flow variables were derived from spirometry data, and then present the results of the GWAS of single variant associations with the aim of ascertaining whether

there is utility in studying these measures of flow, in addition to volumetric lung function measures in genetic studies.

Finally, in Chapter 6, I summarise the findings of the preceding chapters and discuss ongoing and potential future work in the area.

Chapter 2 Association of rare variants with COPD risk and airflow limitation

2.1 Introduction

COPD is characterised by fixed airflow limitation and is a leading cause of morbidity and mortality (89). Approximately 1-2% of COPD cases may be attributed to the rare inherited disorder, alpha1-antitrypsin (AAT) deficiency, which is caused by mutations within the *SERPINA1* gene (129, 135). For the remainder of COPD cases, cigarette smoking is recognised as the most significant risk factor (100); however there is also a genetic component. As discussed in Section 1.3.4, GWAS have successfully identified several genomic regions showing association with COPD or airflow limitation to date (125-127, 129, 130, 137), with many of these regions having also been associated with the quantitative lung function measures FEV_1 and FEV_1/FVC (116-119). These known loci however, only explain a small proportion of the expected heritability (119). GWAS undertaken to date have generally focussed on common variants (typically >5% minor allele frequency [MAF]); one hypothesis is that some of the so-called "missing heritability" might be accounted for by variants of lower frequencies.

This chapter describes the single variant and gene-based association analyses of COPD risk and severity of airflow limitation, which I carried out as part of the UK COPD Exome Chip Consortium. Through these analyses, I primarily aimed to identify low frequency and rare coding variants associated with COPD, through the use of exome array data, thereby uncovering some of the missing heritability. The severity of airflow limitation in COPD is classified using percent predicted FEV_1, as per the Global Initiative for Chronic Obstructive Lung Disease (GOLD) (94). Through these analyses, I also aimed to identify variants associated with percent predicted FEV_1 in COPD cases, as a measure of severity of disease. The results of this chapter were published in Thorax in 2016 (148) and a copy of the manuscript is included in Appendix A.

2.1.1 UK COPD Exome Chip Consortium

The UK COPD Exome Chip Consortium is a collaborative consortium, bringing together COPD cases from twelve COPD disease cohorts and population based studies, who

have been genotyped using a custom exome chip array. The Consortium additionally genotyped approximately 1000 general population controls, to contribute to the wider UK Exome Chip Consortium effort, from which additional controls could also be drawn. Analyses using these samples firstly aimed to investigate the role of rare functional variants in COPD, and secondly to confirm the role of SNPs previously showing association with lung function. This latter aim was facilitated by the inclusion of additional custom content on the array of 2585 SNPs from regions previously showing suggestive association ($P<2.21x10^{-3}$) with lung function in large genome-wide HapMap-imputed meta-analysis of quantitative lung function measures (119). Discovery case-control analyses (analyses of COPD risk) and analyses of percent predicted FEV_1 in COPD cases (analyses of airflow limitation) were carried out, with both the exome chip genotype data (exome analyses) and custom content genotype data (custom content analyses). Replication was undertaken using the UK Biobank Lung Exome Variant Evaluation (UK BiLEVE) study, a subset of 48,943 UK Biobank participants with genome-wide SNP genotyping data which includes substantial overlap with the exome chip. A more powerful discovery strategy was adopted for COPD risk and severity of airflow limitation, by meta-analysing data for the subset of exome chip variants that were in both the COPD exome chip consortium and the UK BiLEVE study.

2.1.2 My role in the study

Samples were first prepared and sent for genotyping by each case collection. I received genotype data for all consortium samples and I carried out thorough quality control of these data. Each case collection provided phenotype information, and I undertook further quality control of these data, to ensure all samples had the required phenotypes available and all met the appropriate inclusion criteria. Subsequently, I undertook the discovery exome analyses, and the meta-analyses of the UK COPD exome chip consortium and UK BiLEVE samples; this included both single variant association analyses and gene-based analyses. This chapter describes the quality control of the data and these analyses fully. I did not lead the analyses of the custom content data, however since the results of these analyses are published alongside the exome analyses (148), some aspects of the custom content analyses are mentioned in this chapter, for completeness.

2.2 Discovery exome analyses of COPD risk and severity

2.2.1 Study participants, phenotypes and genotyping

3487 COPD cases with airflow limitation indicative of GOLD 2-4 COPD were identified from 12 UK collections, listed in Table 2-1. Descriptions of all these studies, including details of spirometry are described in (148) (Appendix A). Individuals met case criteria if they had COPD GOLD 2 (94) or worse ($FEV_1/FVC \leq 0.7$ and percent predicted $FEV_1 < 80\%$, according to the NHANES III spirometric reference equations (85)), did not have a doctor diagnosis of asthma and had reported ever smoking. Five of the sample collections (UKCOPD, COPDBEAT, NottCOPD, EUCOPD and GoTARDIS) were COPD cohorts, with all individuals (total n=1562) having irreversible airflow limitation, and meeting GOLD 2 criteria based on post-bronchodilator spirometry. The remaining cases were taken from general population cohorts. For these samples, only pre-bronchodilator spirometry measures were available; the inclusion of these samples as cases was to increase the power of the association analyses.

Table 2-1: Case collections used in discovery analyses.
N samples prior to quality control of genotype data.

Abbreviation	Study Name	N samples	Study Type
GS:SFHS	Generation Scotland	525	General Population
BRHS	British Regional Heart Study	436	General Population
BWHHS	British Women's Heart and Health Study	261	General Population
UKCOPD	UK COPD Cohort	231	COPD case cohort
HCS	Hertfordshire Cohort Study	340	General Population
COPDBEAT	Biomarkers to target Antibiotic & Systemic Corticosteroid therapy in COPD exacerbations	92	COPD case cohort
NottCOPD	Nottingham COPD Study	77	COPD case cohort
Nott smokers	Nottingham smokers	157	General Population
Gedling	Gedling study	37	General Population
ELSA	English Longitudinal Study of Aging	170	General Population
EUCOPD	EU COPD Gene Scan	292	COPD case cohort
GoTARDIS	GoTARDIS Study	870	COPD case cohort

For the analyses of COPD risk, 4945 healthy controls with exome chip data were selected from: Generation Scotland (GS:SFHS, n=1032); British 1958 Birth Cohort (1958BC, n=1456); Oxford Biobank (OXBB, n=1822) and GoDARTS (n=635). All controls were ever smokers and were free of lung disease, according to available spirometry and/or phenotype information.

2.2.2 Quality Control of phenotype and genotype data

Thorough quality control (QC) checks of the phenotype data from all case and control collections were carried out, and issues with missing or erroneous phenotype information were resolved by liaising with individual sample collections, as necessary. For all samples, FEV_1/FVC and percent predicted FEV_1 were recalculated and all smoking phenotype data examined, with any samples not meeting the appropriate case or control criteria identified for exclusion.

All COPD cases and GS:SFHS controls were genotyped together using a custom version of the Illumina Human Exome BeadChip which included additional custom content for regions which have previously shown modest association with lung function (119). The remaining control samples (1958BC, OXBB and GoDarts) were genotyped separately using the Illumina Human Exome Beadchip.

The UK exome chip consortium is a UK-wide collaborative effort to create a pool of samples with exome chip data available for use as general population controls, and to harmonise genotype calling and quality control of these data. To that end, the consortium developed a standard operating procedure (SOP) (149) which outlines the recommended QC procedure for exome chip data and includes many of the QC steps routinely applied in GWAS studies, as described in Section 1.1.3.1. In the first instance, QC of the exome chip genotype data was carried out based on this SOP, as follows (QC of the custom content genotype data was undertaken separately). The genotype data were received with genotypes that had been called using Illumina's Gencall algorithm in Genomestudio (13). Initial exclusions were carried out to exclude SNPs and samples with >90% missing data. Subsequently, samples with a call rate<98%, an unusually high or low heterozygosity rate (greater than three standard deviations from the mean; sample heterozygosity rates calculated separately for SNPs with MAF>=1% and SNPs with MAF<1%, [Figure 2-1], gender mismatches, and duplicates were excluded. Ancestry principal components analysis (PCA) was carried out with EIGENSTRAT(17), using subset of 3241 ancestry informative markers; any individuals that were more than four standard deviations from the sample mean for either of the first two principal components were excluded (Figure 2-2).

Figure 2-1: Plots of Proportion of SNPs missing versus heterozygosity rate for A. SNPs with MAF≥1% and B. SNPs with MAF<1%.

The vertical lines correspond with the 98% call rate threshold. The horizontal lines indicate 3SDs from the mean heterozygosity rate.

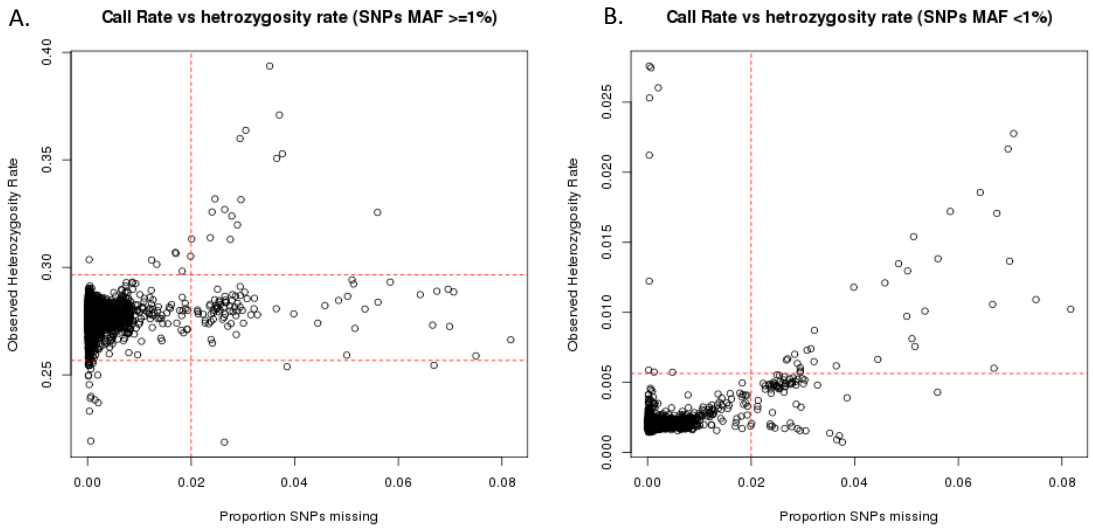
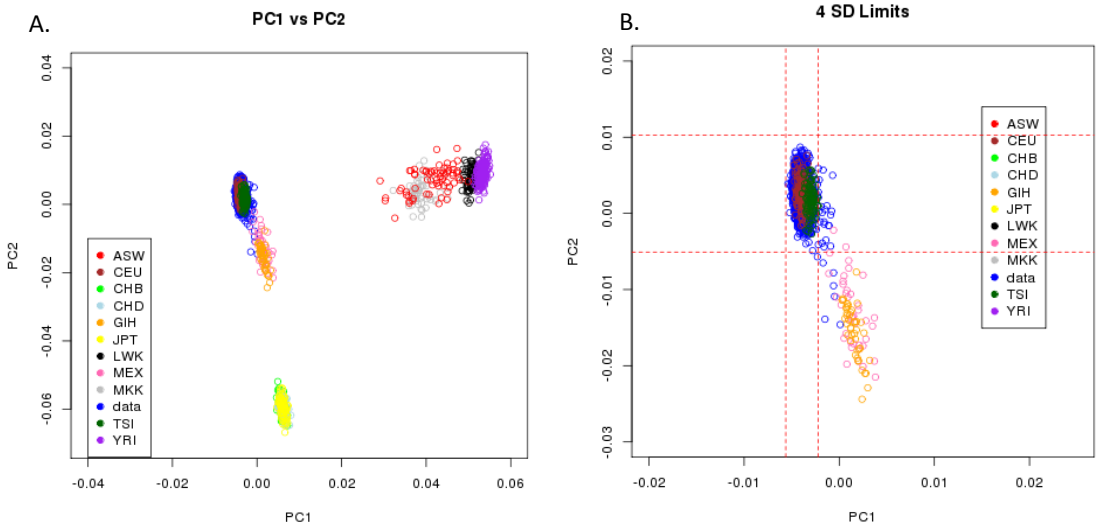


Figure 2-2: Ancestry Principal Components Plots.

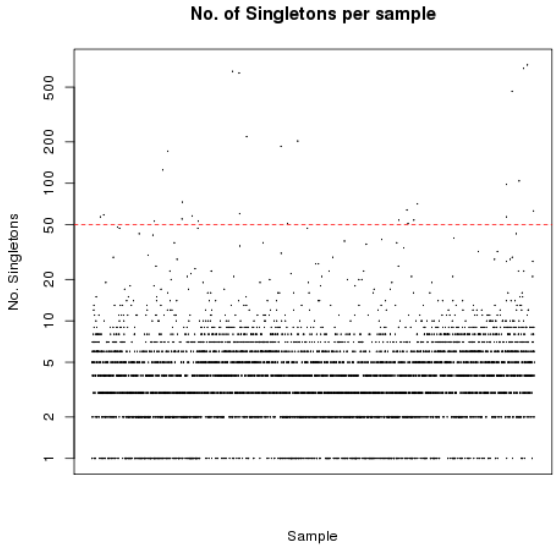
Plots of the first two ancestry principal components. Cases and GS:SFHS controls are labelled data (coloured royal blue) and cluster with Northern European (CEU) and Tuscan (Italian) samples (TSI). Plot A. shows the data in context with all other ancestries. Plot B shows the European cluster, with lines indicating 4SDs from the data sample mean



Following this first stage of QC, missing genotypes were recalled using zCall (51), a software package developed specifically for improving the calling of rare variants (described in Section 1.2.4). A second stage of QC was then carried out using the recalled data; SNPs with call rate<99% or which deviated from Hardy Weinberg Equilibrium ($P<10^{-4}$) were excluded, along with samples with call rate<99%, and heterozygosity outliers. In addition, samples were excluded if they had an excess of

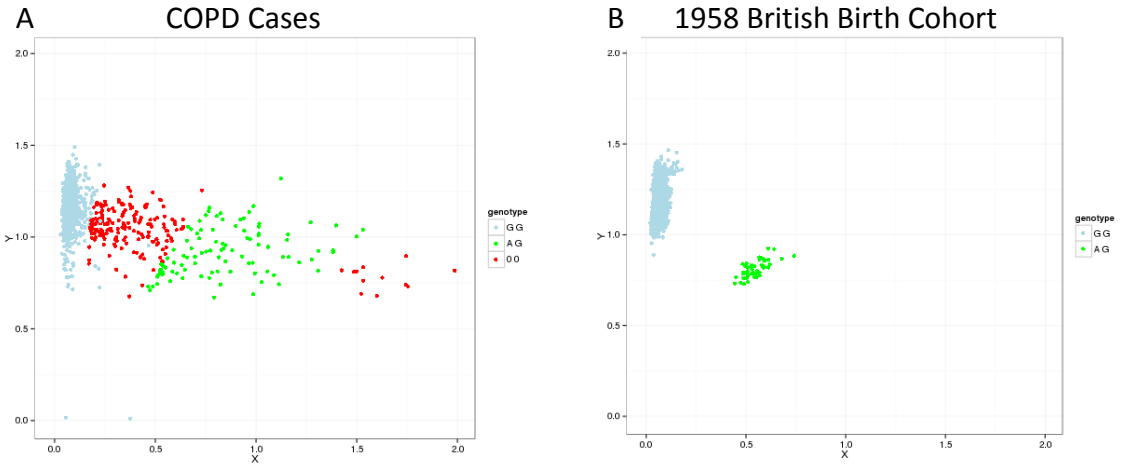
singleton SNPs (samples who were the only individual to have the minor allele for greater than 50 SNPs, Figure 2-3).

Figure 2-3: Plot of the number of SNPs for which each sample is the only one to have the alternate allele for that SNP.



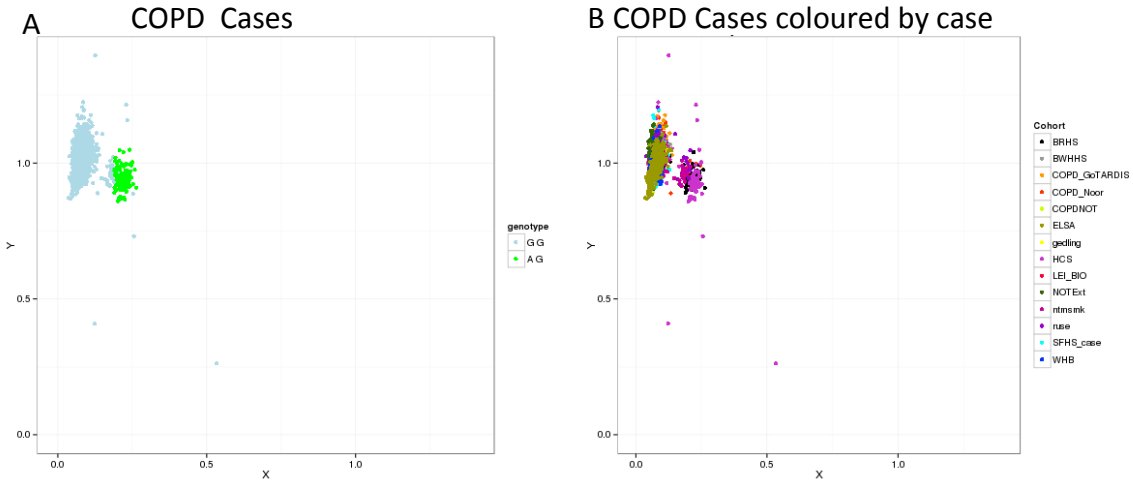
After completing the QC of data, there remained a number of issues regarding the quality of the genotype data of the COPD cases. These issues were evident from the inspection of clusterplots of normalised intensity values, which should result in up to three defined clusters of data, corresponding to genotype groups (Section 1.1.3.4). Firstly, a number of SNPs were poorly clustered in the COPD cases (Figure 2-4 A), whilst showing good clustering in the 1958BC controls (Figure 2-4 B). These SNPs could be identified through testing for differential missingness in cases and controls, using the original Gencall-called data. As a result, an additional genotype QC step was applied, which excluded 9155 SNPs showing differential missingness with $P<10^{-5}$.

Figure 2-4: Example clusterplots of poorly clustered SNP.
SNP is exm462709, with genotypes called by Gencall. A. Clusterplot for COPD cases; genotypes labelled "00" were assigned missing genotypes by Gencall. B. Clusterplot for 1958BC Controls.



Secondly, there were a number of SNPs whose genotypes were subject to batch effects by case collection. Figure 2-5 A shows each sample coloured by genotype; either homozygote GG or heterozygote GA. When each sample is coloured by case collection (Figure 2-5 B), it can be seen that the heterozygote cluster is in fact the result of a case-collection batch effect. For a number of other SNPs, the same case collections appeared to cluster separately from the remainder of the samples. Across these SNPs, Gencall assigned this second cluster as having missing genotypes, with zCall subsequently recalling all samples in the second cluster with heterozygote genotypes. Consequently, it was possible to identify SNPs affected by this batch effect by testing for associations between genotype and sample collection, using the zCall genotypes. I tested for association with each case cohort in turn vs all other cases combined and the 4104 SNPs showing a significant association ($P<10^{-6}$) with any case collection were excluded.

Figure 2-5: Example of SNP that was subject to batch effects by case collection.
SNP was exm578529 with genotypes called by zCall. A. Clusterplot for COPD cases, coloured by zCall assigned genotypes. B. Clusterplot for COPD cases, coloured by case collection.



Finally, there was a small subset of samples which consistently had unusual intensity data values. The SNP shown in Figure 2-6 A (exm1595800) had generally clustered well, however there were a number of samples for which Gencall has assigned a CC genotype, but which were separate from the homozygote CC genotype cluster. Similarly, for rs3097648 (Figure 2-6 B), there are a number of samples assigned an AA genotype, but which are separate from the AA cluster. There were 12 samples in total which were consistently outliers in this way, across many SNPs, but which had passed all quality control filters. I calculated the mean X and Y intensities across all autosomal SNPs for all COPD cases. These 12 outliers all had mean X and/or Y intensities that were greater than four standard deviations from the overall sample mean (Figure 2-7); this criterion was used as a final sample filter, to exclude those 12 outliers. Aside from these 12 samples, there were a further 30 with outlying X and/or Y intensities. 29 of those were excluded as they had a call rate<90%. The final sample was excluded as it had a call rate<98%, was a heterozygosity outlier and had inconsistent sex in the genotype and phenotype data.

Figure 2-6: Example of two SNPs (A. exm1595800 and B. rs3097648) showing samples which consistently had unusual intensity data values.

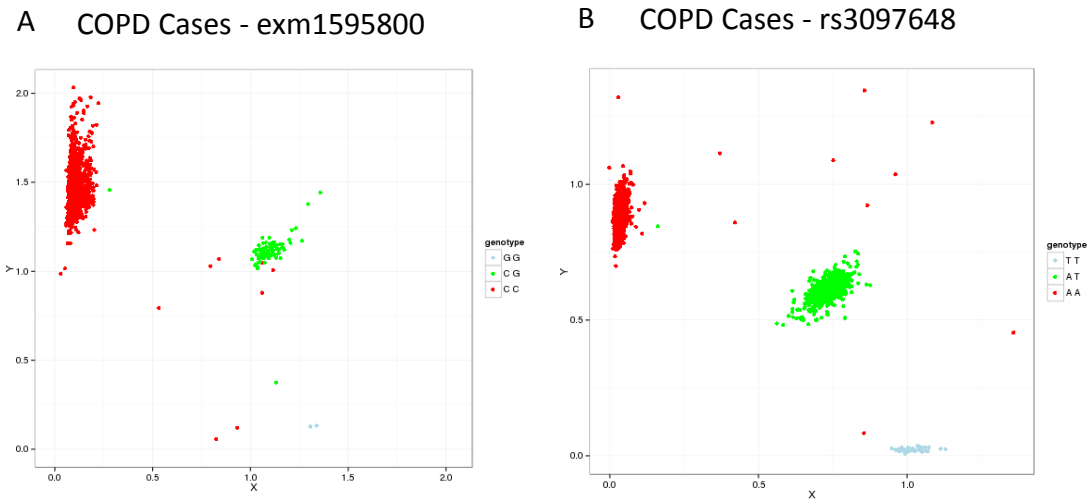
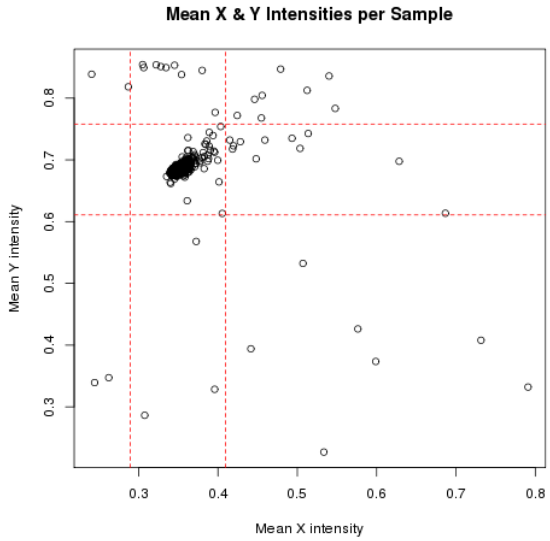


Figure 2-7: Plot of mean X and Y intensities per sample, across all autosomal SNPs.
Red lines indicate 4SDs from overall sample means.



For cases and GS:SFHS controls, I undertook both stages of genotype QC and recalling of genotypes using zCall. The 1958BC, OXBB and GoDARTs samples were shared controls from the UK exome chip consortium and the first stage of QC and recalling of genotypes for these samples were carried out centrally within the consortium. For these data, I undertook the second stage of QC (post-zCall) only. Details of the sample exclusions applied to cases and to each control collection are shown in Table 2-2.

Table 2-2: Genotype QC for samples used in discovery exome analyses.

	Cases	GS:SFHS Controls	1958BC Controls*	OXBB Controls*	GoDARTS Controls*
Initial sample	**3488**	**1032**	**-**	**-**	**-**
Samples failing stage 1 QC: pre zCall					
Call rate<90%	34	1	-	-	-
Sex mismatches	18	2	-	-	-
Heterozygosity outliers (common SNPs MAF≥1%)	41	3	-	-	-
Heterozygosity outliers (rare SNPs MAF<1%)	28	6	-	-	-
Call rate<98%	43	46	-	-	-
Duplicates (PI_HAT>0.95)	56	0	-	-	-
PCA outliers (+/- 3SD of the mean)	12	2	-	-	-
Samples with excess number of singletons SNPs (>50)	15	6	-	-	-
Inconsistency with GWAS data	1	0	-	-	-
XY-intensity outliers (+/- 4SD of the mean)	12	0	-	-	-
Samples passing pre zCall QC	**3302**	**976**	**1456**	**1822**	**635**
Samples failing stage 2 QC: post zCall					
Call rate<99%	0	0	0	0	0
Heterozygosity outliers	76	15	27	52	11
Final Samples Passing both QC stages	**3226**	**961**	**1429**	**1770**	**624**

*Stage 1 QC and recalling of genotypes using zCall carried out for 1958BC, OXBB and GoDARTs controls within UK exome chip consortium

2.2.3 Discovery Analyses: Methods

2.2.3.1 Single Variant Association analyses

Single SNP associations with all 135,818 SNPs passing QC and COPD risk were tested using a logistic regression model, with adjustment for age, sex and pack-years smoking and assuming an additive genetic model, as in equation (2-1). Associations with untransformed percent predicted FEV_1 in COPD cases (analysis of airway limitation) were tested using a linear regression model, with adjustment for pack-years (equation (2-2)). Since not all samples had pack-years data available, secondary analyses were carried out without adjustment for pack-years smoked, for both the COPD risk and airflow limitation analyses. These secondary analyses allowed the inclusion of all samples, thereby giving greater power to detect associations with low frequency SNPs. All single variant associations were carried out using PLINK v1.07 (18). Using a Bonferroni correction for the number of tests undertaken, a significance level of $P<3.7\times10^{-7}$ would be required in the exome single variant analysis to retain a type 1 error of 5% ("exome-wide significant"). SNPs of interest were identified using a less conservative threshold of $P<10^{-5}$ in any of the analyses. For all significant associations, cluster plots were inspected by eye to check genotype calling quality.

$$logit(P_{COPD}) = \beta_0 + \beta_1 SNP + \beta_2 \text{sex} + \beta_3 \text{age} + \beta_4 packyears \quad (2\text{-}1)$$

$$percent\ predicted\ FEV_1 = \beta_0 + \beta_1 SNP + \beta_2 packyears + \varepsilon \quad (2\text{-}2)$$

$$\varepsilon \sim N(0, \sigma_\varepsilon^2)$$

2.2.3.2 Gene-based Association analyses

Gene-based tests assess the pooled effect of several variants within a specified gene region. In these analyses, SKAT-O was utilised, which optimally combines the Weighted Sum Test and SKAT methods (described in Section 1.2.5, equation (1-10)). Variants were annotated using ANNOVAR (150), based on the GRCh37/hg19 database, and all variants located within exons were included in the analyses. Analyses of COPD risk and airflow limitation were undertaken for SNPs with MAF<5% using SKAT-O (69), with covariate adjustments carried out analogously to the single variant analyses. The

default beta distribution weightings were used (weight for jth variant: β(MAF_j; 1, 25)), giving greater weight to rarer variants. Gene-based results were filtered to only include genes with at least two SNPs with a MAF<5%, and with a cumulative MAF>0.05%.

To further evaluate notable gene based signals, a "drop-one" analysis was utilised. This involved recalculating the SKAT-O P-value when individual SNPs were sequentially excluded from the test. If the SKAT-O P-value was considerably attenuated by the removal of a particular SNP, this would indicate that the SKAT-O signal was likely to be largely influenced by that individual SNP, rather than variants within that gene as a whole.

2.2.4 Discovery Analyses: Results

3226 cases and 4784 controls passed all sample genotype QC and were used in the exome analysis. Clinical characteristics of these samples are summarised in Table 2-3. Of the SNPs which passed all quality control criteria in both cases and controls, 135,818 were polymorphic, of which 101,308 (74.6%) had a MAF<1%. In only cases, there were 116,809 polymorphic variants, with 81,347 (69.6%) of those having a MAF<1%. Of the polymorphic variants in cases and controls combined, 120,315 were exonic, with 108,513 annotated as nonsynonymous, 8302 synonymous, 2254 stop-gain, 86 stop-loss and 1160 of unknown function.

Table 2-3: Clinical characteristics of Samples passing Genotype QC.

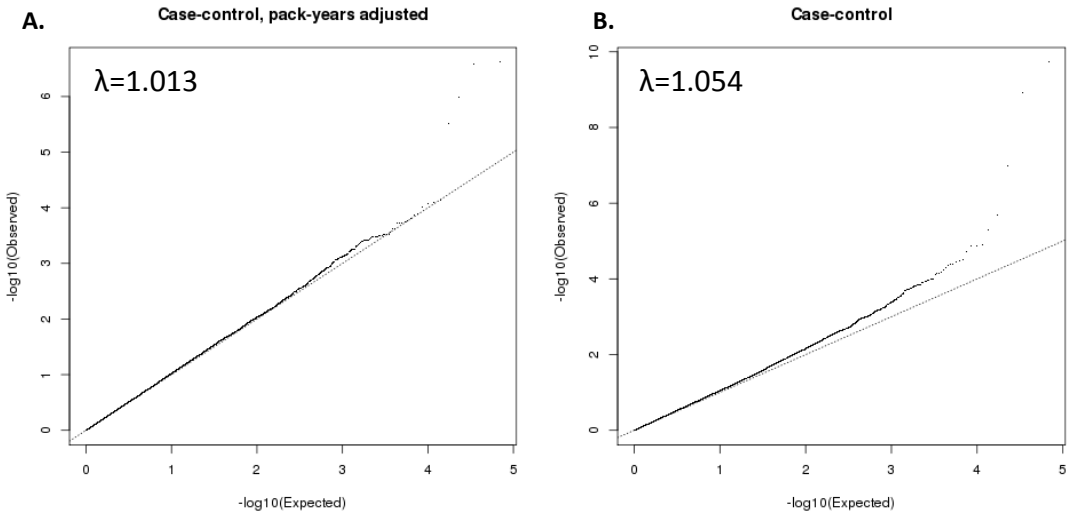
Sample Collection	n	Sex	Age	Percent Predicted FEV_1	FEV_1/FVC	Pack-years	
		Male, n (%)	Mean (SD)	Mean (SD)	Mean (SD)	Samples with data (n)	Mean (SD)
Discovery Analyses COPD cases (total n=3226, with pack-years n=2517)							
Generation Scotland (GS:SFHS)	508	224 (44.1%)	58.9 (8.94)	64.84 (12.64)	0.580 (0.108)	482	29.32 (24.96)
British Regional Heart Study (BRHS)	425	425 (100%)	70.1 (5.46)	59.41 (14.66)	0.597 (0.084)	0	-
British Women's Heart and Health Study (BWHHS)	254	0 (0%)	69.3 (5.46)	64.26 (12.40)	0.603 (0.074)	203	28.1 (18.36)
UK COPD Cohort (UKCOPD) ‡	209	129 (61.7%)	68.7 (8.11)	37.94 (15.29)	0.447 (0.119)	199	50.07 (27.79
Hertfordshire Cohort Study (HCS)	317	203 (64.0%)	66.1 (2.79)	62.89 (13.57)	0.589 (0.101)	312	32.25 (23.37)
Biomarkers to target Antiiotic & Systemic Corticosteroid therapy in COPD exacerbations (COPDBEAT) ‡	87	62 (71.3%)	67.6 (8.77)	45.19 (16.24)	0.480 (0.115)	86	38.69 (21.24)
Nottingham COPD Study (NottCOPD) ‡	76	48 (63.2%)	67.2 (8.97)	50.29 (15.04)	0.482 (0.111)	74	49.02 (26.86)
Nottingham smokers	125	78 (62.4%)	63.1 (8.60)	46.27 (17.65)	0.503 (0.125)	124	41.75 (20.61)
Gedling study	33	26 (78.8%)	69.0 (8.23)	59.67 (16.81)	0.593 (0.103)	31	45.47 (33.40)
English Longitudinal Study of Aging (ELSA)	166	75 (45.2%)	66.0 (8.17)	54.84 (17.24)	0.526 (0.149)	0	-
EU COPD Gene Scan (EUCOPD) ‡	277	155 (56.0%)	67.0 (8.68)	38.51 (14.74)	0.467 (0.120)	277	46.43 (20.56)
GoTARDIS Study‡	749	412 (55.0%)	68.8 (8.97)	52.16 (14.14)	0.509 (0.110)	729	43.26 (21.59)
Discovery Analyses Controls (total n=4784, with pack-years n=3889)							
Generation Scotland (GS:SFHS)	961	552 (57.4%)	54.5 (8.41)	98.18 (10.92)	0.783 (0.051)	961	28.92 (16.86)
British 1958 Birth Cohort (1958BC)	1429	888 (62.1%)	44 (0)	100.90 (13.46)	0.809 (0.060)	1046	14.74 (10.07)
Oxford Biobank (OXBB)	1770	832 (47.0%)	41.6 (5.77)	-	-	1682	9.09 (9.34)
GoDARTS	624	402 (64.4%)	59.0 (10.75)	-	-	200	35.46 (25.89)

‡Sample collection is COPD case cohort

2.2.4.1 Single variant association analyses

Analyses of COPD risk: Due to the limited power to detect single variant associations for very rare SNPs, there was great underinflation of the test statistics in analyses of COPD risk, leading to genomic inflation factors (λ, equation (1-4)) of 0.255 and 0.291 in the pack-years adjusted analysis, and unadjusted analysis, respectively. Removing the very rare SNPs (overall MAF<0.05% ~MAC<8) resulted in a distribution of test statistics closer to expected (quantile-quantile (QQ) plots Figure 2-8), with corresponding λ values of 1.013 (pack-years adjusted) and 1.054 (unadjusted). A total of 4 SNPs in 3 regions met the $P<10^{-5}$ significance threshold in the pack-years adjusted analysis, with 5 SNPs in 4 regions showing $P<10^{-5}$ in the unadjusted analysis.

Figure 2-8: Quantile-quantile plots for analyses of COPD risk A. with and B. without pack-years adjustment. SNPs with MAF>0.05% only shown.



In the pack-years adjusted analysis of COPD risk (2517 cases and 3889 controls; Table 2-4 and Figure 2-9 A), the most significant association was in the 15q25 region. The sentinel SNP in this region was rs8034191 (MAF=34.8%, OR:1.374, $P=2.42\times10^{-7}$). This SNP is highly correlated ($r^2=0.93$) with rs1051730, a SNP previously identified through GWAS as being associated with smoking behaviour (132-134), lung cancer (151, 152), COPD(129) and airflow obstruction (defined as FEV_1 and FEV_1/FVC below the lower limit of normal (LNN)) in ever smokers (130). The pack-years adjusted analysis identified two novel signals of association with COPD ($P<10^{-5}$): a common nonsynonymous SNP, located within *SMPDL3B* (rs3813803, MAF=29.2%, OR:1.370,

P=1.04×10^{-6}) and a rare nonsynonymous SNP located within *MOCS3* (rs7269297, MAF=1.1%, OR:0.251, P=3.08×10^{-6}).

The unadjusted analysis of COPD risk (Table 2-4 and Figure 2-9 B) resulted in similar effect estimates for the 15q25 and *SMPDL3B* SNPs, with a slight augmentation of P-values (the 15q25 SNPs reached genome-wide significance), likely due to the increased sample size. The association with rs7269297 (*MOCS3*) identified in the pack-years adjusted analysis was less significant in the unadjusted analysis.

A further two loci were associated with COPD risk in the unadjusted analysis: a rare nonsynonymous SNP located within *PRICKLE1* (rs3827522, MAF=0.37% OR:0.123 P=1.03×10^{-7}) and rs17368582 (MAF=12.16%, OR:0.712, P=5.01×10^{-6}), a common synonymous SNP located in *MMP12,* a member of the matrix metalloproteinases family of genes which play a role in tissue remodelling. rs2276109, another SNP within *MMP12*, which is strongly correlated with rs17368582 (r^2=0.84), has previously been associated with COPD risk in smokers (137). The associations with the *PRICKLE1* and *MMP12* SNPs did not reach the P<10^{-5} significance threshold in the pack-years adjusted analysis, suggesting that these signals may in part be driven by differences in smoking behaviour between cases and controls.

Table 2-4: Top associations in exome analyses of COPD risk, with and without pack-years adjustment ($P<10^{-5}$).

SNPs ordered by chromosome (Chr) and genomic position (Pos). Only most significant SNP in each region shown. All P-values are two-sided. OR values reflect odds ratios after adjustments for age, sex (and pack-years in pack-years adjusted analysis only).

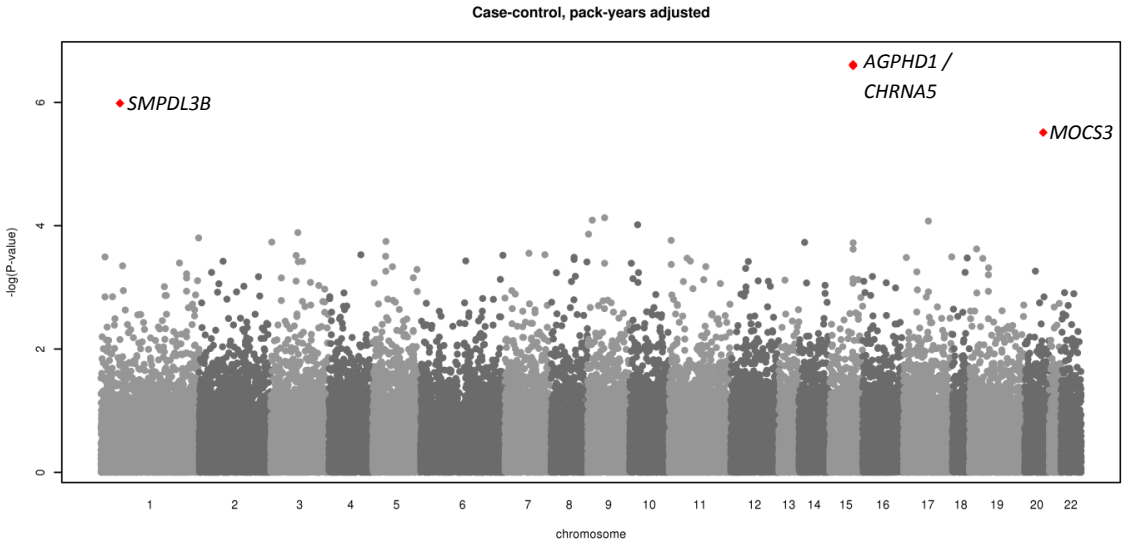
				Pack-years adjusted analysis				Unadjusted analysis				
				MAF (MAC)				MAF (MAC)				
rs no.	Chr	Pos	Effect / noneffect allele	Cases (n=2517)	Controls (n=3889)	OR	P*	Cases (n=3226)	Controls (n=4784)	OR	P*	Gene (Function)
rs3813803	1	28282292	C/T	30.61% (1541)	28.32% (2203)	1.370	**2.41×10^{-6}**	30.32% (1956)	28.45% (2722)	1.288	**2.11×10^{-6}**	*SMPDL3B* (nonsynonymous)
rs17368582	11	102738075	C/T	11.14% (561)	12.87% (1001)	0.767	3.22×10^{-3}	11.14% (719)	12.84% (1229)	0.712	**5.01×10^{-6}**	*MMP12* (synonymous)
rs3827522	12	42853871	A/G	0.22% (11)	0.35% (27)	0.184	1.39×10^{-3}	0.22% (14)	0.48% (46)	0.123	**1.03×10^{-7}**	*PRICKLE1* (nonsynonymous)
rs8034191	15	78806023	C/T	37.98% (1912)	32.73% (2546)	1.374	**2.42×10^{-7}**	37.69% (2432)	32.86% (3144)	1.364	**1.18×10^{-9}**	near *AGPHD*1 (intergenic)
rs7269297	20	49576664	G/T	0.74% (37)	1.41% (110)	0.251	**3.08×10^{-6}**	0.84% (54)	1.45% (139)	0.423	3.98×10^{-4}	*MOCS3* (nonsynonymous)

***P-values in bold significant at $P<10^{-5}$ level**

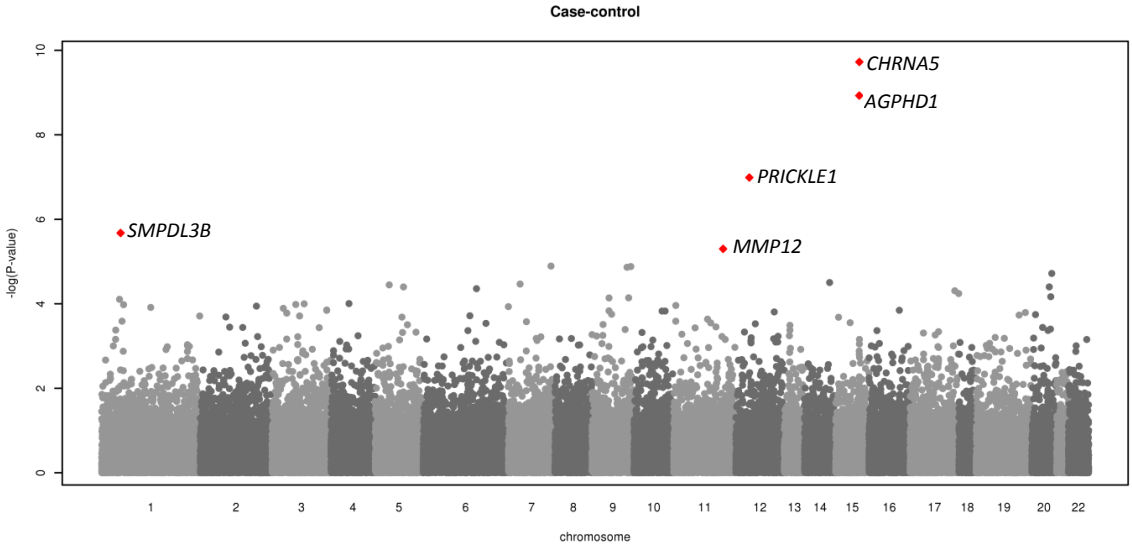
Figure 2-9 A) Analysis of COPD risk, with pack-years adjustment B) Analysis of COPD risk, without pack-years adjustment.

SNPs with MAF>0.05% only; SNPs with P<10^{-5} highlighted.

A.



B.



Analyses of airflow limitation: The genomic inflation factors for all variants in the pack-years adjusted and unadjusted analyses were 1.130 and 1.122, respectively (Figure 2-10). Appropriate genomic control was applied to adjust the standard errors of effect estimates. No SNPs reached the predefined $P<10^{-5}$ significance level in either the pack-years adjusted analysis, or the unadjusted analysis of airflow limitation (Figure 2-11). There were six SNPs which showed associations with $P<10^{-4}$ in these analyses, listed in Table 2-5. The strongest association in both the pack-years adjusted and unadjusted analyses was with rs77108843, a nonsynonymous SNP in RNA exonuclease 1 homolog (*REXO1*) gene (pack-years adjusted analysis: MAF=0.5% , beta=13.37 , $P=7.44\times10^{-5}$). Of note, rs28929474 the z-allele within the *SERPINA1* gene showed association in the unadjusted analysis (MAF=0.2%, Beta=-6.17, $P=2.83\times10^{-5}$); this gene is a well-established cause of alpha1-antitrypsin (AAT) deficiency, a rare inherited disorder that accounts for 1-2% of COPD cases (98, 129, 135). A further signal of association in the unadjusted analysis was with rs11749 (MAF=24.3%, beta=-1.97, $P=4.30\times10^{-5}$), a nonsynonymous SNP located within *DNALI1*¸ a potential candidate for immotile cilia syndrome (153).

Figure 2-10: Quantile-quantile plots for analyses of severity of airflow limitation A. with and B. without pack-years adjustment.
Plots include all SNPs passing genotype QC.

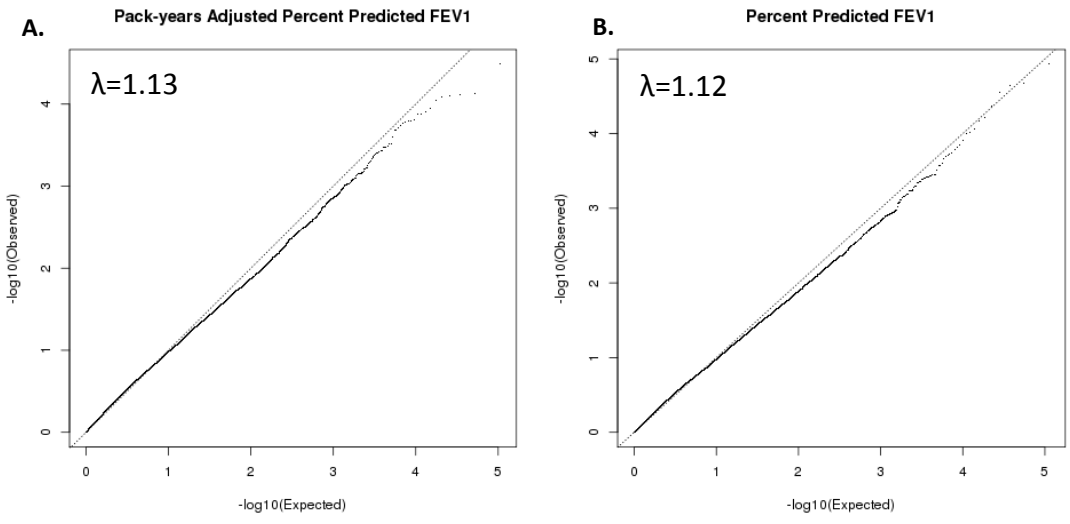
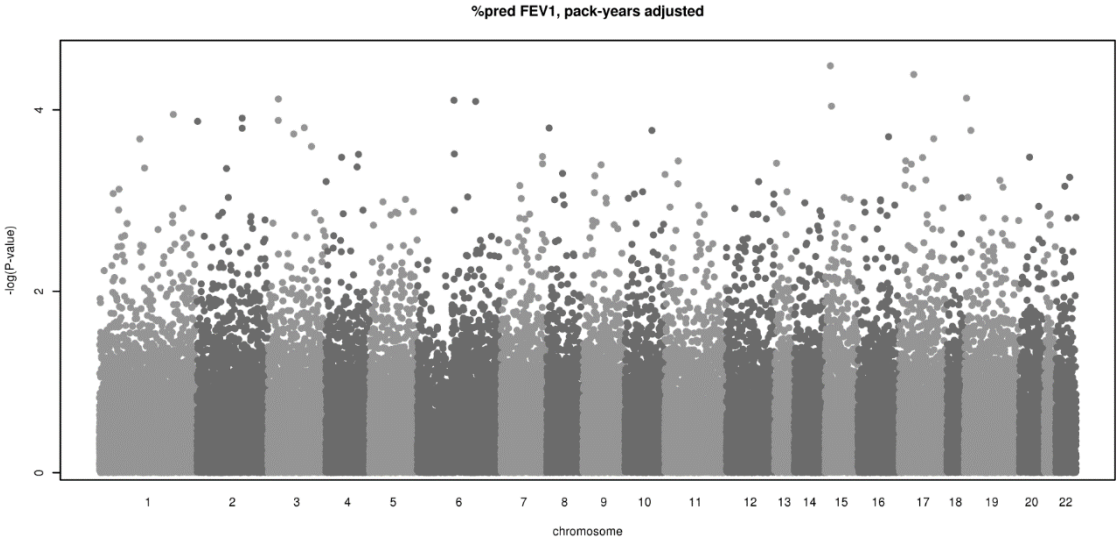


Figure 2-11. A) Manhattan for severity of airflow limitation analysis, adjusted for pack-years smoking, B) Manhattan for severity of airflow limitation analysis without adjustments for pack-years smoking.
Plots include all SNPs passing genotype QC.

A)



B)

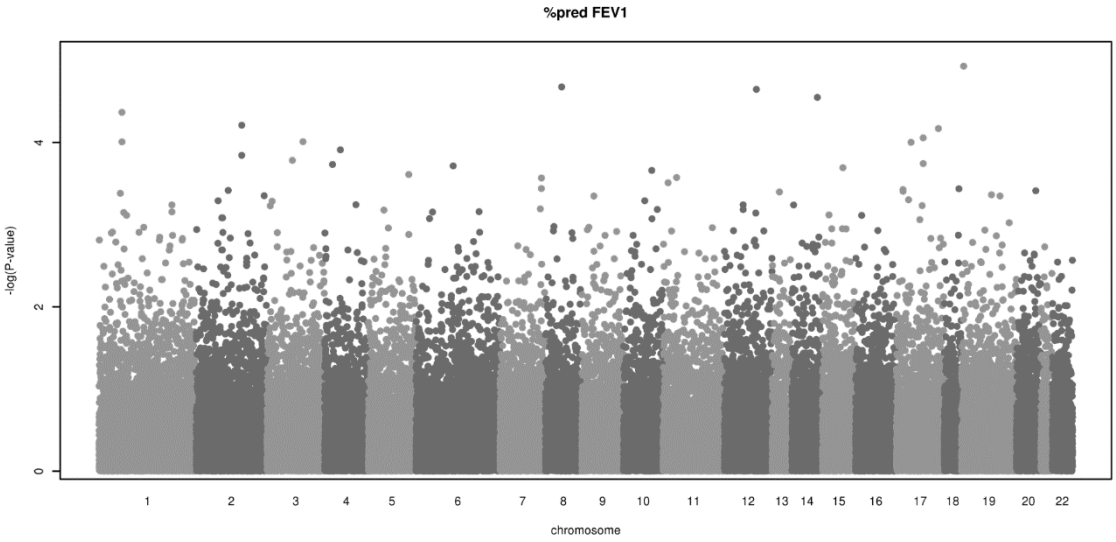


Table 2-5: Top associations in exome analysis of airflow limitation, with and without adjustment for pack-years smoking ($P<10^{-4}$).

SNPs ordered by chromosome (Chr) and genomic position (Pos). Only most significant SNP in each region shown. All P-values are two-sided. Beta values reflect effect size estimates on percent predicted FEV_1 (after adjustments for pack-years in pack-years adjusted analysis only).

				%predicted FEV_1, adjusted for pack-years (n=2517)			%predicted FEV_1 (n=3226)			
rs no.	Chr	Pos	Effect / noneffect allele	MAF (MAC)	Beta	P	MAF (MAC)	Beta	P	Gene (Function)
rs11749	1	38023316	T/C	24.08% (1212)	-1.57	3.99×10^{-3}	24.26% (1565)	-1.97	4.30×10^{-5}	*DNALI1*(nonsynonymous)
rs59035258	8	65527669	T/C	3.72% (187)	4.18	8.75×10^{-4}	3.60% (232)	4.79	2.11×10^{-5}	*CYP7B1*(nonsynonymous)
rs117991621	12	96379884	T/C	0.56% (28)	10.92	6.19×10^{-4}	0.56% (36)	12.02	2.26×10^{-5}	*HAL*(nonsynonymous)
rs28929474	14	94844947	T/C	2.17% (109)	-5.05	1.30×10^{-3}	1.97% (127)	-6.17	2.83×10^{-5}	*SERPINA1*(nonsynonymous)
rs147487857	15	41247629	G/A	1.25% (63)	-8.90	3.27×10^{-5}	1.26% (81)	-5.62	3.19×10^{-3}	*CHAC1*(nonsynonymous)
rs77108843	19	1828148	A/G	0.50% (25)	13.37	7.44×10^{-5}	0.59% (38)	12.09	1.18×10^{-5}	*REXO1*(nonsynonymous)

2.2.4.2 Gene-based Association analyses

In the gene-based analyses of COPD risk, *PRICKLE1* was the only gene to reach the $P<10^{-5}$ significance level ($P=1.968\times10^{-6}$, analysis with no pack-years adjustment). The SKAT-O test utilised three SNPs within this gene (Table 2-6), however "drop-one" analyses, where the SKAT-O P-value is recalculated when individual SNPs are sequentially excluded from the test, showed the signal to be largely driven by rs3827522, the SNP identified in the single variant analysis (COPD risk unadjusted for pack years, $P=1.03\text{x}10^{-7}$). Consequentially, the single SNP, but not the gene was followed up further (Section 2.3.3).

The analyses of airflow limitation identified no associated genes with $P<10^{-5}$. Three genes showed modest association, with P-values$<10^{-4}$, however all were driven by single SNPs which had individual signals of association with $P<10^{-4}$.

Table 2-6: Risk of COPD single variant association results of SNPs included in SKAT-O test of *PRICKLE1*.

		Risk of COPD with pack-years adjustment				Unadjusted analysis of risk of COPD				Drop-one result	
		MAF (MAC)		Association result		MAF (MAC)		Association result			
rs no.	Effect / noneffect allele	Cases (n=2517)	Controls (n=3889)	OR	P	Cases (n=3226)	Controls (n=4784)	OR	P	SKAT-O Analysis utilising all SNPs in *PRICKLE1*	P-value of SKAT-O Analysis if SNP removed
rs3827522	A/G	0.22% (11)	0.35% (27)	0.628	1.38×10^{-3}	0.22% (14)	0.48% (46)	0.123	1.03×10^{-7}	Unadjusted $P=1.968\times10^{-6}$ Pack-years adjusted $P=2.086\times10^{-3}$	unadjusted:P=0.0484; adjusted: P=0.0258
rs146199468	G/T	0.02% (1)	0.00% (0)	-	-	0.02% (1)	0.00% (0)	-	-		unadjusted:$P=1.973\times10^{-6}$; adjusted: $P=4.347\times10^{-3}$
rs79087668	T/C	0.20% (10)	0.32% (25)	0.617	0.999	0.20% (13)	0.30% (29)	0.364	4.44×10^{-2}		unadjusted:$P=1.693\times10^{-9}$; adjusted: $P=1.367\times10^{-3}$

2.3 Meta-Analysis with UK BiLEVE data and look-up of SNPs identified in discovery analysis.

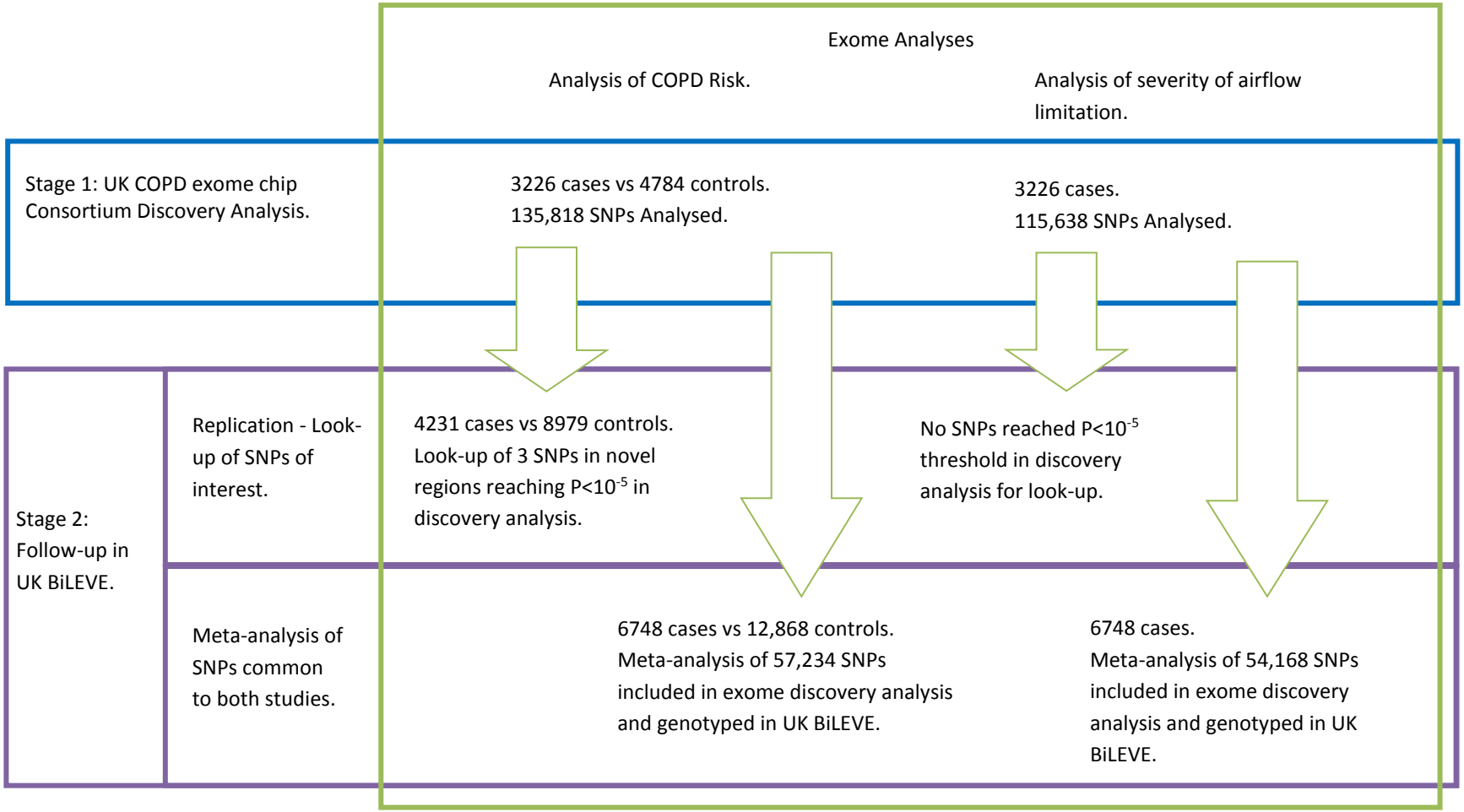
2.3.1 UK BiLEVE study

The UK BiLEVE study is a collection of 48,943 individuals aged 39-72 from UK Biobank with high quality lung function and smoking data. Details of the sample selection for UK BiLEVE are fully described in (120). In brief, 50,008 samples of European ancestry and with acceptable and reproducible spirometry measures, based on European Respiratory Society / American Thoracic Society (ERS/ATS) guidelines, were selected from the extremes and middle of the distributions of percent predicted FEV_1, separately in never smokers and in heavy smokers. DNA was extracted and genotyping was undertaken with the custom-designed Affymetrix Axiom UK BiLEVE array, including variants which were selected from the same sequencing project as the Illumina Human Exome BeadChip. Following thorough variant and sample QC, 48,943 unrelated individuals remained.

I utilised data from the UK BiLEVE study in two ways. Firstly, I carried out a look-up of all SNPs in novel regions that were identified in the single variant analyses, in an attempt to replicate these findings. Secondly, I selected a subset of SNPs that were included on both the Illumina Human Exome Beadchip and the Affymetrix Axiom UK BiLEVE array, and undertook a meta-analysis, as a more powerful discovery study (Figure 2-12).

For these analyses, I selected cases and controls from amongst the 24,457 heavy smokers in UK BiLEVE (average 35 pack-years). 4231 individuals with airflow limitation consistent with GOLD 2 COPD or worse were selected as cases, alongside 8979 controls with $FEV_1/FVC>0.7$, percent predicted $FEV_1>80\%$, and no doctor diagnosis of COPD. All spirometry measures were pre-bronchodilator (reversibility testing was not carried out), and individuals with a doctor diagnosis of asthma or other lung diseases were excluded.

Figure 2-12: Overview of discovery analyses and follow-up in UK BiLEVE.

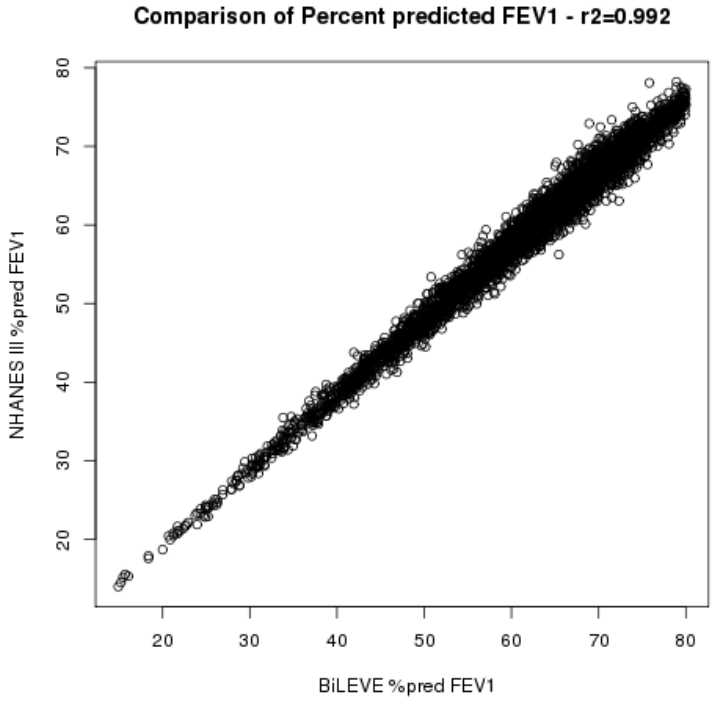


2.3.2 Meta-analyses: Methods

2.3.2.1 Phenotypes

The 4231 cases with airflow limitation indicative of COPD and 8979 controls from UK BiLEVE contributing to the meta-analysis, were selected based on their % predicted FEV_1 values, calculated using reference equations derived using healthy never smokers in the whole of UK Biobank (120). For association testing, percent predicted FEV_1 was recalculated using the NHANES III spirometric reference equations (85) for consistency with the exome discovery analyses. The two methods for calculating percent predicted FEV_1 gave highly correlated values (r^2=0.992) and all selected COPD cases met GOLD 2 criteria (FEV_1/FVC<0.7 and % predicted FEV_1<80%) under both reference equations (Figure 2-13).

Figure 2-13: Comparison of % predicted FEV_1 based on predictive values calculated using healthy never smokers in UK BiLEVE, versus predictive values calculated using NHANES III spirometric reference equations.



2.3.2.2 Single Variant Association analyses

SNP associations with COPD risk were carried out using a logistic regression model, adjusting for age, sex and pack-years and assuming an additive genetic model (as equation (2-1)). For the analysis of severity of airflow limitation, associations with

untransformed percent predicted FEV_1 in cases were tested using a linear regression model, with adjustment for pack-years (as equation (2-2)). All association analyses were undertaken using PLINK v1.07 (18).

2.3.2.3 Meta-analyses of UK COPD exome chip consortium discovery samples and UK BiLEVE samples

The genomic inflation factor (λ, equation (1-4)) was calculated for both the discovery exome analyses and the UK BiLEVE analyses and where λ>1, genomic control was applied, adjusting the standard errors of effect estimates. All SNPs were oriented to the same strand, with consistent effect alleles. Effect estimates were combined across the two analyses using an inverse-variance–weighted meta-analysis:

$$\beta_{meta_j} = \frac{\sum_{k=1}^{2} w_{jk}\beta_{jk}}{\sum_{k=1}^{2} w_{jk}} \qquad (2\text{-}3)$$

where β_{jk} is the estimated effect and w_{jk} is the weight of j^{th} SNP of the k^{th} analysis:

$$w_{jk} = \frac{1}{se_{jk}^2}$$

where se_{jk} is the standard error of the effect estimate for the j^{th} SNP of the k^{th} analysis. Pooled standard errors were estimated as:

$$se_{meta_j} = \frac{1}{\sqrt{\sum_{k=1}^{2} w_{jk}}}$$

λ was calculated for the pooled effect estimates and genomic control was applied again where λ>1. Meta-analysis statistics and figures were produced using R version 3.1.1.

2.3.3 Meta-analyses: Results

The characteristics of the UK BiLEVE samples used in these analyses are summarised in Table 2-7.

2.3.3.1 Replication of signals identified in discovery exome analyses

A total of six SNPs, within five regions were identified in the discovery exome analyses of COPD risk (Table 2-4). For the three SNPs in regions not previously implicated in COPD risk, a look-up within the UK BILEVE single variants association analysis was carried out (Table 2-8). Only rs7269297 in *MOCS3* showed evidence for association with COPD risk in UK BiLEVE, consistent with the discovery stage analysis (OR:0.742, P=0.019); this result is just above the Bonferroni corrected level of significance (P<0.017, corrected for 3 SNPs tested).

Table 2-7: Characteristics of UK BiLEVE samples used in meta-analyses with discovery samples and for replication of signals identified in discovery analyses.

Sample Collection	n	Sex	Age	Percent Predicted FEV_1	FEV_1/FVC	Pack-years	
		Male, n (%)	Mean (SD)	Mean (SD)	Mean (SD)	Samples with data (n)	Mean (SD)
UK BiLEVE Samples (Meta-analysis and replication)							
Airflow limitation cases	4231	2379 (56.2%)	59.54 (6.86)	61.76 (11.8)	0.607 (0.076)	4231	42.41 (21.10)
Controls	8979	4260 (47.4%)	56.19 (7.92)	101.40 (8.1)	0.773 (0.038)	8979	30.43 (14.41)

Table 2-8: Look-up within UK BiLEVE single variant associations, for SNPs in novel regions identified in discovery exome analyses.

				Discovery analysis of COPD risk			Analysis in UK BiLEVE (pack-years adjusted)				
					Association Result		MAF (MAC)		Association Result		
rs no.	Chr	Pos	Coded / other allele	Analysis in which SNP identified	OR	P*	Cases (n=4231)	Controls (n=8979)	OR	P	Gene (Function)
rs3813803	1	28282292	C/T	pack-years adjusted	1.370	2.41×10^{-6}	28.72% (2418)	29.43% (5269)	0.968	0.2984	*SMPDL3B* (nonsynonymous)
rs3827522	12	42853871	A/G	unadjusted	0.123	1.03×10^{-7}	0.25% (21)	0.25% (45)	0.907	0.7309	*PRICKLE1* (nonsynonymous)
rs7269297	20	49576664	G/T	pack-years adjusted	0.2501	3.08×10^{-6}	1.16% (98)	1.41% (252)	0.742	0.0193	*MOCS3* (nonsynonymous)

2.3.3.2 Meta-Analysis of discovery and UK BiLEVE data

Analyses of COPD risk: For the 57,234 SNPs common to both the UK COPD exome chip consortium samples and the UK BiLEVE study, a meta-analysis of the discovery pack-years adjusted analysis and the UK BiLEVE study results was undertaken (total 6748 cases and 12,868 controls). The strongest association in the meta-analysis was with the SNPs in the 15q25 region, identified in the discovery analysis (sentinel SNP rs8034191 near *AGPHD1*). Three further regions, not identified in the discovery analysis, showed association with risk of COPD ($P<10^{-5}$) in the meta-analysis (Figure 2-14 and Table 2-9). The *GYPA/HHIP* and *GPR126* regions have previously been reported as showing association with lung function and COPD or airflow limitation risk (116, 129, 130). The *IFIT3* region signal (rs140549288 in *IFIT3*, MAF=0.67%, OR=1.92 $P=7.49x10^{-6}$) represents a novel rare variant signal of association with COPD.

Figure 2-14: Meta-analysis of COPD risk in discovery and UK BiLEVE samples.

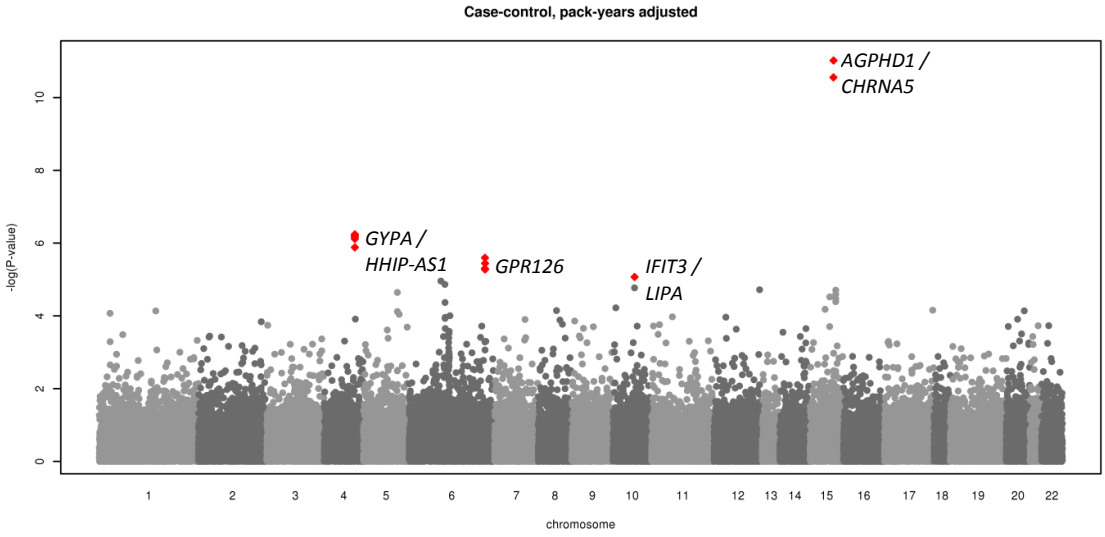


Table 2-9: SNPs with P<10^{-5} in the meta-analysis.

SNPs ordered by chromosome (Chr) and genomic position (Pos). Only most significant SNP in each region shown. All P-values are two-sided. OR values reflect odds ratios after adjustments for age, sex and pack-years.

				Discovery pack-years adjusted analysis				UK BiLEVE pack-years adjusted analysis				Meta-analysis of discovery and UK BiLEVE pack-year adjusted analyses.		
				MAF (MAC)		Association Result		MAF (MAC)		Association Result		Association Result		
rs no.	Chr	Pos	Coded / other allele	Cases (n=2517)	Controls (n=3889)	OR	P*	Cases (n=4231)	Controls (n=8979)	OR	P*	OR	P*	Gene
rs1828591	4	145480780	A/G	35.64% (1794)	39.11% (3042)	0.917	0.153	36.55% (3088)	40.00% (7171)	0.867	**9.88×10^{-7}**	0.876	**5.75×10^{-7}**	*GYPA / HHIP* (intergenic)
rs4896582	6	142703877	A/G	29.26% (1473)	31.73% (2468)	0.859	0.018	28.04% (2349)	30.20% (5344)	0.879	3.87×10^{-5}	0.875	**2.53×10^{-6}**	*GPR126* (intronic)
rs140549288	10	91099466	C/G	0.76% (38)	0.57% (44)	2.156	0.037	0.94% (79)	0.56% (100)	1.880	6.87×10^{-5}	1.924	**8.56×10^{-6}**	*IFIT3* (exonic), *LIPA* (intronic)

***P-values in bold significant at P<10^{-5} level**

Analyses of severity of airflow limitation: 54,168 SNPS were included in the meta-analysis of severity of airflow limitation. One SNP showed association with $P<10^{-5}$: rs140198372, at a splice site in *SERPINA12* (MAF=0.04%, Beta=-33.51, $P=5.72x10^{-6}$, Figure 2-15 and Table 2-10).

Figure 2-15: Meta-analysis of severity of airflow limitation in discovery and UK BiLEVE samples.

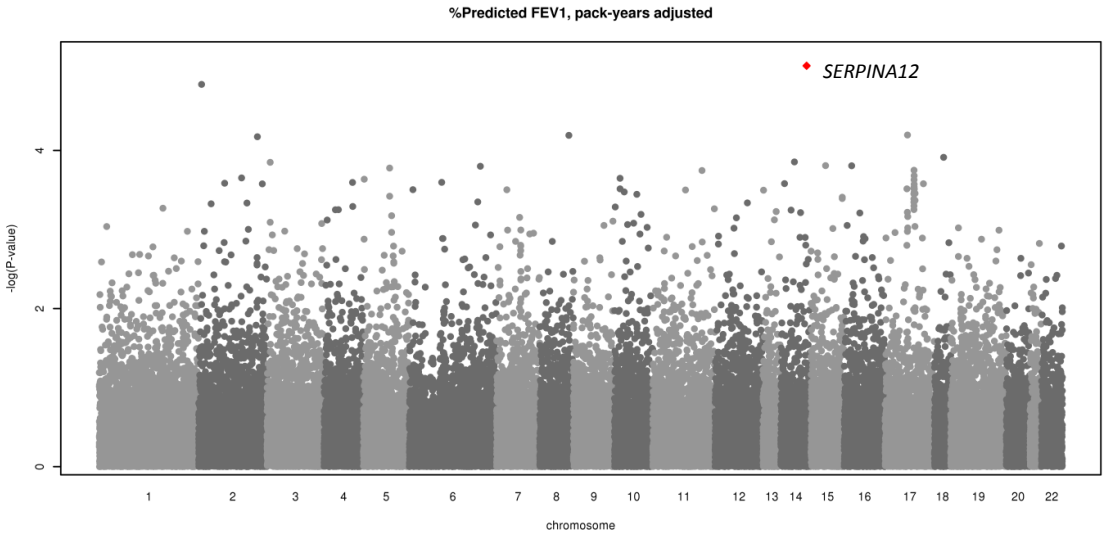


Table 2-10: Top associations ($P<10^{-5}$) in meta-analysis of severity of airflow limitation.

Only most significant SNP in region shown. All P-values are two-sided. Beta values reflect effect size estimates on percent predicted FEV_1 after adjustments for pack-years.

				Severity of airflow limitation, adjusted for pack-years (n=2517)			UK BiLEVE pack-years adjusted analysis (n=4231)			Meta-analysis of discovery and UK BiLEVE pack-year adjusted analyses.		
rs no.	CHR	Position	Coded / other allele	MAF (MAC)	Beta	P	MAF (MAC)	Beta	P	Beta	P	Gene
rs140198372	14	94953832	A/C	0.062% (4)	-26.44	1.80×10^{-3}	0.012% (1)	-38.35	4.11×10^{-4}	-33.51	**5.72×10^{-6}**	*SERPINA12* (splice site)

***P-values in bold significant at $P<10^{-5}$ level**

2.4 Sensitivity analyses to assess COPD case criteria

Clinical diagnosis of airflow limitation in COPD, is based on post-bronchodilator spirometry (94). Of the 3226 COPD cases defined as described above, 1398 had GOLD 2 or worse airflow limitation based on post-bronchodilator spirometry values; for the remainder of samples, these data were unavailable. To assess the potential misclassification bias this could give rise to, a sensitivity analysis was carried out for all SNPs identified in the discovery or meta-analyses of COPD risk, by repeating the discovery analyses including only those 1398 COPD cases who underwent reversibility testing. This sensitivity analysis resulted in estimated effect sizes that were consistent with the original analyses (Table 2-11 and Figure 2-16). In particular, the odds ratios were not substantially altered for rs7269297 in *MOCS3* (sensitivity analysis OR:0.276; original discovery OR:0.251), nor rs140549288 in *IFIT3* (sensitivity analysis OR:2.554; original discovery OR:2.156). The P-values for both of these SNPs were somewhat attenuated in the sensitivity analysis, however this is likely to be largely a result of the reduced sample size.

Table 2-11: Sensitivity analysis to assess COPD case criteria of SNPs identified in either the a. discovery, or b. meta-analyses of COPD risk.

Results for original analyses, and for analyses where cases restricted to include only those with known irreversible airflow limitation

a. SNPs identified in discovery analyses

					Discovery pack-years adjusted analysis (2517 cases, 3889 controls)		Discovery pack-years adjusted analysis (1365 COPD cases with reversibility testing, 3889 controls)		Discovery unadjusted analysis (3226 cases, 4784 controls)		Discovery unadjusted analysis (1398 COPD cases with reversibility testing, 4784 controls)	
rs no.	CHR	Position	Coded / other allele	Gene	OR	P	OR	P	OR	P	OR	P
rs3813803	1	28282292	C/T	*SMPDL3B* (nonsynonymous)	1.37	2.41×10^{-6}	1.46	5.18×10^{-6}	1.288	2.11×10^{-6}	1.382	2.98×10^{-6}
rs17368582	11	102738075	C/T	*MMP12* (synonymous)	0.767	3.22×10^{-3}	0.673	1.08×10^{-3}	0.712	5.01×10^{-6}	0.6567	2.35×10^{-5}
rs3827522	12	42853871	A/G	*PRICKLE1* (nonsynonymous)	0.184	1.39×10^{-3}	0.272	1.08×10^{-3}	0.123	1.03×10^{-7}	0.1836	1.43×10^{-4}
rs8034191	15	78806023	C/T	near *AGPHD1* (intergenic)	1.374	2.42×10^{-7}	1.42	8.14×10^{-6}	1.364	1.18×10^{-9}	1.414	1.33×10^{-7}
rs7269297	20	49576664	G/T	*MOCS3* (nonsynonymous)	0.251	3.08×10^{-6}	0.276	4.05×10^{-4}	0.423	3.98×10^{-4}	0.4502	0.0118

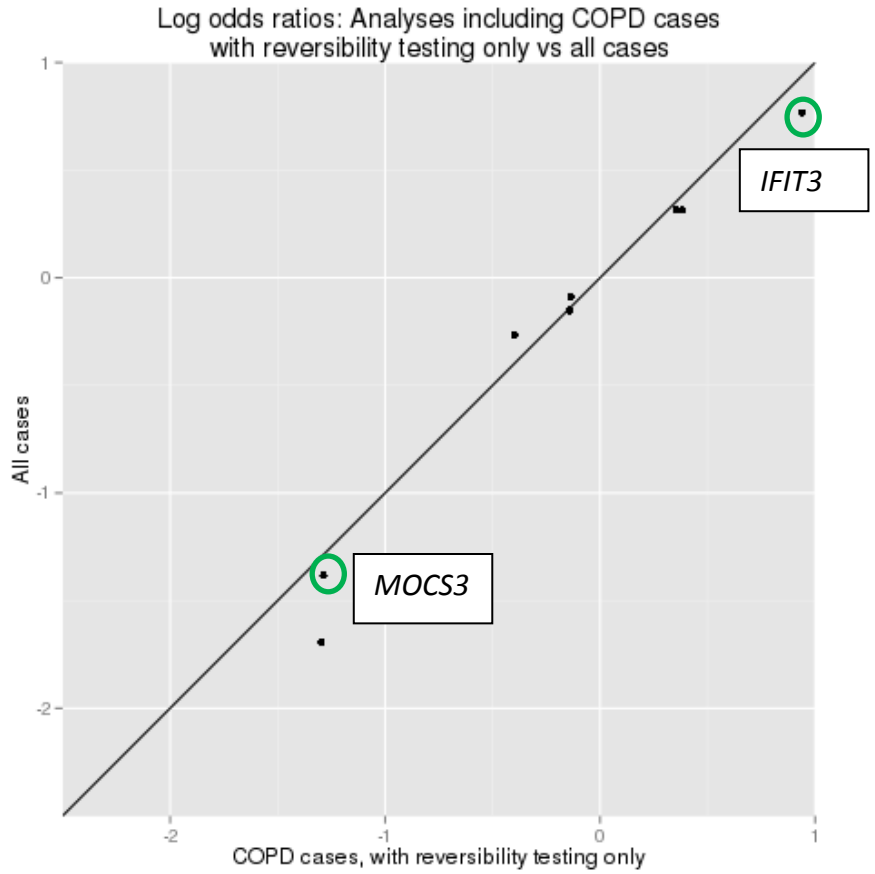
b. SNPs identified in meta-analyses

					Discovery pack-years adjusted analysis (2517 cases, 3889 controls)		Discovery pack-years adjusted analysis (1365 COPD cases with reversibility testing, 3889 controls)		Discovery unadjusted analysis (3226 cases, 4784 controls)		Discovery unadjusted analysis (1398 COPD cases with reversibility testing, 4784 controls)	
rs no.	CHR	Position	Coded / other allele	Gene	OR	P	OR	P	OR	P	OR	P
rs1828591	4	145480780	A/G	*GYPA / HHIP* (intergenic)	0.917	0.153	0.873	0.0866	0.919	0.093	0.8821	0.0566
rs4896582	6	142703877	A/G	*GPR126*	0.859	0.018	0.868	0.0861	0.864	5.95×10^{-3}	0.8702	0.0427
rs140549288	10	91099466	C/G	*IFIT3* (exonic), *LIPA* (intronic)	2.156	0.037	2.554	0.0640	1.823	0.057	2.211	0.0480

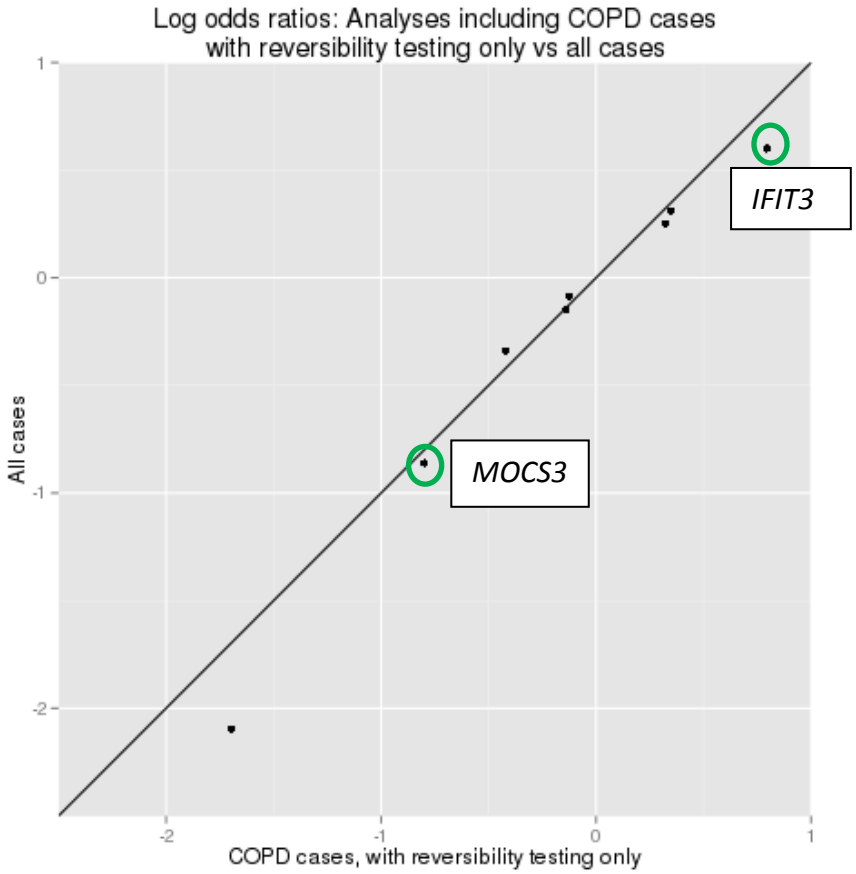
Figure 2-16: Comparison of effect estimates of discovery case-control analysis of COPD risk where the cases were restricted to only include those with known irreversible airflow obstruction, versus the analysis including all COPD cases.

A. Pack-years adjusted analysis of COPD risk. B. Analysis of COPD risk, without adjustment for pack-years. Highlighted are the effect estimates of the two SNPs reported as novel regions (rs7269297 in *MOCS3* and rs140549288 in *IFIT3)*.

A. Discovery analysis of COPD risk, adjusted for pack-years



B. Discovery analysis of COPD risk, without adjustment for pack-years



2.5 Discussion

This chapter describes the analyses of exome chip variants with COPD risk and percent predicted FEV_1 among cases in an attempt to identify low frequency and rare functional variants which may play a role in susceptibility to COPD, and severity of airflow limitation.

Firstly, I carried out discovery analyses utilising 3226 samples from 12 studies in the UK COPD exome chip consortium, and 4784 general population controls. The most significant associations identified in the discovery analysis of COPD risk were with SNPs in the 15q25 region, previously identified through GWAS as being associated with smoking behaviour (132-134), lung cancer (151, 152), COPD (129) and airflow obstruction (130). These discovery analyses also provided independent replication of the previously reported association of *MMP12* with COPD risk (137). Associations with COPD risk were also identified at three loci, not previously implicated in lung function or COPD: *SMPDL3B*, *MOCS3*, and *PRICKLE1*. There was only evidence to support the association at *MOCS3* (rs7269297, Serine to Alanine, MAF=1.3%, $P_{discovery}=3.08x10^{-6}$) in the follow-up analyses in UK BiLEVE, however ($P_{rep}=0.019$). The protein encoded by *MOCS3* adenylates and activates molybdopterin synthase, an enzyme required to synthesize molybdenum cofactor (154).

The discovery analysis of airflow limitation identified no SNPs reaching the predefined significance level of $P<10^{-5}$. Just falling sort of this significance level, the z-allele (rs28929474) within the *SERPINA1* gene was associated with a lower percent predicted FEV_1 in cases (unadjusted analysis: $\beta=-5.053$, $P_{discovery}=2.83x10^{-5}$). As well as being a well-established cause of alpha1-antitrypsin (AAT) deficiency, (98, 129, 135), this SNP has also previously been associated with an increased annual decline in FEV_1 in a general population sample (155) and increased airflow limitation in COPD cases (156). In these analyses, the z-allele was associated with an increased risk of COPD, although this was not statistically significant (OR:1.270, P=0.252). The likely reason for the lack of a significant association with this known COPD locus, is that some of the case collections excluded individuals with AAT deficiency, resulting in selection bias.

For a subset of 57,234 SNPs, I combined association results from the COPD exome chip consortium and UK BiLEVE samples in a meta-analysis. As in the original discovery analyses, the strongest associations in the meta-analysis of COPD risk were with SNPs in the 15q25 region. Associations in two further previously reported COPD regions were also identified, at *HHIP* (116, 129) and *GPR126* (130). There was additionally one SNP in a novel region, *IFIT3,* which showed significant association with COPD risk (rs140549288, Valine to Leucine, MAF=0.7%, $P_{meta}=8.56x10^{-6}$). *IFIT3* is associated with interferon-alpha (IFN-α) antiviral activity and has been found to be up-regulated in respiratory syncytial virus infection (157) and in human lung epithelial cells infected with dengue virus (158). rs140549288 is also located within in an intron of *LIPA*, which is transcribed from the opposite strand; the product of this gene is involved in the hydrolysis of cholesteryl esters and triglycerides and other SNPs within this gene have previously been associated with coronary artery disease (159).

Through the meta-analysis of airflow limitation, an association was identified with a very rare SNP within a serine protease inhibitor gene, *SERPINA12*, that has not previously been associated with COPD (rs140198372, MAF=0.03%, $P_{meta}=5.72x10^{-6}$). *SERPINA12* has been associated with cardiovascular diseases, being implicated in obesity and type 2 diabetes (160).

The lack of strong statistical findings in these analyses might be due to limited statistical power. For example, for a SNP with a MAF of 1% and an OR=2, there is just 54% power to detect an association in the discovery analyses described in this chapter, if a significance level of $P<3.8x10^{-7}$ ("exome-wide significant") is assumed. Due to the limited power of single variant association tests for rare variants, several statistical tests have been developed which test the joint effects of several variants within a gene, as described in Section 1.2.5. In these analyses, I utilised the SKAT-O test, a method which optimally combines both a burden style test, which assumes all variants act in the same direction and with similar magnitude of effects, and the SKAT test, which allows for variants within the gene to act in opposing directions and with different effect sizes (69). Where the underlying genetic architecture is unclear, this optimal test approach should increase the chance of detecting genes influencing disease, without increasing the multiple testing burden. In these analyses, only one

gene was identified through the SKAT-O tests as meeting the elected significance level ($P<10^{-5}$). After subsequent analysis however, it was found that this gene-based signal in *PRICKLE1* was driven by the single SNP which was identified in the single variant discovery analyses, but whose association was not replicated in the UK BiLEVE samples.

The GOLD Global Strategy for Diagnosis, Management, and Prevention of COPD (94) states that airflow limitation in COPD should be determined using post-bronchodilator spirometry (94). In attempt to maximise power, the analyses described in this chapter included some cases that had pre-bronchodilator spirometry measures only; for these samples it could not be determined whether their airflow limitation was reversible, and so a proportion of these cases may not have met the clinical definition of COPD. To assess the effect of the inclusion these samples for whom reversibility testing was not undertaken, a sensitivity case-control analysis was performed, with cases restricted to the subset of 1398 individuals who were taken from COPD cohorts and had known irreversible airflow limitation. In this sensitivity analysis, the effect estimates of the top hits did not substantially change, which suggests that the broader case definition, including samples that did not undergo reversibility testing, did not result in substantial misclassification bias. This is supported by findings from another genetic study which defined cases using pre-bronchodilator spirometry (127); estimated effects for their sentinel SNPs did not substantially differ where the analyses were restricted to individuals with known COPD. Furthermore, this study estimated that 98% of samples defined as having GOLD 2 COPD based on pre-bronchodilator values, also met GOLD 2 criteria based on post-broncholidator spirometry measures.

In summary, the analyses described in this chapter identified potentially interesting associations between COPD risk and low frequency SNPs in *MOCS3* and *IFIT3,* two regions not previously implicated in COPD or lung function. Furthermore, an association was identified with percent predicted FEV_1 in individuals with COPD, and a very rare SNP in *SERPINA12*. These three regions warrant further investigation as they may provide insight into the underlying biological mechanisms of COPD and airflow limitation in smokers. These analyses also independently replicated associations with known COPD loci at *MMP12, HHIP*, *GPR126* and in the 15q25 region, and provided

further evidence that the z-allele within *SERPINA1* may be related to severity of airflow limitation in COPD. These analyses support the hypothesis that low frequency variation, alongside common genetic variants, play a role in COPD; however further follow-up would be required to formally verify the associations identified here and larger sample sizes are necessary to comprehensively assess the genetic contribution of rare variation.

Chapter 3 Investigation of methods: Meta-analysis of gene-based tests

3.1 Introduction

Detection of trait-associated SNPs has generally demanded large sample sizes, a requirement which is of particular importance where associated SNPs are low frequency or rare (161, 162). Such large sample sizes have commonly been attained through meta-analyses of data from a number of studies within a consortium setting. These meta-analyses involve each study undertaking analyses of individual level data according to a centrally agreed analysis plan, with only summary statistics being shared with the central meta-analysts. Methods for combining the summary statistics for single variant associations are well established and have been widely employed for a range of complex traits (14, 163).

With the availability of novel genotyping arrays such as the exome chip (50), and with the falling costs of sequencing, there has developed a greater focus on examining SNPs at the lower end of the frequency spectrum, and in particular an increase in the popularity of gene-based tests, thought to be more powerful than single variant tests for low frequency variants (Section 1.2.5) (162). Until recently, consortia wishing to adopt these gene-based tests in a meta-analysis setting would have required each study to undertake gene-based analyses using individual level data, with the P-values from these gene-based tests being combined centrally using Fisher's (164) or Z-score (165) methods.

Several software packages have now been developed which allow the meta-analysis of gene-based tests, without sharing individual level data (74-78, 166). These packages utilise a common method, which involves each study generating score statistics for each single variant, along with variance-covariance matrices, which describe the correlation between the variants. Gene-based tests may then be constructed centrally, allowing for flexibility in specifying the SNPs used in these tests, in terms of MAF frequency, function and so on. Furthermore, most of these packages also allow for conditional analyses to be carried out, with no additional analyses required at study level; these conditional analyses may be used to determine whether any gene-based associations are a result of an association with a single variant.

In this chapter I describe the methods used for the meta-analysis of gene-based tests and recount a number of the software packages and programs developed to undertake these tests. I then describe in more detail the simulations undertaken to evaluate one of these methods, RAREMETAL (74, 166) (the program used in the analyses described in Chapter 4), as undertaken by the program's authors. I further explore RAREMETAL using real genotype and phenotype data from 48,943 individuals from UK BiLEVE (120), a dataset far larger than has been used previously for the evaluation of these methods (74). The main aim of these analyses was to compare the concordance of the meta-analysis methods with analyses carried out using individual level data. I also examine the Fisher's and Z-score methods of combining P-values, to ascertain the benefit of using the new meta-analysis methods.

3.2 Methods for meta-analysis of gene-based tests

Programs developed to date for the meta-analysis of gene-based tests include both the burden-test methods (weighted sum test (WST) and variable threshold test (VTT)), and variance-component methods which compare genetic similarity (SKAT and SKAT-O), as described in Section 1.2.5. Both the WST and VTT tests broadly involve combining information from all variants under a given MAF threshold within a gene into a single quantity, which can then be used for association testing with a trait. The WST test allows for the incorporation of weights; for example, it can allow greater weightings to rarer SNPs. The VTT test performs the burden test using several MAF thresholds, selecting a threshold for each gene, to give an optimal P-value. These burden tests perform well where the effects of causal variants within a gene are similar in terms of both direction and magnitude. The SKAT test in comparison is powerful where a gene has a combination of protective, deleterious and neutral variants. A further method, SKAT-O combines both the WST test and SKAT, by providing a weighted average of the two tests, optimised for each gene.

In equations (1-7) to (1-10) in Section 1.2.5, it was shown that the WST, VTT, SKAT and SKAT-O could all be constructed using the score statistic (U, equation (1-5)). To construct these gene-based tests for a gene with m variants, in a meta-analysis of K

studies, the MAFs, a vector of score statistics ($\boldsymbol{U_k} = [U_{1k}, U_{2k}, \dots, U_{mk}]^T$) and corresponding variance-covariance matrix ($\boldsymbol{V_k}$, equation (1-6)) is required from each study, *k*.

All gene-based tests may then be constructed as follows.

Meta-WST:

$$Q_{meta-WST} = \sum_{k=1}^{K}\sum_{j=1}^{m} w_j U_{jk} \sim N\left(0, \sum_{k=1}^{K}\sum_{j=1}^{m} w_j^2 V_{jjk}\right) \quad \textbf{(3-1)}$$

where U_{jk} is the score statistic and w_j is the weight for the j^{th} variant in the k^{th} study and V_{jjk} is the variance of U_{jk}.

Meta-VTT:

$$Q_{meta-T}(F) = \sum_{k=1}^{K}\sum_{\mathrm{j}=1}^{m} T_F U_{jk} \quad \textbf{(3-2)}$$

$$Q_{meta-VTT} = \max(Q_{meta-T}(F)) \sim N\left(0, \sum_{k=1}^{K}\sum_{j=1}^{m} T_F \mathrm{V_{jjk}}\right)$$

where T_F is an indicator variable that equals 1 if the MAF of the j^{th} variant is less than *F*, or 0 otherwise.

The meta-analysis of SKAT may be undertaken either assuming homogeneous genetic effects (meta-SKAT-hom), or heterogeneous effects (meta-SKAT-het) across studies. Meta-SKAT-hom first sums the weighted score of a variant across studies, and then sums the squared summed statistic across all variants in the gene (equation (3-3)), whilst meta-SKAT-het sums the squared weighted score statistic across studies and variants in the gene (equation (3-4)).

Meta-SKAT-hom:

$$Q_{meta-SKAT-hom} = \sum_{j=1}^{m}\left(\sum_{k=1}^{K} w_j U_{jk}\right)^2 \sim \sum_{j=1}^{m} \lambda_j \chi_{1j}^2 \quad \textbf{(3-3)}$$

Meta-SKAT-het:

$$Q_{meta-SKAT-het} = \sum_{k=1}^{K}\sum_{j=1}^{m} w_j^2 U_{jk}^2 \sim \sum_{j=1}^{m} \lambda_j \chi_{1j}^2 \quad \textbf{(3-4)}$$

where $\lambda_{1, \ldots,} \lambda_M$ are eigenvalues of $(\sum_{k=1}^{K} \boldsymbol{V_k})^{\frac{1}{2}} \boldsymbol{W} \left((\sum_{k=1}^{K} \boldsymbol{V_k})^{\frac{1}{2}}\right)^T$ and $\boldsymbol{W} = diag[w_1^2, \ldots, w_m^2]$ is a diagonal matrix of squared weights and χ_{11}^2, …, χ_{1m}^2 are independent χ_1^2 variables.

The meta-analysis of SKAT-O may be undertaken using the meta-WST test statistic in conjunction with the meta-SKAT statistic assuming either homogeneous or heterogeneous effects.

Meta-SKAT-O:

$$Q_{meta-p}(\rho) = (1-\rho) Q_{meta-SKAT} + \rho Q_{meta-WST} \quad \textbf{(3-5)}$$

$$Q_{meta-SKAT-O} = \max\left(Q_{meta-p}(\rho)\right) \sim \sum_{j=1}^{m} \lambda_j \chi_{1j}^2$$

3.3 Summary of meta-analysis software packages

There have been several software packages developed to implement the methods described in Section 3.2; all packages vary in the specific tests they run, the traits which they can analyse (quantitative, binary, survival), and whether they can handle studies of related individuals. These software packages and their characteristics are summarised in Table 3-1.

Table 3-1: Summary of software packages developed for the meta-analysis of gene-based tests.

Meta-analysis software	Study-level software to generate score statistics and covariance matrices	Gene-based tests performed	Phenotypes that may be tested	Allows for homogeneous and heterogeneous effects across studies?	Can handle related samples?
MetaSKAT (75)	MetaSKAT	WST, SKAT, SKAT-O	Quantitative and binary	Yes	No
RAREMETAL (74, 166)	Rvtests RAREMETALWORKER	WST, VTT, SKAT	Quantitative and binary	Only homogeneous	Yes
SeqMeta (78)	SeqMeta	WST, SKAT, SKAT-O	Quantitative, binary and survival	Only homogeneous	Yes (Quantitative traits only)
MASS (76, 77)	SCORE-Seq SCORE-SeqTDS	WST, VTT, SKAT, SKAT-O	Quantitative and binary	Yes	No

These four meta-analysis packages all require the studies contributing to the meta-analysis to generate score statistics and variance-covariance matrices using individual level data and specific study-level packages. These summary statistics may be generated either using the packages themselves (MetaSKAT and SeqMeta), or by using companion software (RAREMETALWORKER or Rvtests for RAREMETAL and SCORE-Seq or SCORE-SeqTDS for MASS, Table 3-1). Each meta-analysis package may then be used by a central meta-analyst to combine the study-level summary statistics and undertake gene-based tests, as described in Section 3.2. Recently a program called PreMeta has been developed which allows the conversion of summary statistics generated by any of the study-level packages listed in Table 3-1 to the format required for any of the other meta-analysis packages listed (167).

A further package (MAGA) has also been developed which allows the meta-analysis of gene-based tests using the single variant summary statistics only. MAGA collates the results of single variant analyses based on Wald, score or likelihood ratio tests to estimate the score vector from either the test statistics or the betas and P-values. A covariance matrix may then be estimated using individual level data from one study, or if unavailable, an external reference panel, and together with the score vector may be used to construct WST, VTT and SKAT tests, as in Section 3.2 (168).

The methods implemented by these packages have been evaluated through a number of simulations and empirical analyses (74-78, 166-168). The power of the meta-analysis methods was consistently found to be equivalent to that of the joint analysis of individual level data (74, 75, 78) and type 1 error was generally well controlled (75,

77). Other findings of these simulations were as expected, given the properties of each statistical test. Meta-SKAT and meta-SKAT-O were more powerful where the effects of variants in a gene were bidirectional, and where a smaller proportion of variants in a region were causal. Conversely, meta-WST and meta-VT tests were more powerful where larger proportions of variants in a region were causal, and where all variants were acting in the same direction (75). The meta-SKAT-hom method was more powerful where genetic effects were homogeneous across studies, with meta-SKAT-het becoming more powerful as the heterogeneity of genetic effects increased (75, 77). In general, the power of all tests improved with an increasing proportion of causal variants (75) and larger overall sample sizes (74).

Overall these methods have been most extensively evaluated using simulated data, which made several assumptions regarding the underlying genetic architecture. Evaluation of these methods in real genome-wide genetic data have generally used relatively small samples (up to approximately 10,000 samples) and only quantitative traits have been considered. Assessment of the concordance of the meta-analysis methods with equivalent analyses using individual level data has also been less thoroughly examined in a real dataset.

3.4 Empirical investigation of RAREMETAL in UK BiLEVE

Chapter 4 of this thesis describes a meta-analysis of exome array data and lung function traits. In that analysis, the meta-analysis software RAREMETAL is utilised. In this section, I shall describe in more detail the simulations undertaken by the RAREMETAL authors. I then use data from UK BiLEVE to examine the performance of RAREMETAL in a real dataset that is considerably larger than was used in the published RAREMETAL evaluation (74). To this end, I utilise real phenotype data and genotypes from 48,943 individuals, and test whether the meta-analyses of WST and SKAT gene-based tests are equivalent to the corresponding analysis using individual level data.

3.4.1 Evaluation of simulation methods undertaken in RAREMETAL paper

The RAREMETAL authors undertook several simulations to evaluate the software's methods. Genes were simulated as 5000 base pair regions, using a coalescent approach and a demographic model based on European population history. They generated "studies" of 1000 individuals each from one of several related populations, such that some variants were shared between studies, and other variants were study specific.

Firstly, test statistics and P-values were compared from a meta-analysis and a joint analysis of 10,000 simulated genes in samples from 3 studies. They carried out this comparison, assuming three different trait models:

1. 50% of all variants were causal, with each causal variant increasing trait values by 0.25 standard deviations (SDs).
2. 50% of all variants were causal, with 80% causal variant increasing trait values by 0.25 SDs and 20% decreasing trait values by 0.25 SDs.
3. 50% of all variants were causal, with the effect of each causal variant following a normal distribution with mean 0 and SD of 0.25.

For the meta-WST and meta-SKAT tests, test statistics and P-values from the meta-analysis were consistent to those obtained through a joint analysis of individual level data. For the meta-VTT analysis, the P-values, both estimated asymptotically and empirically using Monte-Carlo methods, were slightly less concordant.

Secondly, type 1 error was assessed through 50 million null simulations, based on meta-analyses of 3, 6 and 9 studies, and was found to be well controlled for all gene-based tests at significance levels $\alpha=1x10^{-3}$, $1x10^{-4}$ and $2.5x10^{-6}$.

Finally, power was investigated in scenarios of between 2 and 100 studies, under several phenotype models:

1. 50% of variants with MAF<0.5% were causal, with each causal variant increasing trait values by 0.25 SDs.

2. 50% of all variants were causal, with each causal variant increasing trait values by 0.25 SDs.
3. 50% of all variants were causal, with 80% of causal variants increasing trait values by 0.25 SDs and 20% decreasing trait values by 0.25 SDs.

The power of RAREMETAL was compared to the power of combining study-level P-values using Fisher's method and by a minimum P-value approach, using a type 1 error rate of $\alpha=2.5x10^{-6}$. RAREMETAL was found in all scenarios to be the most powerful approach, with increasing power as number of studies increases; however all methods were found to be fairly underpowered, until the sample size became very large samples (~60% power achieved in the analysis of 100 cohorts, n=100,000 samples).The RAREMETAL methods were then applied using real data, in a meta-analysis of blood lipid traits in 18,699 samples from 7 sample collections, with exome array genotype data. Associations with high density lipoprotein (HDL), low density lipoprotein (LDL) and triglycerides levels were tested in each study, with covariate adjustment and inverse normalisation of traits. Meta-WST, meta-SKAT and meta-VT tests were then undertaken using RAREMETAL. These analyses identified several associations, although many of these appeared to be driven by SNPs which also showed association in single variant analysis. One gene (*LDLR*) was identified, whose signal did appear to be driven by several rare variants and would not have been identified in single variant association analysis.

Type 1 error and power were also assessed using the real genotype data from three case collections (total n=10,361) and simulated phenotypes; similar trends were observed as were in the simulations. Finally, the P-values of a RAREMETAL analysis using two sample collections (total n=7862) were compared to the joint analysis of individual level data and were found to be highly concordant.

The properties of RAREMETAL methods were largely tested using simulated genetic data, which made a number of assumptions regarding the underlying genetic architecture of the trait and causal variants. All simulated scenarios assumed that 50% of variants under a given MAF threshold within a gene were causal. Different assumptions were also made regarding direction and magnitude of variant effects; in

some scenarios, all causal variants acted in the same direction, whilst other simulations considered a scenario where 80% of causal variants acted in one direction, (increased disease risk / trait value), whilst the other 20% acted in the opposite direction (decreased risk / trait value). The simulations which tested the concordance of the RAREMETAL and joint analysis methods also included a scenario where variant effects were normally distributed, with a mean effect of zero, although this was not considered in the simulations examining type 1 error and power. In reality, the distribution and effects of causal variants are unknown in advance, so it is not clear how realistic the considered scenarios are. A further consideration is whether causal variants might vary across studies; it is not clear to what extent this was the case in these simulations.

The genetic data simulated was all based on genes of 5kb in length, with each having an average of 100 variants within them, with 80% having a MAF<1%, and 49% as singletons. Across the genome, genes vary greatly in length, with some less than 100bp, and the longest genes up to 2.4Mb (169); it is not clear whether the results of the simulations would be applicable to genes with varying lengths and numbers of variants. The simulations included all variants within the MAF inclusion threshold, and so would be akin to analyses of sequencing data. In exome chip studies, not all rare variants in a gene will be measured, so there are likely to be far fewer variants included within each test.

The application of RAREMETAL to real data in the meta-analysis of blood lipid traits addresses some of these issues. The real genotype data from a subset of the studies included in the meta-analysis was also used to assess type 1 error, power and concordance with joint analysis of individual level data, in an exome chip study. These investigations were limited to only 2 or 3 sample collections, so it is unclear whether these findings would be the same where the meta-analysis consisted of many more studies, as is more common in practice. Only quantitative traits were considered throughout, so it is not clear if similar trends would be seen for binary traits.

3.4.2 Investigation of RAREMETAL in UK BiLEVE: Aims

The most extensive evaluation of the meta-analysis methods employed by RAREMETAL were undertaken using simulated data, that were akin to sequence data, and limited in terms of the underlying genetic architecture. Whilst the RAREMETAL authors also assessed the methods in real exome chip data, this was in fewer studies and samples than are likely to be included in meta-analyses of exome chip data in practice. In addition, the RAREMETAL methods were only evaluated using a quantitative trait. I therefore aimed to further evaluate the methods in a more realistic setting, using data from 48,943 samples from the UK BiLEVE study. Through these analyses, I primarily sought to investigate whether the results of a RAREMETAL meta-analysis were in agreement with a joint analysis undertaken with the individual level data (mega-analysis) for the following:

1. A quantitative trait
2. A binary trait, where each study has a balanced (1:1) ratio of cases to controls
3. A binary trait where each study has an unbalanced ratio of cases to controls

To simulate a meta-analysis of gene-based tests, I split the UK BiLEVE data, by randomly allocating samples to sub-studies, and performed meta-SKAT and meta-WST tests using summary statistics generated for each study, as would be done in a meta-analysis setting. To address each aim, I used real phenotypes: FEV_1 (quantitative trait) and smoking status (binary trait: ever/never smoker). I compared the results of these meta-analyses to the results of the equivalent mega-analysis, performed using individual level data. I also undertook a comparison of results from the mega-analysis with those obtained through a meta-analysis combining P-values from gene-based tests carried out at study level, using both Fisher's method (164) and Stouffer's Z-score method (165), to confirm the advantage of using the RAREMETAL methods.

3.4.3 Investigation of RAREMETAL in UK BiLEVE: Methods

3.4.3.1 UK BiLEVE data

The UK BiLEVE study selected 50,008 samples from UK Biobank, based on their smoking history and FEV_1 measurements. Samples were selected from the extremes

and middle of the distributions of percent predicted FEV_1, separately in never smokers and heavy smokers and were genotyped using a custom-designed Affymetrix Axiom UK BiLEVE array. The UK BiLEVE array includes rare coding variants selected from the same sequencing project as the exome array, alongside additional genome-wide coverage. Following genotype and sample quality control (QC), 48,943 unrelated individuals remained for analysis. The number of samples passing QC that were included in each smoking-FEV_1 stratum are summarised in Table 3-2.

Table 3-2: Number of samples selected in each smoking- FEV_1 stratum.

	Never smokers	Heavy smokers
Low FEV_1	9750	9750
Average FEV_1	9831	9803
High FEV_1	4902	4907
Total	24,483	24,460

The UK BiLEVE samples were genotyped in 11 batches; only those SNPs which passed QC in all batches were included in these analyses. Furthermore, SNPs were filtered to include those with overall call rate>99% and MAF<5% only. The genotype data were annotated using VEP based on the GRCh37/hg19 database and gene files were created based on these annotations. For computational efficiency, I restricted the analyses to a single chromosome (chromosome 6).

3.4.3.2 Gene-based tests mega-analyses

Mega-analyses using the individual level data from all samples were carried out using the SKAT R package (68). Analyses were undertaken for FEV_1 as a quantitative trait and for smoking as a binary trait with a balanced ratio of cases and controls overall; the results of these analyses were used for comparison with all meta-analyses scenarios. For both traits, the null model of no association was fitted, with adjustment for covariates, as listed in Table 3-3. The residuals of the null models were then used for two gene-based tests: 1. SKAT with default weighting (weight for j^{th} variant: $\beta(MAF_j; 1, 25)$); 2. WST with Madsen Browning (66) weightings, achieved by including

the arguments `r.corr=1, weights.beta=c(0.5,0.5)` in the SKAT command.

Table 3-3: Covariate adjustments undertaken for each trait.

Trait	Covariate adjustments
FEV_1	Sex, age, height, pack-years and 10 principal components (PCs)
Smoking	10PCs

3.4.3.3 Meta-analyses: meta-analysis of gene-based tests using RAREMETAL

For the meta-analyses of the quantitative trait and the binary trait with a balanced ratio of cases to controls, the 48,943 UK BiLEVE samples were randomly split into the following scenarios:

I. 2 studies: 1 study with n=24,471; 1 study with n=24,472.
II. 5 studies: 4 studies with n=9789; 1 study with n=9787.
III. 10 studies: 9 studies with n=4894; 1 study with n=4897.
IV. 49 studies: 48 studies with n=999; 1 study with n=991.
V. Mixed studies: 1 study with n=15,000; 1 study with n=10,000; 3 studies with n=5000; 3 studies with n=2000; 2 studies with n=1000, 1 study with n=943.

Scenarios I-IV were intended to allow the evaluation of the RAREMETAL meta-analysis methods as samples came from an increasing number of smaller studies. In scenario V, the samples were split into 11 studies of varying sample sizes, which is more typical to what is seen in practice in a consortium meta-analysis effort.

For the meta-analyses of the binary trait with unbalanced ratio of cases to controls, the samples were randomly split into 10 studies (9 studies with n=4894; 1 study with n=4897) as follows:

a. 1:1 ratio: 10 studies with case-control ratio 1:1.
b. 2:1 ratio: 5 studies with case-control ratio 2:1; 5 studies with case-control ratio 1:2.

c. 5:1 ratio: 5 studies with case-control ratio 5:1; 5 studies with case-control ratio 1:5.
d. 10:1 ratio: 5 studies with case-control ratio 10:1; 5 studies with case-control ratio 1:10.

Whilst the ratio of cases to controls in each study for scenarios b-d was unbalanced, the overall ratio of cases to controls was 1:1. This allowed direct comparison between each scenario and the same mega-analysis undertaken using all samples and with a balanced ratio of cases and controls.

Within each scenario, RAREMETALWORKER was run for each study using both smoking and FEV_1 (scenarios I-V only). Traits were adjusted for the covariates listed in Table 3-3 and score statistics and variance-covariance matrices were calculated for each study using the resulting residuals.

The score statistics and variance-covariance matrices from each study were then combined using RAREMETAL to perform both meta-SKAT-hom (equation (3-3)) and meta-WST (equation (3-1)) analyses. Default weightings were used for meta-SKAT, with Madsen-Browning weightings used for the meta-WST tests.

3.4.3.4 Meta-analyses: Combining P-values of gene-based tests

Samples were split into studies as per scenarios I-V, above. Within each study, SKAT and WST tests were performed for both FEV1 and smoking using the SKAT R package, analogously to the mega-analysis. The P-values (P_k) from the gene-based tests of each study k, with sample size n_k were then combined, by two methods:

1. Z-score meta-analysis:

$$Z = \frac{\sum_{k=1}^{K} w_k Z_k}{\sqrt{\sum_{k=1}^{K} w_k^2}} \sim N(0,1), \qquad \textbf{(3-6)}$$

where $w_k = \sqrt{n_k}$ and $Z_k = \Phi^{-1}\left(1 - \frac{P_k}{2}\right)$.

2. Fisher's method:

$$X^2 = -2\sum_{k=1}^{K}\ln(P_k) \sim \chi^2_{2K}, \quad \textbf{(3-7)}$$

where K is the number of studies.

Unlike the output from RAREMETAL, the SKAT package does not provided an effect estimate for the WST test, so the estimated direction of effect could not be taken into account in the Z-score meta-analysis.

3.4.4 Investigation of RAREMETAL in UK BiLEVE: Results

3.4.4.1 Comparison of RAREMETAL meta-analysis methods versus mega-analysis

Figure 3-1 and Figure 3-2 show comparisons of the SKAT P-values resulting from a mega-analysis of all samples and each meta-analysis scenario (I-V) for the SKAT and WST analyses of FEV_1, respectively. Figure 3-3 and Figure 3-4 show the equivalent plots for the analysis of smoking, with a balanced ratio of cases and controls.

Figure 3-1: Comparison of -$\log_{10}$ P-values from the mega-analysis, and RAREMETAL meta-analysis scenarios for the SKAT analysis of FEV_1.

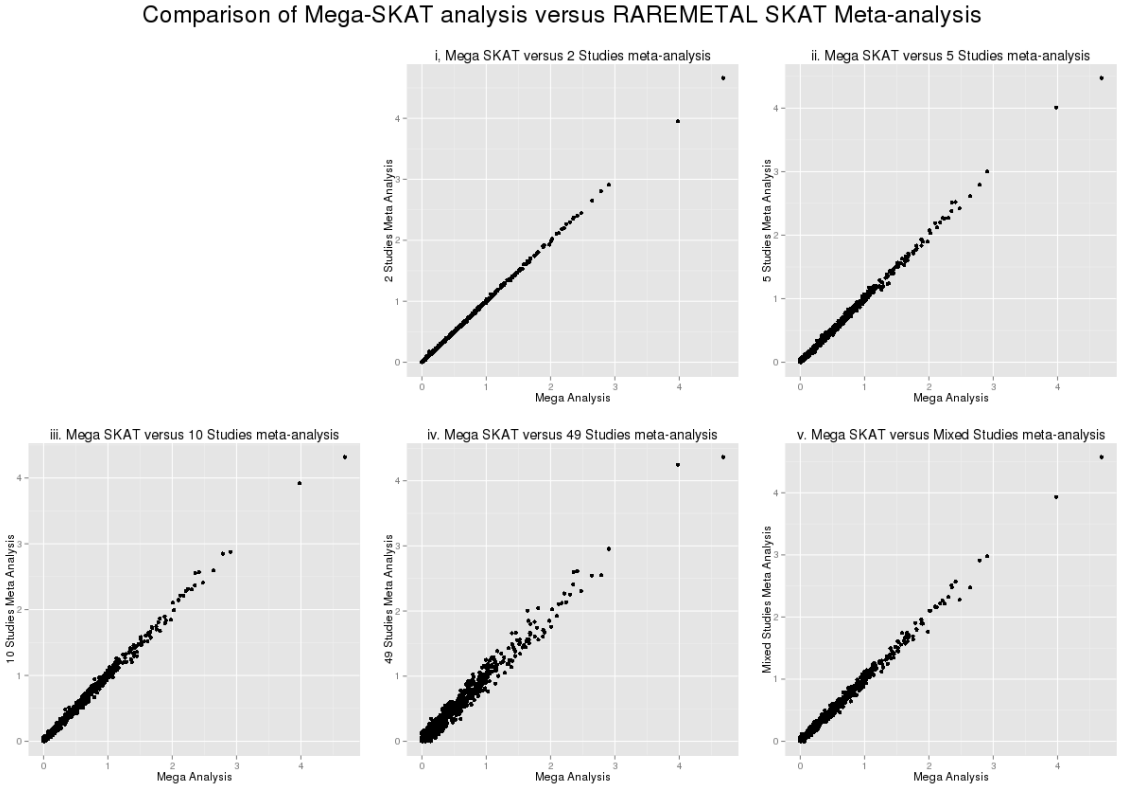


Figure 3-2: Comparison of -log$_{10}$ P-values from the mega-analysis, and RAREMETAL meta-analysis scenarios for the WST analysis of FEV$_1$.

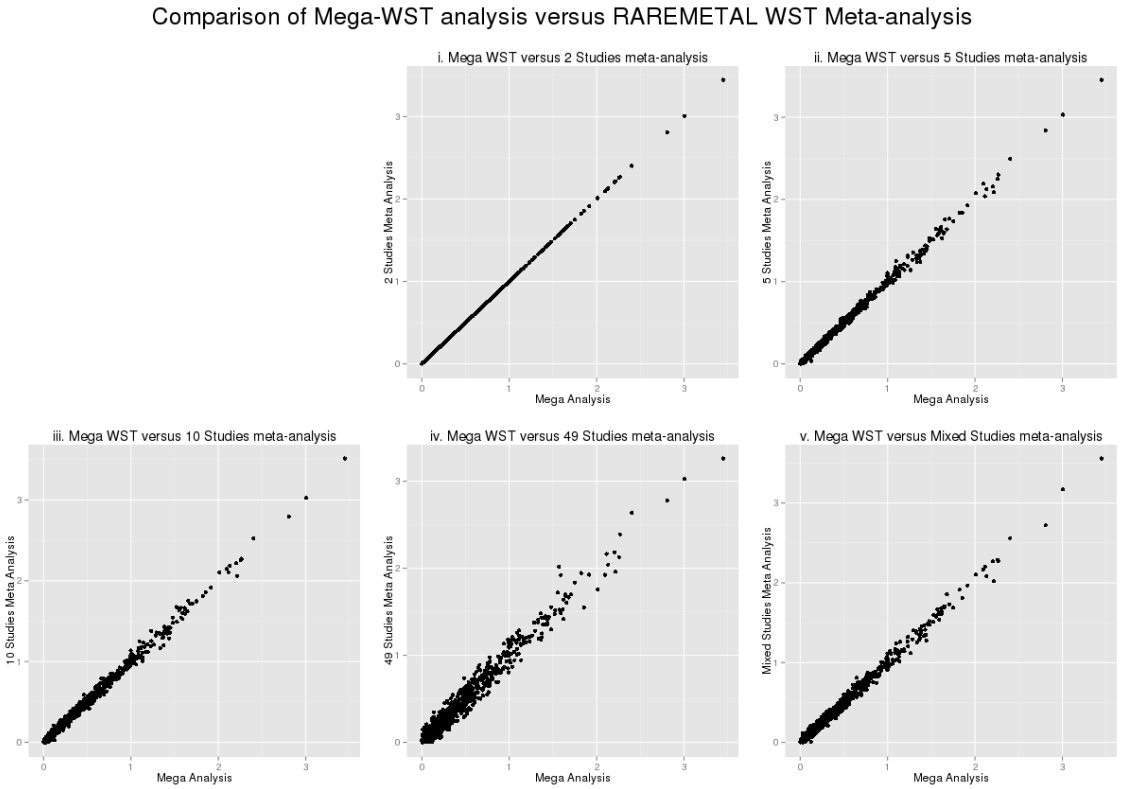


Figure 3-3: Comparison of -log$_{10}$ P-values from the mega-analysis, and RAREMETAL meta-analysis scenarios for the SKAT analysis of smoking with a balanced ratio of cases and controls.

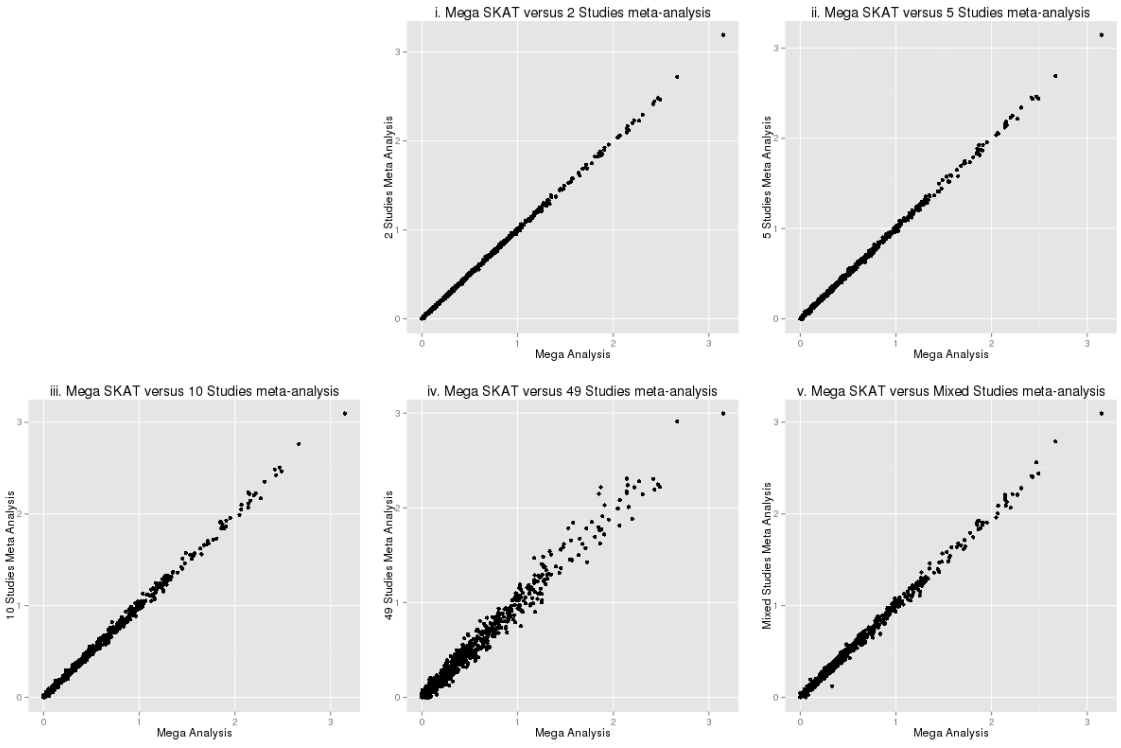
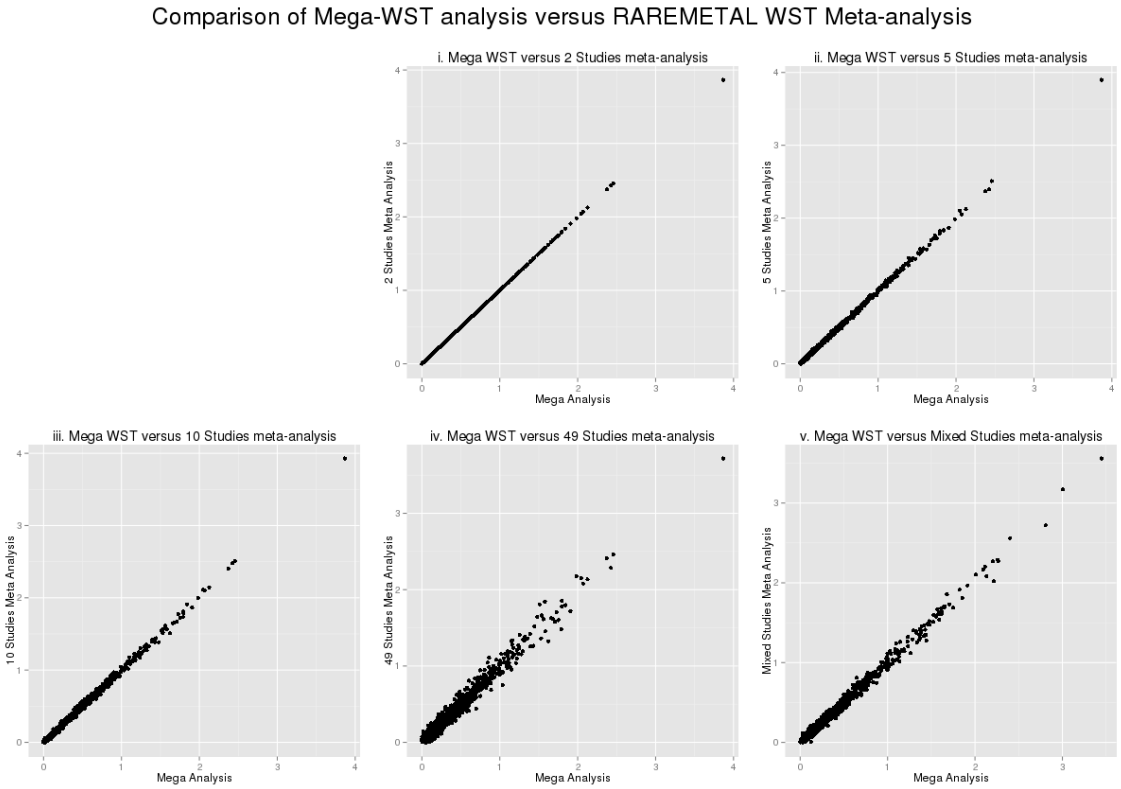
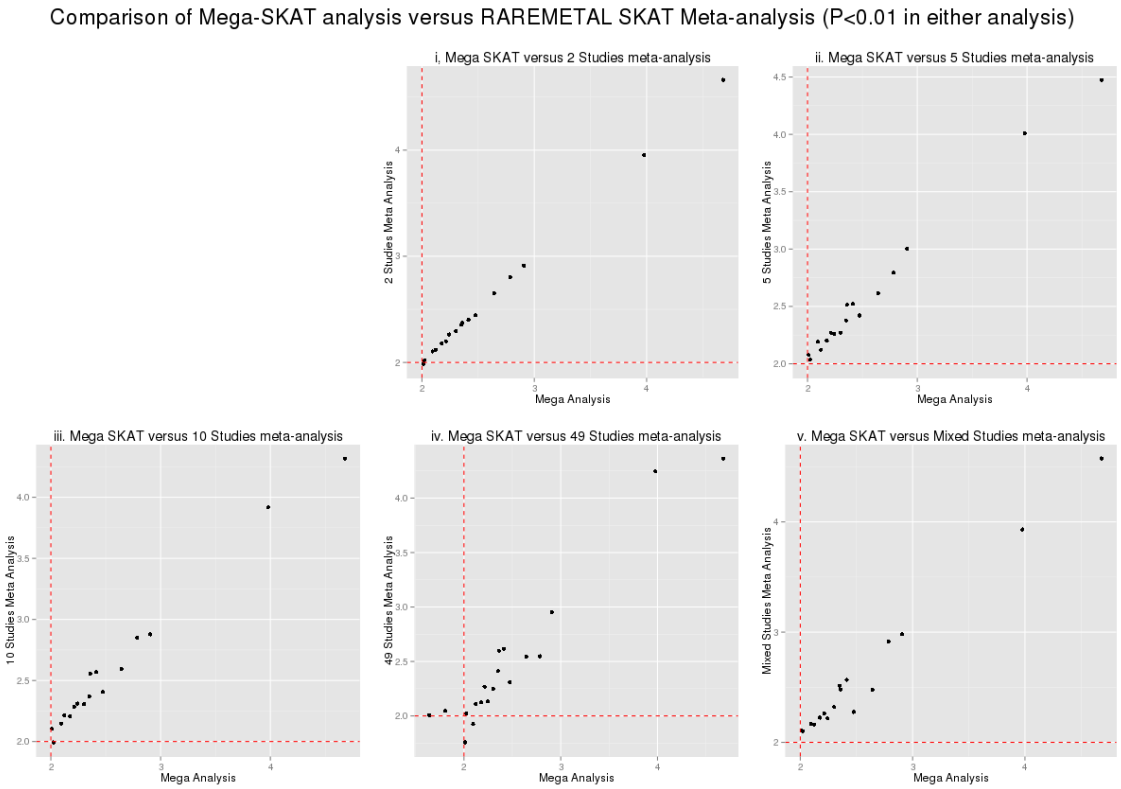


Figure 3-4: Comparison of -log$_{10}$ P-values from the mega-analysis, and RAREMETAL meta-analysis scenarios for the WST analysis of smoking with a balanced ratio of cases and controls.



The RAREMETAL results became less closely correlated with the mega-analysis results as the data were split into an increasing number of studies; however all meta-analyses were highly concordant overall with the mega-SKAT and mega-WST results, even when split into 49 studies (scenario IV, concordance correlation coefficient=0.993 [FEV_1 SKAT analysis]). The meta-analysis with 11 studies of mixed sizes (scenario V) showed similar results to the meta-analysis with 10 equal sized studies (scenario III). Figure 3-5 shows the P-value comparisons for all genes meeting the P<0.01 (-log$_{10}$P>2) threshold in either the mega-SKAT or each meta-SKAT analysis of FEV1, for the five scenarios. These plots show that a similar, if not identical set of top genes would be selected through all meta-analyses compared to the mega-analysis, if a significance level of P<0.01 was assumed. Whilst there are some genes which would be identified at the P<0.01 significance level in the meta-analysis, but not in the mega-analysis (or vice versa), these genes all have P-values close to the 0.01 threshold in both analyses.

Figure 3-5: Comparison of -log$_{10}$ P-values from the mega-analysis, and RAREMETAL meta-analysis scenarios for the SKAT analysis of FEV$_1$: genes with P<0.01 in either analysis only.



Where RAREMETAL analyses were carried out using studies with unbalanced ratios of cases to controls (scenarios a-d), the resulting P-values were not so highly concordant with the mega-analysis P-values, as can be seen in Figure 3-6 and Figure 3-7. The mega-SKAT analysis had a balanced ratio of cases to controls (1:1) and identified a total of 18 genes with P<0.01. Table 3-4 compares the number of genes meeting this significance threshold in the mega-SKAT analysis and in each meta-SKAT analysis scenario. In scenario a, where the ratio of cases to controls in each study was 1:1, the RAREMETAL analysis resulted in the same 18 genes meeting the P<0.01 significance level. For the remaining scenarios, as the imbalance of cases to controls increased, fewer of those 18 genes met the P<0.01 significance level and conversely, there were an increasing number of genes meeting this level, which did not have P<0.01 in the mega-analysis. Similar trends were seen for the meta-WST analyses for unbalanced studies.

Figure 3-6: Comparison of -log$_{10}$ P-values from the mega-analysis, and RAREMETAL meta-analysis scenarios for the SKAT analysis of smoking with unbalanced ratios of cases and controls.

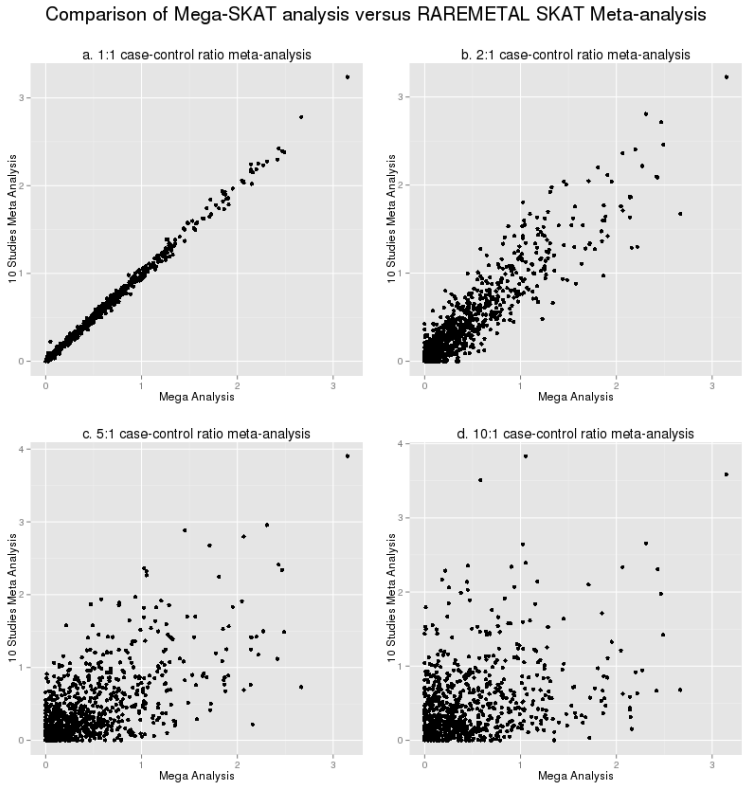


Figure 3-7: Comparison of -log$_{10}$ P-values from the mega-analysis, and RAREMETAL meta-analysis scenarios for the WST analysis of smoking with unbalanced ratios of cases and controls.

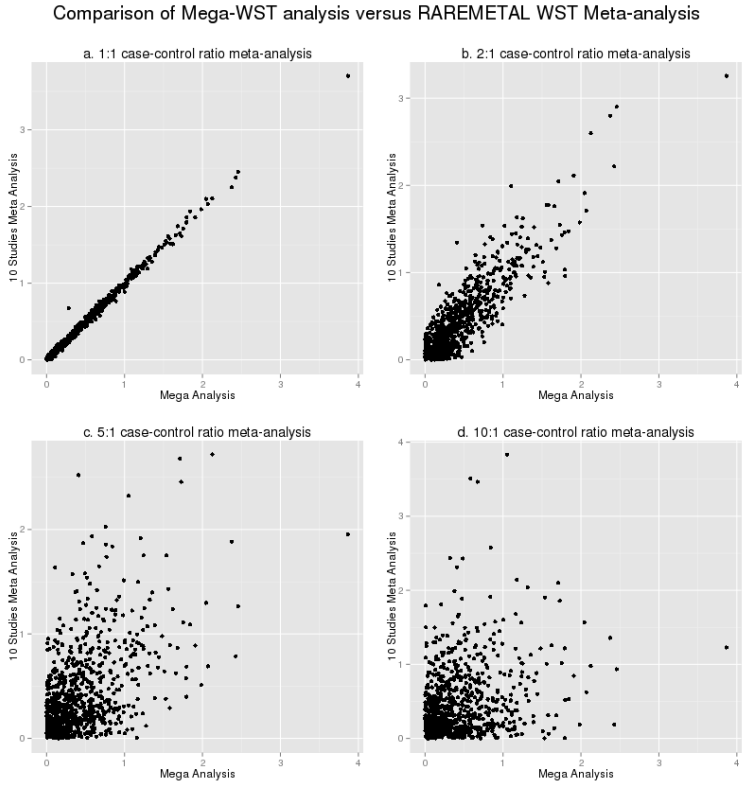


Table 3-4: No. genes meeting the P<0.01 in either the mega-analysis, or each RAREMETAL meta-analysis scenario for the SKAT analysis of Smoking.

Scenario a

		1:1 ratio meta-analysis	
		P≥0.01	P<0.01
Mega-analysis	P≥0.01	789	0
	P<0.01	0	18

Scenario b

		2:1 ratio meta-analysis	
		P≥0.01	P<0.01
Mega-analysis	P≥0.01	783	6
	P<0.01	9	9

Scenario c

		5:1 ratio meta-analysis	
		P≥0.01	P<0.01
Mega-analysis	P≥0.01	783	6
	P<0.01	13	5

Scenario d

		10:1 ratio meta-analysis	
		P≥0.01	P<0.01
Mega-analysis	P≥0.01	775	14
	P<0.01	14	4

3.4.4.2 Comparison of combining P-values of gene-based tests versus mega-analysis

Figure 3-8 and Figure 3-9 show comparisons of the SKAT P-values resulting from a mega-analysis of all samples and five meta-analysis scenarios for the combined SKAT P-values analyses of FEV_1. Similar trends for the WST and for smoking were seen; equivalent figures are in appendix B.

Figure 3-8 shows the comparison of P-values from the mega-analysis, and each meta-analysis scenario (I-V) where P-values were combined using Fisher's Method. Overall, the meta-analysis P-values become less concordant with P-values from the mega-SKAT analysis as the data were split into increasingly smaller studies. Genes with a low cumulative MAF (<0.1%) are highlighted in orange; as the number of studies increased, the meta-SKAT P-values for these genes with these low cumulative MAFs tended to 1 ($-\log_{10}P=0$). Table 3-5 outlines the number of genes associated with FEV_1 with P<0.01 in either the mega-analysis, or in each of the meta-analyses. The mega-SKAT analysis identified 17 genes with P<0.01 in total. Where the data were split into two studies (scenario I), the meta-SKAT analysis resulted in 12 genes meeting P<0.01; of those 6 also had P<0.01 in the mega-analysis. When the data were split into 49 studies

(scenario IV), only 7 genes had P<0.01, with none of those genes meeting P<0.01 in the mega-analysis. If we assume that the mega-analysis resulted in "true" P-values, then the number of false negatives increased as there was an increasing number of smaller studies, whereas the number of false positives remained fairly consistent in all scenarios.

Figure 3-8: Comparison of -log$_{10}$ P-values from the mega-analysis, and each Fisher's Method meta-analysis scenario for the SKAT analysis of FEV$_1$. Genes with Cumulative MAF<0.1% highlighted in orange.

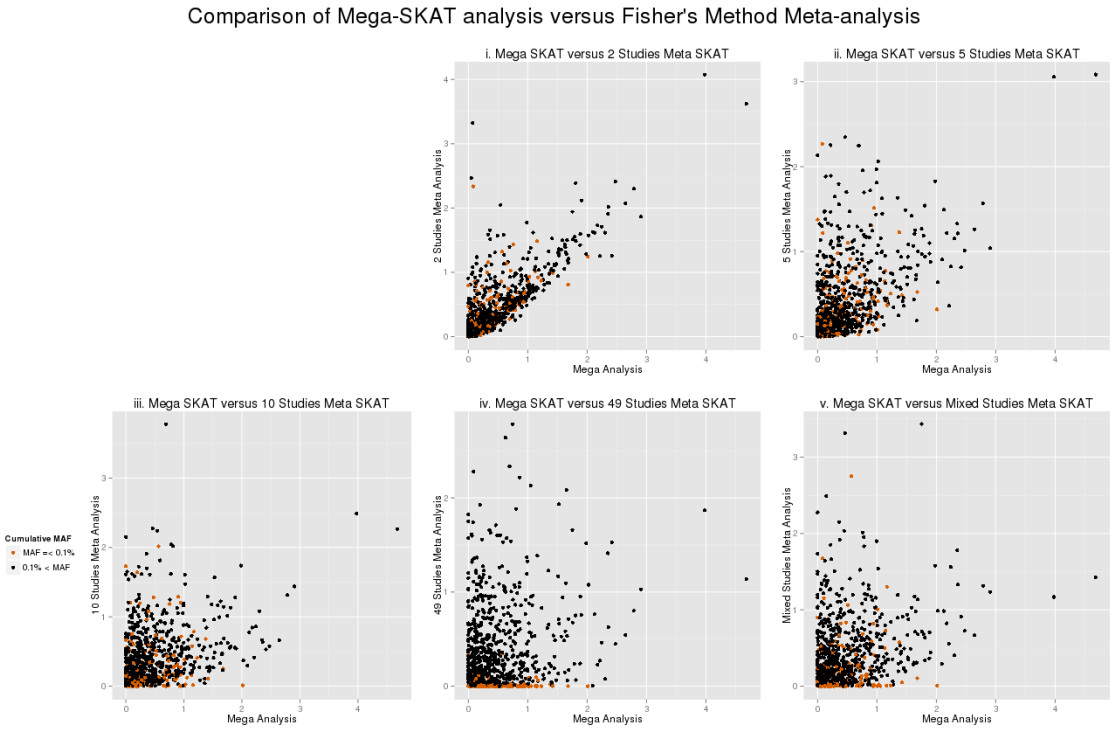


Table 3-5:No. genes meeting the P<0.01 in either the mega-analysis, or each Fisher's Method meta-analysis scenario for the SKAT analysis of FEV_1.

Scenario I

		2 studies meta-analysis	
		P≥0.01	P<0.01
Mega-analysis	P≥0.01	784	6
	P<0.01	11	6

Scenario II

		5 studies meta-analysis	
		P≥0.01	P<0.01
Mega-analysis	P≥0.01	784	6
	P<0.01	15	2

Scenario III

		10 studies meta-analysis	
		P≥0.01	P<0.01
Mega-analysis	P≥0.01	783	7
	P<0.01	15	2

Scenario IV

		49 studies meta-analysis	
		P≥0.01	P<0.01
Mega-analysis	P≥0.01	783	7
	P<0.01	17	0

Scenario V

		Mixed studies meta-analysis	
		P≥0.01	P<0.01
Mega-analysis	P≥0.01	782	8
	P<0.01	17	0

Figure 3-9 shows the P-value comparison where SKAT P-values were combined using Z-score meta-analysis. This meta-analysis method is mostly anti-conservative, an issue that worsens as the data are split into increasingly smaller studies. As the number of studies increased, the meta-SKAT P-values for genes with a low cumulative MAF tended to get closer to 1 ($-\log_{10}P=0$); whereas for those genes with a higher cumulative MAF, the P-values became more significant.

Figure 3-9: Comparison of -log$_{10}$ P-values from the mega-analysis, and each Z-score meta-analysis scenario for the SKAT analysis of FEV$_1$.

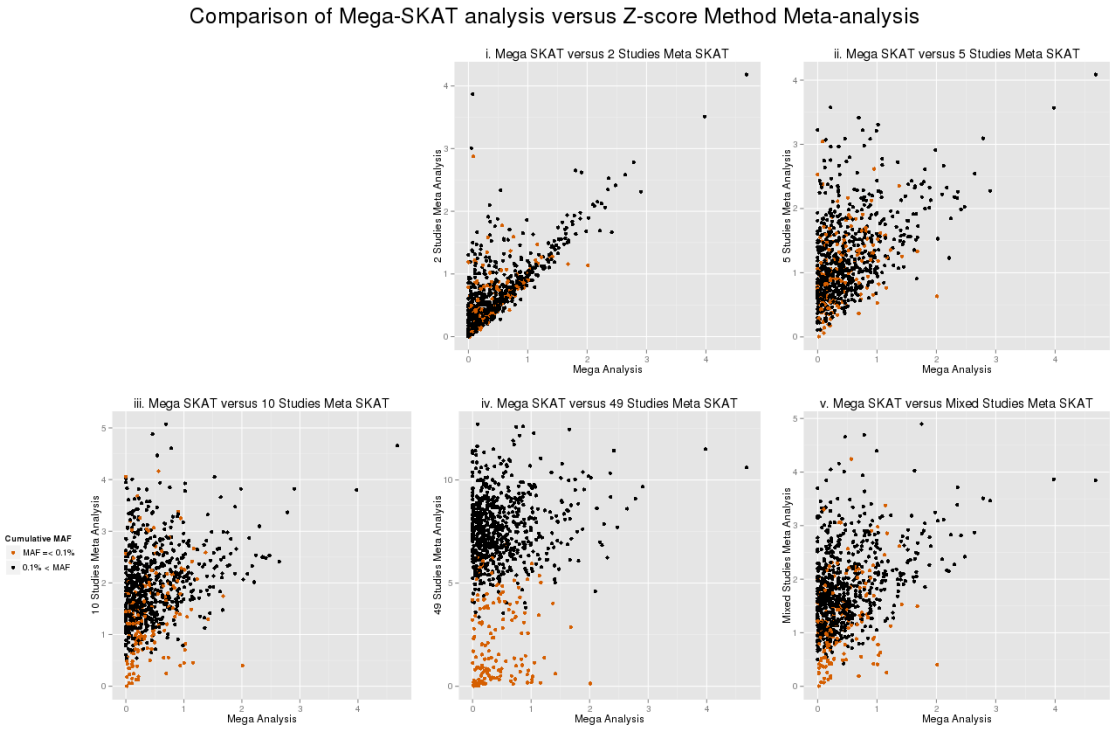


Table 3-6 also illustrates the anti-conservative results of this method. Where the data were split into two studies (scenario I), the meta-SKAT analysis identified 13 genes with P<0.01 that also had P<0.01 in the mega-analysis, along with an additional 7 genes. When the data were split into 49 studies (scenario IV), the majority of genes (709 of 807) had P<0.01. Overall, the number of false positives significantly increased as there were an increasing number of smaller studies, whilst the number of false negatives tended to decrease.

Table 3-6: No. genes meeting the P<0.01 in either the mega-analysis, or each Z-score Method meta-analysis scenario for the SKAT analysis of FEV_1.

Scenario I

		2 studies meta-analysis	
		P≥0.01	P<0.01
Mega-analysis	P≥0.01	783	7
	P<0.01	4	13

Scenario II

		5 studies meta-analysis	
		P≥0.01	P<0.01
Mega-analysis	P≥0.01	711	79
	P<0.01	5	12

Scenario III

		10 studies meta-analysis	
		P≥0.01	P<0.01
Mega-analysis	P≥0.01	460	330
	P<0.01	1	16

Scenario IV

		49 studies meta-analysis	
		P≥0.01	P<0.01
Mega-analysis	P≥0.01	81	709
	P<0.01	1	16

Scenario V

		Mixed studies meta-analysis	
		P≥0.01	P<0.01
Mega-analysis	P≥0.01	544	246
	P<0.01	1	16

To summarise the findings of this section, the performance of each gene-based meta-analysis method, in terms of concordance with the equivalent mega-analysis using individual level data, is described in Table 3-7.

Table 3-7: Summary of the performance of meta-analysis methods.

Meta-analysis method (Traits)	Summary of performance compared to mega-analysis
RAREMETAL (Quantitative Trait and Binary Trait with balanced case:control ratio)	Meta-analysis P-values were overall highly concordant with those obtained for mega-analysis, even where data split into a large number of small studies.
RAREMETAL (Binary Trait with unbalanced case:control ratio)	As the imbalance of cases to controls in each study increased, the P-values became less concordant with those from the mega-analysis, likely resulting in an increase of both false positive and false negative findings.
Fisher's Method to combine P-values (Quantitative trait and Binary trait with balanced case:control ratio)	Concordance to mega-analysis P-values decreased as the number of studies increased. Generally, P-values became conservative (less significant) as the numbers of studies increases, especially for genes with a low cumulative MAF, likely resulting in an increase in false negative findings.
Z-score Method to combine P-values (Quantitative trait and Binary trait with balanced case:control ratio)	Concordance to mega-analysis P-values decreased as the number of studies increased. P-values tended to get more anti-conservative (highly significant) as the number of studies increased, likely resulting in an increase in false positive findings; however this trend was not so distinct in genes with a low cumulative MAF.

3.5 Discussion

The analyses conducted in this chapter have shown that the meta-analysis of SKAT and WST gene-based tests using RAREMETAL result in equivalent P-values to an analysis undertaken using individual level data, for both quantitative trait and a binary trait with a balanced ratio of cases to controls. The RAREMETAL meta-analysis methods did not give such consistent results to the mega-analysis of a binary trait with an unbalanced ratio of cases to controls however. If it is assumed that the mega-analysis result gave the "true" P-value, then with an increasing magnitude of imbalance, the number of false positives increased, along with the number of false negatives. An imbalanced ratio of cases to controls has previously been found to be problematic in single variant analyses of low frequency variants, leading to highly inflated type 1 errors (170).

The analyses undertaken in this chapter primarily aimed to determine empirically the consistency of the results of a RAREMETAL meta-analysis, compared to a joint analysis of individual level data, in real array-based genotype data from the UK BiLEVE study. These analyses did not address other properties of the meta-analysis methods, such as power and type 1 error; this would have required the use of simulated phenotype data. For the analyses of quantitative traits and binary traits with a balanced ratio of cases and controls, these properties have been extensively examined by the authors of RAREMETAL, and the other software packages (74-78, 166, 167) and have found type 1 error to generally be well controlled (75, 76), although the power of these tests was somewhat limited, except for very large samples sizes (74, 75, 78). For the analyses of binary traits with an unbalanced ratio of cases and controls, power and type 1 error have not been fully assessed, to date. The benefit of using real genotype and phenotype data from UK BiLEVE is that a sample size more realistic to a real meta-analysis of GWAS or exome array data could be achieved, and no assumptions regarding the underlying genetic architecture of a trait were required to be made.

The UK BiLEVE samples used in these analyses all came from UK Biobank, which is likely to be a relatively homogeneous population compared to many meta-analyses, which often use samples from several geographical regions (163), and sometimes include multiple ancestries (171). Other between study differences may add to the

heterogeneity, including differences in phenotype definitions and differences in study level genotyping and QC or analyses (163). It is possible that meta-analyses using studies from more diverse populations may affect the performance of the gene-based meta-analysis methods. In the meta-analyses described in this chapter, the simulated studies consisted of individuals selected at random. To introduce heterogeneity to the studies, thereby creating more realistic meta-analysis scenarios, studies could have been generated according to geographical region (eg by study centre), or alternatively some studies could have been selected based on secondary phenotypes (eg samples could be identified for disease case collections). Furthermore, the UK BiLEVE samples were restricted to unrelated individuals only; meta-analyses often include studies with related individuals and it is not clear whether the results of this chapter will also be applicable to analyses which include studies with related samples.

As well as investigating the concordance of the RAREMETAL meta-analysis results to those from a mega-analysis, alternative meta-analysis methods which combine gene-based P-values were additionally considered. The analyses combining SKAT and WST P-values using Fisher's or Z-score methods performed very poorly, giving inconsistent results to the equivalent mega-analysis. Other evaluations of the Z-score and Fisher's method for combining SKAT or burden test P-values have found the method to have similar or less power to the equivalent mega-analysis (75, 172). These previous evaluations have considered meta-analysis scenarios with only 2 or 3 studies. In terms of using Fisher's method for combining SKAT and WST P-values, the results of this chapter are fairly consistent with these previous findings, and furthermore demonstrate that the method performs more poorly as the number of studies in a meta-analysis increases. The results of this chapter showed that the Z-score method resulted in a large number of false positive findings. The reason for this is the Z-score method effectively assigned a direction of effect to each study; since no effect size was given for SKAT or the WST (when implemented in the SKAT R package), the assumed direction of effect was consistent across all studies, therefore resulting in an overestimate of the meta-analysis Z statistic. This issue is likely to be compounded where direction of effects differed across studies, and as the number of studies included in the meta-analysis is increased. Some software packages, for example

RAREMETAL provide gene effect estimate for the WST; where this is the case, it is probable that the Z-score method would perform better, as directions of effect would be taken into account, however this method for combining P-values is still likely to be less powerful than the RAREMETAL method.

Overall, these results show there is a clear advantage to using the RAREMETAL methods for the meta-analysis of gene-based tests for the analysis of a quantitative trait, or a binary trait with a balanced ratio of cases and controls. Chapter 4 of this thesis describes the meta-analyses of three quantitative traits utilising these methods. The findings of this chapter give confidence that the results of these meta-analyses should be consistent with a mega-analysis of the data. The gene-based meta-analysis methods appear less suitable for the analysis of a binary trait where contributing studies have a large imbalance of cases to controls however; in this instance, the RAREMETAL method appears likely to result in a number of both false positives and false negatives.

Chapter 4 Meta-analysis of exome array data and quantitative lung function traits

4.1 Introduction

Lung function measures are an important predictor of mortality and morbidity and are used in the diagnosis of a number of diseases, including chronic obstructive pulmonary disease (COPD), a leading cause of death globally (89). Lung function is largely affected by environmental factors such as smoking and exposure to air pollution; however there is also a genetic component, with heritability estimates ranging up to 66% (105, 108, 110, 173). A number of large-scale genome-wide association studies (GWAS) of lung function have successfully identified single nucleotide polymorphisms (SNPs) influencing lung function in over 50 regions (43, 116-120, 136), yet these identified regions only account for a small proportion of the estimated heritability. Low frequency (minor allele frequency (MAF) 1-5%) and rare (MAF<1%) variants, have been largely underexplored by GWAS to date.

Many of the GWAS which have identified lung function loci have been meta-analyses carried out within the SpiroMeta and CHARGE consortia. Large consortia such as these allow information from a number of studies to be combined, thereby maximising sample size, without sharing individual level data. In this context, a meta-analysis involves studies generating summary statistics, according to a pre-determined analysis plan. The meta-analysis of single variant associations has been found to be equivalent to analyses using individual level data (174) and these methods have been widely used to identify SNPs associated with a range of complex traits. More recently, methods for meta-analysing gene-based associations have been developed (74-76, 168), which I explored further in Chapter 3.

In this chapter I describe a meta-analysis of exome array data and three lung function measures: forced expiratory volume in one second (FEV_1), forced vital capacity (FVC) and the ratio of FEV_1 to FVC (FEV_1/FVC) in 23,398 individuals of European ancestry, from 11 studies from the SpiroMeta Consortium. Analyses of both single variant associations and gene-based associations were carried out. The most significant single variant and gene associations were then followed up in up to 93,390 independent

samples. My role in this study was to generate analysis plans describing the analyses to be carried out by each study, and to undertake the study level analyses for the 1958 British Birth Cohort (1958BC). Once the study-level analyses were completed, I carried out thorough quality control (QC) checks of the study-level summary statistics, performed the meta-analyses and associated follow-up. This chapter describes all aspects of this work.

4.2 Discovery stage samples, study-level analyses and quality control of data

4.2.1 Discovery Stage Samples and Phenotypes

The discovery stage analyses were carried out using data from 11 studies from the SpiroMeta consortium, listed in Table 4-1. All samples were genotyped using the Illumina Human Exome BeadChip v1. Analyses were carried out for three lung function phenotypes: FEV_1; FVC; FEV_1/FVC.

Table 4-1: Studies included in discovery analyses.

Study name (abbreviation)	**n**
1958 British Birth Cohort (1958BC)	5270
Generation Scotland: Scottish Family Health Study (GS:SFHS)	8164
Cooperative Health Research in the Region of Augsburg (KORA F4)	1447
CROATIA-Korcula cohort (KORCULA)	791
Lothian Birth Cohort 1936 (LBC1936)	974
Study of Health in Pomerania (SHIP)	1681
Northern Swedish Population Health Study (NSPHS)	880
Prospective Investigation of the Vasculature in Uppsala Seniors (Pivus)	836
Swiss study on Air Pollution and Lung Disease in adults (SAPALDIA)	2707
The Cardiovascular Risk in Young Finns Study (YFS)	434
Finnish Twin Cohort (FIN)	214
Total Discovery Sample Size	23,398

4.2.2 Study design and analysis plans

From the outset of this study, it was intended that both single variant and gene-based associations would be tested. At the start of this study, a number of the gene based tests described in Section 1.2.5 had been developed; however it was unclear how these tests might be performed in a meta-analysis setting. For this reason, I first generated an analysis plan for studies to undertake single variant analyses (analysis 1), with the equivalent gene-based analysis plan to be generated at a later date. Some months into the project, several software packages for the meta-analysis of gene-based tests had been developed, as described in Chapter 3. Within the consortium, it was decided that we would utilise the RAREMETAL package (74) for the gene-based analyses, largely as this software package had been adopted by other consortia, and this would limit extra work for study analysts. I therefore generated a second analysis plan (analysis 2) to generate study level summary statistics to be used in the RAREMETAL analyses. The analyses carried out by each study is described in Sections 4.2.4 (study-level analysis 1: single variant associations) and 4.2.5 (study-level analysis 2: RAREMETAL) and both analysis plans are included in appendix C.

As well as performing the meta-analysis of gene-based tests, the RAREMETAL package additionally allows meta-analyses of single variant associations to be carried out. I therefore performed two single variant meta-analyses: in the first analysis I combined the summary statistics generated according to the first analysis plan using the statistical package R; the second meta-analysis was carried out using RAREMETAL and the summary statistics generated according to the second analysis plan. I subsequently carried out a comparison of the results generated by these two meta-analyses, described in Section 4.3.2. The final single variant association analysis results presented in this chapter are based on the analyses carried out using RAREMETAL. This method was selected as there were some discrepancies in the number of samples and QC procedures included in the two study level analyses for some studies; the utilisation of RAREMETAL for both the single variant association analyses and gene-based analyses meant the samples included in both analyses were consistent.

4.2.3 Study Level Quality Control

All QC of genotype data was carried out at study level. For 1958BC, KORA F4, NSPHS, PIVUS, SAPALDIA, SHIP, FIN and YFS, genotype calling and QC were carried out in accordance with the Exome Chip Quality Control SOP Version 5, as developed within the UK exome chip consortium (149). Genotypes were initially called using Gencall in Illumina's Genome Studio software (13) and the following QC of SNPs and samples was performed. Initial filters applied excluded SNPs with very low call rate (<90%) and samples with low call rate (<98%), heterozygosity outliers, duplicates, gender mismatches and ancestral outliers. SNPs with missing data were then recalled using genotype calling software zCall (51). All alleles were mapped to the forward strand of human genome build 37 and secondary exclusions were applied to remove SNPs with low call rate (<99%) or deviations from HWE ($P<10^{-4}$). Samples with call rate <99% and heterozygosity outliers were then also excluded.

The KORCULA and LBC1936 samples underwent genotype calling and QC as above before study level analysis 1 (Section 4.2.4) was carried out. Prior to study level analysis 2 (Section 4.2.5), these two studies underwent an alternative genotype calling and QC process: genotypes were called using Gencall in Illumina's Genome Studio software (13) via the CHARGE Consortium joint calling cluster file (http://www.chargeconsortium.com/main/exomechip) and quality control of the genotype data was undertaken according to the CHARGE exome chip best practices, described elsewhere (53). Genotype calling and QC for GS:SFHS samples was carried out according to CHARGE best practices for both study level analyses.

The QC procedures carried out by all studies, prior to both study level analyses are summarised in Figure 4-1.

Figure 4-1: Summary of QC procedures carried out, prior to the two study-level analyses.
Analysis 1: single variant associations; Analysis 2: RAREMETAL.

	Genotype calling and QC according to UK Exome Chip Consortium SOP.	Genotype calling and QC according to CHARGE best practices.
Study level analysis 1	1958BC, KORA F4, NSPHS, PIVUS, SAPALDIA, SHIP, FIN, YFS, KORCULA, LBC1936	GS:SFHS
Study level analysis 2	For 1958BC, KORA F4, NSPHS, PIVUS, SAPALDIA, SHIP, FIN, YFS	GS:SFHS, KORCULA, LBC1936

4.2.4 Study Level analyses 1: Single Variant Associations

Within each study, the following association analyses were carried out for FEV_1, FVC and FEV_1/FVC: all traits were adjusted for sex, age, age^2 and height; the resulting residuals were converted to ranks and then to normally distributed z-scores. These inverse rank normalised traits were used for all subsequent association testing. Studies of unrelated individuals tested for single SNP associations assuming additive genetic effects, using a linear model (equation (1-1)), with adjustment for principal components, implemented in PLINK (18) or EPACTS (175). Studies with related individuals tested associations using a mixed model to account for relatedness (equation (1-3)), implemented in GEMMA (176). Software used by each study is listed in Table 4-2.

Table 4-2: Summary of software used by each cohort for study level analysis 1.

Study	Related	Association Analysis Software
1958BC	No	PLINK
KORCULA	Yes	GEMMA
KORA F4	No	PLINK
LBC1936	No	PLINK
PIVUS	No	EPACTS
SHIP	No	PLINK
GS:SFHS	Yes	GEMMA
FIN	Yes	GEMMA
NSPHS	Yes	GEMMA
SAPALDIA	No	PLINK
YFS	No	PLINK

4.2.5 Study Level analyses 2: RAREMETAL

Within each study, single-variant score statistics were calculated for each SNP (equation (1-5)), along with a variance-covariance matrix (equation (1-6)), describing correlations between variants, using RAREMETALWORKER or rvtests (74). For each trait, these summary statistics were generated separately in ever and never smokers, with adjustment for sex, age, age^2, height, and with each trait being inverse normally transformed prior to association testing. Further adjustments were made for the first 10 principal components and for familial relationships, as appropriate. Software used by each study listed in Table 4-3.

Table 4-3: Summary of software used by each cohort for study level analysis 2.

Study	Related	Association Analysis Software
1958BC	No	RAREMETALWORKER
KORCULA	Yes	RAREMETALWORKER
KORA F4	No	RAREMETALWORKER
LBC1936	No	rvtests
PIVUS	No	RAREMETALWORKER
SHIP	No	RAREMETALWORKER
GS:SFHS	Yes	RAREMETALWORKER
FIN	Yes	RAREMETALWORKER
NSPHS	Yes	RAREMETALWORKER
SAPALDIA	No	RAREMETALWORKER
YFS	No	rvtests

4.2.6 1958BC Study level analysis

The individual level data for 1958BC was available centrally, and I undertook the QC and analyses for this study. The quality control of 1958BC genotype data was undertaken as described in Section 4.2.3. The first stage of the QC and re-calling of genotypes by zCall was undertaken within the UK exome chip consortium. I undertook sample and SNP exclusions in a second stage of QC (post-zCall), as summarised in Table 4-4.

Table 4-4: SNP and sample exclusions: second stage of genotype QC (post-zCall) only.

Total number of SNPs		*247,849*
SNP Exclusions (N failing QC)	Deviates from HWE ($P<10^{-4}$)	2587
	Call rate< 90%	170
Total SNPs passing genotype QC		**245,249**
Total number of samples		*5963*
Sample Exclusions (N failing QC)	Call rate<99%	0
	Heterozygosity outlier (>3SDs from mean heterozygosity rate for SNPs with MAF≥5%)	66
	Heterozygosity outlier (>3SDs from mean heterozygosity rate for SNPs with MAF<5%)	54
Total samples passing genotype QC		5844

Quality control of phenotype data was also undertaken to exclude samples with outlying or unusual values of FEV_1 or FVC (Figure 4-2). 10 samples with FEV_1 measures greater than their FVC measure were excluded along with 26 samples with extreme values of FVC (samples FVC<1L or FVC>8L examined for plausibility given their sex and height). There was additionally a cluster of samples with unusually low FEV_1 values, given their measure of FVC; these 167 samples with FVC >3L and FEV_1 <1L were also identified for exclusion (FEV_1/FVC outliers). Whilst it is plausible a small number of the measurement from these individuals with outlying values may be in fact accurate, many of these outlying measurements were undertaken by a limited subset of nurse and spirometer combinations, suggesting issues with spirometer calibration or measurement.

Figure 4-2: Plot of FEV_1 versus FVC with samples identified as outliers highlighted.

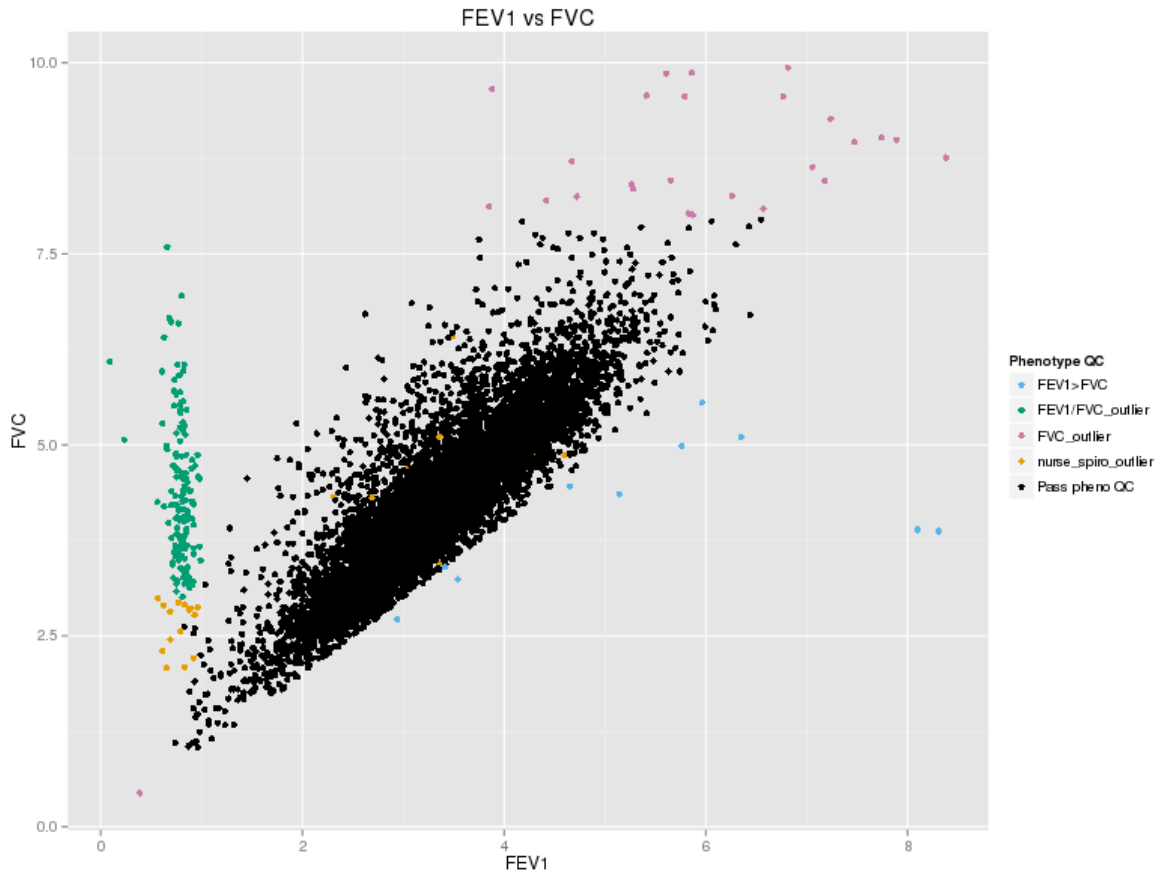


Table 4-5 shows the nurse-spirometer combinations which had more than 50% samples identified as outliers. Consequentially, the remaining samples with measurements undertaken by these nurse/spirometer combinations were additionally excluded.

Table 4-5: Number of outliers per nurse-spirometer combination.
Only nurse-spirometer combinations which had more than 2 (or ≥50%) samples as outliers shown.

Nurse_Spirometer ID	N outlier samples	Total samples measured by Nurse_Spirometer combination.
120_53	11	13
123_1	1	2
123_11	15	15
162_22	7	11
162_61	18	19
176_38	39	42
179_33	26	26
179_40	11	13
179_49	9	10
216_36	7	7

A total of 5270 samples passing genotype and phenotype QC and with complete data for smoking history (ever/never), sex and height were then selected for analyses.

Study level analyses 1 were undertaken using PLINK v1.07 (18) and study level analyses 2 were undertaken using RAREMETALWORKER (74). Both analyses were undertaken for FEV_1, FVC and FEV_1/FVC, with adjustment for sex and height (no adjustment for age was made as all individuals were the same age) and 10 principal components. Analyses were undertaken in 2805 ever smokers and 2465 never smokers separately and traits were inverse normally transformed. Secondary analyses for a subset of 2489 ever smokers was undertaken with additional adjustment for pack-years smoked (pack-years data unavailable for remaining ever smokers).

The QQ plots and Manhattan plots for these analyses are shown for never smokers in Figure 4-3 and Figure 4-4 respectively, and for ever smokers in Figure 4-5 and Figure 4-6. The results from study level analyses 1 and 2 were comparable so for brevity, the results for study level analyses 2 are shown only.

Figure 4-3: QQ Plots and genomic inflation factors (GC) for 1958BC never smokers.
Analysis of A. FEV_1 B. FVC and C. FEV_1/FVC.

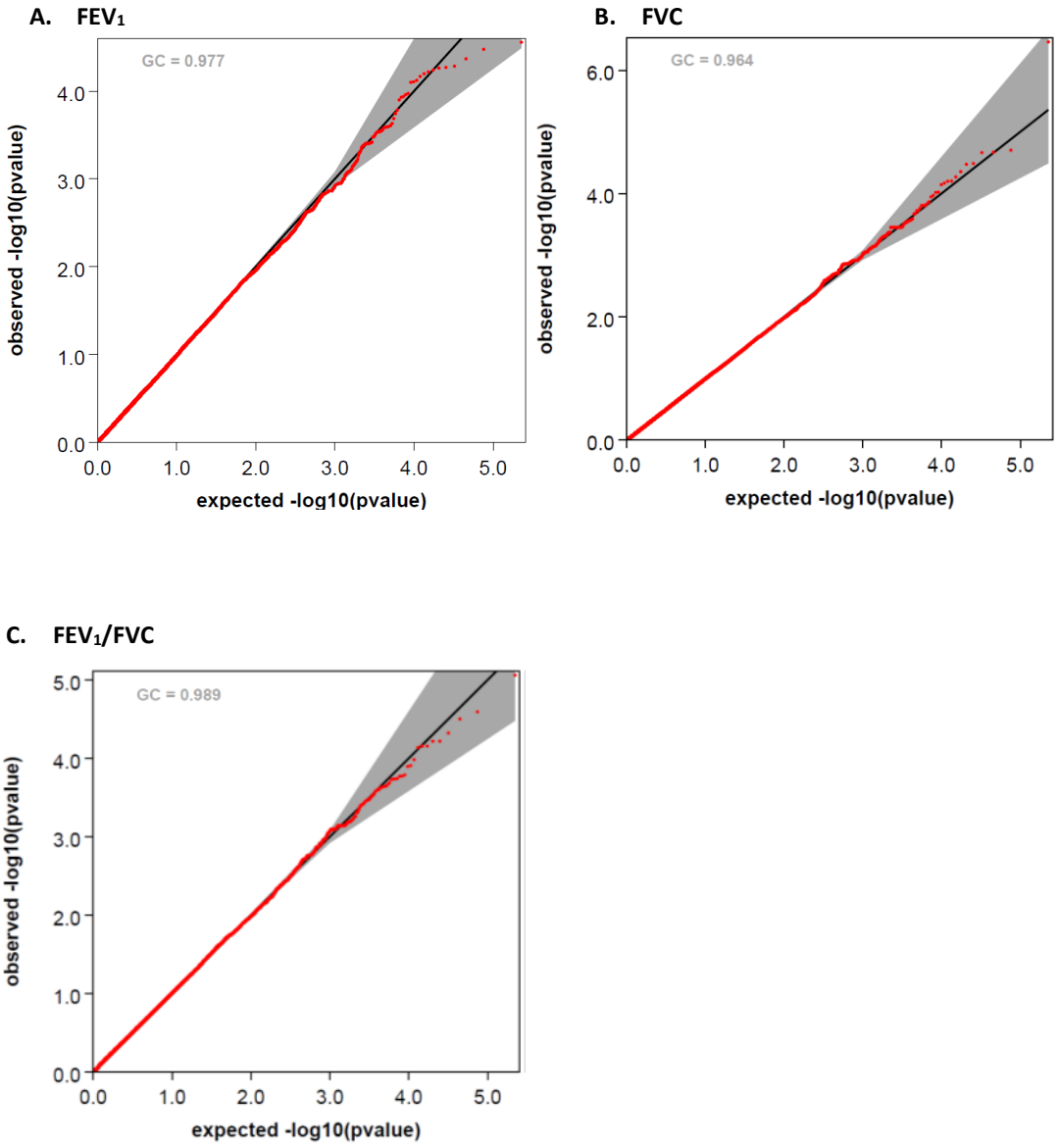


Figure 4-4: Manhattan plots for 1958BC never smokers.
Analysis of A. FEV_1 B. FVC and C. FEV_1/FVC.

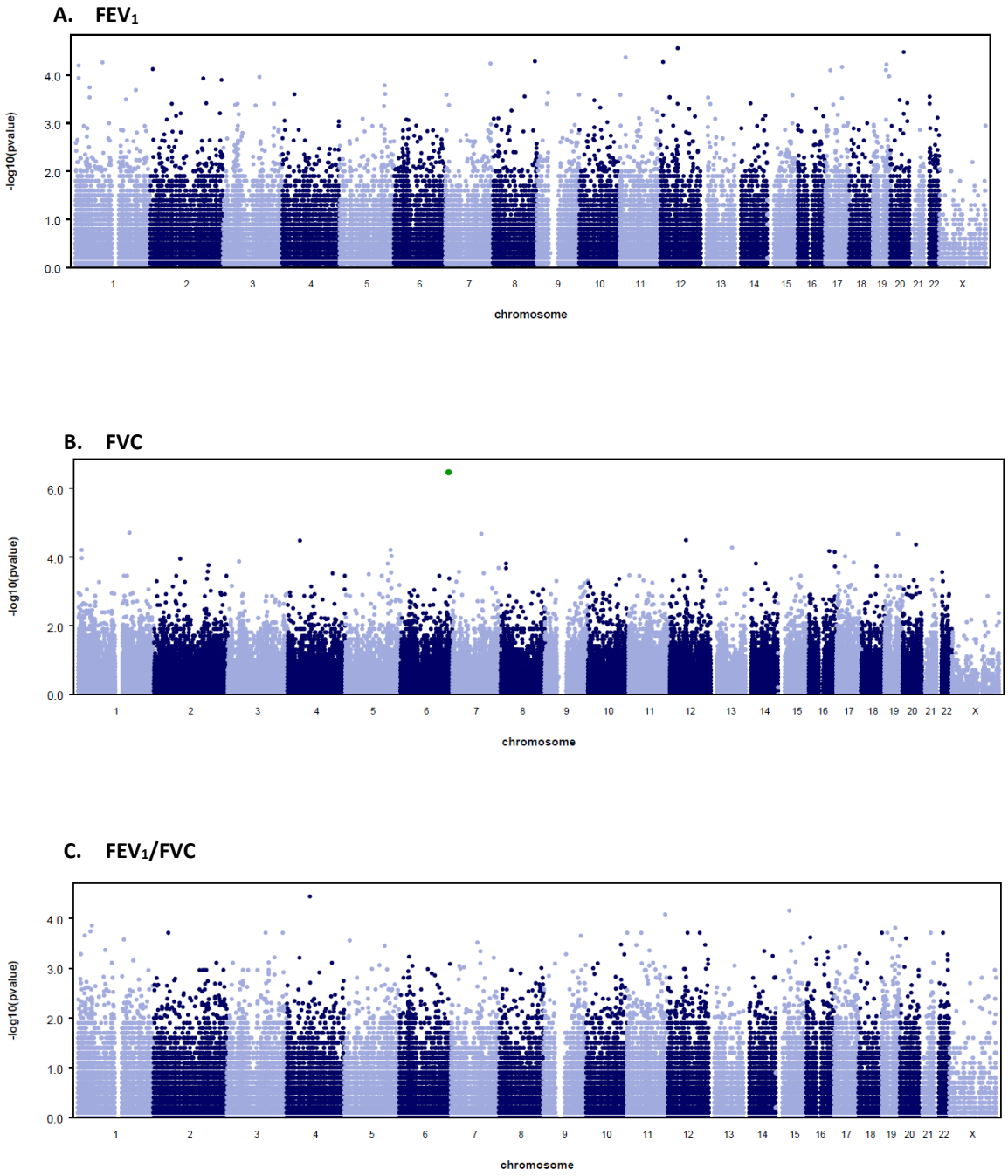


Figure 4-5: QQ Plots and genomic inflation factors (GC) for 1958BC ever smokers.
Analysis of A. FEV_1 B. FVC and C. FEV_1/FVC.

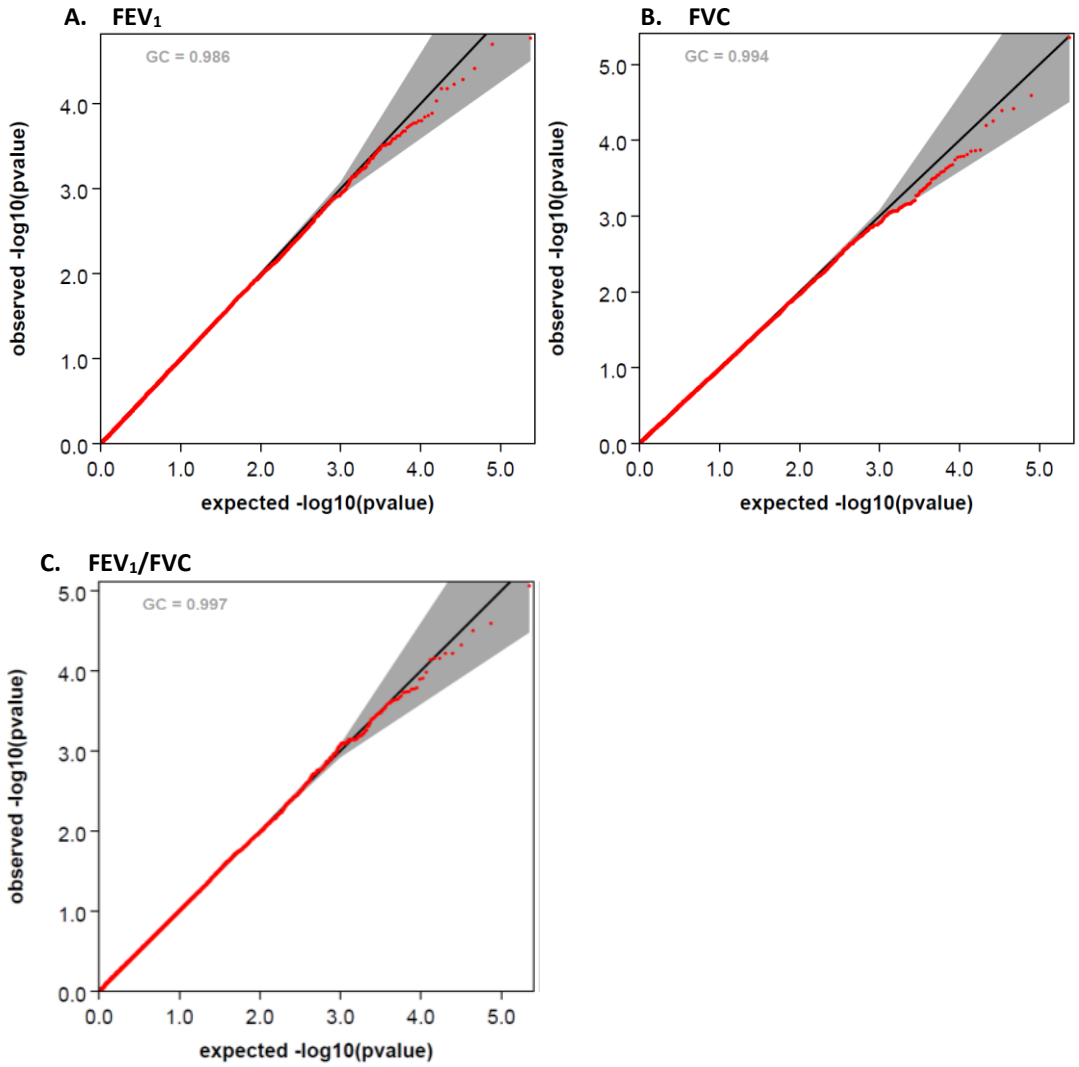
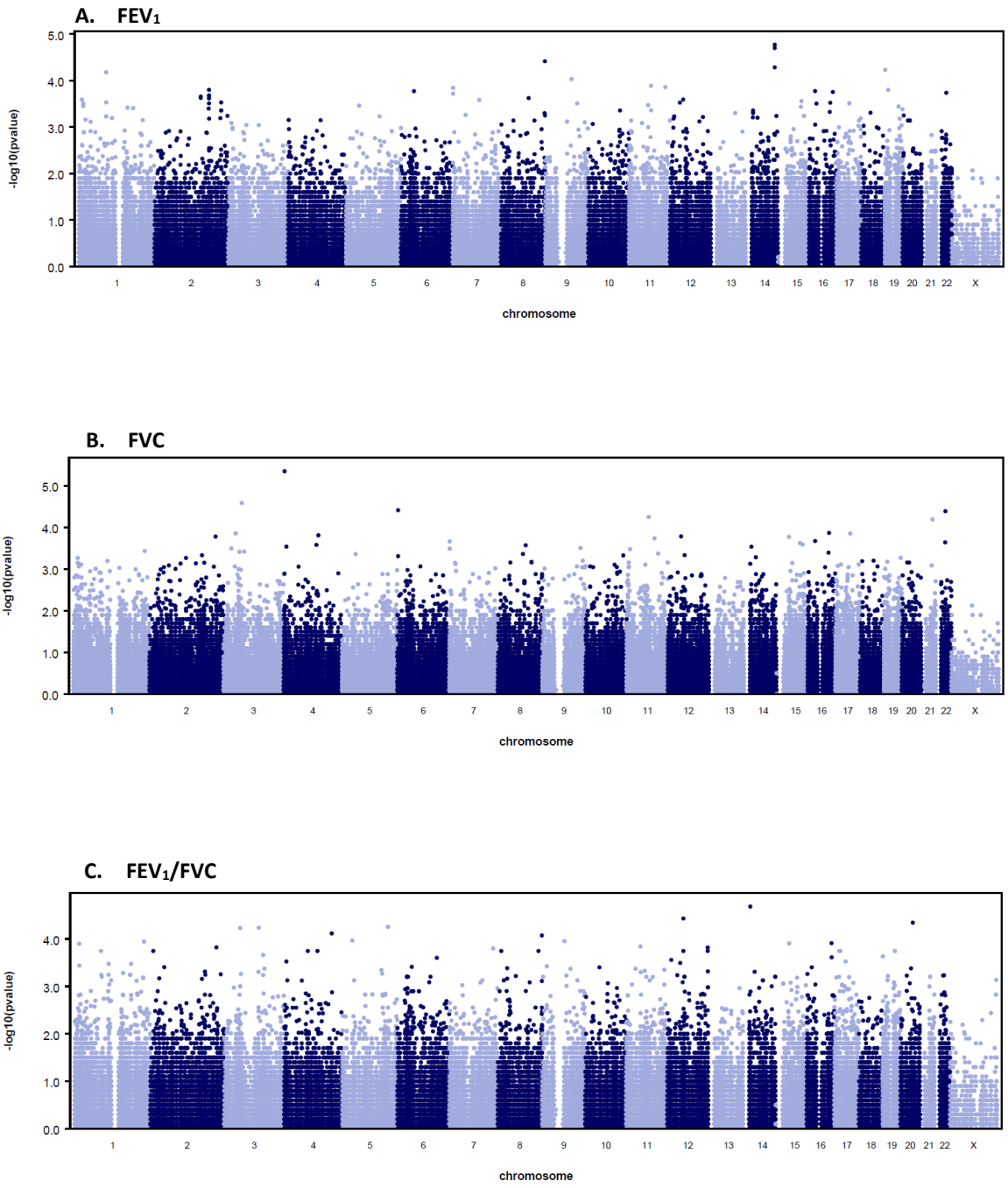


Figure 4-6: Manhattan plots for 1958BC ever smokers.
Analysis of A. FEV_1 B. FVC and C. FEV_1/FVC.



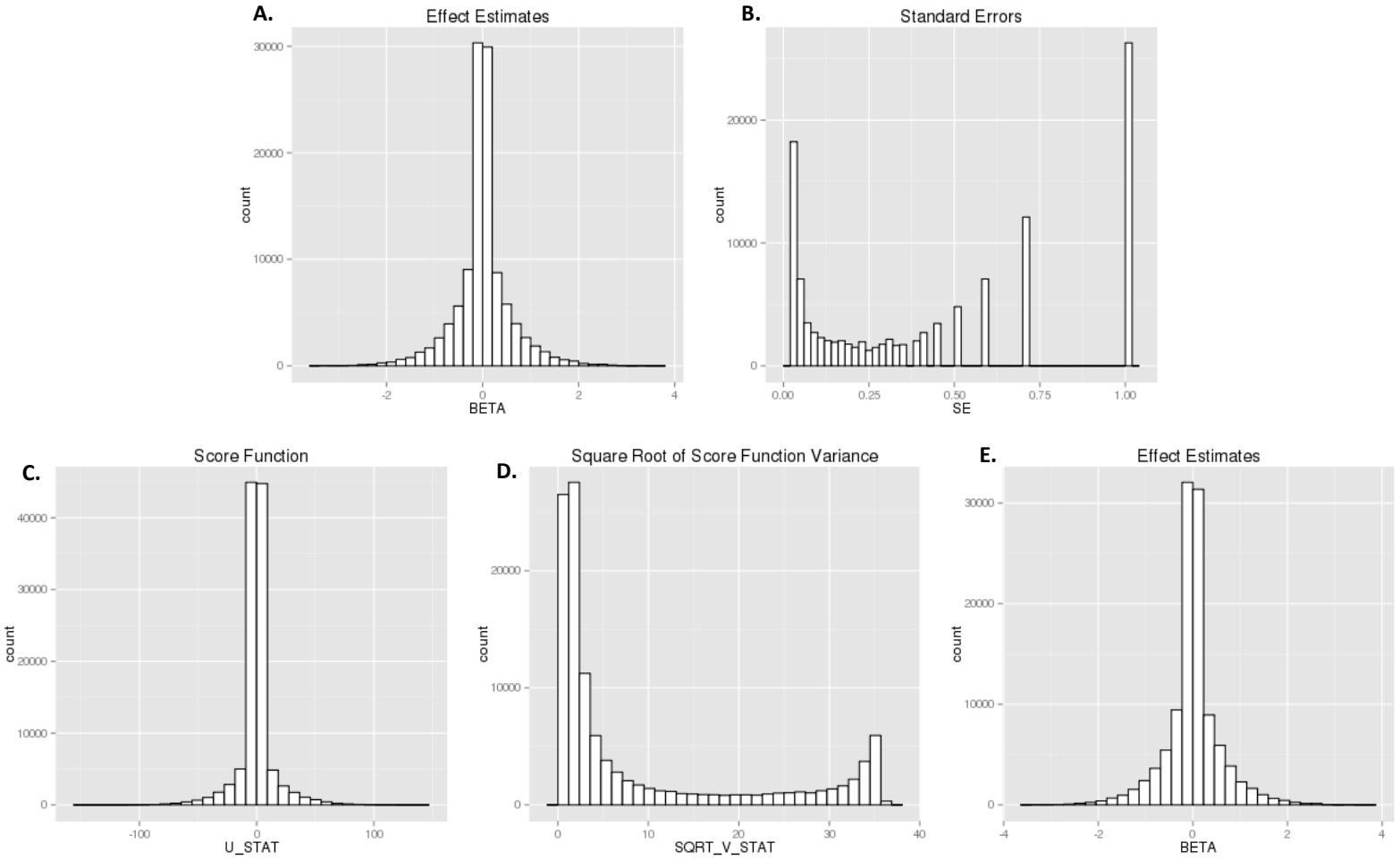
4.2.7 Quality control of study level data

The results from the analyses from each study were uploaded and a number of QC checks were undertaken on the study level data, to identify any unusual results, which were suggestive of errors in the data or analyses. As a basic initial check, for each study the effect estimates (β), and standard errors (se_β) from study level analysis 1 were plotted, along with the effect estimates (β), score function (U) and the square root of the variance of the score function ($\sqrt{V}$) from study level 2. Examples of the expected distributions of all these statistics are shown in Figure 4-7.

Figure 4-7: Examples of expected distributions of study level results (study level analysis of FEV_1 in never smokers).

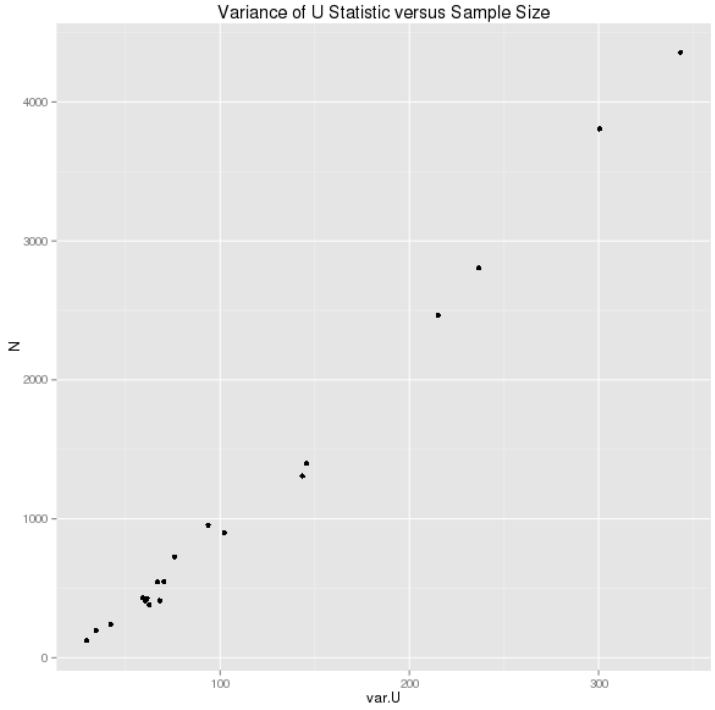
Top: A. Effect estimates and B. standard errors from study level analysis 1.

Bottom: C. score function, D. square root of the score function variance and E. effect estimates from study level analysis 2.



Due to the trait transformation, the effect estimates from each analysis should approximately follow a standard normal distribution. The distribution of U statistics from each analysis was also expected to follow a normal distribution, with mean 0 and variance increasing with samples size (Figure 4-8).

Figure 4-8: Variance of the distribution of score function (U) statistics within each analysis, versus the sample size. Analysis of FEV_1, Statistics for ever smokers and never smokers from each study plotted separately.



The distributions of the standard errors from study analysis 1 and the $\sqrt{V}$ statistics from study analysis 2 were related to the distribution of MAFs (Figure 4-9): the $\sqrt{V}$ statistics increased with increasing MAF, whilst the standard errors decreased with increasing MAF (Figure 4-10).

Figure 4-9: Example distribution of allele frequencies.
Study level analysis of FEV_1 in never smokers

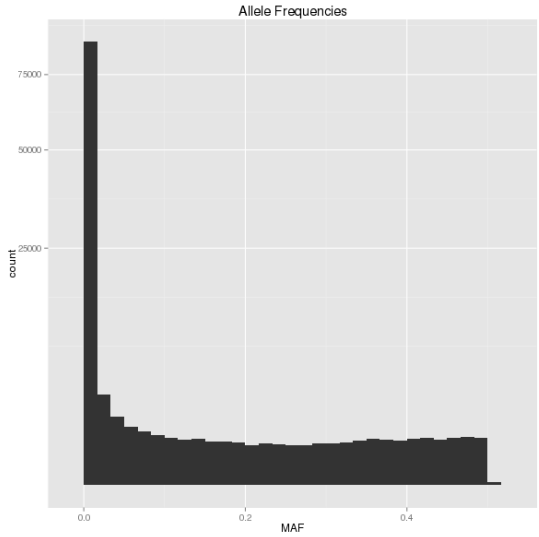
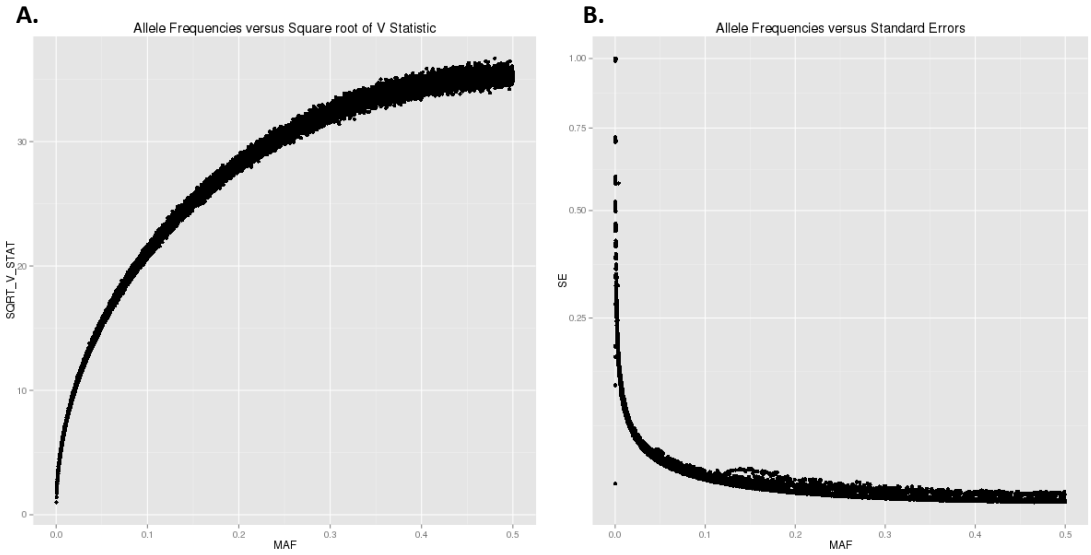


Figure 4-10: Example trends from a study level analysis of FEV_1 in never smokers.
A. allele frequencies versus the square root of V statistics. B. allele frequencies versus the standard errors of effect estimates.

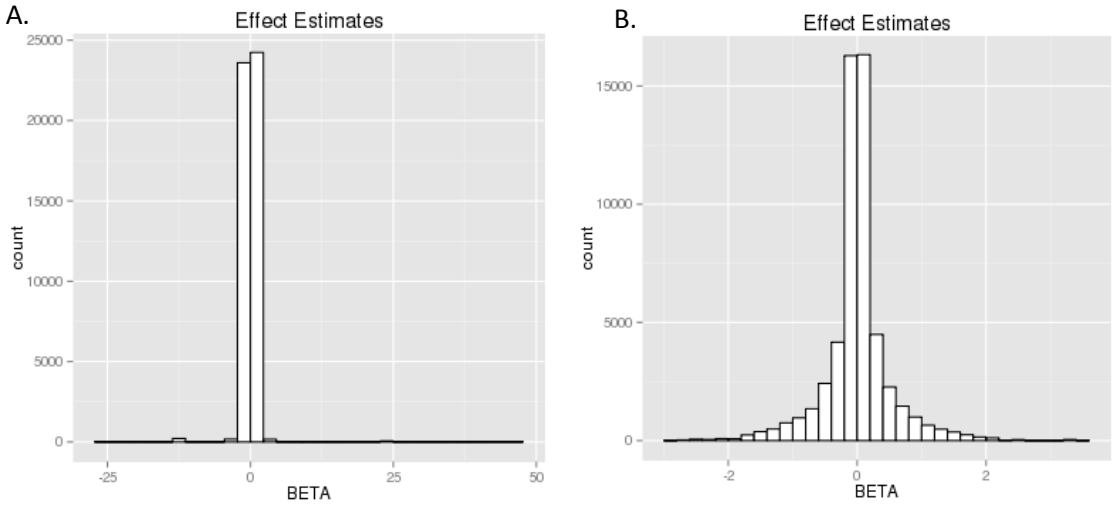


In one study, plots of the effect estimates showed that there were a small number of SNPs for which there were extreme effect estimate (β) values (Figure 4-11 A). In this study, the principal components analysis had been carried out using a set of rare exome chip variants, which led to some extreme principal components values, and in turn, extreme effect estimates. The study analyst repeated the principal components analysis using a set of common variants and re-ran all analyses using the updated principal components. The resulting effect estimates were in line with what would be

expected, and was seen in the remaining studies (Figure 4-11 B). No other data QC issues were identified from these plots.

Figure 4-11: Distributions of study level result for a study with extreme values for the effect estimates for some SNPs.

A. Distribution of effect estimates with extreme values, due to incorrect principal components. B. Distribution of effect estimates after analyses re-ran using the updated principal components.



A comparison of the results of study level analyses 1 (Single variant associations) and 2 (RAREMETAL) allowed the identification of other issues. For all analyses undertaken in each study, the P-values, effect estimates and test statistics from the two study level analyses were plotted against each other. In study level analysis 1, the Wald test statistic is calculated as

$$Z = \frac{\beta}{se}$$

In study level analysis 2, the score test statistic may be calculated as

$$S = \frac{U}{\sqrt{V}}$$

Both the Wald test statistic (Z) and the score test statistic (S) follow a standard normal distribution and are asymptotically equivalent.

These comparative plots helped to identify a number of other issues. Figure 4-12 shows a comparison of the results from study level analyses 1 and 2, from a typical study of unrelated individuals. Overall, the effect estimates were highly concordant, although the test statistics and P-values were more conservative from study level

analysis 2. There were a small number of SNPs for which the summary statistics varied greatly in the two study level analyses, although there was no obvious reason for these discrepancies. These highly discordant SNPs were flagged as potentially erroneous results.

Figure 4-13 shows a comparison of a typical study with related individuals. In studies of related individuals, summary statistics from the two study level analyses were less closely correlated than in studies of unrelated individuals. All studies of related individuals used GEMMA for study level analysis 1 and RAREMETALWORKER for study level two. GEMMA and RAREMETALWORKER both utilise mixed models with empirically estimated kinship matrices to account for relatedness (74, 176).The two software packages adopt different methods for estimating the relationship matrices however and this is likely the cause of the differences in summary statistics from the two analyses.

Figure 4-12: Example comparison of results from study level analysis 2 versus study level analysis 1 - Study of unrelated individuals.

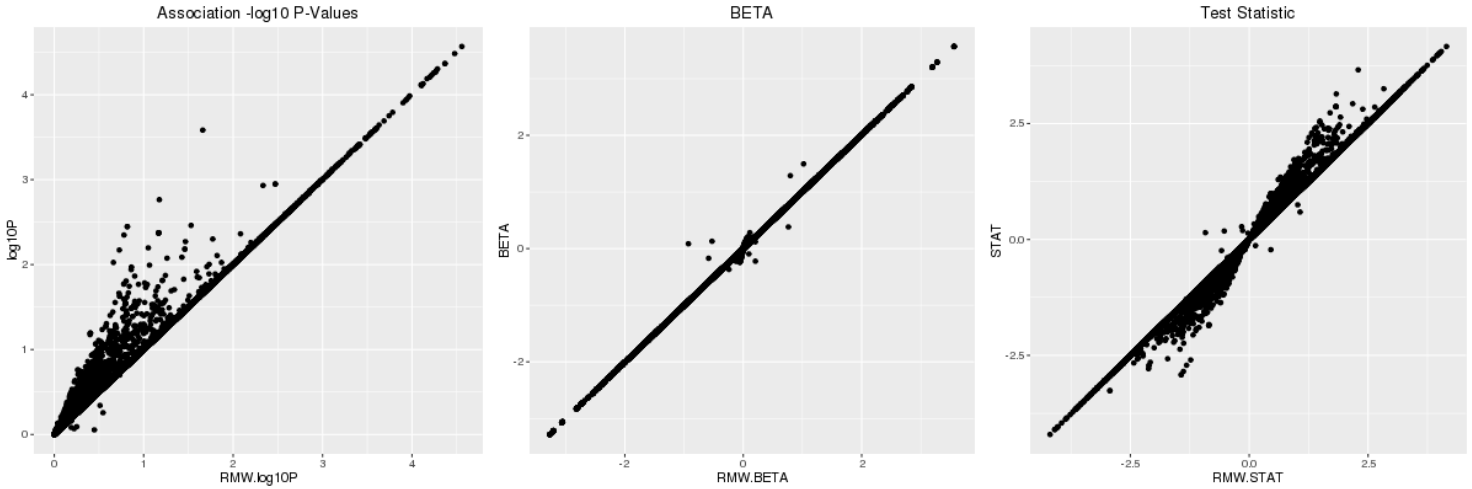
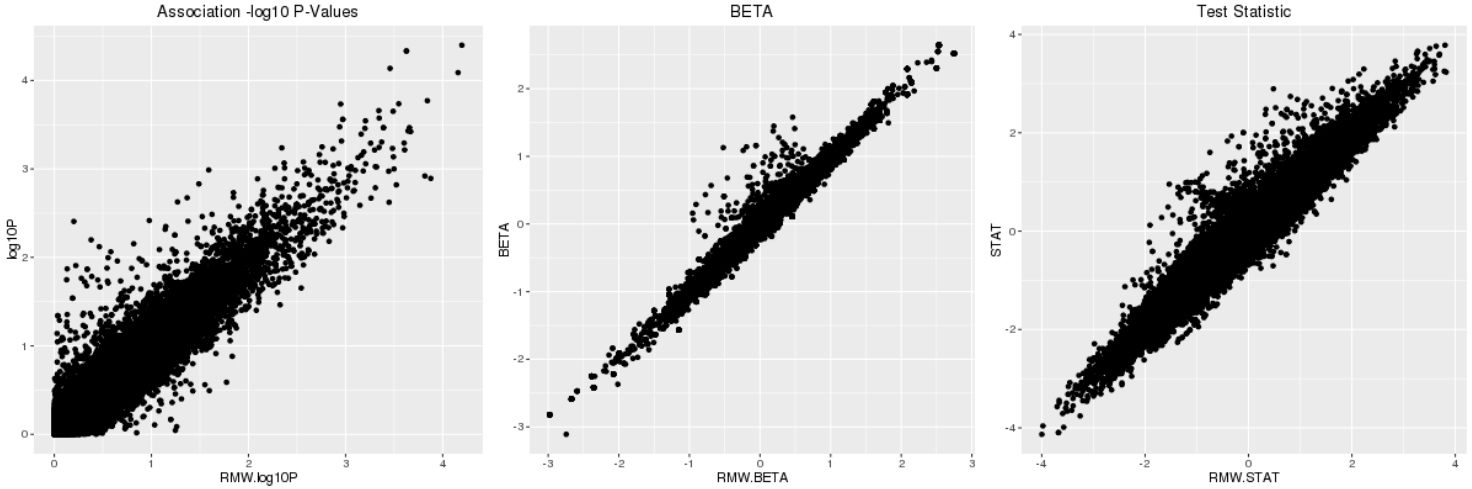
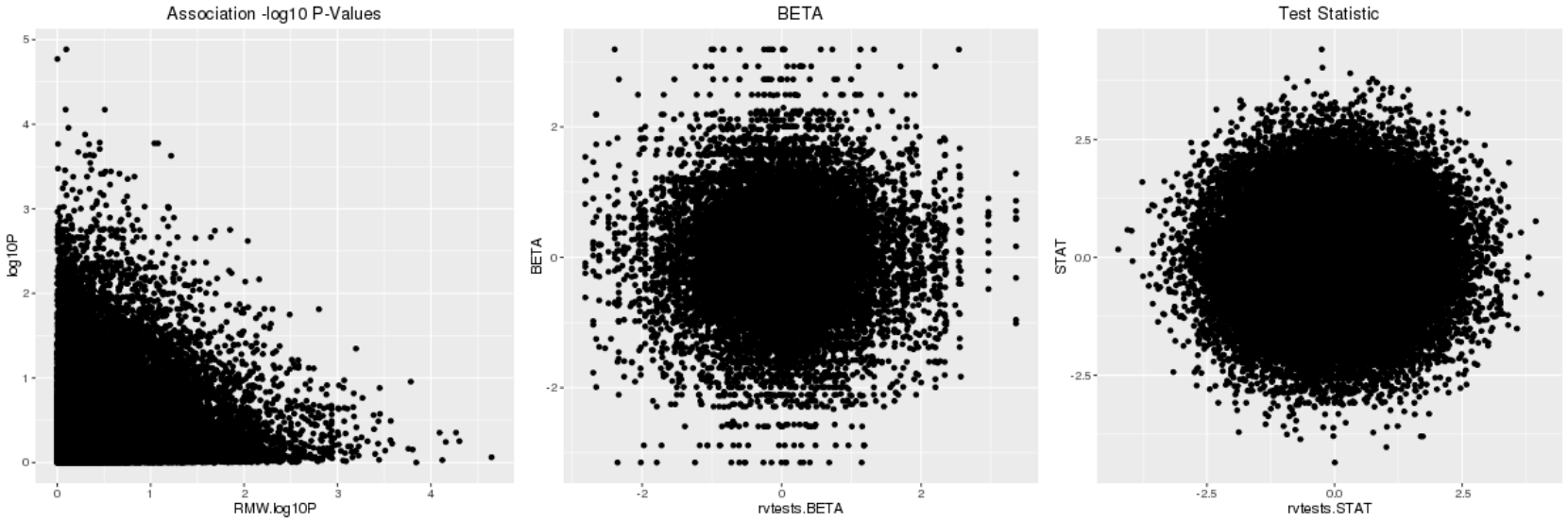


Figure 4-13: Example comparison of results from study level analysis 2 versus study level analysis 1 - Study of related individuals.



There was one study, for which the comparative plots of the analyses of FVC and FEV_1/FVC in never smokers showed the results of study level analyses 1 and 2 were entirely discordant (Figure 4-14). After contacting the original analyst, it transpired the study level analysis 2 for never smokers was in fact conducted using the study's ever smokers. These analyses were repeated using the correct samples, and the resulting summary statistics showed good correlation between the two study-level analyses.

Figure 4-14: Study with discordant summary statistics from study level analyses 1 and 2 (never smokers analysis of FVC).

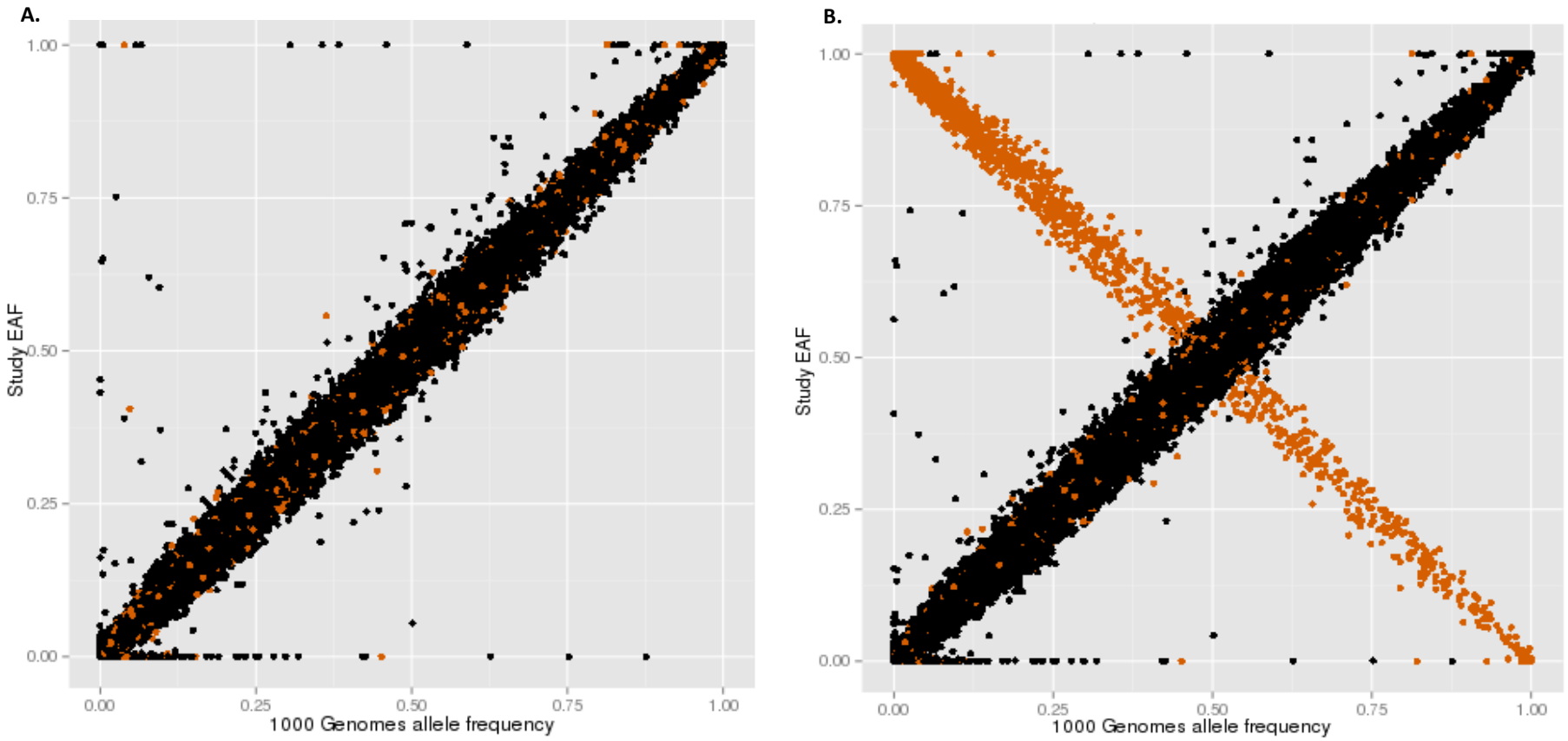


The effect allele frequencies from each study were also plotted against allele frequencies from European samples from the 1000 Genomes reference panel. This was useful, firstly to check the allele frequencies in all studies did not differ significantly than would be expected in a population of European ancestry individuals, and secondly to check that consistent effect alleles were used across all studies. All studies were asked to report alleles on the + strand and Figure 4-15 shows the effect allele frequencies of two studies, plotted against frequencies from 1000 Genomes, based on alleles on the + strand. Figure 4-15 A shows generally good correlation between allele frequencies and is typical of what was seen in most studies. In one study, shown in Figure 4-15 B, there was high agreement for the majority of SNPs although for a subset of SNPs, the allele frequencies suggested the wrong effect allele was reported. All of the inconsistent alleles were either C/G or A/T SNPs and so the inconsistencies were a result of alleles being on the incorrect strand. For the affected SNPs in this study, the direction of the effect estimates and effect allele frequencies were changed so that the effect allele was consistent with other studies. Where the effect allele frequencies of the A/T and C/G SNPs were close to 50%, it was difficult to ascertain which strand was reported and as a result, these SNPs with 45%<EAF<55% were excluded.

Figure 4-15: Plots of study specific effect allele frequencies, against frequencies in European 1000 Genomes samples.

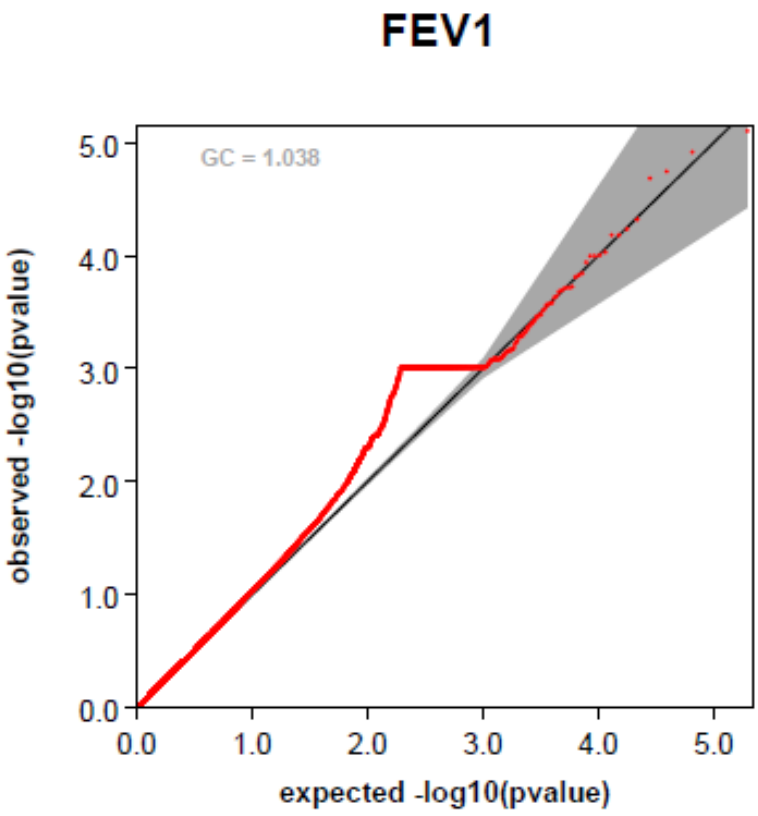
A/T and C/G highlighted in orange.

A. Study where allele frequencies where consistent with 1000 Genomes. B. Study where a number of A/T and C/G SNPs were reported on the incorrect strand.



Finally, for each analysis from all studies, the QQ plots and the genomic inflation factor (λ, equation (1-4)) were inspected. For one study, in the analyses of never smokers, there were a large number of SNPs with identical P-values, manifesting as a horizontal line on the QQ plots (Figure 4-16). It was suggested to the study analyst that this unusual distribution of P-values might be due to one individual with an excess of singletons, that is where there were a large number of SNPs for which that individual was the only one to have the alternative allele. In the example illustrated in Figure 4-16, it appears that individual might also have a fairly extreme FEV_1, resulting in a number of identical and significant associations. Since this distribution of P-values was only seen in the analyses of never smokers, this additionally supported the idea that this anomaly was due to a single sample.

Figure 4-16: QQ plot of study with an individual with an excess of singletons.
The effect of this individual was a large number of identical P-values (indicated by horizontal line on plot).



The study analyst found that the unusual distribution of P-values in the never-smokers analyses was indeed due to a single individual, and re-ran the analyses without this individual. Once this issue had been resolved, λ values for all studies were not

considerably greater than 1, indeed for many analyses, there was deflation of the test statistics (λ values for study level analysis 2 of FEV_1 shown in Table 4-6 as an illustrative example).The QQ plots showed that the distributions of test statistics did not significantly deviate from the null, suggesting there were no issues of underlying population structure and that familial relationships were correctly accounted for in the analyses.

Table 4-6: Genomic inflation factor (λ) values for study level analysis 2 of FEV_1.

	Genomic Inflation factor (λ)	
Study	**Smokers**	**Non-smokers**
1958BC	0.986	0.977
KORCULA	1.021	0.958
KORA F4	1.039	1.034
LBC1936	1.036	1.018
PIVUS	0.976	0.979
SHIP	0.842	0.942
GS:SFHS	0.991	0.978
FIN	-	1.026
NSPHS	0.823	0.997
SAPALDIA	0.985	1.043
YFS	1.052	0.945

4.3 Meta-Analyses Methods

4.3.1 Discovery meta-analysis of single variant associations

4.3.1.1 Meta-analysis 1 (R)

Using the summary level data from study level analyses 1, an inverse variance weighted meta-analysis (as in equation (2-3)) was firstly carried out using R, to estimate pooled effect estimates (β_{meta_j}) and standard errors (se_{meta_j}), as follows. After ensuring that effect estimates across all studies corresponded with consistent effect alleles, the standard errors of all study level effect estimates were adjusted using genomic control where λ>1, using study-specific genome inflation factors (equation (1-4)). Results for ever and never smokers were combined within each study, using inverse variance weighted meta-analysis. Overall effect estimates and standard errors from each study were then combined, again through meta-analyses

using inverse variance weights. The standard errors of the resulting pooled effect estimates were then adjusted using genomic control (λ calculated using pooled results) and P-values were calculated, using the Wald test:

$$Z_{meta_j} = \frac{\beta_{meta_j}}{se_{meta_j}} \sim N(0,1)$$

4.3.1.2 Meta-analysis 2 (RAREMETAL)

The summary statistics from study level analyses 2 were utilised in a second meta-analysis using RAREMETAL (74). Score statistics from each study were combined using a Cochran-Mantel-Haenzsel meta-analysis:

$$S_{meta_j} = \frac{\sum_{k=1}^{k} U_{jk}}{\sqrt{\sum_{k=1}^{k} V_{jjk}}} \sim N(0,1) \quad \text{(4-1)}$$

where U_j is the score statistic and V_{jjk} is the variance of the score statistic of the j^{th} SNP of the k^{th} study. The genomic inflation factor was calculated using the pooled results and where λ (equation (1-4)) was greater than 1, genomic control adjustment was applied to the resulting test statistic and P-values.

4.3.2 Comparison of R and RAREMETAL methods

After completing both meta-analyses, I compared the results, to ensure their consistency. For brevity, only the comparison of the analyses of FEV_1 is shown here, however comparisons for all traits were similar. Figure 4-17 shows the $-\log_{10}$ P-values (A) and effect estimates (B), from the two meta-analyses. Overall, the results from these two analyses were highly correlated, although not identical. Crucially, the P-values and effect estimates were highly concordant in the two analyses for the most strongly associated SNPs (Figure 4-18). A total of 25 SNPs had $P<10^{-4}$ in either meta-analysis: 21 SNPs in meta-analysis 1 and 24 SNPs in meta-analysis 2. 20 SNPs had $P<10^{-4}$ in both analyses, with the remaining 5 SNPs falling just short of this significance threshold in one analysis.

Figure 4-17: Comparison of Meta-Analysis methods for the analysis of FEV_1.
A. Comparison of $-\log_{10}$ P-values. B. Comparison of effect estimates (Betas).

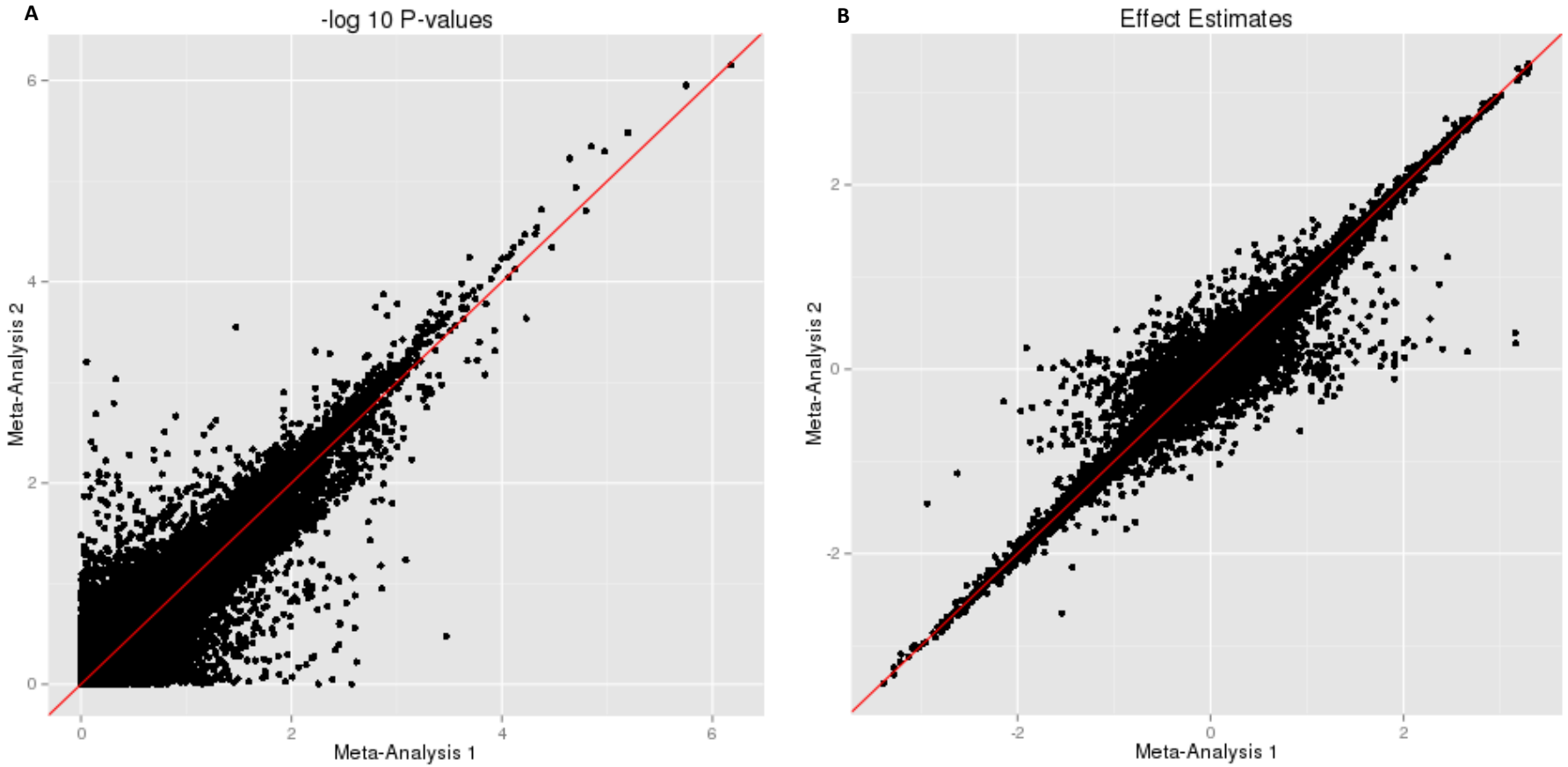
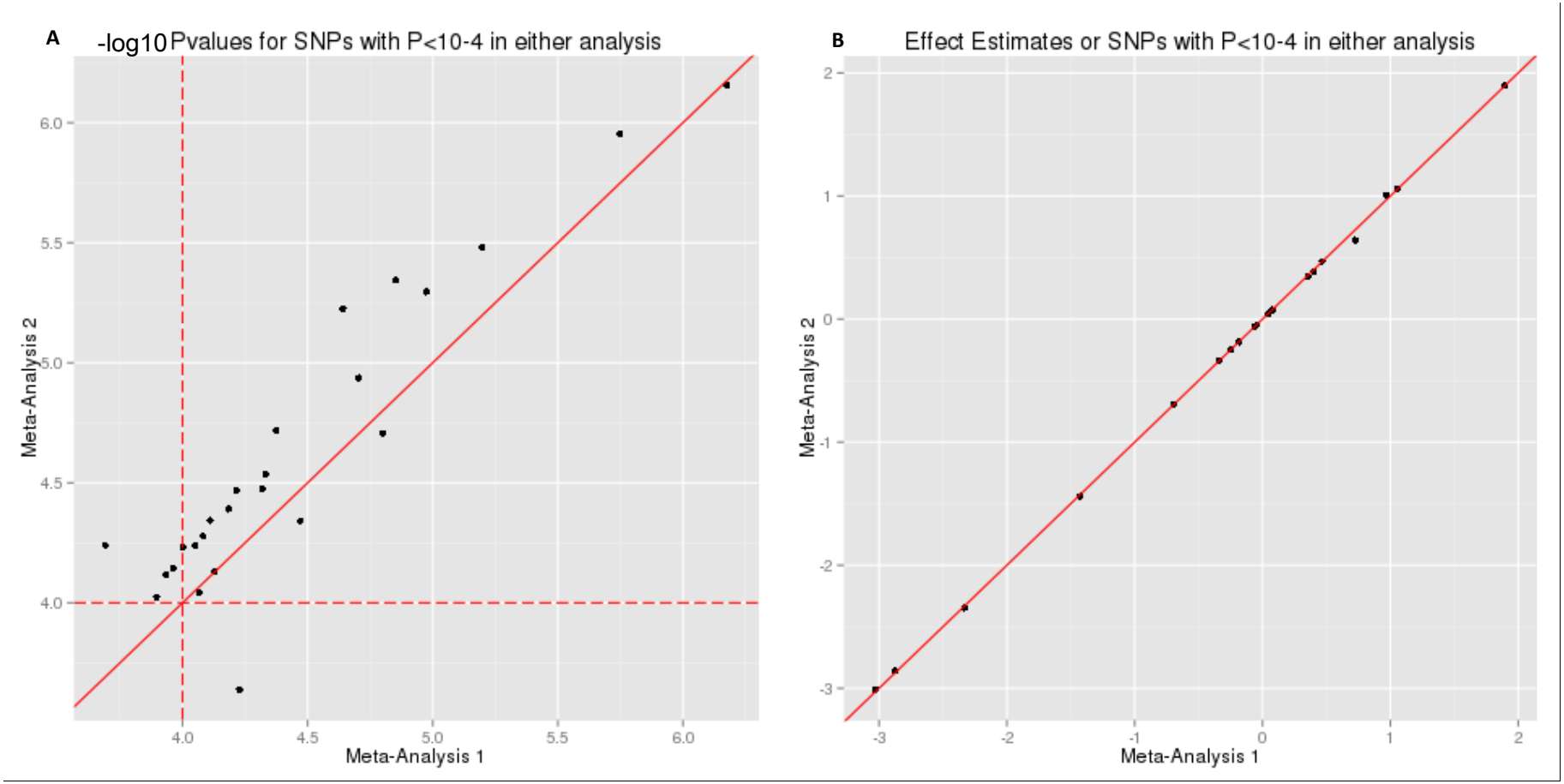


Figure 4-18: Comparison of Meta-Analysis methods for the analysis of FEV_1, restricted to SNPs with $P<10^{-4}$ in either analysis.

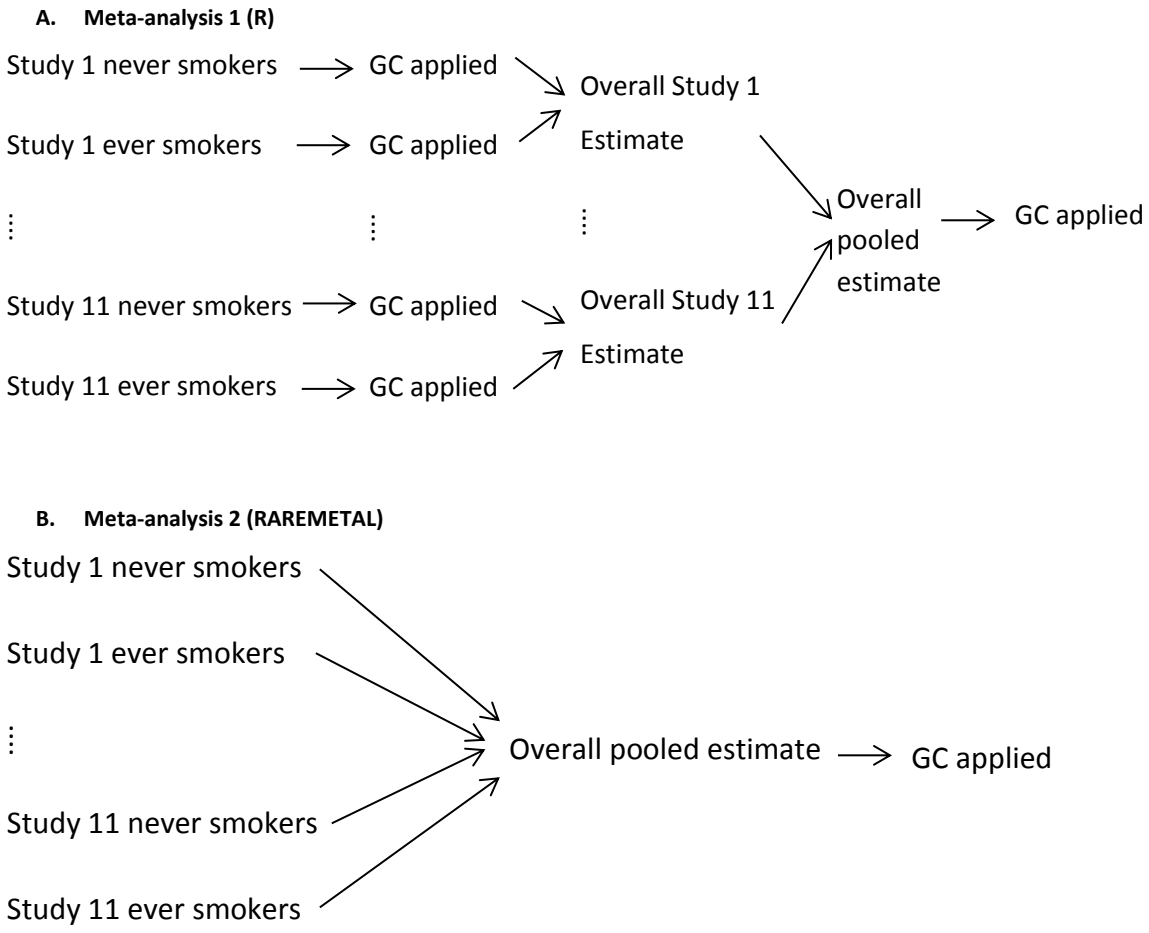
A. Comparison of $-\log_{10}$ P-values. B. Comparison of effect estimates (Betas).



There are a number of explanations for the discrepancies of the results of the less strongly associated SNPS. Firstly, as was seen in Section 4.2.6, the two sets of summary statistics generated at study-level for the two meta-analyses were not identical, particularly in studies of related individuals. Furthermore, the number of samples and SNPs included from each study was not always consistent for the two meta-analyses: there were two studies where the number of samples included in each of the two study level analyses differed and in many studies, there were some SNPs which had missing summary statistics in one study level analyses, and not the other.

Secondly, the meta-analysis methods were not identical. In meta-analysis 1, non-smokers and smokers within each study were first meta-analysed to give overall study estimates, with these overall study estimates then being combined in an overall meta-analysis. This two-stage meta-analysis is consistent with previous large scale meta-analyses of lung function (43). The RAREMETAL software used in meta-analysis 2 did not allow this two-step meta-analysis, rather non-smokers and smokers are treated as individual studies (Figure 4-19). A further difference in the two meta-analyses was the way adjustments for genomic control were implemented. In meta-analysis 1, standard errors of the study level effect estimates for never smokers and ever smokers were adjusted using analysis specific λ values before meta-analysis, with further genomic control adjustment carried out on the final pooled standard errors. RAREMETAL does not adjust study level results, instead correction for genomic control was applied only to the final standard errors, after meta-analysis. This likely resulted in the less conservative P-values for meta-analysis 2 (Figure 4-18) and indeed why 24 SNPs reached the $P<10^{-4}$ significance level in meta-analysis 2, compared to 21 SNPs in meta-analysis 1.

Figure 4-19: Summary of two meta-analyses.



Based the comparison of the two meta-analyses overall, I was confident that the results from both study level analyses were sufficiently consistent. For the final analyses the results from meta-analyses 2 were used as this would allow for consistency in the study level data contributing to the meta-analyses of both single variant associations and gene-based associations. The remainder of this chapter is therefore based on the results of meta-analysis 2.

4.3.3 Replication of single variant associations

4.3.3.1 Selection of SNPs for follow-up

Since low frequency and rare variants were of particular interest in this analysis, no MAF or minor allele count (MAC) filter was applied. SNPs of interest were identified as

those with P <10^{-4}. Where a SNP in close proximity to a previously identified lung function SNP was identified, the SNP was deemed to represent an independent signal if it had r^2<0.2 with the known SNP, and if it retained P <10^{-4} when conditional analyses were carried out with the known SNP, or a genotyped proxy, where this was possible. Conditional analyses were carried out using the RAREMETAL package, and utilising linkage disequilibrium estimates based on the variance-covariance matrices of single variant score statistics, similarly to the method described by Yang et al (177). Where a SNP was associated with more than one trait, the SNP was followed up with the trait for which it was most significantly associated only.

4.3.3.2 Two Stage Replication Design

I undertook a two-stage replication analysis, utilising samples from UK BiLEVE and UKHLS (described below), alongside a look-up from the CHARGE Consortium. The results of the discovery and replication analyses were combined using a sample-size weighted Z-score meta-analysis:

$$Z_{meta_j} = \frac{\sum_{k=1}^{K} w_k Z_{jk}}{\sqrt{\sum_{k=1}^{K} w_k^2}} \sim N(0,1) \qquad \textbf{(4-2)}$$

where w_k is the weight of the k^{th} study, with sample size $n_{k:}$

$$w_k = \sqrt{n_k}$$

and Z_{jk} is the Z-score, estimated using the P-value (P_{jk}) for the j^{th} SNP from the k^{th} study:

$$Z_{jk} = \phi^{-1}(1 - P_{jk})$$

This method was adopted as it required only the P-value and direction of effect from each replication analysis, thus allowing the inclusion of the results from the CHARGE analysis as these were based on untransformed traits, as opposed to the inverse normally transformed traits which were used in all other studies.

4.3.3.3 Replication Stage Samples

The UK BiLEVE samples (n=48,943) were genotyped using the Affymetrix Axiom UK BiLEVE array, which has substantial overlap with the Illumina Human Exome BeadChip. The QC and imputation procedure of the UK BiLEVE genotype data is described elsewhere (120). In brief, thorough sample and genotype QC was undertaken before imputation to a combined 1000 Genomes (16) and UK10K Project (40) reference panel. Following imputation, SNPs were excluded if they had imputation (INFO) score ≤0.5 or MAC <3.

UKHLS samples (n=7449) were genotyped using the Illumina Infinium HumanCoreExome-12 v1.0 BeadChip and genotype calling was performed using Illumina's GenCall software (2). Samples were excluded using the following filters: call rate<98%, heterozygosity outliers (>3 SD), gender mismatches, duplicates and ancestral outliers. SNPs were mapped to the forward strand of human genome build 37 and QC was performed as follows: SNPs with HWE $P<1\times10^{-4}$, a call rate < 98% and poor genotype clustering values (<0.4) were removed.

Details of the QC undertaken by the CHARGE Consortium are described elsewhere (53). The CHARGE consortium pulmonary function analysis includes 44,719 samples (36,998 European ancestry and 7721 African ancestry). Only the results for European samples are included in these analyses.

All study level QC and analysis of single variant associations in the replication samples were carried out by individual study analysts. I undertook the central meta-analyses of these data.

4.3.3.4 Stage 1 Replication in UK BiLEVE

All SNPs of interest identified in the discovery analysis were taken forward to a first stage of replication, using samples from UK BiLEVE. Analyses with all traits were carried out separately in never smokers and heavy smokers. All traits were adjusted for age, age^2, height, sex and 10 principal components, with additional pack-years adjustment for heavy smokers. Residuals were inverse normally transformed separately for each smoking stratum and used as the trait for association testing.

Associations were carried out using the score test as implemented in SNPTEST v2.5b4 (19), with results for never and heavy smokers combined using sample size weighted Z-score meta-analysis. Overall UK BiLEVE estimates and SpiroMeta discovery estimates were then meta-analysed.

Any SNP still meeting the $P<10^{-4}$ significance threshold in the meta-analysis, and with consistent direction of effect in both SpiroMeta and UK BiLEVE were selected for stage 2 replication. Any SNP not available in UK BiLEVE (either genotyped or imputed) was also selected for stage 2 replication.

4.3.3.5 Stage 2 replication

The second stage of replication utilised samples from UKHLS, and look-ups from a concurrent analysis undertaken by the CHARGE consortium. For the subset of SNPs taken forward to the second replication stage, UKHLS generated study level summary statistics for never smokers and ever smokers, analogously to the discovery studies, as per the analysis plan (study level analysis 2) in appendix C. These summary statistics from never and ever smokers were then centrally meta-analysed using RAREMETAL to give overall study estimates. The CHARGE Consortium carried out a meta-analysis of FEV_1, FVC and FEV_1/FVC. In their analyses, all traits were adjusted for former smoking, current smoking and pack-years of smoking, age, age^2, sex, height, height^2, centre and principal components. FVC was additionally adjusted for weight. For each SNP in the stage 2 replication, effect sizes, standard errors and P-values were provided from the CHARGE meta-analysis, for all traits in ever and never smokers and in all samples combined.

The results from UKHLS and the CHARGE consortium look-ups were combined with the results from the discovery analysis and stage 1 replication in a sample size weighted Z-score meta-analysis (equation (3-6)). SNPs with overall exome-wide significance of $P<2.7\times10^{-7}$ (Bonferroni corrected for 189,962 SNPs tested) are reported as novel loci.

4.3.4 Discovery meta-analyses of gene-based associations

Using combined score statistics and variance-covariance matrices, two gene-based tests were constructed using RAREMETAL: firstly the weighted sum test (WST,

equation (3-1)) using Madsen Browning weightings (66), which performs well when variants within a gene have similar directions and magnitude of effect; and secondly SKAT (equation (3-3)), which is more powerful where there are both protective and deleterious variants within a gene (68). Variants were annotated to genes using ANNOVAR (150) on the basis of the GRCh37 gene database. Analyses were restricted to include only exonic SNPs with MAF<5%, and only genes with at least two such variants were included. For any gene with $P<10^{-4}$, additional analyses were carried out, which conditioned on the most significantly associated individual SNP within that gene, to determine whether this was a true gene-based signal, or whether the association could be ascribed to the single SNP.

4.3.5 Replication of gene-based associations

All genes of interest ($P<10^{-4}$) were followed up using data from UK BiLEVE and UKHLS. I undertook the study-level RAREMETALWORKER analysis for UK BiLEVE and the central meta-analyses of these data.

Summary statistics were generated in UK BiLEVE using RAREMETALWORKER and including directly genotyped SNPs only. Equivalent summary statistics were generated by UKHLS, using RAREMETALWORKER. Firstly, gene-based tests were constructed in RAREMETAL using the summary statistics from UK BiLEVE and UKHLS only. Secondly, the results from the discovery cohorts with the two replication cohorts were combined in an overall combined meta-analysis using RAREMETAL. Any genes with overall $P<2.4x10^{-6}$ (Bonferroni corrected for 14,865 genes tested) in our combined meta-analysis was declared statistically significant. Further supporting evidence for the genes of interest was sought through a look-up of gene-based associations within the CHARGE consortium.

4.3.6 Smoking stratum specific analyses

Secondary discovery meta-analyses were additionally undertaken in ever smokers (n=11,632) and never smokers (n=11,766) separately. For the single variant associations, a two-stage replication was carried out, similarly to the analyses utilising all samples. Any SNP with $P<10^{-4}$ was selected for a first stage of replication, using the

corresponding smoking stratum from UK BiLEVE (n=24,460 ever smokers and n=24,483 never smokers). Any SNP still meeting the $P<10^{-4}$ significance threshold in SpiroMeta and UK BiLEVE combined, along with all SNPs not genotyped or imputed in UK BiLEVE, were selected for a second stage of replication using UKHLS samples only (n=4509 ever smokers and n=2940 never smokers).

Gene-based association analyses in never smokers and ever smoker separately were also undertaken. For all genes of interest from these analyses ($P<10^{-4}$), replication was carried out in UK BiLEVE and UKHLS, analogously to the analyses of all individuals, described above.

4.3.7 Tests of heterogeneity

For the SNPs identified in the single variant association analyses within novel regions, heterogeneity was tested for amongst the discovery stage samples, using Cochran's Q statistic, calculated as follows:

$$Q_j = \sum_{k=1}^{K} w_{jk}\left(\beta_{metaj} - \beta_{jk}\right)^2 \sim \chi^2_{K-1}$$

where β_{metaj} is the pooled effect size estimate and , β_{jk} and w_{jk} are the study-specific effect estimates and weights, respectively for the j^{th} SNP. The Q statistic follows a chi-squared distribution with *K-1* degrees of freedom, where K is the number of studies (ever smokers and never smokers from each sample collection treated as separate studies).

4.3.8 Functional characterisation of novel loci

The sentinel SNPs, and proxies ($r^2>0.3$) within newly identified regions were assessed in three eQTL data sets.

Firstly, associations with blood eQTLs were searched for within a publicly available blood eQTL dataset with results from the analysis of 5,311 individuals, imputed to HapMap 2 (178). Association testing was undertaken both for cis (+/- 250Kb distance between the SNP and the probe midpoint) and trans (distance between the SNP and the probe midpoint >5Mb) eQTLs. All eQTL signals detected at a false-discovery rate

(FDR) of 50% were available in the dataset; signals meeting the 10% FDR genome-wide significant threshold were identified.

Secondly, the selection of SNPs were assessed in lung tissue expression data from 124 samples from the GTEx project (179). Only cis-eQTLs (+/- 1Mb distance between the SNP and transcription start site) meeting the 5% FDR genome-wide significant threshold were available in the dataset.

Finally, selected SNPs were assessed in a lung eQTL resource based on lung tissues of 1,111 individuals. The descriptions of the lung eQTL dataset and subject demographics have been published previously (180-182). Briefly, non-tumor lung tissues were collected from patients who underwent lung resection surgery at three participating sites: Laval University (Quebec City, Canada, n=409), University of Groningen (Groningen, The Netherlands, n=363), and University of British Columbia (UBC, Vancouver, Canada, n=339). Whole-genome gene expression and genotype data imputed to the 1000 Genomes Project (16) reference panel were available for all samples. eQTLs were identified as either cis (within 1 Mb of transcript start site) or in trans (all other eQTLs) and meeting the 10% FDR genome-wide significant threshold.

4.4 Meta-Analyses Results

The meta-analysis of single variant associations consisted of a discovery analysis including 23,398 samples from 11 studies, followed up with two replication stages, including 48,943 (replication stage 1) and up to 44,447 (replication stage 2) samples, respectively. Moreover the joint effects of rare variants were investigated through a meta-analysis of gene-based associations. Study-specific characteristics of the samples included in these analyses are described in Table 4-7.

Table 4-7: Characteristics of samples from 11 SpiroMeta cohorts contributing to the discovery analyses and 2 replication cohorts.

Discovery Cohorts							
Study Name	**n**	**n (%) Male**	**Ever Smokers, n (%)**	**Age, mean (SD)**	**FEV_1, litres. mean (SD)**	**FVC, litres. mean (SD)**	**FEV_1/FVC, mean (SD)**
1958 British Birth Cohort (1958BC)	5270	2961 (56.19%)	2866 (53.29%)	44.00 (0.00)	3.35 (0.79)	4.29 (1.03)	0.788 (0.09)
Generation Scotland (GS:SFHS)	8164	3413 (41.81%)	3806 (46.62%)	51.59 (13.33)	2.78 (0.87)	3.91 (1.01)	0.710 (0.12)
Cooperative Health Research in the Region of Augsburg (KORA F4)	1447	701 (48.45%)	900 (62.20%)	54.82 (9.66)	3.24 (0.85)	4.20 (1.04)	0.771 (0.07)
CROATIA-Korcula cohort (KORCULA)	791	296 (36.82%)	418 (51.99%)	55.56 (13.69)	2.72 (0.83)	3.29 (0.95)	0.829 (0.10)
Lothian Birth Cohort 1936 (LBC1936)	974	501 (50.55%)	554 (55.90%)	69.55 (0.84)	2.38 (0.67)	3.04 (0.87)	0.787 (0.10)
Study of Health in Pomerania (SHIP)	1681	831 (49.43%)	955 (56.81%)	52.25 (13.43)	3.29 (0.88)	3.88 (1.03)	0.848 (0.07)
Northern Swedish Population Health Study (NSPHS)	880	407 (46.25%)	122 (13.86%)	49.13 (19.96)	2.93 (0.90)	3.53 (1.06)	0.831 (0.09)
Prospective Investigation of the Vasculature in Uppsala Seniors (Pivus)	836	413 (49.4%)	426 (50.96%)	70.20 (0.17)	2.44 (0.68)	3.20 (0.87)	0.76 (0.10)
Swiss study on Air Pollution and Lung Disease in adults (SAPALDIA)	2707	1379 (50.9%)	1399 (51.7%)	40.86 (10.92)	3.65 (0.83)	4.62 (1.04)	0.794 (0.07)
The Cardiovascular Risk in Young Finns Study (YFS)	434	198 (47.3%)	186 (44.4%)	38.88 (5.07)	3.73 (0.75)	4.68 (0.99)	0.80 (0.06)
Finnish Twin Cohort (FIN)	214	0 (0%)	0 (0%)	68.73 (3.31)	2.18 (0.47)	2.79 (0.58)	0.786 (0.08)
Total Discovery Sample Size	23,398						
Replication Cohorts							
Study Name	**n**	**n (%) Male**	**Ever Smokers, n (%)**	**Age, mean(SD)**	**FEV_1, litres. mean (SD)**	**FVC, litres. mean (SD)**	**FEV_1/FVC, mean (SD)**
UK Biobank Lung Exome Variant Evaluation study (UK BiLEVE) (Replication Stage 1)	48,943	24,489 (50.0%)	24,460 (50.0%)	56.93 (7.89)	2.65 (0.87)	3.59 (1.05)	0.73 (0.08)
UKHLS (Replication Stage 2)	7449	3293 (44.2%)	4509 (60.5%)	53.10 (15.94)	2.89 (0.90)	3.83 (1.08)	0.753 (0.09)

4.4.1 Meta-Analyses of single variant associations

Firstly single variant associations were evaluated between FEV_1, FVC and FEV_1/FVC and the 189,962 SNPs which passed study level quality control and were polymorphic in at least one of the 11 discovery stage studies. The QQ plots and genomic inflation factors (λ) for each meta-analysis are shown in Figure 4-20. The λ values for FEV_1 and FEV_1/FVC were both less than one. In the meta-analysis of FVC, λ was slightly above one, at 1.043; the $-\log_{10}$ P-values shown in the QQ plot have been adjusted accordingly.

The analyses of associations across all discovery samples identified a total of 50 SNPs in 49 regions not previously associated with lung function, showing association with at least one trait at $P<10^{-4}$ (Table 4-8).

Figure 4-20: Quantile-quantile plots for the meta-analyses of A. FEV_1, B. FVC and C. FEV_1/FVC.

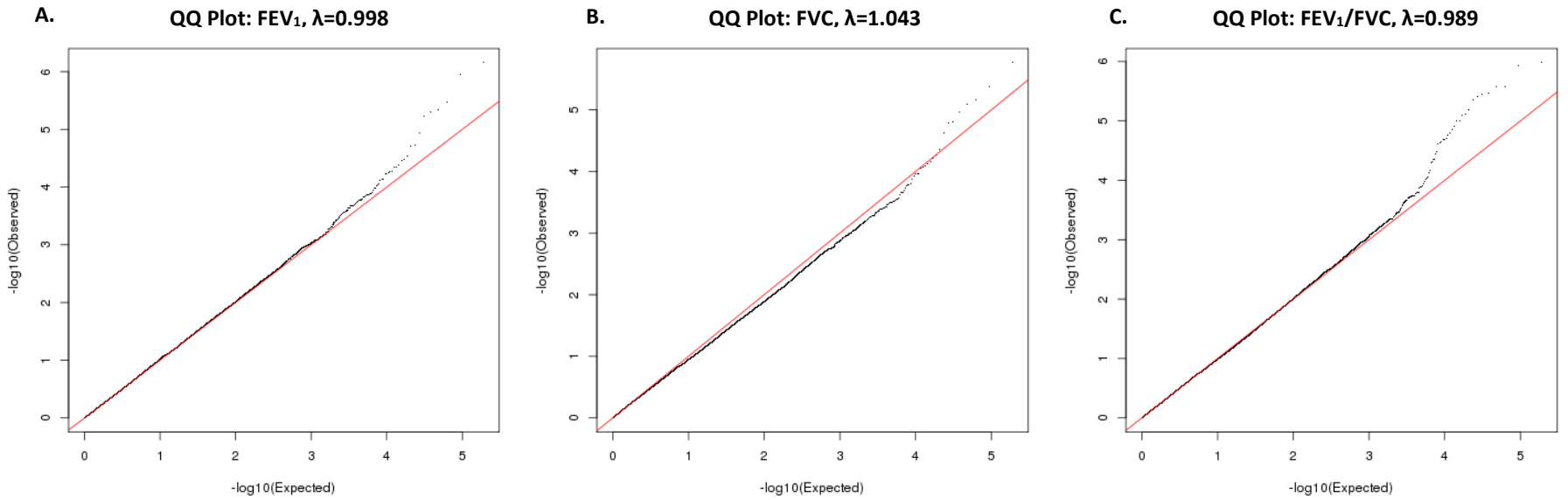


Table 4-8: All SNPs showing association ($P<10^{-4}$) with FEV_1, FVC or FEV_1/FVC in the discovery stage meta-analysis.

SNPs ordered by chromosome (Chr) and genomic position (Pos). Only variants in novel loci shown, and only the trait for which each SNP was most significantly associated is shown. All P-values are two-sided. Beta values reflect effect-size estimates on an inverse-normal transformed scale after adjustments for age, age^2, sex, height and ancestry principal components, and stratified by ever smoking status.

SNP	Chr:Pos	Gene(s)	Function	Effect / other allele	Trait	N	Effect allele frequency (EAF)	Beta	P-value
rs201163722	1:28661302	*MED18*	exonic	T/C	FEV_1	23386	0.006%	-2.347	$5.29x10^{-5}$
rs200090958	1:86951099	*CLCA1*	exonic	A/G	FEV_1/FVC	23397	0.004%	2.783	$9.67x10^{-5}$
rs141436979	1:89844088	*GBP6*	exonic	T/C	FEV_1/FVC	23396	0.017%	-1.486	$2.44x10^{-5}$
rs1192415	1:92077097	*CDC7, TGFBR3*	intergenic	A/G	FVC	23376	81.24%	-0.059	$6.11x10^{-7}$
rs11204697	1:150658971	*PTCD3*	exonic	T/C	FEV_1	23393	0.030%	1.060	$7.62x10^{-5}$
rs35608243	2:174131392	*ZAK*	exonic	C/T	FVC	23395	8.17%	0.081	$1.60x10^{-6}$
rs148627602	2:209309610	*PTH2R*	exonic	A/G	FEV_1	23397	0.034%	1.009	$5.75x10^{-5}$
rs144052038	3:49720010	*APEH*	exonic	G/A	FVC	23392	0.066%	0.755	$2.59x10^{-5}$
rs141921900	3:74334458	*CNTN3*	exonic	A/G	FEV_1/FVC	23384	1.32%	-0.178	$1.21x10^{-5}$
rs62290268	3:194790799	*XXYLT1*	exonic	G/C	FEV_1	23394	0.079%	0.681	$5.47x10^{-5}$
rs3733250	4:77192868	*FAM47E, FAM47E-STBD1*	exonic	A/G	FVC	23392	41.02%	-0.044	$2.71x10^{-6}$
rs142127543	4:90833153	*MMRN1*	exonic	A/G	FEV_1/FVC	23395	0.143%	0.522	$2.09x10^{-5}$
rs17037102	4:107845794	*DKK2*	exonic	T/C	FEV_1/FVC	23394	10.34%	-0.062	$4.78x10^{-5}$
rs79300690	4:122250654	*QRFPR*	exonic	A/G	FEV_1/FVC	23385	1.77%	0.143	$4.44x10^{-5}$
rs147517729	4:147561147	*POU4F2*	exonic	A/C	FVC	18294	1.46%	-0.202	$4.56x10^{-6}$
rs772835	5:944298	*TRIP13, LOC100506688*	intergenic	G/A	FVC	23395	9.95%	-0.067	$2.03x10^{-5}$
rs17648108	5:177831556	*COL23A1*	intronic	C/T	FVC	12820	28.38%	0.046	$6.93x10^{-6}$
rs1294421	6:6743149	*LY86, RREB1*	intergenic	G/T	FEV_1/FVC	23395	61.47%	0.037	$9.14x10^{-5}$
rs3749903	6:42992825	*RRP36*	exonic	G/C	FVC	23389	12.38%	-0.072	$2.95x10^{-5}$

SNP	Chr:Pos	Gene(s)	Function	Effect / other allele	Trait	N	Effect allele frequency (EAF)	Beta	P-value
rs9784763	6:109624937	*CCDC162P*	ncRNA intronic	A/G	FEV_1	15224	39.24%	0.039	$4.07x10^{-5}$
rs143974258	6:136552493	*MTFR2*	exonic	A/G	FEV_1/FVC	22559	0.064%	0.740	$7.35x10^{-5}$
rs57658073	8:24775940	*NEFM*	exonic	A/G	FEV_1	23297	0.279%	0.344	$9.04x10^{-5}$
rs146520900	8:145667730	*TONSL*	exonic	A/G	FEV_1	22392	0.375%	0.386	$7.00x10^{-7}$
rs141834891	9:12694063	*TYRP1*	exonic	T/C	FEV_1/FVC	23397	0.077%	0.755	$6.45x10^{-6}$
rs2773347	9:100388197	*TSTD2*	exonic	T/C	FEV_1	23391	67.49%	0.043	$1.16x10^{-5}$
rs143386455	9:107533244	*NIPSNAP3B*	exonic	C/G	FEV_1/FVC	23139	0.080%	0.648	$8.81x10^{-5}$
rs41278437	9:113170060	*SVEP1*	exonic	A/G	FEV_1/FVC	23397	0.066%	-0.769	$2.24x10^{-5}$
rs17578859	9:139879170	*LCNL1*	exonic	A/G	FEV_1	23387	25.90%	0.048	$5.95x10^{-6}$
rs141541697	10:92635830	*RPP30*	exonic	T/C	FEV_1	23361	0.013%	1.897	$3.30x10^{-6}$
rs61736639	11:14891141	*PDE3B*	exonic	C/G	FEV_1	23396	0.583%	0.238	$8.58x10^{-5}$
rs188851356	11:125647897	*PATE2*	exonic	A/G	FEV_1/FVC	23397	0.088%	-0.718	$1.00x10^{-5}$
rs187124232	11:126144859	*FOXRED1*	exonic	G/C	FEV_1/FVC	23392	0.077%	-0.780	$7.61x10^{-6}$
rs35639297	12:56142553	*GDF11*	exonic	T/G	FEV_1/FVC	23396	0.479%	-0.278	$3.26x10^{-5}$
rs142653430	12:121469271	*OASL*	exonic	A/G	FEV_1	23395	0.0006%	-2.858	$1.12x10^{-6}$
rs201930455	12:129360559	*GLT1D1*	exonic	A/G	FEV_1	23396	0.0004%	-3.011	$3.35x10^{-5}$
rs7984952	13:31231806	*USPL1*	exonic	C/T	FVC	23394	40.72%	-0.039	$4.14x10^{-5}$
rs3742302	13:31233063	*USPL1*	exonic	A/G	FVC	23358	40.72%	-0.039	$3.64x10^{-5}$
rs149470963	13:67477723	*PCDH9*	exonic	T/G	FEV_1	12633	0.146%	-0.690	$2.90x10^{-5}$
rs11558436	14:32257065	*NUBPL*	exonic	C/A	FEV_1	23397	0.611%	-0.247	$3.40x10^{-5}$
rs1952153	14:87775721	*LOC283585, GALC*	intergenic	C/A	FVC	23390	57.67%	-0.039	$3.29x10^{-5}$
rs61991737	14:93712290	*BTBD7*	exonic	A/C	FEV_1	23397	0.154%	0.469	$5.85x10^{-5}$

SNP	Chr:Pos	Gene(s)	Function	Effect / other allele	Trait	N	Effect allele frequency (EAF)	Beta	P-value
rs118125046	15:79586782	*ANKRD34C*	exonic	G/C	FEV_1/FVC	23376	0.753%	0.219	$4.38x10^{-5}$
rs3751093	17:25958304	*LGALS9*	exonic	A/G	FVC	12633	20.90%	-0.070	$6.62x10^{-6}$
rs144042976	19:37975803	*ZNF570*	exonic	A/G	FEV_1	23386	0.021%	-1.439	$5.06x10^{-6}$
rs149178822	19:40540724	*ZNF780B*	exonic	C/A	FEV_1/FVC	12631	1.26%	-0.233	$5.05x10^{-5}$
rs146608853	20:49225233	*FAM65C*	exonic	A/G	FEV_1/FVC	23397	0.036%	-1.085	$7.90x10^{-6}$
rs200373931	20:62193999	*HELZ2*	exonic	T/C	FVC	23381	0.024%	1.237	$4.46x10^{-5}$
rs140025782	21:28216862	*ADAMTS1*	exonic	A/C	FEV_1	23378	0.299%	-0.338	$7.20x10^{-5}$
rs35946782	21:40763754	*WRB*	exonic	A/G	FEV_1/FVC	23397	0.021%	1.458	$4.37x10^{-6}$
rs77543787	22:33264982	*SYN3*	exonic	T/C	FVC	23396	0.015%	1.547	$4.07x10^{-5}$

A first stage of replication using samples from UK BiLEVE was carried out for a subset of 38 SNPs (the remaining 12 SNPs were not available in the UK BiLEVE data). 7 SNPs still met the $P<10^{-4}$ significance level in the meta-analysis of the discovery stage result and the UK BiLEVE replication results and these 7 SNPs, along with 12 SNPs identified in the discovery analysis, but not present in the UK BiLEVE data, were taken forward to a second stage of replication. The second stage of replication combined results of a look-up within the CHARGE consortium, and analyses within UKHLS (Figure 4-21). Combining the results from the discovery and both stages of replication in a meta-analysis identified one intergenic SNP close to *LY86*, showing association with FEV_1/FVC (rs1294421, MAF=38.5%, $N_{\mathbf{meta}}$= 116,772, $P_{meta}=1.12x10^{-13}$, Table 4-9). The replication results for all SNPs identified in the discovery analyses are in appendix D (Table D-1).

Figure 4-21: Overview of discovery analysis and 2 stage replication, annotated with the number of SNPs analysed in each stage.

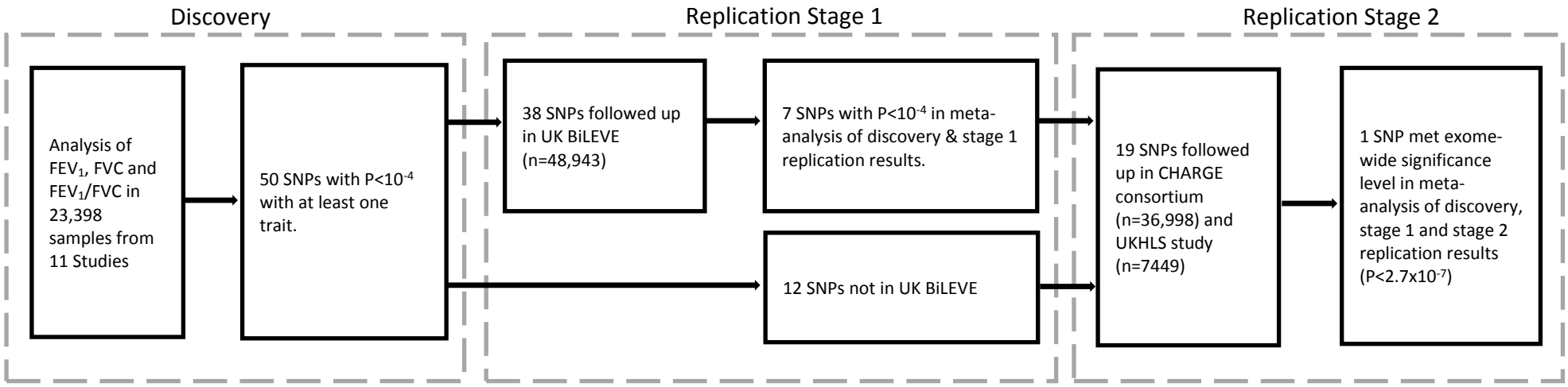


Table 4-9: Novel locus identified in meta-analysis of FEV_1/FVC.

Two-sided P-values are given for the discovery analysis, stage 1 and stage 2 replication analyses, alongside the combined P-value of all three analyses (discovery + replication stage 1 + replication stage 2). Beta values reflect effect-size estimates on an inverse-normal transformed scale after adjustments for age, age^2, sex, height and ancestry principal components, and stratified by ever smoking status. Beta values for the stage 2 replication are from UKHLS only, as the CHARGE consortium effect estimates are based on untransformed traits. Total Combined sample size N=116,772.

						Discovery Analysis			Stage 1 Replication			Stage 2 Replication			Combined
SNP	Chr:Pos	Genes	Trait	Effect / other allele	Effect allele frequency (Discovery)	N	β_{disc}	P_{disc}	N	β_{rep1}	P_{rep1}	N	β_{rep2}	P_{rep2}	P_{meta}
rs1294421	6:6743149	LY86(dist=87,933), RREB1(dist=364,681)	FEV_1/FVC	G/T	61.47%	23,395	0.0373	9.14×10^{-5}	48,943	0.0314	1.59×10^{-6}	44,434	0.0284	3.17×10^{-5}	1.12×10^{-13}

Meta-analyses of single variant associations were additionally carried out in ever smokers and never smokers separately. These analyses identified an additional 16 SNPs associated with at least one trait ($P<10^{-4}$) in never smokers and 37 SNPs in ever smokers (Table 4-10). These SNPs were taken forward to a two stage replication using data from UK BiLEVE (stage1) and UKHLS only (stage 2), as shown in Figure 4-22. These replication analyses provided no evidence to support the associations with the 16 SNPs identified in the never smokers only analyses. Of the 37 SNPs identified in the ever smokers discovery analyses, one intergenic SNP, close to *FGF10* attained exome-wide significance in the meta-analysis of the discovery and replication data (rs1448044, MAF=31.5%, N_{meta}=40,447, $P_{meta}=1.90x10^{-8}$, Table 4-11). The replication results for all SNPs identified in the smoking stratum specific discovery analyses are in appendix D (Table D-2).

Table 4-10: All SNPs showing association ($P<10^{-4}$) with FEV_1, FVC or FEV_1/FVC in the discovery stage meta-analysis, in ever smokers and never smokers separately.
Only variants in novel loci and that were not identified in the analyses of ever and never smokers combined are shown. Only the trait for which each SNP was most significantly associated is shown. Chromosome (Chr) and position (Pos) in build 37 are given for each SNP. Beta values reflect effect-size estimates on an inverse-normal transformed scale after adjustments for age, age^2, sex, height and ancestry principal components, and stratified by ever smoking status.

SNP	Chr:Pos	Gene(s)	Function	Effect / other allele	Trait	Smoking subset (Ever / never smokers)	N	Effect allele frequency (EAF)	Beta	P-value
rs201163722	1:28661302	*MED18*	nonsynonymous	T/C	FEV_1/FVC	ever	11558	0.0087%	-2.9956	$2.35x10^{-5}$
rs1414896	1:95692310	*TMEM56-RWDD3*	intronic	A/G	FVC	ever	11562	60.84%	-0.0561	$4.14x10^{-5}$
rs2764504	1:119234198	*SPAG17, TBX15*	intergenic	C/T	FVC	ever	11560	5.89%	0.1133	$5.78x10^{-5}$
rs199581193	1:201175658	*IGFN1*	nonsynonymous	G/A	FVC	never	11418	0.013%	-2.8726	$7.45x10^{-7}$
rs76611705	1:223177974	*DISP1*	nonsynonymous	A/G	FEV_1	never	11834	1.55%	-0.2065	$8.09x10^{-5}$
rs200091857	2:97008504	*NCAPH*	nonsynonymous	T/C	FEV_1/FVC	never	11832	0.072%	-1.0461	$1.91x10^{-5}$
rs202022630	2:234750666	*HJURP*	nonsynonymous	T/C	FEV_1/FVC	ever	11562	0.009%	2.7461	$9.20x10^{-5}$
rs147184138	3:25833094	*OXSM*	nonsynonymous	G/C	FVC	ever	11563	0.087%	0.9031	$7.47x10^{-5}$
rs61747991	3:56653424	*CCDC66*	nonsynonymous	T/A	FEV_1/FVC	ever	11557	4.90%	0.1249	$4.52x10^{-5}$
rs201934751	3:150404106	*FAM194A*	nonsynonymous	T/A	FEV_1/FVC	ever	11563	0.009%	2.7604	$6.75x10^{-5}$
rs7652177	3:171969077	*FNDC3B*	synonymous	G/C	FEV_1	ever	11558	46.38%	-0.0549	$6.90x10^{-5}$
rs7627615	3:183818416	*HTR3E*	synonymous	A/G	FEV_1	ever	11555	58.72%	-0.0561	$3.54x10^{-5}$
rs144473454	5:9136617	*SEMA5A*	nonsynonymous	A/G	FVC	ever	10440	0.273%	0.5816	$8.92x10^{-5}$
rs4634319	5:27418887	*CDH9, LINC01021*	intergenic	G/A	FVC	ever	11564	6.89%	0.1058	$5.43x10^{-5}$
rs1448044	5:44296986	*NNT, FGF10*	intergenic	A/G	FVC	ever	11550	31.53%	0.0563	$8.41x10^{-5}$
rs255888	5:111103258	*NREP*	intronic	T/C	FEV_1	ever	11562	55.70%	0.0550	$3.56x10^{-5}$
rs147752980	5:130791507	*RAPGEF6*	synonymous	C/T	FEV_1	never	11834	0.123%	0.8385	$6.44x10^{-6}$
rs148279287	5:140768844	*PCDHGB4*	nonsynonymous	T/C	FVC	ever	10165	0.551%	-0.3905	$4.16x10^{-5}$
rs6941356	6:87967636	*ZNF292*	nonsynonymous	G/A	FEV_1/FVC	never	11833	10.97%	0.0833	$7.48x10^{-5}$

SNP	Chr:Pos	Gene(s)	Function	Effect / other allele	Trait	Smoking subset (Ever / never smokers)	N	Effect allele frequency (EAF)	Beta	P-value
rs144830879	6:129649451	*LAMA2*	nonsynonymous	A/G	FEV_1/FVC	never	11832	0.017%	1.9884	$8.17x10^{-5}$
rs41298397	6:132891977	*TAAR6*	nonsynonymous	C/T	FEV_1	ever	11562	0.398%	-0.4721	$6.28x10^{-6}$
rs35839363	6:132909838	*TAAR5*	nonsynonymous	A/G	FVC	ever	11561	0.035%	1.4025	$7.42x10^{-5}$
rs13286541	9:113251951	*SVEP1*	nonsynonymous	C/T	FEV_1/FVC	ever	11559	9.87%	0.0870	$8.11x10^{-5}$
rs5030723	9:120476694	*TLR4*	nonsynonymous	A/G	FEV_1/FVC	ever	11553	0.307%	0.4750	$7.12x10^{-5}$
rs2296957	9:134401335	*UCK1*	synonymous	T/C	FEV_1/FVC	never	10069	95.42%	-0.1336	$8.65x10^{-5}$
rs7871194	9:139544437	*MIR4674, EGFL7*	intergenic	C/A	FEV_1/FVC	never	11830	57.04%	-0.0525	$6.44x10^{-5}$
rs141660796	10:72360577	*PRF1*	nonsynonymous	A/G	FEV_1/FVC	ever	11563	0.056%	1.1473	$3.65x10^{-5}$
rs821205	10:107727810	*SORCS3, SORCS1*	intergenic	C/A	FVC	ever	11564	52.01%	0.0522	$7.93x10^{-5}$
rs5006889	11:5373104	*OR51B6*	nonsynonymous	G/A	FEV_1	never	11827	26.30%	-0.0696	$3.51x10^{-6}$
rs142159415	11:5776626	*OR52N4*	nonsynonymous	T/C	FVC	ever	11564	0.951%	0.2677	$7.98x10^{-5}$
rs199618034	11:114182882	*NNMT*	nonsynonymous	C/G	FEV_1	ever	11559	0.022%	1.8491	$3.59x10^{-5}$
rs1982528	12:132237848	*SFSWAP*	synonymous	C/T	FVC	ever	10147	98.35%	-0.2244	$5.25x10^{-5}$
rs140930007	13:51854595	*FAM124A*	nonsynonymous	A/G	FVC	never	11419	0.031%	-1.6837	$8.63x10^{-6}$
rs144854034	13:98829388	*RNF113B*	nonsynonymous	G/T	FEV_1/FVC	ever	11560	0.108%	-0.7922	$8.43x10^{-5}$
rs140501662	14:20711665	*OR11H4*	nonsynonymous	T/C	FEV_1/FVC	never	11831	0.148%	0.6755	$7.66x10^{-5}$
rs200081065	14:91755506	*CCDC88C*	nonsynonymous	T/C	FEV_1	never	11834	0.080%	0.9679	$2.47x10^{-5}$
rs200614333	15:42143077	*SPTBN5*	stopgain	T/C	FVC	never	11832	0.021%	1.7770	$6.41x10^{-5}$
rs138439412	15:52017135	*LYSMD2*	nonsynonymous	T/C	FVC	ever	11561	0.917%	-0.2737	$7.75x10^{-5}$
rs79030022	15:75941897	*SNX33*	nonsynonymous	T/C	FEV_1/FVC	ever	11553	0.052%	1.1613	$5.71x10^{-5}$
rs144617499	16:21073933	*DNAH3*	nonsynonymous	A/G	FEV_1	ever	11561	2.15%	-0.1813	$8.23x10^{-5}$
rs77439178	16:31091757	*ZNF646*	nonsynonymous	A/G	FEV_1/FVC	ever	11479	0.009%	2.9418	$3.61x10^{-5}$
rs141225776	16:66547713	*TK2*	nonsynonymous	A/T	FVC	never	11833	0.021%	1.7422	$9.80x10^{-5}$
rs146239773	17:1387496	*MYO1C*	synonymous	A/G	FEV_1	ever	11563	0.022%	2.0179	$1.11x10^{-5}$

SNP	Chr:Pos	Gene(s)	Function	Effect / other allele	Trait	Smoking subset (Ever / never smokers)	N	Effect allele frequency (EAF)	Beta	P-value
rs7207403	17:47210506	*B4GALNT2*	synonymous	A/C	FEV_1	ever	11562	56.42%	0.0591	$6.02x10^{-5}$
rs143270448	17:74274071	*QRICH2*	nonsynonymous	A/G	FEV_1	never	11834	0.152%	0.7305	$1.20x10^{-5}$
rs201979657	18:77171089	*NFATC1*	nonsynonymous	A/G	FEV_1/FVC	ever	11557	0.017%	-1.9564	$9.00x10^{-5}$
rs200123506	19:38375738	*WDR87*	nonsynonymous	C/T	FEV_1	ever	11563	0.018%	0.6251	$7.17x10^{-5}$
rs61737337	19:40197267	*LGALS14*	nonsynonymous	A/G	FEV_1/FVC	ever	11137	0.009%	2.9111	$4.18x10^{-5}$
rs201361713	19:48737800	*CARD8*	synonymous	T/C	FVC	ever	11564	0.087%	0.8992	$6.58x10^{-5}$
rs143501994	19:51870712	*CLDND2*	synonymous	T/C	FEV_1/FVC	ever	11563	0.372%	0.4191	$9.25x10^{-5}$
rs200402559	21:44838346	*SIK1*	nonsynonymous	A/G	FVC	never	11833	0.008%	2.9643	$2.54x10^{-5}$
rs201423754	22:37893171	*CARD10*	nonsynonymous	T/C	FEV_1/FVC	ever	11563	0.013%	-2.5360	$1.26x10^{-5}$
rs12841259	X:118893390	*SOWAHD*	nonsynonymous	G/A	FEV_1/FVC	ever	8744	0.480%	0.3479	$4.25x10^{-5}$

Figure 4-22: Overview of discovery analysis and 2 stage replication in ever smokers and never smokers separately, annotated with the number of SNPs analysed in each stage.

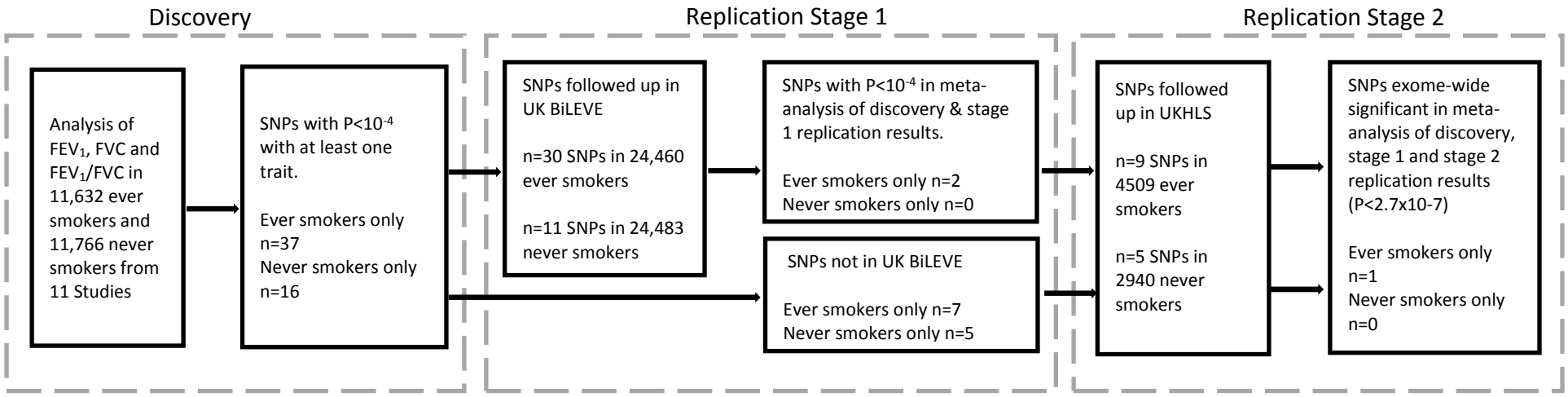


Table 4-11: Novel locus identified in meta-analysis of FVC in ever smokers.

Two-sided P-values are given for the discovery analysis, stage 1 and stage 2 replication analyses, alongside the combined P-value of all three analyses (discovery + replication stage 1 + replication stage 2). Beta values reflect effect-size estimates on an inverse-normal transformed scale after adjustments for age, age^2, sex, height and ancestry principal components, and stratified by ever smoking status. Total Combined sample size N=40,447.

							Discovery Analysis			Stage 1 Replication			Stage 2 Replication			Combined
SNP	Chr:Pos	Genes	Trait	Analysis	Effect / other allele	Effect allele frequency (Discovery)	N	β_{disc}	P_{disc}	N	β_{rep1}	P_{rep1}	N	β_{rep2}	P_{rep2}	P_{meta}
rs1448044	5:44296986	*FGF10*(dist=8111), *NNT*(dist=591,318)	FVC	Ever Smokers	A/G	31.53%	11,550	0.0563	8.41×10^{-5}	24,460	0.0402	3.31×10^{-5}	4437	0.0199	0.3796	1.90×10^{-8}

4.4.2 Associations in known lung function regions

In addition to the two novel loci, associations ($P<10^{-4}$) were identified in a number of regions previously associated with one or more of FEV_1, FVC and FEV_1/FVC: *MCL1-ENSA; MECOM; FAM13A; GSTCD; HTR4; PTCH1; PTHLH-CCDC91; LRP1; THSD4; LTBP4* and 2 signals in the *GPR126* region (Table 4-12). In the analysis of ever smokers only, there was an association with rs8034191 (*HYK- AGPHD1*) and FEV_1/FVC; this SNP is in strong LD (r^2=0.93) with a smoking behaviour associated SNP in *CHRNA5* (rs1051730) (132) and has previously been implicated in COPD (129), as well as lung cancer (183).

Table 4-12: Associations identified in discovery analysis in known lung function (and related traits) regions.

							Discovery Analysis			
SNP	Chr:Pos	Gene(s)	Function	Effect / other allele	Trait	Smoking subset	Effect Allele Frequency	Beta	P-value	Details
rs11204697	1:150658971	*GOLPH3L*	intronic	T/C	FEV_1	all	41.10%	0.0377	$7.39x10^{-5}$	r^2=0.633 with rs6681426 (*MCL1-ENSA*, previously associated with FEV_1/FVC) (43)
rs1344555	3:169300219	*MECOM*	intronic	T/C	FEV_1	all	19.78%	-0.0535	$4.52x10^{-6}$	SNP previously associated with FEV_1 (119)
rs7671167	4:89883979	*FAM13A*	intronic	T/C	FEV_1/FVC	all	50.27%	-0.0428	$3.94x10^{-6}$	r^2=0.664 with rs2045517 (previously associated with FEV_1/FVC) (119)
rs10516526	4:106688904	*GSTCD*	intronic	G/A	FEV_1	all	6.49%	0.0756	$5.79x10^{-5}$	SNP previously associated with FEV_1 (119)
rs11168048	5:147842353	*HTR4*	intronic	C/T	FEV_1/FVC	all	41.58%	0.0445	$2.59x10^{-6}$	r^2=1 with rs1985524 (previously associated with FEV_1 & FEV_1/FVC) (119)
rs17280293	6:142688969	*GPR126*	exonic	G/A	FEV_1/FVC	all	2.53%	0.1253	$2.35x10^{-5}$	r^2=0.85 with rs148274477 (previously associated with FEV_1/FVC) (43)
rs7763064	6:142797289	*GPR126, LOC153910*	intergenic	A/G	FEV_1/FVC	all	29.91%	0.0437	$2.04x10^{-5}$	r^2=0.961 with rs262129 (previously associated with FEV_1/FVC) (119)
rs16909898	9:98231008	*PTCH1*	ncRNA exonic	G/A	FVC	all	9.74%	0.0729	$3.23x10^{-6}$	SNP previously associated with FEV_1/FVC (119)
rs10843206	12:28722756	*CCDC91, FAR2*	intergenic	T/C	FVC	all	48.55%	-0.0412	$1.05x10^{-5}$	Not independent to indel rs11383346 (*PTHLH-CCDC91*, previously associated with FVC) (43)
rs1564374	12:58010163	*ARHGEF25*	exonic	G/A	FEV_1/FVC	all	59.11%	0.0393	$3.52x10^{-5}$	r^2=0.022 with rs1172113 in (*LRP1*, previously associated with FEV_1/FVC) (119)
rs12899618	15:71645120	*THSD4*	intronic	A/G	FEV_1/FVC	all	15.37%	-0.0628	$1.19x10^{-6}$	r^2=0.688 with rs8033889 (previously associated with FEV_1/FVC) (119)
rs8034191	15:78806023	*HYKK, AGPHD1*	intronic	C/T	FEV_1	ever	34.38%	-0.0645	$5.23x10^{-6}$	SNP previously associated with smoking and COPD (129, 133)
rs34093919	19:41117300	*LTBP4*	exonic	A/G	FEV_1/FVC	all	1.14%	0.1910	$1.58x10^{-5}$	r^2=0.99 with rs113473882(previously associated with FEV_1/FVC) (43)

4.4.3 Meta-Analyses of gene-based associations

WST and SKAT tests were undertaken to assess the joint effects of multiple low frequency variants within a gene on lung function traits. In the discovery analyses of all 23,398 samples, 14,865 genes were tested, all of which had at least two exonic variants with MAF<5%.

The SKAT analyses identified 8 genes associated ($P<10^{-4}$) with FEV_1, FVC or FEV_1/FVC. Single SNPs within seven of these eight genes were identified in the meta-analyses of single variant associations (Table 4-8). The only exception was *PTPRA;* within this gene there was a SNP whose association with FEV_1 was just below the $P<10^{-4}$ significance level (rs148576102, chr20:3016535, $P_{disc}=1.25x10^{-4}$). Additional analyses were carried out, conditioning for the most significantly associated single SNP within each gene, and these analyses showed that all gene associations were driven by these single SNPs (Table 4-13). Consequently, none of these gene-based associations were followed up further.

Table 4-13: SKAT test association results for all genes identified in discovery SKAT test analyses ($P<10^{-4}$).
For each gene, a conditional analysis was carried out, conditioning on the most significantly associated individual SNP within that gene. Chromosome: Position of individual SNP shown, along with conditional SKAT P-value when that SNP has been conditioned on. Only the trait for which each gene was most significantly associated is shown.

		Discovery Analysis - SpiroMeta Consortium		
Gene Name	Trait	No. variants included in test (n_{snp})	P-value	Conditional Analyses
NUBPL	FEV_1	8	$5.61x10^{-5}$	Signal driven by chr14:32257065. Conditional P=0.4520
PTPRA	FEV_1	2	$9.27x10^{-5}$	Signal driven by chr20:3016535. Conditional P=0.1633
TONSL	FEV_1	27	$1.89x10^{-6}$	Signal driven by chr8:145667730. Conditional P=1.0
POU4F2	FVC	5	$8.15x10^{-6}$	Signal driven by chr4:147561147. Conditional P =0.5076
PATE2	FEV_1/FVC	2	$9.51x10^{-6}$	Signal driven by chr11:125647897. Conditional P =0.1732
ANKRD34C	FEV_1/FVC	3	$5.88x10^{-5}$	Signal driven by chr15:79586782. Conditional P =0.3165
QRFPR	FEV_1/FVC	12	$7.71x10^{-5}$	Signal driven by chr4:122250654. Conditional P =0.6188

The WST analyses identified one significant association between *SEMA7A* and FVC, and this did not appear to be driven by a single SNP (n_{SNPs}=12, β_{disc}=0.0062, P_{disc}=2.55x10^{-5}). Initially, this gene-based signal was followed up using samples from UK BiLEVE and UKHLS; however the signal did not replicate (n_{SNPs}=7, β_{rep}=0.0014, P_{rep}=0.156, Table 4-14). Further evidence for replication was sought from the CHARGE consortium; the WST result for this gene showed inconsistent direction of effect on FVC to the discovery analysis (n_{SNPs}=14, β_{rep}= -3.174 [effect estimate on untransformed scale], P_{rep}=0.047, Table 4-14). A common intronic SNP within *SEMA7A* (rs8036030) has previously been associated with airflow obstruction, however this association did not meet genome-wide significance (130).

Table 4-14: WST association results for all genes identified in discovery WST analyses ($P<10^{-4}$).

Discovery + Stage 1 replication meta result calculated using RAREMETAL. Discovery + Stage 1 + Stage 2 replication meta result calculated by combining Discovery + Stage 1 replication result with Stage 2 (CHARGE Consortium) result, by sample size weighted z-score meta-analysis. Betas values for Discovery and Stage 1 replication results reflect effect-size estimates on an inverse-normal transformed scale. For the CHARGE Consortium (replication Stage 2), beta values represent untransformed trait effect estimates.

		Discovery Analysis SpiroMeta Consortium			Stage 1 Replication UK BiLEVE & UKHLS			Discovery + Stage 1 replication meta		Stage 2 Replication CHARGE			Discovery + Stage 1 + Stage 2 replication meta
Gene Name	Trait	No. variants included in test (n_{snp})	Effect size(β_{disc})	P-value (P_{disc})	No. variants included in test (n_{snp})	Effect size(β_{rep})	P-value (P_{rep})	Effect size(β_{meta})	P-value (P_{meta})	No. variants included in test (n_{snp})	Effect size(β_{rep})	P-value (P_{rep})	P-value (P_{meta})
SEMA7A	FVC	12	0.0062	2.55×10^{-5}	7	0.0014	0.1564	0.0020	8.44×10^{-3}	14	-3.1740	0.0471	0.2878

Gene-based tests were also carried out in ever smokers and never smokers separately; these analyses identified a further eight genes associated ($P<10^{4}$) with at least one trait, that were not driven by single SNP associations (Table 4-15 and Table 4-16). There was some independent evidence of replication of one of the WST associations in the replication samples: *LRPPRC* showed suggestive evidence of association with FEV_1/FVC in ever smokers (replication analysis: $n_{SNPs}=7$, $\beta_{rep}=-0.0009$, $P_{rep}=3.49x10^{-3}$, Table 4-17). A combined meta-analysis of the discovery and replication samples was then undertaken using RAREMETAL; the resulting WST result for LRPPRC and FEV1/FVC in ever smokers fell short of the predefined $P<2.4x10^{-6}$ significance threshold ($\beta_{meta}=-0.0010$, $P_{meta}=3.65x10^{-3}$, Table 4-17). The replication results for all gene-based associations listed in Table 4-15 and Table 4-16 are given in Appendix D (SKAT: Table D-3; WST: Table D-4).

Table 4-15: SKAT association results for all genes identified in discovery SKAT analyses in ever smokers and never smokers separately ($P<10^{-4}$).

			Discovery Analysis - SpiroMeta Consortium	
Gene Name	Trait	Smoking subset (Ever / never smokers)	No. variants included in test (n_{snp})	P-value (P_{disc})
C12orf77	FEV_1/FVC	Ever	4	$5.33x10^{-5}$
NFATC1	FEV_1/FVC	Never	10	$8.41x10^{-5}$

Table 4-16: WST association results for all genes identified in discovery WST analyses in ever smokers and never smokers separately ($P<10^{-4}$).

Betas values reflect effect-size estimates on an inverse-normal transformed scale.

			Discovery Analysis - SpiroMeta Consortium		
Gene Name	Trait	Smoking subset (Ever / never smokers)	No. variants included in test (n_{snp})	Effect size(β_{disc})	P-value (P_{disc})
NPEPL1	FEV_1	Never	2	0.0180	$8.46x10^{-5}$
PGBD1	FEV_1	Never	13	0.0067	$2.63x10^{-5}$
FAM45A	FVC	Ever	6	-0.0134	$3.78x10^{-5}$
GPR123	FVC	Ever	4	0.0170	$8.40x10^{-5}$
WRB	FVC	Never	2	-0.0177	$9.73x10^{-5}$
C12orf77	FEV_1/FVC	Ever	4	0.0189	$3.29x10^{-6}$
LRPPRC	FEV_1/FVC	Ever	16	-0.0102	$8.03x10^{-5}$

Table 4-17: Replication of WST association results for *LRPPRC* and FEV_1/FVC in ever smokers.

Discovery, replication and combined meta results calculated using RAREMETAL. Beta values for Discovery and Stage 1 replication results reflect effect-size estimates on an inverse-normal transformed scale.

			Discovery Analysis - SpiroMeta Consortium			Replication Analysis - UK BiLEVE & UKHLS			Combined Meta-analysis	
Gene Name	Trait	Smoking subset (Ever / never smokers)	No. variants included in test (n_{snp})	Effect size(β_{disc})	P-value (P_{disc})	No. variants included in test (n_{snp})	Effect size(β_{rep})	P-value (P_{rep})	Effect size(β_{meta})	P-value (P_{meta})
LRPPRC	FEV_1/FVC	Ever	16	-0.0102	8.03×10^{-5}	7	-0.0009	3.49×10^{-3}	-0.0010	3.65×10^{-3}

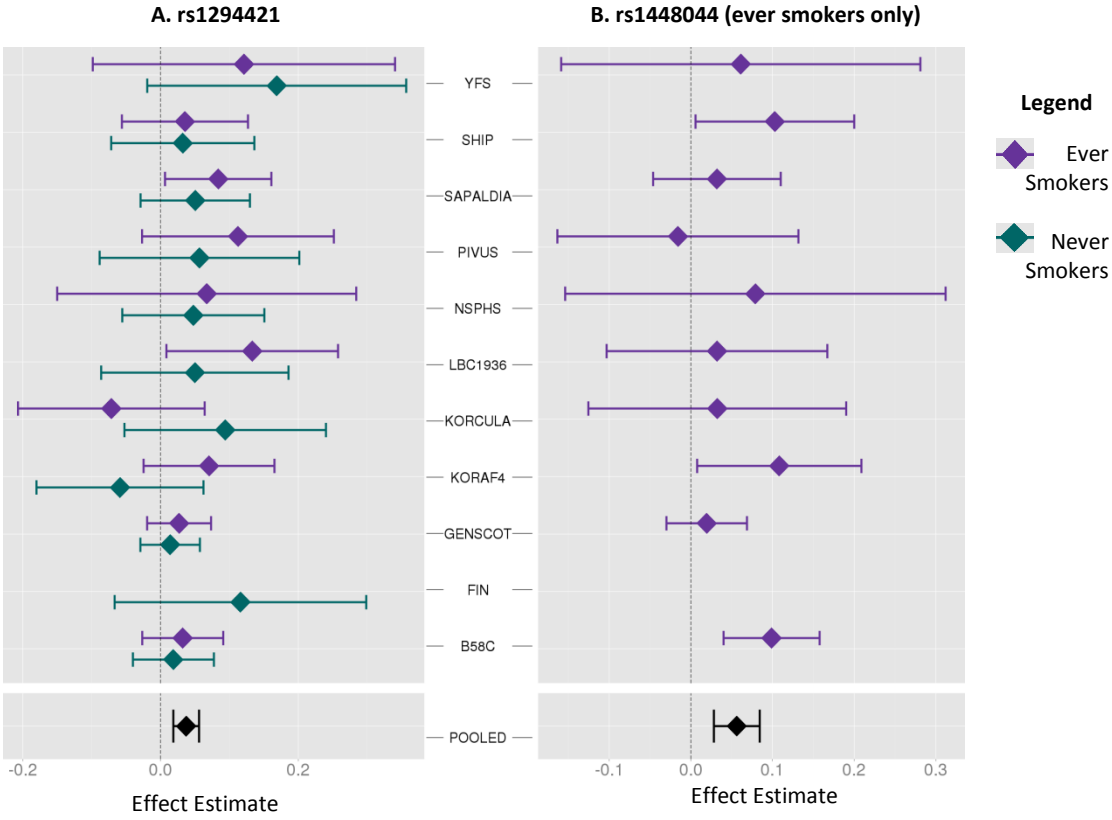
4.4.4 Heterogeneity of signals identified in single variant association analyses

The heterogeneity of the results across studies from the two novel loci identified in the single variant association analyses was tested using Cochran's Q Statistic. No significant heterogeneity was identified for either signal (Table 4-18). Forest plots for the two SNPs (Figure 4-23) also show that the results from each study were broadly consistent, and neither was driven by extreme results from one or two studies.

Table 4-18: Test of Heterogeneity, using Cochran's Q statistic for the two SNPs in novel regions, identified through the single variant association analyses.

SNP (Gene(s))	Analysis	Q Statistic	N Studies	P_{het}
rs1294421 (*LY86, RREB1*)	FEV_1/FVC, all samples	15.913	21	0.722
rs1448044 (*FGF10, NNT*)	FVC, Ever smokers	7.618	10	0.573

Figure 4-23: Forest plots for two SNPs in novel regions, identified through the single variant association analyses. A. rs1294421 associated with FEV_1/FVC. B. rs1448044 associated with FVC in ever smokers only.



4.4.5 Functional characterization of novel loci

In order to gain further insight into the two loci identified in the analyses of single variant associations, it was assessed whether these regions were associated with gene expression levels. This was done by carrying out a look-up of the two sentinel SNPs, and all proxies (r^2>0.3) within the newly identified regions in a number of publically available eQTL data sets, as described in Section 4.3.7.

Firstly, the Blood eQTL database, a dataset with results from the analysis of 5,311 individuals, was searched for cis (+/- 250Kb distance between the SNP and the probe midpoint) and trans (>5Mb distance between the SNP and the probe midpoint) eQTLs (178). Secondly, lung tissue expression data from 124 samples from the GTEx project was searched for cis-eQTLs (+/- 1Mb distance between the SNP and transcription start site) (179). In these two datasets, no significant eQTLs were found for the sentinel SNPs, or any proxy.

SNPs were further assessed in a lung eQTL resource based on lung tissues of 1,111 individuals from three sites: Laval, Groningen and UBC (180). Lung eQTLs were identified as associated with mRNA expression in either cis (+/- 1Mb distance between the SNP and transcription start site) or in trans (all other eQTLs). A proxy of rs1448044 in the *FGF10* region (rs6892212, r^2=0.464 with rs1448044) was identified as a cis-eQTL for *MRPS30* (P=1.71x10^{-5}), and proxies of rs1294421 in the *LY86* region (strongest associated SNP rs1294416, r^2=0.69 with rs1294421) were associated with the expression of a cDNA clone (BC039678, P=1.85x10^{-8}), which does not map to a gene (Table 4-19).

Table 4-19: Evidence for the role of novel variants identified in single variant association analyses as eQTLs in lung.

Z-score Laval, Z-score Groningen and Z-score UBC are the per-study estimates which were then meta-analysed to give the overall P-value. For each probeset variants were ranked firstly according to their correlation with the sentinel SNP (r2) and secondly by eQTL P-value, and only the top ranked SNP for each probeset is presented (rs id, chromosome and position is shown). The total number of significant SNP-probeset associations for each gene is also given.

Sentinel SNP	eQTL SNP (chr:pos)	LD (r2) with sentinel	Effect / other allele	Z-score Laval	Z-score Groningen	Z-score UBC	P-value	Sequence Source id	Gene Symbol	Gene Name	Total no. Proxy SNP-probeset associations
rs1448044	rs6892212 (5:44382842)	0.464	A/C	2.4640	2.6080	2.3746	$1.71x10^{-5}$	DA746332	*MRPS30*	mitochondrial ribosomal protein S30	1
rs1294421	rs1294416 (6:6741327)	0.69	A/G	-4.4911	-1.3252	-3.6964	$1.85x10^{-8}$	BC039678	-	-	9

Finally evidence of protein expression for *LY86* and *FGF10* in the respiratory system was searched for by querying the Human Protein Atlas (184). The protein products of both these genes were expressed in bronchial epithelial cells, pneumocytes and lung macrophages (Table 4-20).

Table 4-20: Protein expression results fornovel regions identified in single variant association analyses.
Expression levels in the respiratory system for epithelial cells, pneumocytes and macrophages were assessed for identified genes, in the Human Protein Atlas. Two results are for 1.HPA044895 staining; 2.CAB025000 staining. Expression level abbreviations: ND=Not Detected; NA=Not available.

SNP	Gene	Trait (analysis)	P-value	Respiratory epithelial cells, Nasopharynx	Respiratory epithelial cells, Bronchus	Pneumocytes, lung	Macrophages, lung
rs1448044	*FGF10* (dist=8111)	FVC (ever smokers)	1.90×10^{-8}	Medium ND	Low ND	Low ND	Low High
rs1294421	*LY86* (dist=87933)	FEV_1/FVC (all samples)	1.12×10^{-13}	NA Medium	NA Medium	NA Low	NA Low

4.4.6 Association of novel loci with smoking behaviour

It was further examined whether the signals at the two novel loci might be driven by smoking behaviour, or a result of a gene-smoking interaction. Firstly, a Z-test was carried out to compare the effect estimates for never smokers and ever smokers in the discovery samples, as a simple test for gene-smoking interaction (Table 4-21). For rs1294421 (*LY86*), there was no evidence for a different effect in never smokers compared to ever smokers (P=0.550); however, for rs1448044 (*FGF10)*, there was a significant interaction ($P=1.44 \times 10^{-3}$). It is not possible to definitively conclude that this locus represents a gene-smoking interaction however. In the stage 1 replication (UK BiLEVE samples), a similar effect was seen for this SNP in both ever smokers ($\beta_{rep1}=0.040$, $P_{rep1}=3.31 \times 10^{-5}$) and in never smokers ($\beta_{rep1}=0.038$, $P_{rep1}=7.23 \times 10^{-5}$, Table 4-22). Furthermore, the combined analysis of discovery and replication samples for this SNP, including both ever and never smokers, met the exome-wide significance level ($P=2.35 \times 10^{-9}$) and showed no statistically significant evidence of interaction overall (P=0.06).

Table 4-21: Discovery analysis results for novel loci in ever smokers and never smokers separately, and in ever smokers with additional adjustment for pack-years.

Betas and P-values from discovery meta-analysis only. Interaction P-value is for test for difference in effect sizes for ever smokers vs never smokers.

SNP	Gene	Trait	Analysis	N	Beta	P-value	Interaction P-value
rs1448044	*FGF10*	FVC	never	11809	-0.0077	0.5851	1.44×10^{-3}
			ever	11550	0.0563	8.41×10^{-5}	
			ever (pack-years)	10368	0.0553	2.71×10^{-4}	
rs1294421	*LY86*	FEV_1/FVC	never	11827	0.0294	0.0279	0.550
			ever	11561	0.0454	8.26×10^{-4}	
			ever (pack-years)	10378	0.0408	4.57×10^{-3}	

Table 4-22: Association of rs1448044 and FVC in ever smokers and never smokers separately, and in all individuals combined.

		Discovery Analysis		Stage 1 Replication		Stage 2 Replication		Discovery + Stage 1 + 2 Replication meta
		SpiroMeta Consortium		UK BiLEVE		UKHLS		
Name	Smoking subset	Beta	P-value	Beta	P-value	Beta	P-value	P-value
rs1448044	never	-0.0077	0.5851	0.0384	7.23×10^{-5}	0.0047	0.8640	3.93×10^{-3}
	ever	0.0563	8.41×10^{-5}	0.0402	3.31×10^{-5}	0.0199	0.3796	1.90×10^{-8}
	all	0.0238	0.0229	0.0393	9.39×10^{-9}	0.0138	0.4312	2.35×10^{-9}

The main meta-analyses did not adjust for smoking amount in the ever smokers; sensitivity analyses were therefore also carried out in ever smokers with additional adjustment for pack-years where available. Effect estimates for rs1448044 (*FGF10)* were unchanged when pack-years were adjusted for, and only slightly attenuated for rs1294421 (*LY86*), suggesting that these associations are not driven by smoking behaviour (Table 4-21).

Finally, a look-up of this SNP in the publicly available results of a GWAS of smoking behaviour by the Tobacco and Genetics Consortium (132) was undertaken (Table 4-23). There was some evidence that rs1448044 (*FGF10*) was associated with ever vs never smoking (P=0.04), and rs1294421 (*LY86*) was associated with pack-years (P=0.01).

Table 4-23: Look-up of effect of novel variants on smoking phenotypes in the Tobacco and Genetics (TAG) consortium.

		Cigarettes per day		Ever vs Never smoker		Current vs former smoker		Log age of starting smoking	
SNP	Gene	Beta	P	OR	P	OR	P	Beta	P
rs1448044	*FGF10*	0.1511	0.097	0.9734	0.0388	0.9885	0.511	-0.0026	0.3029
rs1294421	*LY86*	0.2187	0.0101	1.0055	0.6518	1.0079	0.6355	-0.0001	0.9777

4.5 Discussion

This chapter describes the analysis of 23,398 samples from 11 studies with exome array data and three lung function traits, with follow up of the most significant single SNP and gene-based associations in up to 93,390 independent samples. The combined analyses of the discovery and replication single variant associations identified two SNPs meeting the pre-defined exome-wide significance level ($P<2.7x10^{-7}$), in regions not previously implicated in lung function.

The first of these was a common (MAF=38.5%), intergenic SNP close to *LY86* (lymphocyte antigen 86*)*, associated with FEV_1/FVC ($P_{meta}=1.12x10^{-13}$). *LY86* interacts with the toll-like receptor signalling pathway, when bound with RP105 to form a heterodimer (185). The sentinel SNP rs1294421 has previously been associated with waist-hip ratio (186), and an intronic SNP within *LY86* (rs7440529, LD with rs1294421: $r^2=0.005$) has previously been implicated in asthma in two studies of individuals of Han Chinese ancestry (187, 188).

The second identified association was another common (MAF=31.5%) intergenic SNP, close to *FGF10* ($P_{meta}=1.90x10^{-8}$) associated with FVC. *FGF10* is a member of the fibroblast growth factor family of proteins, involved in a number of biological processes, including embryonic development, cell growth, morphogenesis, tissue repair, tumor growth and invasion. Specifically, the *FGF10* signalling pathway plays an essential role in lung development and lung epithelial renewal (189). A study in mice found a deficiency in *FGF10* resulted in a fatal disruption of branching morphogenesis during lung development (190). A further study in mice with bleomycin-induced lung fibrosis found overexpression of *FGF10* to result in attenuated fibrosis and increased survival (191). A proxy of the sentinel SNP in the *FGF10* region was identified as a cis-

eQTL for *MRPS3,* a constituent of the mitochondrial ribosome which plays a role in oxidative phosphorylation (192).

The relationship with the two novel loci and smoking was also examined. rs1924421 in *LY86* showed suggestive association with amount smoked in the publically available TAG meta-analysis. Whilst the primary analyses did not adjust for smoking quantity, a secondary analyses of ever smokers with adjustment for pack-years was performed. In this adjusted analysis, the estimated effect of rs1924421 on FEV_1/FVC was not substantially changed. rs1448044 in *FGF10* also showed weak association with ever vs never smoking, and there was some evidence for a gene-smoking interaction in the discovery analysis of FVC. In the stage 1 replication samples, similar effects were seen for this SNP in both ever and never smokers however, so further evidence would be required to conclude that this signal is driven by smoking behaviour.

Through the use of the exome array, it was hoped that associations with low frequency and rare functional variants would be identified. Whilst the discovery analyses identified single SNP associations with a number of low frequency variants, these findings were not replicated in the replication samples. There are a number of possible explanations for the lack of replication of these findings. Firstly, many of the rarest of these SNPs were not genotyped, or were monomorphic in our follow-up samples, so replication was not possible. In some instances, a failure to replicate might be due to the phenomenon of "winner's curse", whereby the effects identified in the discovery samples may have been overinflated.

Overall, the lack of convincing associations with rare variants is likely due to a limited statistical power for identifying single variant associations, particularly if those variants exhibit only modest effects. So far, low frequency variants in only 2 regions have been associated with quantitative lung function. In the present analysis, the most strongly associated SNPs in these regions were rs17280293 in *GPR126* (MAF=2.5%, β=0.125) and rs34093919 in *LTBP4* (MAF=1.1%, β=0.191). For a SNP with MAF=1% and an effect size of β=0.2, there is only 22% power to detect an association in the present discovery analysis (n=23,398) at exome-wide significance ($P<2.7x10^{-7}$). Figure 4-24 shows that for low frequency SNPs (MAF<5%), the present study generally has limited power to

detect associations at this significance level, unless effect sizes are much larger than observed SNP effects identified to date. Figure 4-25 further shows that for low frequency SNPs with modest effects (β=0.1), even analyses undertaken with a sample size of 100,000 would have very low power to detect significant associations (power of 24%% and 2% for SNPs with MAF=1% and 0.5%, respectively).

Figure 4-24: Power to detect exome-wide significant associations ($P<2.7x10^{-7}$) with low frequency variants in the SpiroMeta discovery data (n=23,398).
Beta estimates are on the inverse-normal transformed scale.

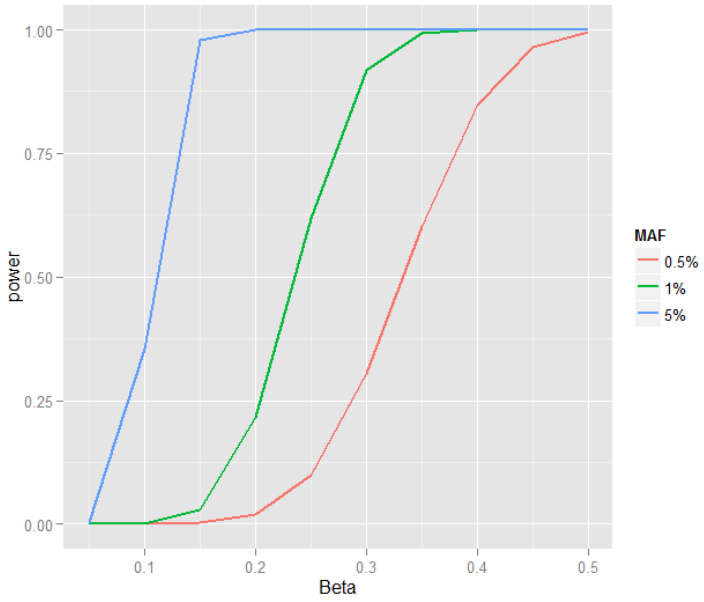
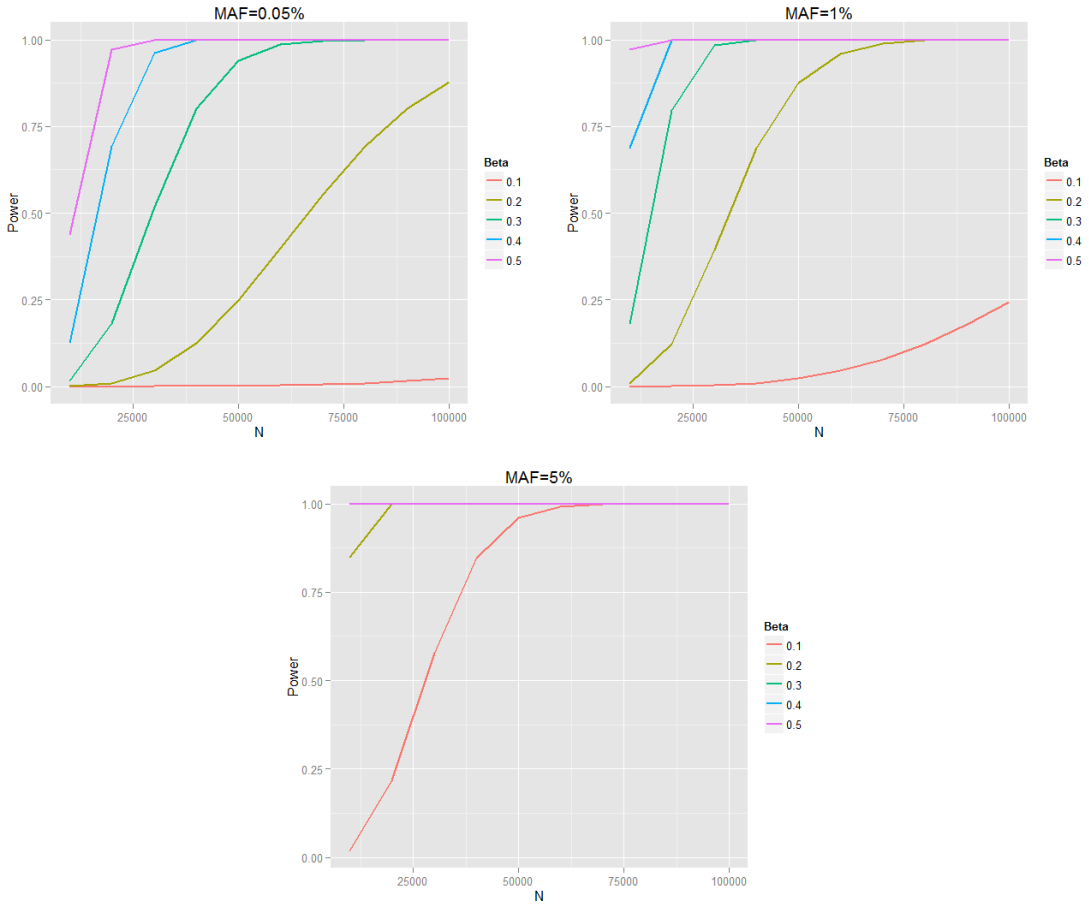


Figure 4-25: Power to detect exome-wide significant associations ($P<2.7x10^{-7}$) with variants of varying effect sizes (Beta) and sample sizes (N).
Power shown for SNPs with MAFs of A. 0.05%, B. 1%, C. 5%. Beta estimates are on the inverse-normal transformed scale.



SKAT and burden gene-based tests were additionally employed to investigate the joint effects of low frequency and rare variants within a gene on lung function traits. Many of the genes identified through these analyses were in fact a result of single SNPs, which were themselves identified through the meta-analyses of single variant associations. For those genes that did not appear to be driven by single SNPs, replication was sought. This proved challenging however, as again many SNPs included within these analyses in the discovery samples were not genotyped, or were monomorphic in the replication samples. This often meant a disparity in the gene unit being tested in the discovery and replication samples; hence the interpretation of these results was not straightforward.

These analyses also identified associations in several regions that have previously been implicated in lung function and related traits. These included associations the two

previously mentioned two low frequency SNPs (*GPR126* and *LTBP4*) associated with FEV_1/FVC in these analyses. SNPs in strong LD (r^2>0.85) with these two SNPs were first identified in separate meta-analyses of lung function within the SpiroMeta Consortium, using data imputed to the 1000 Genomes reference panel (43). The two SNPs identified in these analyses are both nonsynonymous, and could potentially represent the causal SNP in each region.

Through the analyses described in the chapter, I have identified two common, intergenic SNPs in regions not previously implicated in lung function. These associations were both replicated in the follow-up analyses and met the pre-specified exome-wide significance level overall, as well as the well the more stringent genome-wide significant level ($P<5x10^{-8}$). Further interrogation of these loci could lead to greater understanding of lung function and lung disease, and could provide novel targets for therapeutic interventions. Whilst these analyses had limited success in identifying novel low frequency and rare variants associated with lung function, we cannot rule out the possibility that more of these variants do make a contribution to the underlying genetic basis of lung function. Future studies, of increasingly larger sample sizes, will be required in order to fully evaluate the effect of variants across the allele frequency spectrum.

Chapter 5 Analysis of flow lung function measures PEF and FEF_{25-75} in UK Biobank

5.1 Introduction

Genome-wide association studies (GWAS) of lung function to date have generally focussed on three volumetric lung function measures: FEV_1, FVC and FEV_1/FVC. Other measures that may be derived from spirometry include measures of flow, such as the peak expiratory flow (PEF) and the forced expiratory flow between 25% and 75% of vital capacity (FEF_{25-75}). PEF and FEF_{25-75} are correlated with each other and with FEV_1, FVC and FEV_1/FVC, however each measure varies in terms of clinical significance.

PEF is the maximum measure of expiratory flow measured during a forced expiratory manoeuvre. PEF is determined by a number of physiological factors including lung volume and dimensions of the large airways, the elastic properties of the lung and the power of the expiratory muscles (193). PEF is a valid indicator of lung function impairment and can be helpful in clinical practice for monitoring airflow limitation, particularly in individuals with asthma (193, 194). It is less useful however for distinguishing between obstructive and restrictive impairment and has a low sensitivity to measuring obstruction in the small airways (194). FEF_{25-75} is the average forced expiratory flow rate over the middle 50 percent of the FVC. FEF_{25-75} is considered a measure of the calibre of the small airways, with a reduction in FEF_{25-75} thought to be an early indicator of airflow limitation, even where an individual has normal FEV_1 (195, 196). However, a reduction in FEF_{25-75} is not a specific measure of small airway disease (197) and is highly effort dependent (194).

Similarly to the volumetric lung function measures, PEF and FEF_{25-75} vary with height, age, sex and ethnicity, and various reference equations for these measures exist (85, 87). PEF is also subject to diurnal variation, with maximal PEF usually occurring during the afternoon to early evening, with lowest values occurring during the night and early morning (193).

As discussed in Section 1.3.4, there have been a limited number of GWAS studies undertaken for PEF and FEF_{25-75} (121-123), with only one signal from these studies (a

signal in *CDH13* associated with FEF_{25-75} decline) meeting genome-wide significance and with independent replication (121). This chapter describes the largest GWAS of flow lung function measures carried out to date with the aim of determining whether investigating these measures of lung function can reveal any regions of the genome which might be influencing lung function, which would not be identified through studies of FEV_1, FVC and FEV_1/FVC alone. The analyses described in this chapter, utilise a subset of individuals from UK Biobank who have both genotype and high quality spirometry data available. Firstly, the process of deriving various phenotype and blow quality measures from spirometric curves is described, followed by a summary of the QC of phenotype data and the selection of samples from UK Biobank. Summaries of all lung function measures are then shown and correlations between the traits described. The results of GWAS of PEF and FEF_{25-75} are then presented. For all signals identified as showing association with one or both of PEF and FEF_{25-75}, associations with three volumetric lung function traits (FEV_1, FVC and FEV_1/FVC) were tested, and the overlap in signals for all traits examined.

5.2 Quality Control of phenotype data and sample selection

5.2.1 Derivation of variables from blow curves

Each individual in UK Biobank underwent spirometry using a Vitalograph Pneumotrac 6800. Each individual performed two blows; the reproducibility of the two blows was then checked by the Vitalograph software (defined by software as <5% difference in FEV_1 and FVC). If the blows were reproducible, no further blows were undertaken. Where the blows were not reproducible, a third blow was performed. Within the UK Biobank data, measures of FEV_1, FVC and PEF recorded by the Vitalograph software were provided for each blow, along with a number of indices of blow quality. Additionally, for each blow, the volume measures in mililitres, recorded at 10 milisecond (ms) intervals were provided (blow data points), from which blow curves could be plotted, and a series of measures derived.

Firstly, from the 502,682 individuals in the whole of UK Biobank, 427,222 individuals of self reported white European ancestry and with at least 2 FEV_1 and FVC measures were

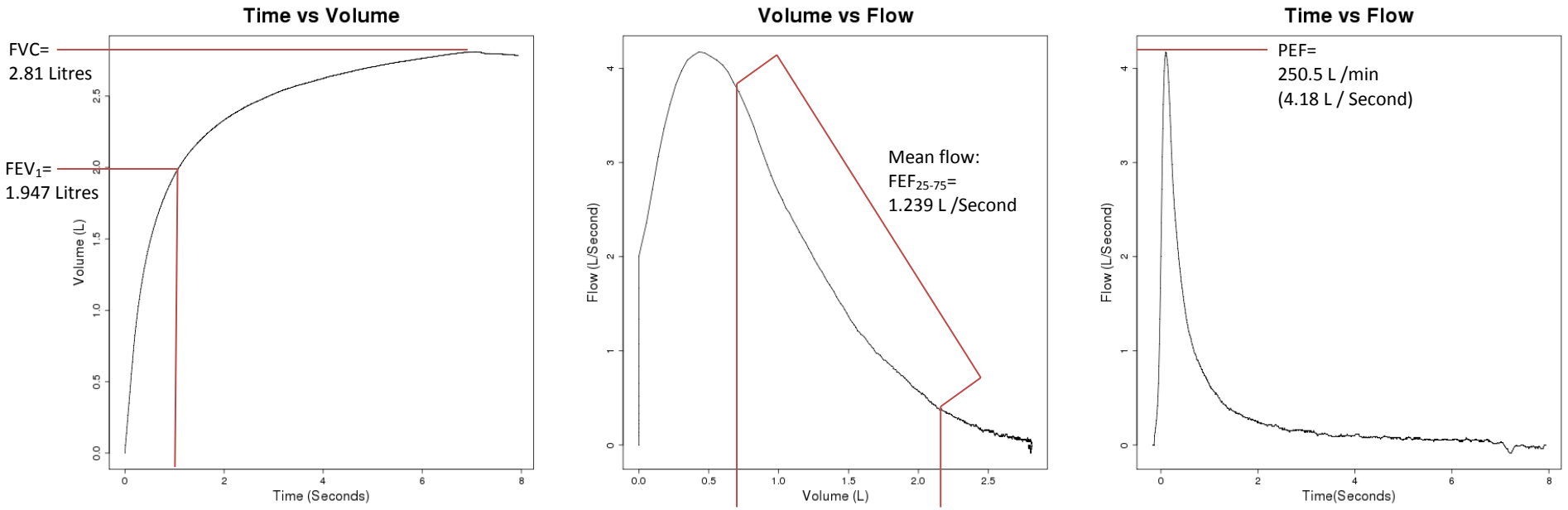
selected as the sampling frame. For these individuals, a number of measures (listed in Table 5-1) were derived from the blow data points, for all of the blows carried out by each individual. Whilst the lung function measures FEV_1, FVC and PEF were already available in the UK Biobank data, deriving these variables from the blow data points and comparing these to the recorded measure was used as an indicator of data quality (Section 5.2.2). The other phenotype included in these analyses, FEF_{25-75}, was not included in the measures recorded by the Vitalograph software, so this had to be derived from the blow data points. The remaining variables were derived for quality control purposes.

Table 5-1: Variables derived from the blow data points.
Each variable was generated for up to 3 blows from each individual.

Variable type	Description
Phenotype	FEV_1 derived from blow data points
Phenotype	FVC derived from blow data points
Phenotype	PEF - peak expiratory flow over 80ms interval (L/min), derived from blow data points
Phenotype	FEF_{25-75} - 80ms flow rate between 25% and 75% of FVC (L/s)
QC measure	Time of PEF
QC measure	Back-extrapolated volume (ml)
QC measure	FET - Time recorded after back-extrapolated zero (s).
QC measure	Flow in last 1.0 second of FET (ml/s)

All flow values were derived using the mean flow over 80 ms intervals. PEF was then derived as the highest of these flow values. A new "time zero" was also estimated, which back extrapolated from the time of PEF, assuming constant flow; this back-extrapolated time zero was then used as the start for all timed measurements. The back extrapolated volume was derived as the volume of air already exhaled by the new time zero (79). Figure 5-1 shows an example of blow curves, generated for one blow from one individual, with the derived values of FEV_1, FVC, PEF and FEF_{25-75} indicated.

Figure 5-1: Example blow curves for an individual, with derived volume and flow measures.



5.2.2 Selection of acceptable and reproducible blows

There were a total of 1,281,666 blows recorded from the 427,222 individuals included in the sampling frame. For each blow there was a field which described the acceptability of the blow, as determined by the Vitalograph software. This field indicated whether there were any problems with the blow, for example if a cough was detected during the manoeuvre, the field indicated "*COUGHING*", or if the blow had a duration of less than 6 seconds, the field indicated "*TEST_DURATION*". At the end of the manoeuvre, the healthcare professional had the opportunity to accept or reject the blow; if this was done, the fields indicated "*USER_ACCEPTED*" or "*USER_REJECTED*", respectively. The UK Biobank protocol specified that all blows should be automatically accepted or rejected by the software however; where there were no issues with the blow and the blow was automatically accepted, the field was blank. Blows were first excluded if in this UK Biobank (Vitalograph) acceptability field they had anything other than "*USER_ACCEPTED*", "*USER_ACCEPTED* + *TEST_DURATION*", "*TEST_DURATION*", or were blank. Blows were then excluded based on the derived variables: if they had poor start of blow quality (back extrapolated volume>5% of FVC, or 150ml, whichever was greater), or were without a terminal plateau (flow rate >25ml/s in last second), or had a duration of less than four seconds (FET<4). Table 5-2 summarises the number of blows failing each quality criterion.

Table 5-2: Quality control of blows.

Quality Criterion	**Total no. blows**
Poor blow based on Vitalograph acceptability field	532,057
Poor start of blow quality	183,098
No terminal plateau	395,489
Duration <4 seconds	164,481
No. blows failing one or more QC criteria	**671,976**

671,976 blows failed at least one QC criterion and were excluded, leaving a total of 609,690 acceptable blows. The numbers of individuals with one, two and three acceptable blows were 124,227, 175,953 and 44,519, respectively. Following these exclusions, each acceptable blow was then checked for reproducibility (within 250ml) with any other blow (acceptable or not).

For each individual, the "best" FEV_1 and FVC were selected as the highest acceptable and reproducible measure for each, not necessarily from the same blow. The best flow measures (PEF and FEF_{25-75}) were derived from the blow with the highest acceptable measure of FEV_1+FVC; again this was not necessarily from the same blow as FEV_1 and/or FVC. 311,762 individuals in total had acceptable and reproducible measures for all phenotypes.

The best FEV_1, FVC and PEF measures derived from the blow curve data were then compared to the values recorded in UK Biobank (Figure 5-2). The FEV_1 and FVC values were overall highly concordant, with a few exceptions: for 58 individuals the best derived FEV_1 and/or FVC differed to the measure recorded in UK Biobank by greater than 5%. These 58 individuals were excluded. For PEF, there were many more individuals whose PEF derived from the blow curves significantly differed from the value recorded in the data. Plots of the PEF versus FEV_1 values from the UK Biobank dataset and the derived variables (Figure 5-3), show that these two measures are correlated for the derived variables, but for the values in UK Biobank, there are a large number of individuals with very low values of PEF given their FEV_1. This indicated that the PEF values recorded in UK Biobank were erroneous; the derived PEF values therefore were chosen for all further analyses.

Figure 5-2: Comparison of FEV_1, FVC and PEF values recorded in UK Biobank with the equivalent measures derived from the blow curve data.

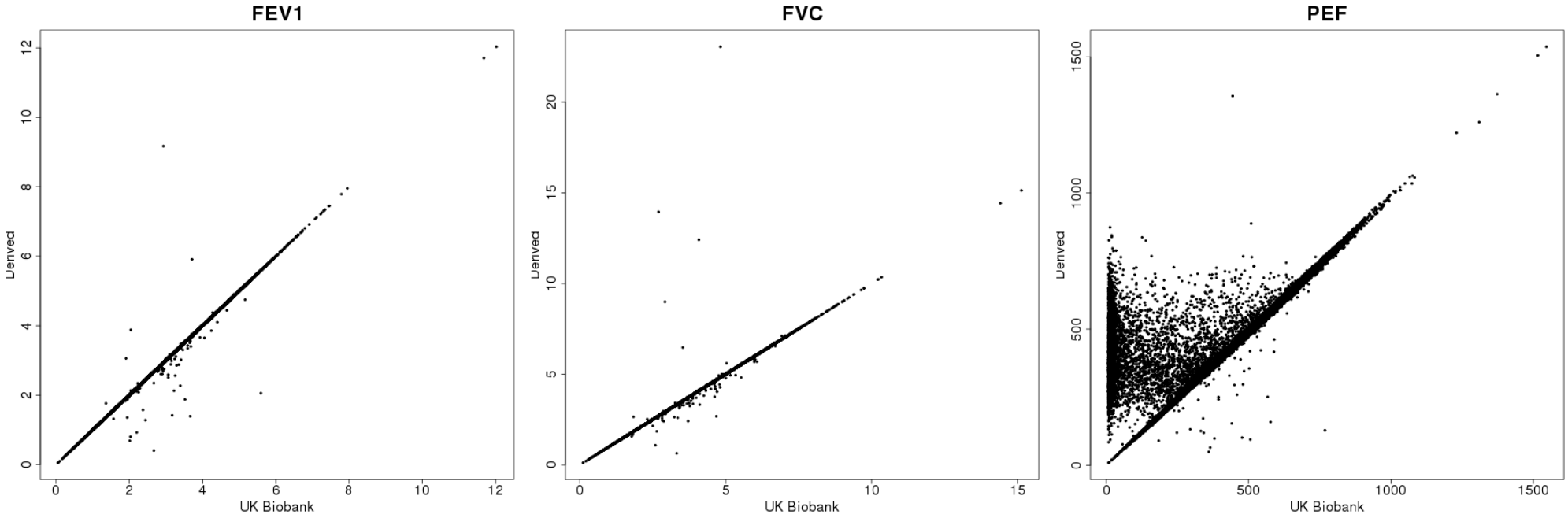
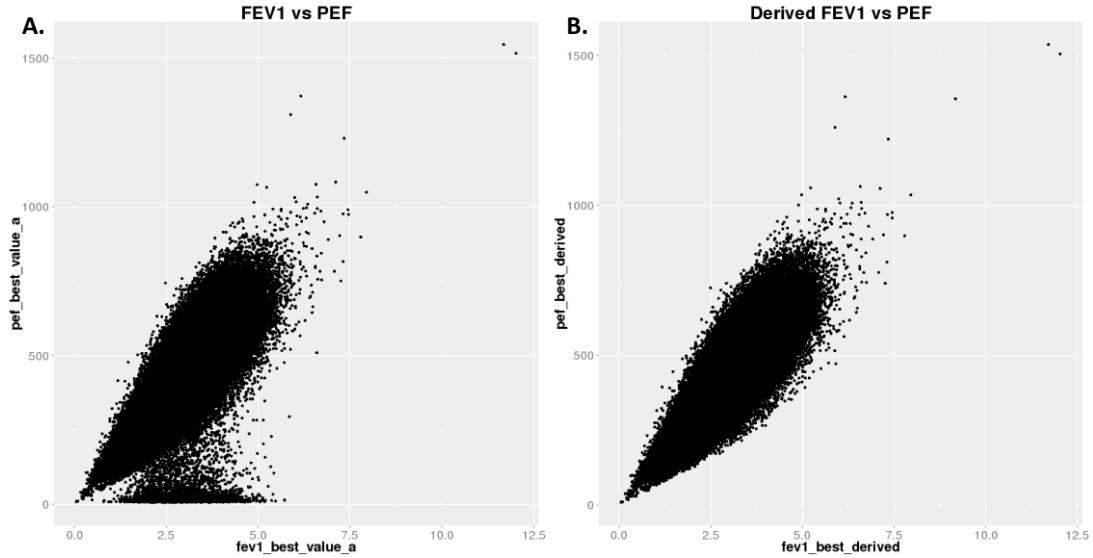


Figure 5-3: Comparison of FEV_1 versus PEF.
A. Measures from UK Biobank. B. measures derived from blow curves.

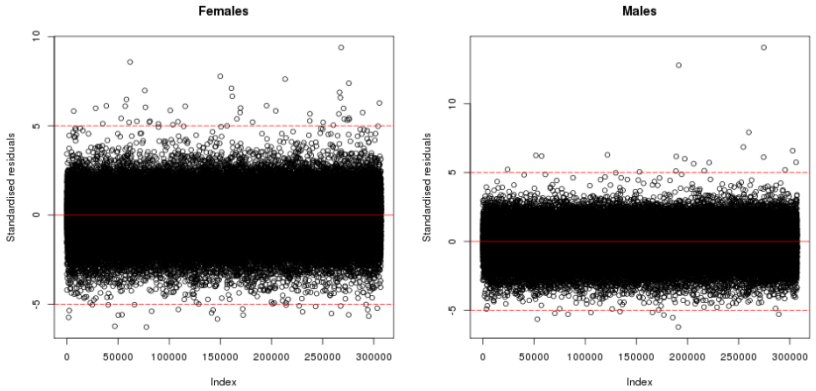


Further exclusions were then made to remove individuals with extreme lung function measures or missing covariates. Firstly 4554 individuals with missing phenotype data for height and smoking status were excluded. Linear models were then fitted for FEV_1, FVC, PEF and FEF_{25-75}, with adjustment for age, age^2, height and height^2 and ever smoking, separately in males and females. 188 individuals with studentised residuals greater than 5, or less than -5 for any trait were identified as outliers (Figure 5-4).

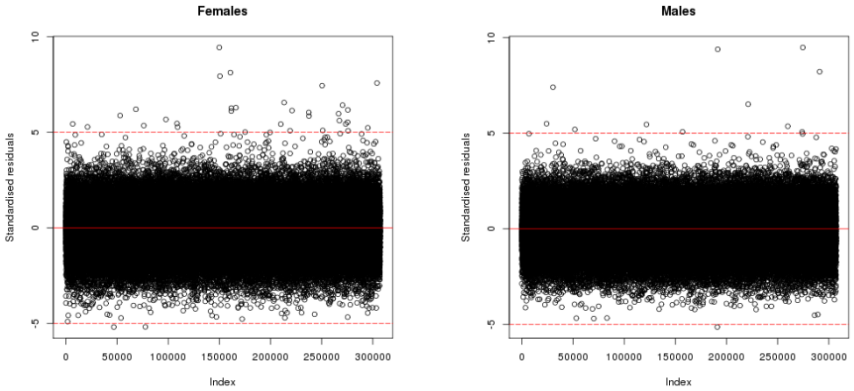
Following all exclusions, 306,962 individuals with full lung function and smoking data remained, of which 105,547 have been genotyped.

Figure 5-4: Studentised residuals of FEV_1, FVC, PEF and FEF_{25-75}, after adjustment for age, age^2, height and height^2 and ever smoking, separately in males and females.
Red lines indicate 5 SDs from the sample mean.

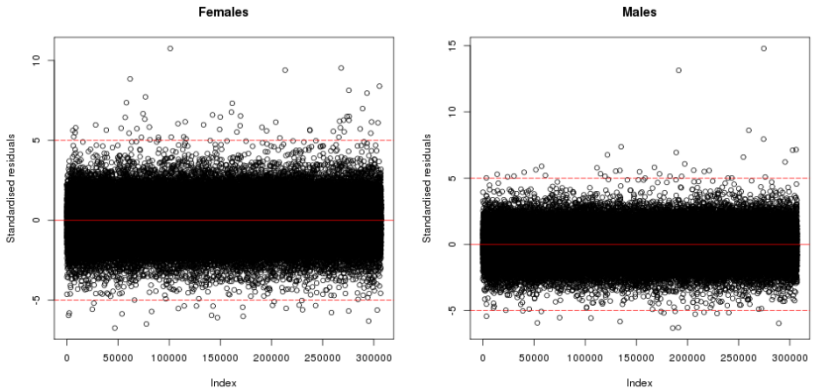
FEV_1



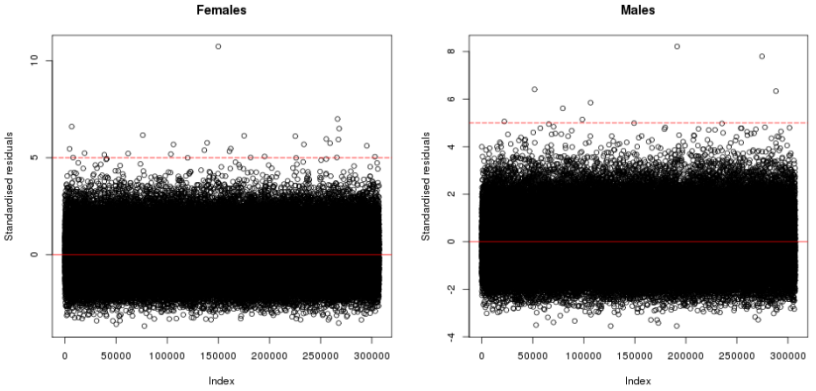
PEF



FVC



FEF_{25-75}



5.2.3 Relation of sample selection process to the UK BiLEVE study

The quality control of the UK Biobank phenotype data and selection of the samples described in the preceding sections was undertaken similarly to in the UK BiLEVE study. The UK BiLEVE study was the first genetic study to be carried out in UK Biobank, and involved the selection of 48,943 individuals, carried out as follows: firstly, the sampling frame was defined as individuals of white European ancestry (self-reported) who had at least two recorded spirometric measures meeting ATS/ERS criteria (blows acceptable and within 150ml of another blow). The smoking status of these individuals was then determined, and those who were classified as either never smokers, or heavy-smokers were retained (heavy smokers defined using the measure pack-years as a proportion of adult lifespan). Predicted values of FEV_1 were then calculated using an internal reference sample of healthy never smokers, who reported no respiratory disease. Using these predicted values, percent predicted FEV_1 was then calculated for all individuals. Within the never smokers and heavy smokers separately, samples were selected for high FEV_1, low FEV_1 and middle FEV_1 groups, based on these percent predicted values. This process was undertaken by myself, in parallel to a second analyst; this was to ensure consistency of sample selection. A full description of the sample selection for the study and results of this study were published in the Lancet Respiratory Medicine in 2015 (120).

Since the UK BiLEVE study was undertaken, approximately 100,000 further samples have been genotyped in UK Biobank. For the analyses described in this chapter, I repeated the quality control of the phenotype data and defined a new selection of samples, in order to utilise as many of the 150,000 genotyped individuals as possible. To this end, I adapted the sample selection process I undertook for the UK BiLEVE study, informed by an evaluation of UK Biobank spirograms and derived indices from the volume-time curves, carried out independently (D. P. Strachan, personal communication, 02/09/2015), to ensure the most suitable flow measures were selected.

5.3 Quality Control of genotype data

The 105,547 samples included in these analyses were amongst 152,256 individuals from UK Biobank genotyped using the Affymetrix Axiom UK BiLEVE or UK Biobank arrays. These two arrays are very similar, sharing 95% of their content. Once genotyped, all samples were imputed to a combined 1000 Genomes Phase 3 (16) and UK10K (40) reference panel. A series of sample genotype QC metrics were provided by UK Biobank along with the imputed data. These metrics included sex mismatches, missingness, heterozygosity rate, and ancestry based on principal component analysis. For pairs of related samples, kinship coefficients, estimated using KING's robust estimator (198) were also provided. Sample QC was firstly undertaken to exclude samples, based on the QC metrics in Table 5-3.

Table 5-3: Sample exclusions based on genotype QC metrics provided by UK Biobank.

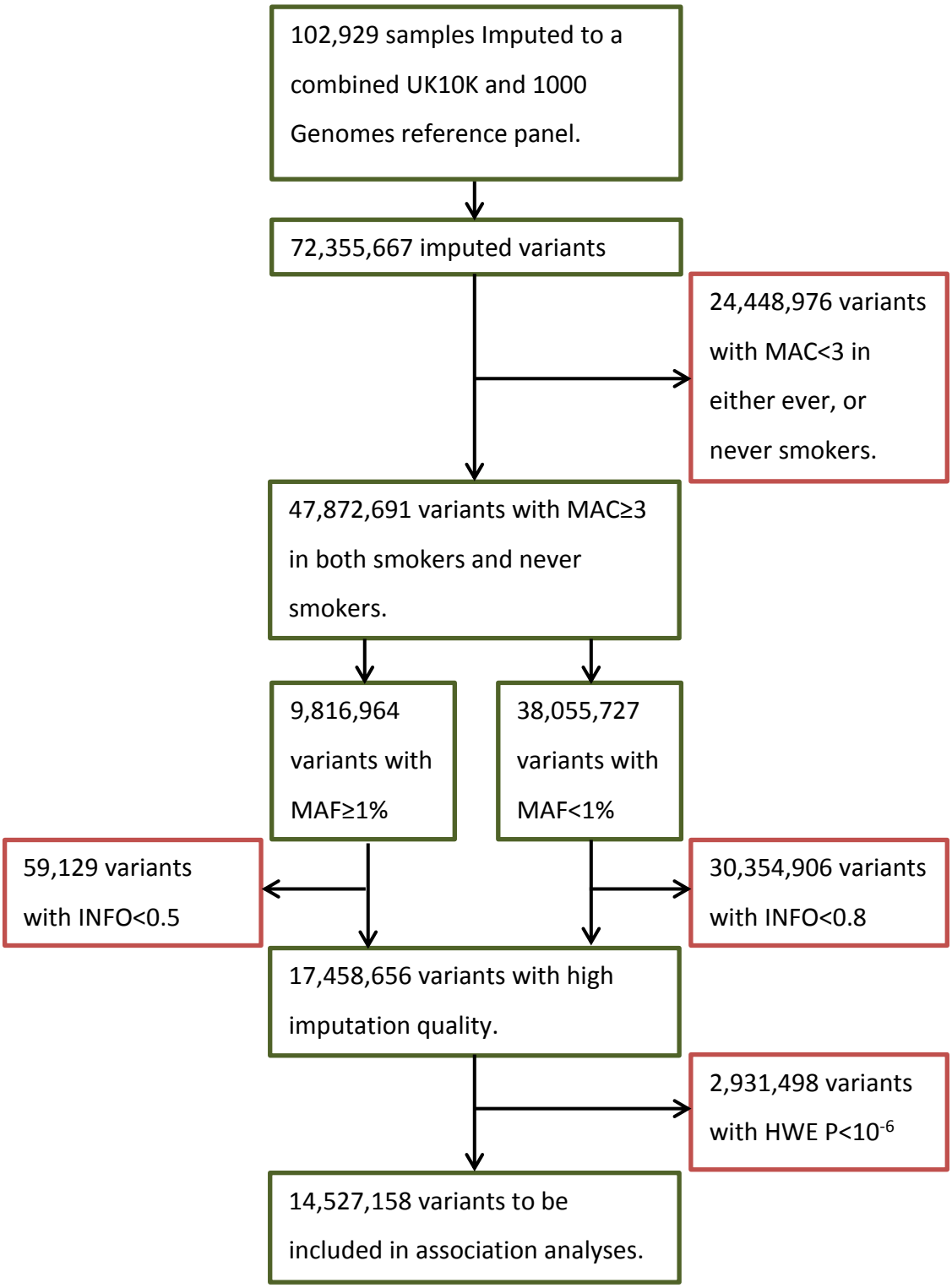
Reason for exclusion	**N samples**
Withdrawn consent	4
Sex mismatch	131
Heterozygosity outlier / high missingness	303
Non-European ancestry (PCA outlier)	272
Total samples excluded	**689**

In the remaining 104,858 samples, pairs of related individuals were identified as those with kinship coefficients >0.088, equivalent to at least 2nd degree relatives. For each related pair, the sample with the highest rate of missingness was selected for exclusion. In total, 1929 related samples were removed, leaving a final sample of 102,929 unrelated individuals (54,538 never smokers and 48,391 ever smokers).

There were a total of 72,355,667 imputed variants available for the 102,929 selected individuals. The majority of those variants were SNPs, but some short indels were also included. Stringent genotype QC was carried out to exclude variants with a very low minor allele count (MAC) (MAC<3 in either the ever smokers or never smokers), variants which had a low imputation information (INFO) score, or which deviated from HWE ($P<10^{-6}$). The INFO score is a metric which ranges from 0 to 1 and acts as a measure of uncertainty of the imputed genotype (INFO=1 indicates no uncertainty of genotype, whilst INFO=0 indicates complete uncertainty of the genotype). For variants

with a MAF>1%, variants with INFO<0.5 were excluded. For rare variants with MAF≤1%, INFO<0.8 was used for exclusions. These INFO score filters are fairly conservative; however in these analyses, there are no independent samples in which to verify identified associations, so a strict INFO score threshold was utilised to limit the number of false positive findings. The numbers of variants excluded due to a low MAC or INFO score, or as they deviated from HWE are summarised in Figure 5-5.

Figure 5-5: Summary of variant exclusions, prior to association testing.



5.4 Analysis of PEF and FEF_{25-75}: Methods

5.4.1 Statistical Analyses

For both PEF and FEF_{25-75}, linear models were fitted separately in never and ever smokers, with age, age^2, sex, height, and 10PCs as covariates. The residuals resulting from these models were converted to ranks and then to normally distributed z-scores. These inverse rank normalised traits were used for all subsequent association testing. Associations were carried out using the score test as implemented in SNPTEST v2.5b4 (19), assuming an additive genetic model, and using genotype doses (continuous from 0 to 2). The genomic inflation factor (equation (1-4)) was calculated for the genome-wide results for never and ever smokers separately, and the standard errors and P-values for each smoking stratum were adjusted accordingly. The results for ever and never smokers were then combined, using inverse variance weighted meta-analysis (equation (2-3)). Finally the overall genomic inflation factor was calculated using the genome-wide results for all samples combined, and the standard errors and P-values adjusted.

5.4.2 Selection of signals

In the remainder of this chapter, the results shall be described in terms of signals, sentinel SNPs and regions. Regions are defined based on genomic position only, with each region potentially including multiple signals. Signals refer to each independent association, and often include several variants which are associated with the trait, due to the LD structure. The sentinel SNP is the most highly significant SNP from each signal.

Firstly, all SNPs associated with either PEF or FEF_{25-75} with $P<5x10^{-8}$ were identified. For both traits, selection of sentinel SNPs was carried out as follows: Amongst the identified SNPs, the most statistically significant association was selected as the first sentinel SNP. All SNPs located +/-1MB from that sentinel SNP were then excluded. Of the remaining SNPs, the next most significant was selected as a second sentinel SNP, with all SNPs +/-1MB from that sentinel SNP subsequently excluded. This process was repeated until all sentinel SNPs, each representing a 2Mb region were selected.

Following the generation of this initial list of sentinel SNPs, secondary signals were identified, by undertaking further analyses within each 2MB region, conditioning on the sentinel SNP in that region. Conditional analysis was undertaken using SNPTEST, where the sentinel SNP was included as an additional covariate in the model. All SNPs with $P<5x10^{-8}$ in these conditional analyses were identified, with the most strongly associated SNP in each region selected as a secondary sentinel.

Where a secondary signal existed, regions were redefined such that the region spanned both sentinels in the region, +/-1MB. Where these newly defined regions overlapped, the regions were merged to form larger regions with multiple sentinels. In these larger regions, an additional conditional analysis was undertaken, for all sentinels, conditioning on all other sentinels in the region, to ensure each sentinel represented an independent signal. Any sentinel with $P>5x10^{-8}$ in these conditional analyses was removed from the sentinel list, and the region redefined, if appropriate.

A final conditional analysis was undertaken in all newly defined regions with more than one sentinel, conditioning on all sentinel SNPs in the region, in order to identify any further signals. Finally, to ensure all selected sentinel SNPs did in fact represent independent signals, linkage disequilibrium between all SNPs was estimated, to ensure none were correlated ($r^2>0.1$).

5.4.3 Analysis of top findings and volumetric lung function traits

For all SNPs identified as showing association with PEF and/or FEF_{25-75}, associations were tested for three volumetric measures of lung function: FEV_1, FVC and FEV_1/FVC. Each volumetric trait was inverse normally transformed, separately in ever and never smokers, with adjustments made for age, age^2, sex, height, and 10PCs. The inverse rank normalised traits were used for association testing, with the results for ever and never smokers combined using inverse-variance weighted meta-analysis. The results for each volumetric trait were compared with those for the flow measures. SNPs which were associated with flow measures with $P<5x10^{-8}$, but which were not associated with any volumetric measure with $P<5x10^{-5}$, were selected as PEF or FEF_{25-75} specific signals.

5.4.4 Quality Control of results

For all sentinels, cluster plots (Section 1.1.3.4) were generated to assess the accuracy of genotype calling. Where a sentinel SNP was not directly genotyped, clusterplots were generated for the strongest genotyped proxy SNP (based on LD) to the imputed SNP and checked for calling accuracy. For all identified sentinel SNPs, additional association analyses were undertaken to determine whether associations might be a result of a chip effect. Associations were tested between each SNP and genotyping array (UK BiLEVE array vs UK Biobank array) using the score test as implemented in SNPTEST v2.5b4 (19) and SNPs that were significantly associated with genotyping array ($P<10^{-5}$) were identified.

5.5 Analysis of PEF and FEF_{25-75}: Results

5.5.1 Lung function phenotypes in UK Biobank

Phenotype summaries for all 102,929 samples included in these analyses are in Table 5-4.

Table 5-4: Phenotype summaries for all samples included in analyses, by smoking status.

	Ever smokers (n=48,391)	Never smokers (n=54,538)
Sex, N (%) Male	24,336 (50.3%)	22,920 (42.0%)
Age, Mean (SD)	57.1, (7.9)	56.0 (8.0)
FEV_1, Litres, Mean (SD)	2.714 (0.798)	2.815 (0.796)
FVC, Litres, Mean (SD)	3.563 (0.973)	3.684 (0.999)
FEV_1/FVC, Mean (SD)	0.740 (0.077)	0.764 (0.063)
PEF (Litres/min), Mean (SD)	393.6 (123.5)	401.5 (119.5)
FEF_{25-75} (Litres/sec), Mean (SD)	2.252 (0.993)	2.499 (0.984)
Pack-years, Mean (SD) (n=38,377 ever smokers with pack-years)	27.29 (18.14)	-

PEF and FEF_{25-75} and the three volumetric lung function traits are all closely related, as can be seen in Figure 5-6 and Figure 5-7. PEF can be seen to be most strongly correlated with FEV_1 (inverse-normally transformed traits r^2=0.741), with slightly

weaker correlations with FVC (r^2=0.601) and FEV_1/FVC (r^2=0.512). FEF_{25-75} shows an even stronger correlation with FEV_1 (r^2=0.824) and is also highly correlated with FEV_1/FVC (r^2=0.85), whilst showing weaker correlation with FVC (r^2=0.539). PEF and FEF_{25-75} are also moderately correlated with each other (r^2=0.676).

Figure 5-6: Comparisons of PEF versus other lung function trait values (FEV_1, FVC, FEV_1/FVC and FEF_{25-75}) in the n=102,929 samples included in the association analyses.
A. raw trait values B. inverse-normally transformed trait values.

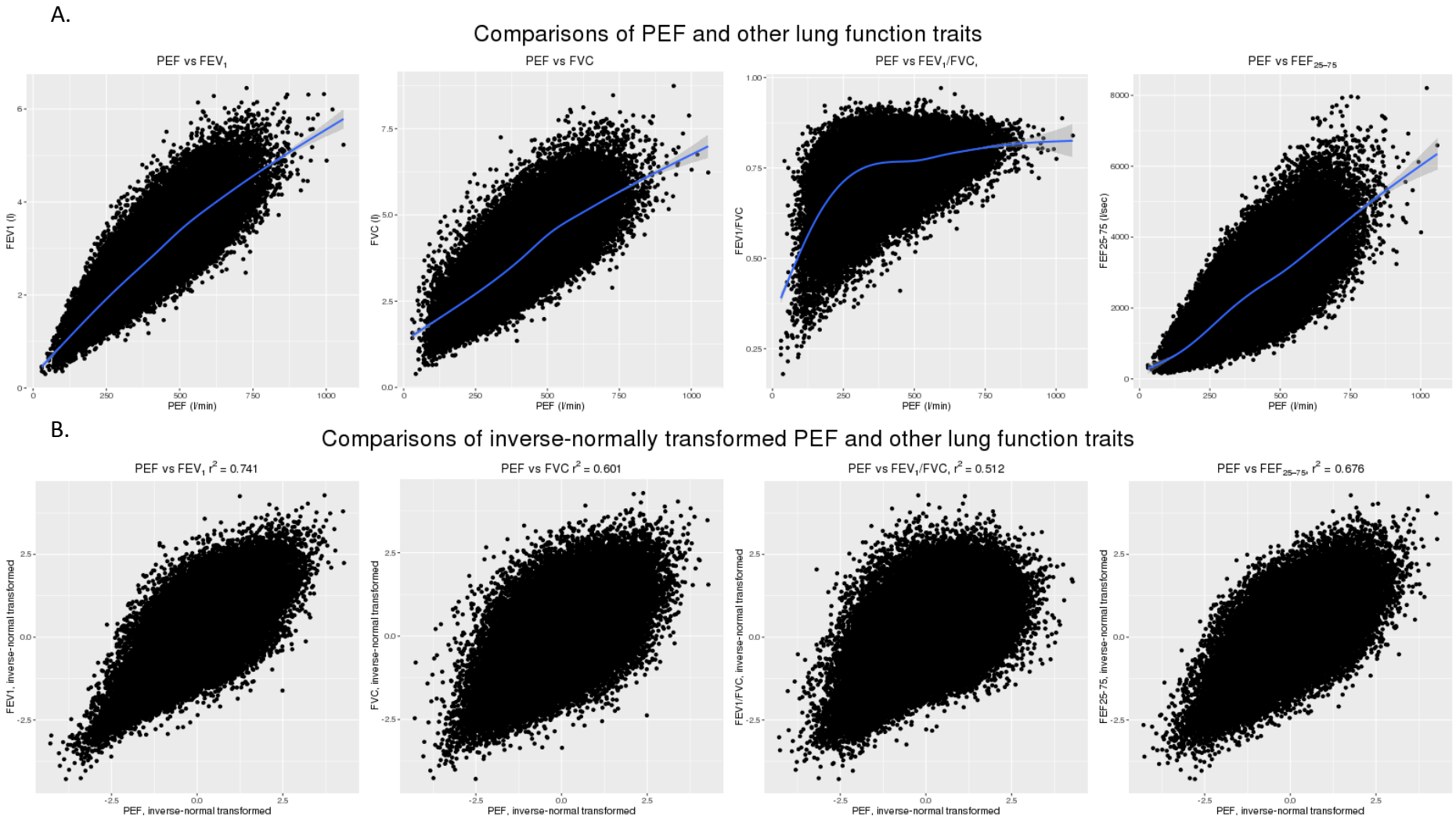
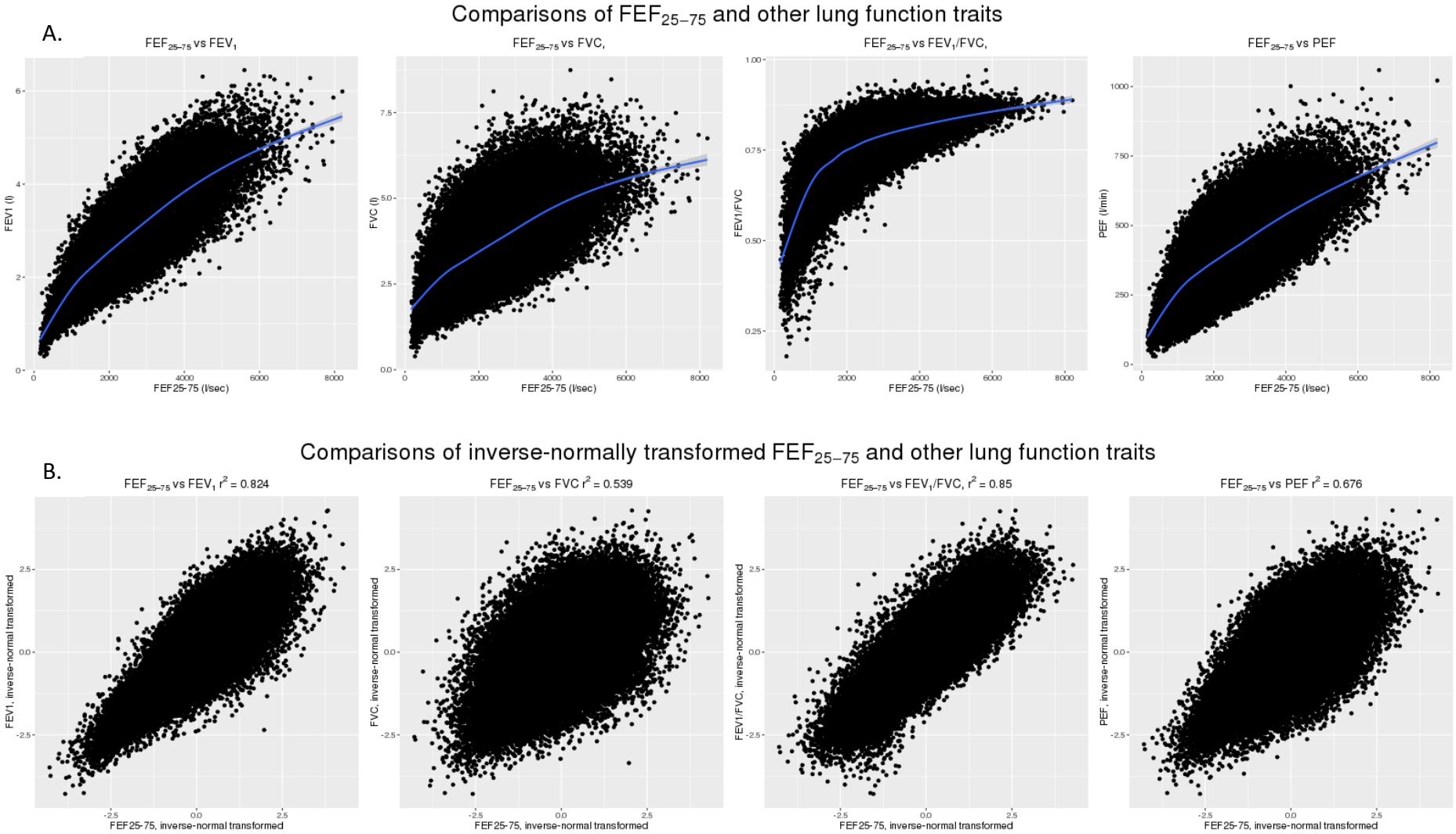


Figure 5-7: Comparisons of FEF_{25-75} versus other lung function trait values (FEV_1, FVC, FEV_1/FVC and PEF) in the n=102,929 samples included in the association analyses.
A. raw trait values B. inverse-normally transformed trait values.

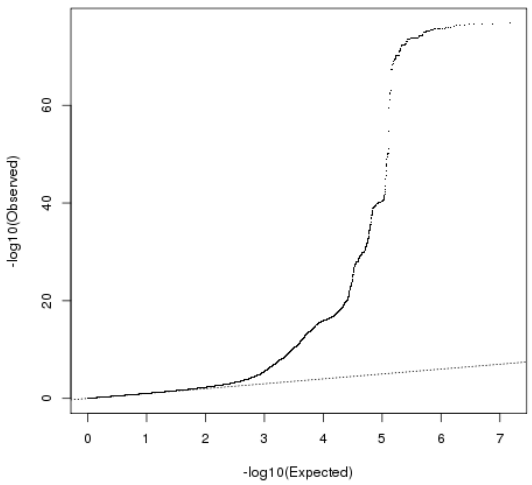


5.5.2 Single variant association analyses

A total of 14,527,158 variants and 102,929 samples passed all QC and were included in the association analyses. The genomic inflation factorsl (λ, equation (1-4)) for the analyses of PEF and FEF_{25-75} were 1.0617 and 1.0646, respectively. There were a considerable number of strongly associated variants for both traits, as can be seen in the QQ plots (Figure 5-8), which show substantial deviation from the null, and in the Manhattan plots (Figure 5-9).

Figure 5-8: QQ plots of analysis of A. PEF and B. FEF_{25-75}

A. PEF: λ=1.0617



B. FEF_{25-75}: λ=1.0646

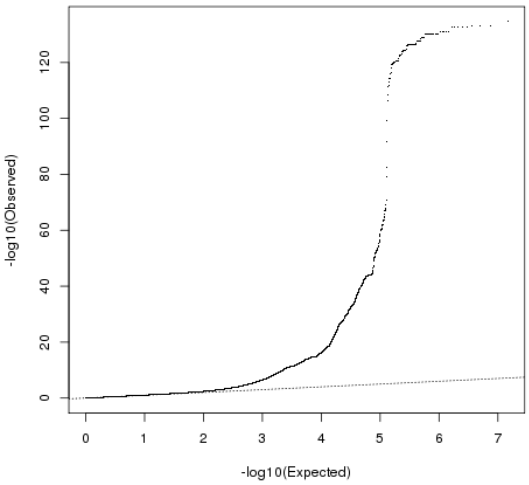
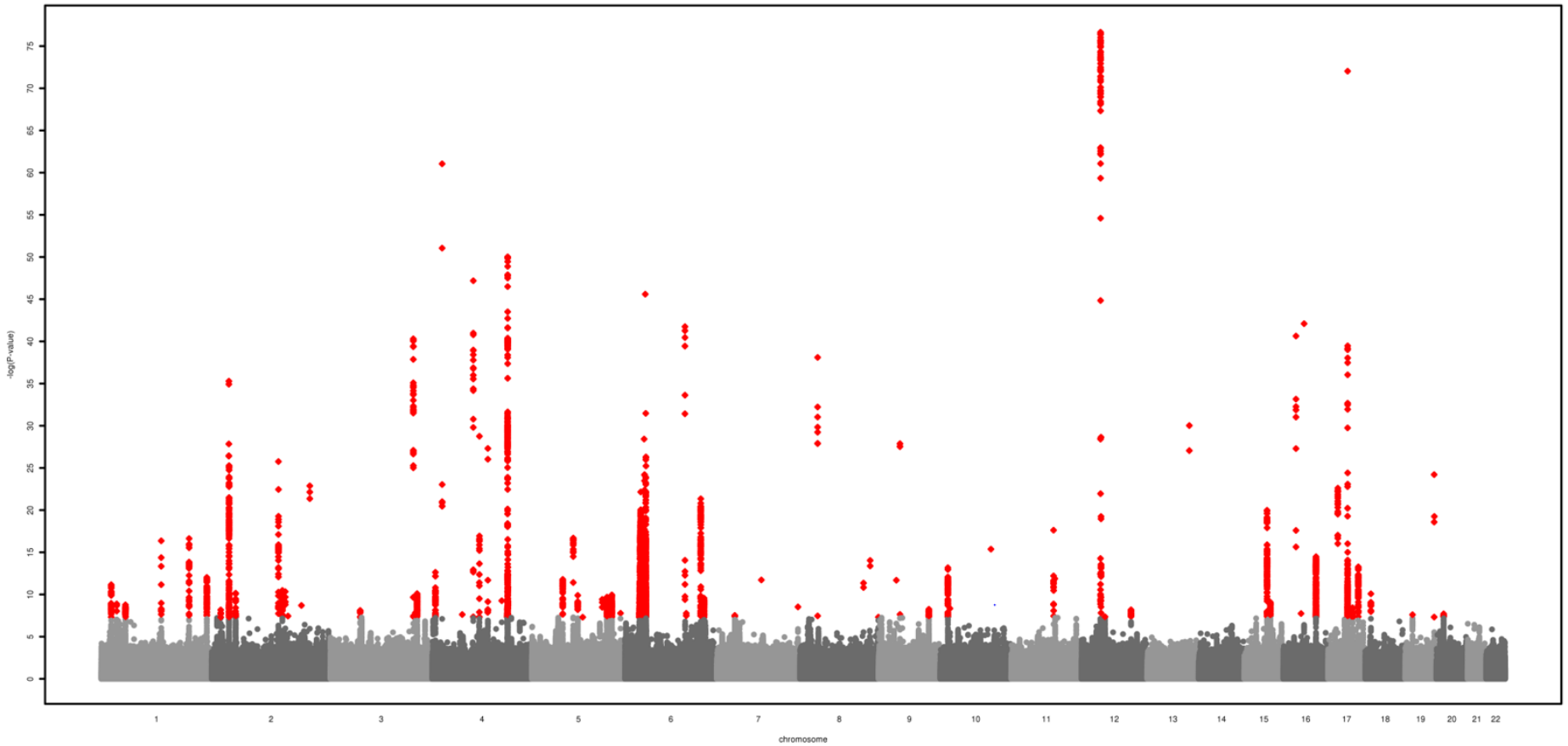
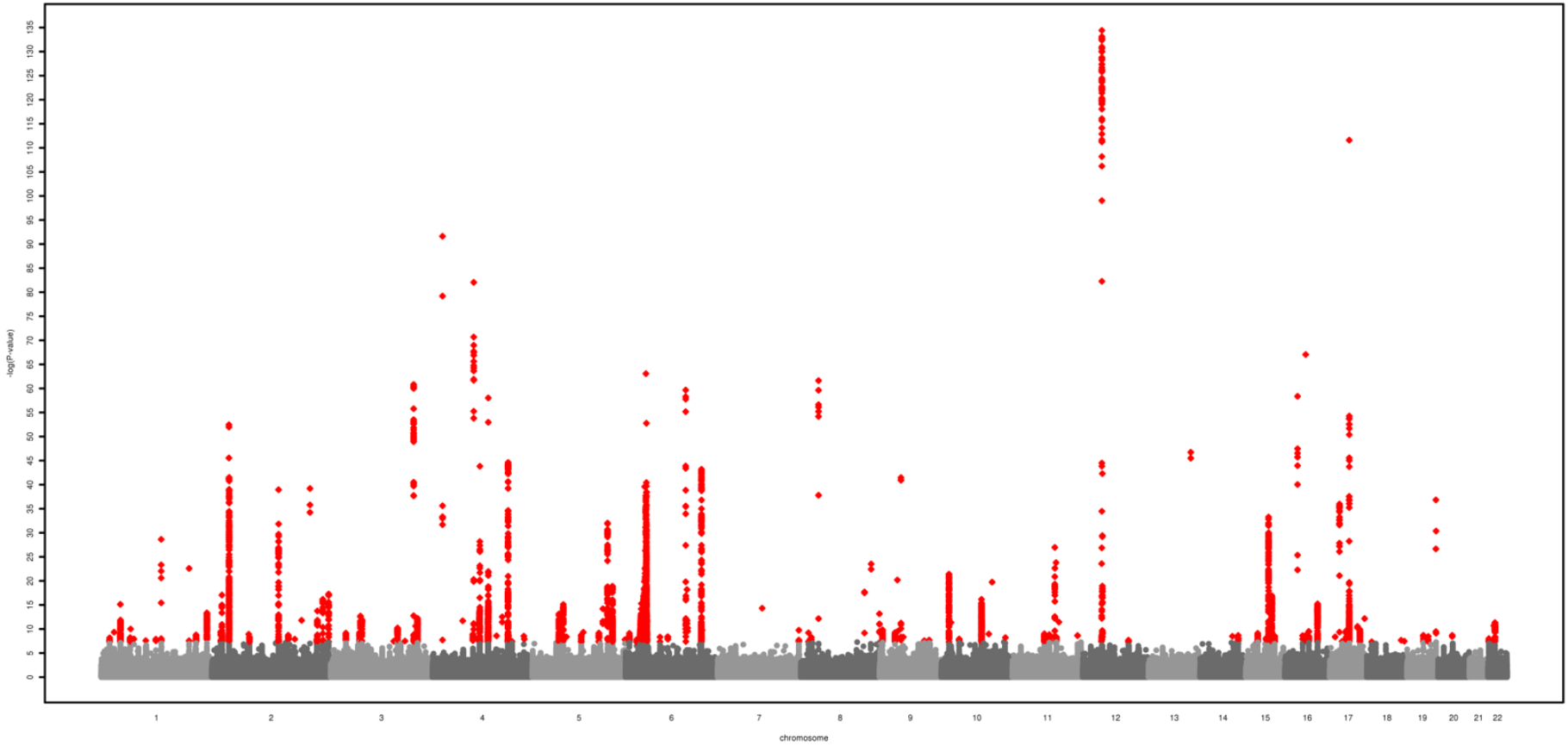


Figure 5-9: Manhattan plots for the analysis of A. PEF and B. FEF_{25-75}.
Highlighted SNPs significant $P<5x10^{-8}$

A. Analysis of PEF



B. Analysis of FEF_{25-75}



5.5.3 Selection of Signals

Initially, 94 sentinel SNPs were identified as showing association with PEF with $P<5x10^{-8}$, each SNP chosen as the most significantly associated SNP within a 2MB region. Within each region, secondary signals were tested for, through conditional analyses, conditioning on the sentinel SNP; these conditional analyses identified a further 24 SNPs independently associated with PEF with $P<5x10^{-8}$. Final conditional analyses were undertaken in regions with more than one sentinel to identify tertiary signals; an additional 12 SNPs were identified in these final analyses. In total, 127 SNPs in 93 regions were identified as significantly associated with PEF. For FEF_{25-75}, a total of 215 SNPs in 153 regions were identified as showing association $P<5x10^{-8}$. The process of identifying secondary and tertiary signals for both traits is shown in Figure 5-10. Figure 5-11 shows an example of this process for a region on chromosome 4 with 3 signals identified in the analysis of FEF_{25-75}. This example is in a region spanning several genes including *TET2*, *INST12* and *GSTCD*, and within this region multiple independent signals have been previously been identified for FEV_1 and FEV_1/FVC (43).

Figure 5-10: Selection of sentinel SNPs.

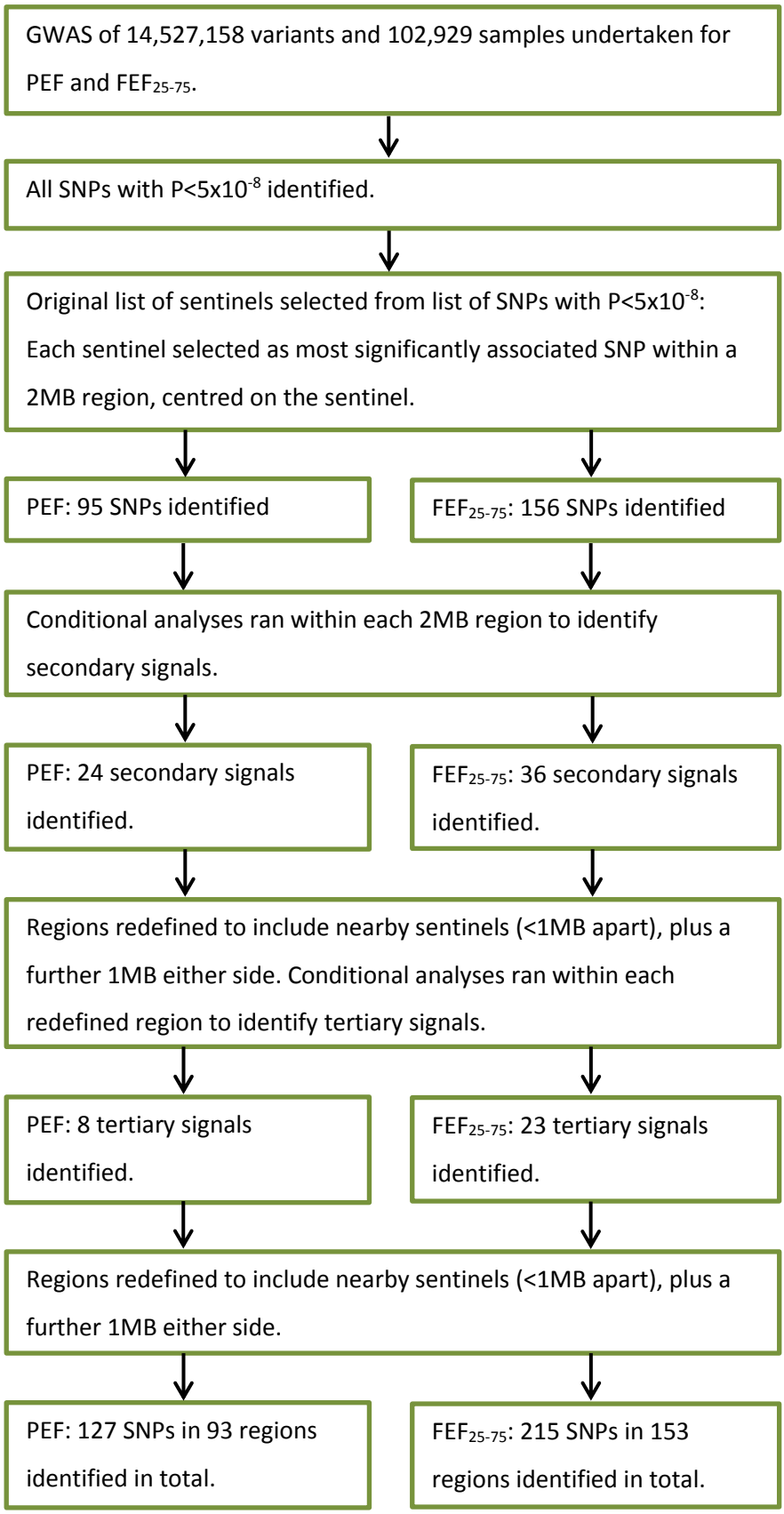
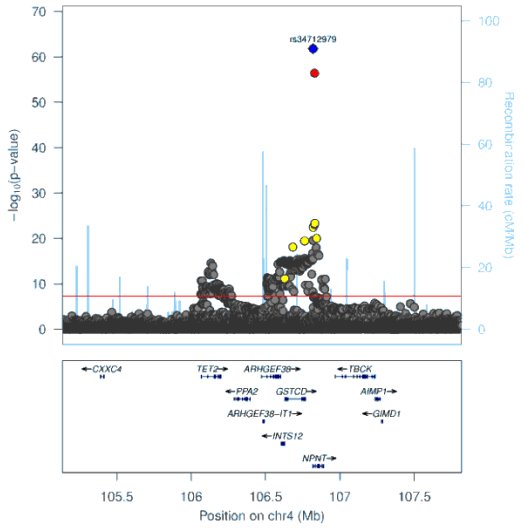


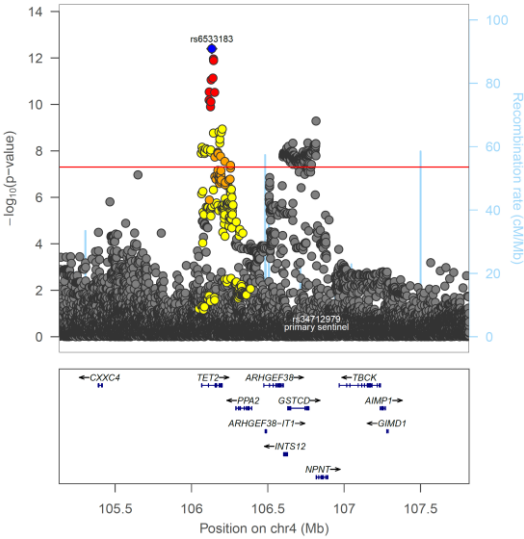
Figure 5-11: Identification of primary, secondary and tertiary signals in region chr4:105133184-107819053, associated with FEF_{25-75}.

Plots (right) show the sentinel SNP in each region is highlighted in blue with the LD (r^2) of nearby SNPs to the sentinel indicated by colour (red: r^2>0.8, orange: 0.8≥r^2>0.5, yellow:0.5≥r^2>0.2, grey: r2<0.2). The fine scale recombination rate is shown in light blue.

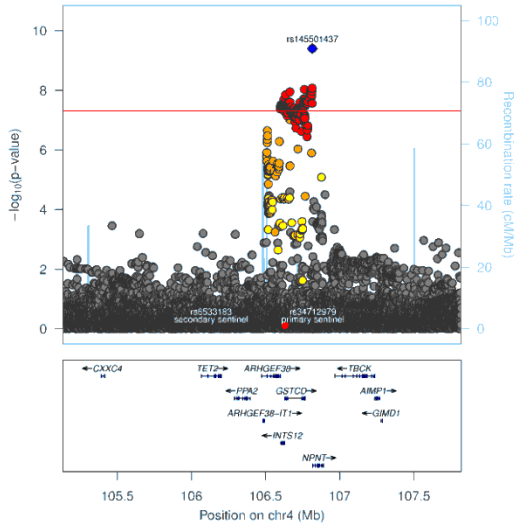
rs34712979 (chr4:106819053, P=1.61x10^{-62}) selected as primary sentinel SNP in 2MB region (chr4: 105819053-107819053).



Analysis carried out in region (chr4: 105819053-107819053), conditional on rs34712979. rs6533183 (chr4: 106133184, P=4.02x10^{-13}) identified as secondary signal. Region redefined, to include both sentinels +/- 1MB (chr4:105133184-107819053).



Analysis carried out in region (chr4:105133184-107819053), conditional on both rs34712979 and rs6533183. rs145501437 (chr4:106815984, P=3.46x10^{-9}) identified as tertiary signal. Region defined, to include all three sentinels +/- 1MB (chr4:105133184-107819053).



5.5.4 Top findings

The 10 most significantly associated SNPs for each trait are listed in Table 5-5 (PEF) and Table 5-6 (FEF_{25-75}). The full results for each trait are in Appendix E (PEF: Table E-1; FEF_{25-75}: Table E-2). There was significant overlap in the SNPs and regions identified for PEF and FEF_{25-75} and 9 of the 10 most most significantly associated regions overlapped for the two traits.

In both analyses, the most strongly associated signals were within *CPNE8* (PEF: rs150950471, MAF=2.4%, $P_{PEF}=3.65x10^{-82}$; FEF_{25-75}: rs115903505, MAF=2.5%, $P_{FEF25-75}=5.68x10^{-144}$), rs73314997 in *MAPT* (MAF=2.3%, $P_{PEF}=3.00x10^{-77}$; $P_{FEF25-75}=1.24x10^{-119}$) and rs191050570 in *CC2D2A* (MAF=1.9%, $P_{PEF}=1.42X10^{-65}$; $P_{FEF25-75}=2.40x10^{-98}$). *CPNE8* and *CC2D2A* have not previously been implicated in lung function, or related traits. Other signals in regions not previously associated with respiratory traits include SNPs in *SYT17, CASC16*, *STF6B* and an intergenic SNP near *CCDC15* and *SHROOM3*.

The signal identified in *MAPT* (microtubule-associated protein tau), is within a common inversion locus, which shows marked differences in allele frequencies across European populations (199). SNPs within this gene have previously shown associations with idiopathic pulmonary fibrosis (IPF) (200) and a common (MAF=24%) SNP in the nearby gene *KANSL1,* also within the inversion locus was associated with extremes of FEV_1 (high FEV_1 versus low FEV_1) in the UK BiLEVE study (120). The SNP identified in these analyses is the first low frequency SNP in this region to be associated with lung function.

Table 5-5: Analysis of PEF; 10 most strongly associated SNPs.

Chromosome (Chr) and position (Pos) in build 37 are given for each SNP. Effect estimates are on an inverse-normal transformed scale after adjustments for age, age^2, sex, height and ancestry principal components, and stratified by ever smoking status.

SNP (Chr:Pos)	Effect/ noneffect allele	INFO	MAF	Effect estimate	Standard error	P-value	Annotation(s)
rs79105080 (chr3:160779029) *	T/C	0.8474	2.28%	-0.2397	0.0173	$1.61x10^{-43}$	*PPM1L*, intron_variant
rs191050570 (chr4:15582642) *	T/C	0.6451	1.88%	-0.3890	0.0227	$1.42X10^{-65}$	*CC2D2A*, intron_variant
rs74936215 (chr4:77349465)*	G/A	0.7627	1.51%	-0.3133	0.0225	$3.07x10^{-44}$	Intergenic, near *CCDC158, SHROOM3*
rs7658614 (chr4:145445694)	T/A	0.9987	46.81%	0.0707	0.0046	$7.55x10^{-54}$	Intergenic, near *HHIP*
rs538489083 (chr6:32095727) *	C/T	0.5303	1.70%	-0.3768	0.0256	$3.83x10^{-49}$	*ATF6B*, intron_variant
rs138535200 (chr6:108633740) *	C/T	0.7463	2.54%	-0.2482	0.0176	$4.73x10^{-45}$	*LACE1*, intron_variant
rs150950471 (chr12:39134817)*	C/G	0.6038	2.41%	-0.4098	0.0213	$3.65x10^{-82}$	*CPNE8,* intron_variant
rs74930371 (chr16:19273328) *	G/T	0.9151	1.95%	-0.2447	0.0176	$7.15x10^{-44}$	*SYT17*, intron_variant
rs3104770 (chr16:52627368) *	A/T	0.7563	2.77%	-0.2379	0.0168	$1.98x10^{-45}$	*CASC16,* intron_variant, non_coding_transcript_variant
rs73314997 (chr17:44061123) *	C/T	0.6376	2.31%	-0.3898	0.0210	$3.00x10^{-77}$	*MAPT*, missense_variant

* SNP in one of the 10 most strongly associated regions with FEF_{25-75} (Table 5-6)

Table 5-6: Analysis of FEF_{25-75}; 10 most strongly associated SNPs.

Chromosome (Chr) and position (Pos) in build 37 are given for each SNP. Effect estimates are on an inverse-normal transformed scale after adjustments for age, age^2, sex, height and ancestry principal components, and stratified by ever smoking status.

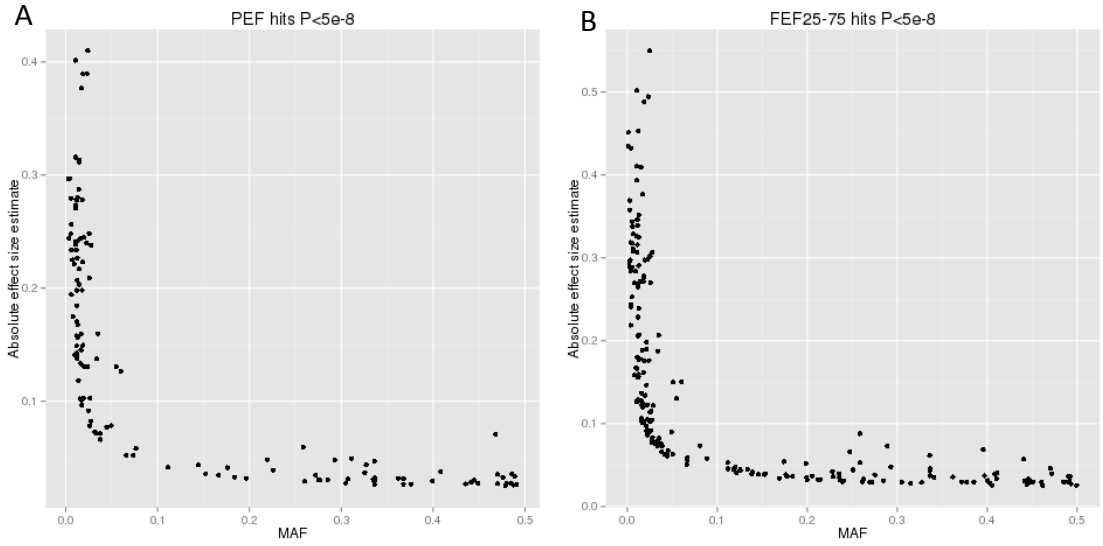
SNP (Chr:Pos)	Effect/ noneffect allele	INFO	MAF	Effect estimate	Standard error	P-value	Annotation(s)
rs79105080 (chr3:160779029) *	T/C	0.8474	2.28%	-0.2985	0.0175	$1.67x10^{-65}$	*PPM1L,* intron_variant
rs191050570 (chr4:15582642) *	T/C	0.6451	1.88%	-0.4881	0.0232	$2.40x10^{-98}$	*CC2D2A,* intron_variant
rs116291420 (chr4:77346196)*	G/C	0.7825	1.52%	-0.4094	0.0222	$5.10x10^{-76}$	Intergenic, near *CCDC158, SHROOM3*
rs538489083 (chr6:32095727) *	C/T	0.5303	1.70%	-0.4520	0.0260	$6.70x10^{-68}$	*ATF6B,* intron_variant
rs138535200 (chr6:108633740) *	C/T	0.7463	2.54%	-0.3021	0.0178	$2.88x10^{-64}$	*LACE1,* intron_variant
rs6981627 (chr8:22530061)	G/C	0.9132	2.56%	-0.2699	0.0157	$2.31x10^{-66}$	*BIN3*, upstream_gene_variant
rs115903505 (chr12:39146668)*	T/A	0.6039	2.48%	-0.5497	0.0215	$5.86x10^{-144}$	*CPNE8*, intron_variant
rs74930371 (chr16:19273328) *	G/T	0.9152	1.95%	-0.2974	0.0178	$7.04x10^{-63}$	*SYT17*, intron_variant
rs3104770 (chr16:52627368) *	A/T	0.7563	2.77%	-0.3065	0.0171	$3.84x10^{-72}$	*CASC16*, intron_variant non_coding_transcript_variant
rs73314997 (chr17:44061123) *	C/T	0.6376	2.31%	-0.4944	0.0213	$1.24x10^{-119}$	*MAPT,* missense_variant

* SNP in one of the 10 most strongly associated regions with PEF (Table 5-5)

The only common SNP amongst the top ten hits for each trait was an intergenic SNP, rs7658614 (MAF=46.81%), associated with with PEF (P_{PEF} =7.55x10^{-54}). This SNP is upstream of *HHIP* and in LD with SNPs which have previously been associated with FEV_1/FVC (117) and COPD related traits (201, 202). The remaining top hits were all low frequency (1%≤MAF<5%) and indeed, the majority of of identified variants had a MAF<5%.

Of the 127 SNPs identified as associated with PEF, 64 had a MAF of 1-5%, and 11 had a MAF<1%. For FEF_{25-75}, 96 of the 215 identified SNPs had a MAF between 1-5%, and 27 had a MAF <1%. Figure 5-12 shows the MAFs and effect estimates of all identified SNPs; low frequency and rare SNPs can be seen to have far larger effect size estimates than common SNPs. This result supports the notion that rare variants are likely to exhibit larger effects than do common variants on a complex trait (37, 203); however very low frequency SNPs are only likely to be detected if they have large effect sizes. It is probable that many more low frequency SNPs with more modest effects could be influencing these traits, however there is not enough statistical power to detect these associations. Furthermore, the estimated effect sizes of these SNPs may also be inflated due to the winner's curse phenomenon, and it is likely that observed effects in independent follow-up samples would be more modest.

Figure 5-12: Comparison of MAF and effect sizes for all SNPs identified in the analyses of A. PEF and B.FEF_{25-75}.



It should also be noted that these low frequency SNPs may be more prone to false positive associations as lower frequency SNPs are more likely to be less well imputed

than are common SNPs. Indeed the majority of the most strongly associated low frequency SNPs (1%≤MAF<5%, Table 5-5 and Table 5-6) had INFO<0.8, indicating some uncertainty of the imputed genotypes (INFO<0.5 used as filter for SNPs with MAF≥1%).

In terms of common SNPs (MAF≥5%), the 10 most significant associations for each trait are listed in Table 5-7 (PEF) and Table 5-8 (FEF_{25-75}). Some of the SNPs listed in these tables are secondary or tertiary signals in regions where the sentinel SNPs for the primary signals were low frequency. The majority of these common SNP associations are within regions previously associated with FEV_1 or FEV_1/FVC, including *HTR4, THSD4, NPNT* (117), *HHIP* (116), *GPR126* (118), *CFDP1 (119).* There were also several SNPs within the major histocompatibility complex (MHC) region associated with PEF and FEF_{25-75} in these analyses; a number of genes within this region have previously been implicated in lung function (117-120). Associations in regions not previously implicated in lung function included common SNPs in, or near to *SLC26A9*, *HAPLN1*, *BIN3* and *BIRC6*.

None of the regions previously identified in GWAS of PEF and FEF_{25-75} (121-123) were identified in the present analyses. rs2325934 in *CDH13* is the only genome-wide significant association with FEF_{25-75} (decline) identified to date, however in this analyses, this SNP showed no association with FEF_{25-75} ($P_{FEF25-75}$=0.398). Other SNPs within *CDH13* did show modest association with FEF_{25-75} however (strongest association with rs552901786, MAF=0.20%, INFO=0.851, $P_{FEF25-75}=1.45x10^{-4}$).

Table 5-7: Analysis of PEF; 10 most strongly associated common SNPs (MAF≥5%).

Chromosome (Chr) and position (Pos) in build 37 are given for each SNP. Effect estimates are on an inverse-normal transformed scale after adjustments for age, age^2, sex, height and ancestry principal components, and stratified by ever smoking status.

SNP (Chr:Pos)	Effect/ noneffect allele	INFO	MAF	Effect estimate	Standard error	P-value	Conditioned SNP(s)	Annotation(s)
rs1342062 (chr1:205912786)	G/T	0.9514	32.70%	0.0437	0.0050	2.33×10^{-18}	NA	*SLC26A9*, upstream_gene_variant
rs34712979 (chr4:106819053)	G/A	1.0000	25.85%	-0.0592	0.0052	1.04×10^{-29}	NA	*NPNT*, splice_region_variant,intron_variant
rs7658614 (chr4:145445694)	T/A	0.9987	46.81%	0.0707	0.0046	7.55×10^{-54}	NA	Intergenic, near *HHIP*
rs4466136 (chr5:82985576)	T/G	0.9962	21.94%	-0.0485	0.0055	2.05×10^{-18}	*NA*	*HAPLN1, intron_variant*
rs9273229 (chr6:32613914)	A/C	0.8656	36.45%	-0.0502	0.0051	8.46×10^{-23}†	rs532524051, rs538489083, rs560438058	*HLA-DQA1*, downstream_gene_variant
rs560438058 (chr6:32670158)	T/G	0.8307	5.48%	-0.1304	0.0113	1.22×10^{-30}†	rs538489083	Intergenic, near *HLA-DQB1, HLA-DQA2*
rs190516 (chr6:142813761)	T/C	0.9947	31.09%	0.0494	0.0050	2.29×10^{-23}	NA	Intergenic, near *GPR126*, *ADGRG6*
rs34249114 (chr8:22535398)	G/A	0.9461	6.01%	-0.1264	0.0100	1.01×10^{-36}	rs6981627	Intergenic, near *BIN3*, *EGR3*
rs1441358 (chr15:71612514)	T/G	1.0000	33.60%	-0.0468	0.0049	6.39×10^{-22}	NA	*THSD4*, intron_variant
rs11149827 (chr16:75435143)	A/G	0.9878	40.82%	-0.0380	0.0047	4.54×10^{-16}	NA	*CFDP1*, intron_variant

† P-value conditional on SNP(s) listed in Conditioned SNP(s) column.

Table 5-8: Analysis of FEF_{25-75}; 10 most strongly associated common SNPs (MAF≥5%).

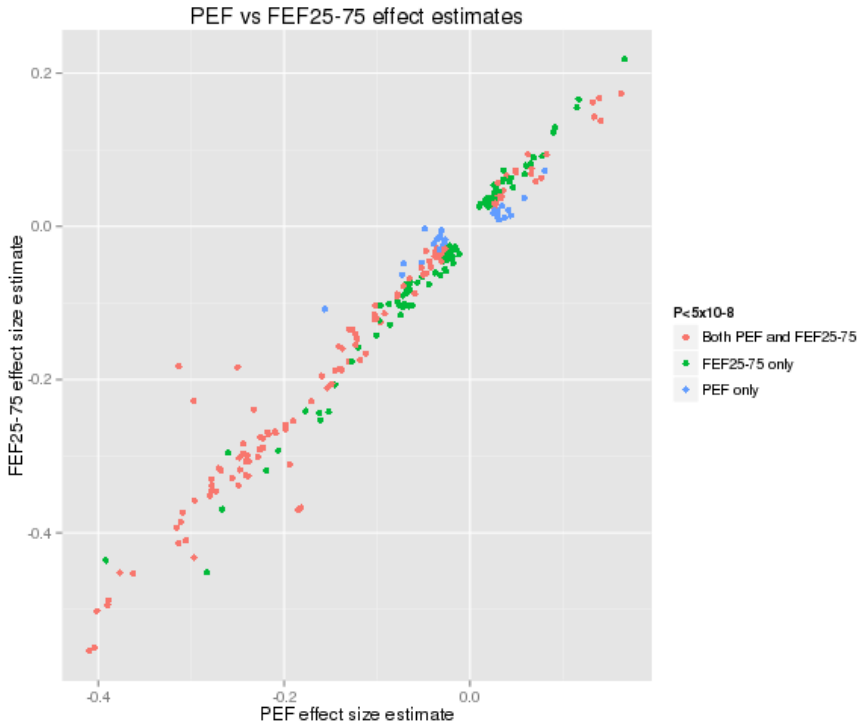
Chromosome (Chr) and position (Pos) in build 37 are given for each SNP. Effect estimates are on an inverse-normal transformed scale after adjustments for age, age^2, sex, height and ancestry principal components, and stratified by ever smoking status.

SNP (Chr:Pos)	Effect/ noneffect allele	INFO	MAF	Effect estimate	Standard error	P-value	Conditioned SNP(s)	Annotation(s)
rs73922761 (chr2:32816432)	T/C	0.889	5.06%	-0.1403	0.0115	$1.62x10^{-34}$†	rs143880252,rs139999372	*BIRC6*, intron_variant
rs34712979 (chr4:106819053)	G/A	1.000	25.85%	-0.0879	0.0053	$1.61x10^{-62}$	NA	*NPNT*, splice_region_variant,intron_variant
rs6817273 (chr4:145492003)	T/C	0.998	39.56%	0.0686	0.0047	$3.16x10^{-48}$	NA	Intergenic, near *HHIP*
rs7733410 (chr5:147856522)	G/A	1.000	44.04%	0.0571	0.0046	$8.49x10^{-35}$	NA	*HTR4*, intron_variant
rs9270377 (chr6:32558260)	G/T	0.729	44.95%	-0.0685	0.0054	$5.77x10^{-37}$†	rs149405105 ,rs532524051, rs538489083, rs560438058	Intergenic, near *HLA-DRB1*
rs560438058 (chr6:32670158)	T/G	0.831	5.48%	-0.1633	0.0114	$2.58x10^{-46}$†	rs538489083	Intergenic, near *HLA-DQB1, HLA-DQA2*
rs3748069 (chr6:142767633)	A/G	1.000	28.91%	0.0729	0.0051	$1.02x10^{-46}$	NA	Intergenic, near *GPR126*, *ADGRG6*
rs34249114 (chr8:22535398)	G/A	0.946	6.01%	-0.1503	0.0100	$1.25x10^{-50}$†	rs6981627	Intergenic, near *BIN3*, *EGR3*
rs2271804 (chr10:12252217)	G/A	0.996	47.03%	0.0462	0.0046	$1.73x10^{-23}$	NA	*CDC123*, intron_variant
rs1441358 (chr15:71612514)	T/G	1.000	33.60%	-0.0614	0.0049	$3.89x10^{-36}$	NA	*THSD4*, intron_variant

† P-value conditional on SNP(s) listed in Conditioned SNP(s) column.

There was a significant overlap in the SNPs and regions identified through the analyses of PEF and FEF_{25-75}. Of the 127 sentinel SNPs for PEF, 98 were significantly associated with FEF_{25-75} ($P<5x10^{-8}$). Similarly, of the 215 sentinel SNPs identified for FEF_{25-75}, 102 were significantly associated with PEF. The estimated effect sizes for the two traits were also highly correlated, as can be seen in Figure 5-13, which shows a comparison of the effect estimates for all SNPs identified as sentinels for one or both of PEF and FEF_{25-75}. This overlap in the signals identified for these two traits is unsurprising, given the correlation of these two flow measures of lung function, as was seen in Section 5.5.1 (inverse-normally transformed traits, $r^2=0.676$, Figure 5-6).

Figure 5-13: Comparison of estimated effect sizes for PEF and FEF_{25-75}.
All SNPs identified as sentinel SNPs for one or both of PEF and FEF_{25-75} shown.

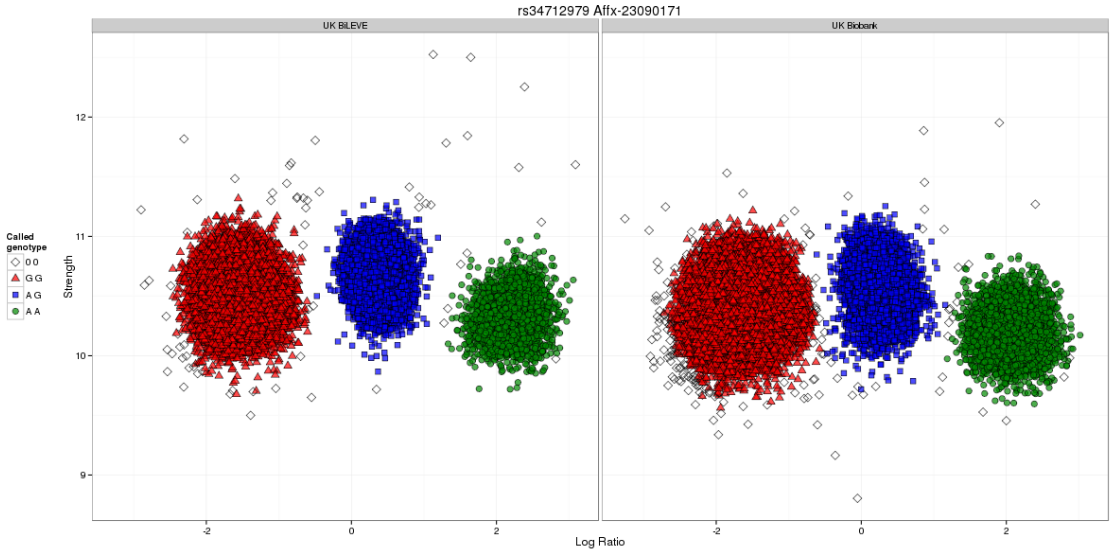


5.5.5 Quality Control of results

There are a number of steps which should be taken to limit the number of false positive associations identified in these analyses, which are outlined in this section. Firstly, clusterplots should be generated for all identified SNPs. Where SNPs have been directly genotyped, for example rs34712979 in *NPNT*, which was associated with both PEF and FEF_{25-75} (MAF=25.85%, $P_{PEF}=1.04x10^{-29}$, $P_{FEF25-75}=1.61x10^{-62}$), the clusterplot for that SNP can be inspected. Figure 5-14 shows two clusterplots for this SNP: the left

shows genotype clusters for samples genotyped using the UK BiLEVE array, whilst the right plot shows clustering for samples genotyped using the UK Biobank array. Genotype calling for samples genotyped on both arrays appears accurate in this example, so this signal is not an artefact of genotyping error, nor of poor genotype calling.

Figure 5-14: Clusterplot for rs34712979, identified as a sentinel in the analyses of PEF and FEF_{25-75}.
The left plot shows samples genotyped on the UK BiLEVE array; the right plot shows samples genotyped on the UK Biobank array.



Where sentinel SNPs have been imputed and not directly genotyped, clusterplots for SNPs which are in LD with the sentinel SNP and have been directly genotyped should be inspected, where possible. rs6817273 near *HHIP* (MAF=25.85%, $P_{FEF25-75}=3.16x10^{-48}$) was not directly genotyped, but appeared well imputed with an INFO score of 0.998. The region plot showing the association of this SNP with FEF_{25-75} shows that there are a number of nearby SNPs in high LD ($r^2>0.8$) with the sentinel SNP, which also show strong association with FEF_{25-75} (Figure 5-15). One of the SNPs in strong LD with the sentinel SNP was rs1980057 ($r^2=0.976$ with rs6817273), which had been directly genotyped; the clusterplots for this SNP (Figure 5-16) appear to show accurate clustering and genotype calling, providing supporting evidence that this does not represent a false positive association due to genotyping error.

Figure 5-15: Region plot for the association of rs6817273 and FEF_{25-75}.
The sentinel SNP in each region is highlighted in blue with the LD (r^2) of nearby SNPs to the sentinel indicated by colour. The fine scale recombination rate is shown in light blue.

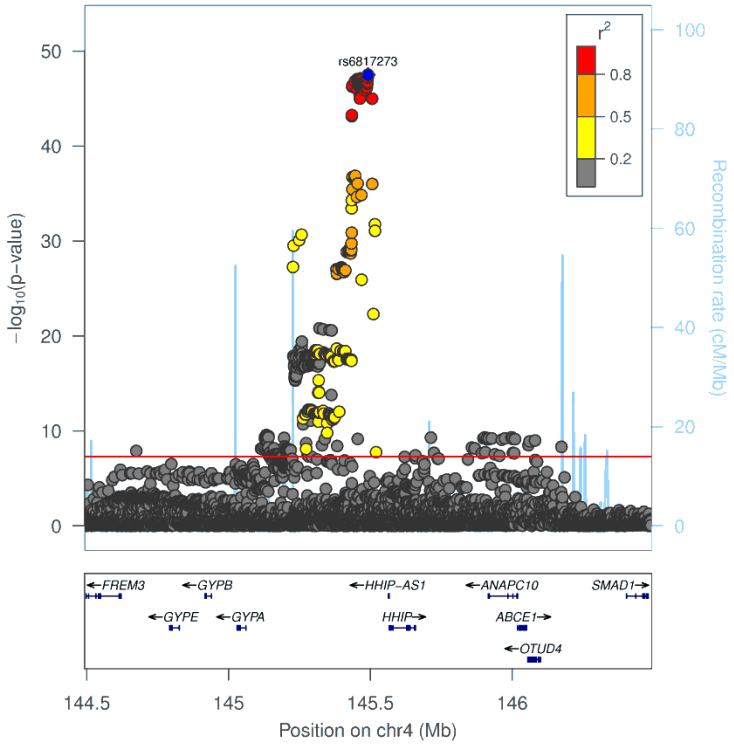
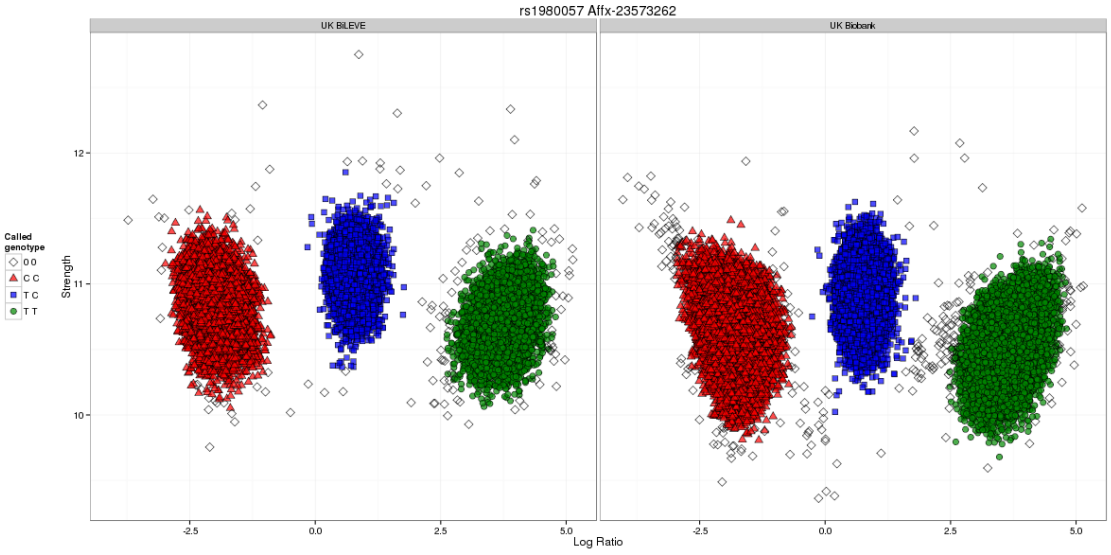


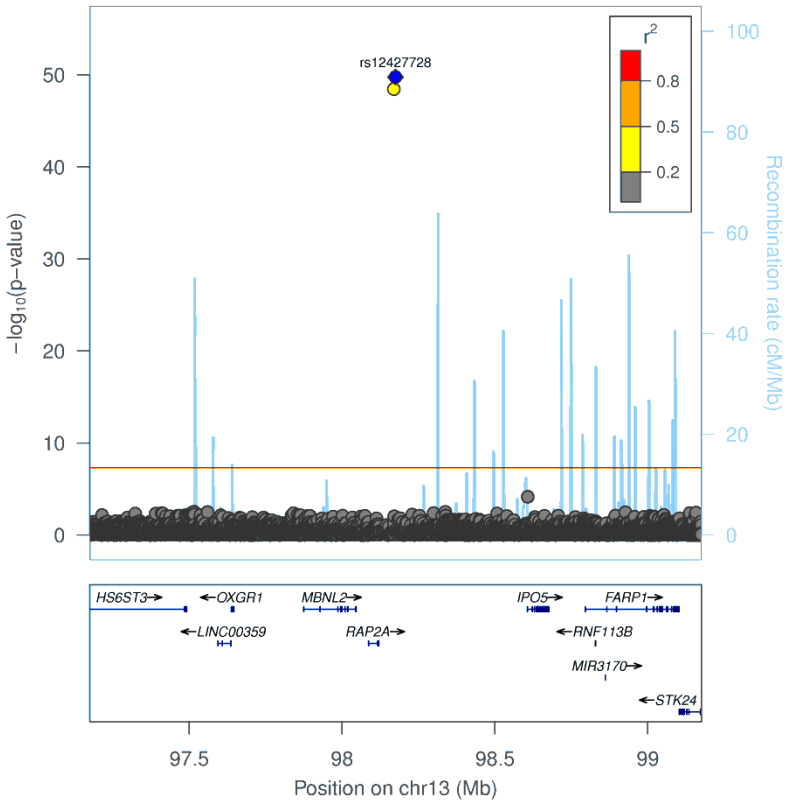
Figure 5-16: Clusterplot for rs1980057, in strong LD with the sentinel SNP rs6817273 (r^2=0.976) identified in the analysis of FEF_{25-75}.
The left plot shows samples genotyped on the UK BiLEVE array; the right plot shows samples genotyped on the UK Biobank array.



Region plots for some of the other sentinel SNPs show less supporting evidence of association from nearby SNPs in the region, such as intergenic SNP rs12427728

(MAF=1.05%, INFO=0.565 $P_{FEF25-75}$=1.74x10^{-50}, Figure 5-17). This SNP is not in LD with any genotyped SNP (LD between rs12427728 and all genotyped SNPs $r^2<10^{-7}$), and has an INFO score close to the exclusion threshold (INFO<0.5 for SNPs with MAF≥1%; INFO<0.8 for SNPs with MAF<1%), so is therefore more likely to be a false positive.

Figure 5-17: Region plot for the association of rs12427728 and FEF_{25-75}.
The sentinel SNP in each region is highlighted in blue with the LD (r2) of nearby SNPs to the sentinel indicated by colour. The fine scale recombination rate is shown in light blue.



In total, 20 (15.7%) SNPs associated with PEF were directly genotyped and 37 (29.1%) SNPs had a proxy (r^2≥0.1) that was genotyped and for which cluster plots could be inspected. For the remaining 70 (55.1%) SNPs there were only very weak genotyped proxies (r^2<0.1). For the FEF_{25-75} associated SNPs, 57 (26.5%) were directly genotyped, 59 (27.4%) had a genotyped proxy (r^2≥0.1) with the remaining 99 (46%) SNPs having only a very weak proxy that was genotyped. For both traits, it was mostly the low frequency and rare variants for which no good genotyped proxy was available (Table 5-9); the assessment of genotype calling errors for these SNPs in particular is challenging.

Table 5-9: Summary of the number of SNPs that were genotyped, or had a genotyped proxy, for which clusterplots could be generated to assess genotype calling.

SNPs categorised by MAF. Strength of proxies determined by linkage disequilibrium (r^2).

A. SNPs associated with PEF

	Directly Genotyped	Strong proxy ($r^2 \geq 0.5$)	Weak proxy ($0.5 > r^2 \geq 0.1$)	No good proxy ($r^2 < 0.1$)
MAF≥5%	6	26	10	10
5%≥MAF>1%	14	1	0	49
MAF<1%	0	0	0	11

B. SNPs associated with FEF_{25-75}

	Directly Genotyped	Strong proxy ($r^2 \geq 0.5$)	Weak proxy ($0.5 > r^2 \geq 0.1$)	No good proxy ($r^2 < 0.1$)
MAF≥5%	15	42	16	19
5%≥MAF>1%	41	1	0	54
MAF<1%	1	0	0	26

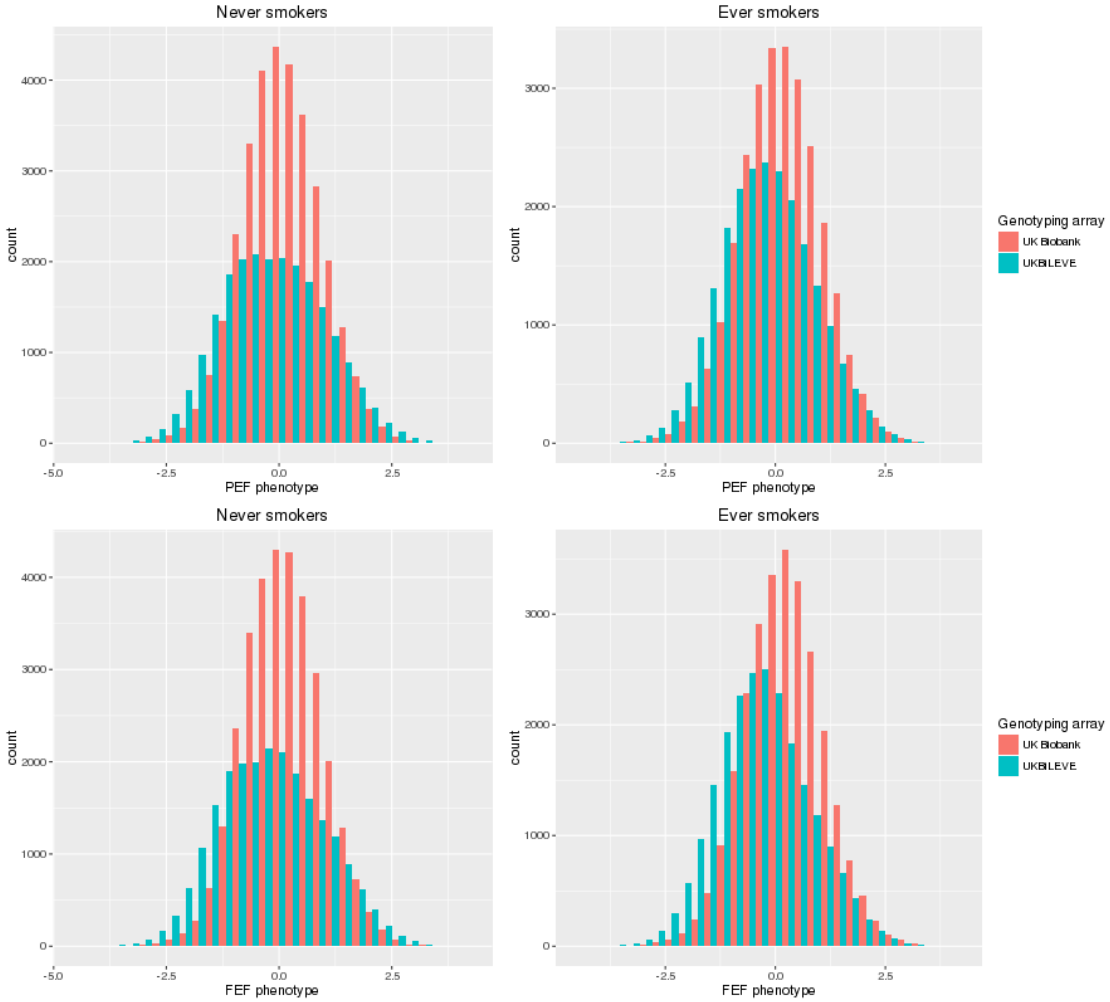
A further potential cause of false positive associations in the present analyses is from the use of two different genotyping arrays. To investigate this issue, analyses were undertaken in which associations between all identified sentinel SNPs and genotyping array were tested (UK BiLEVE array [n=44,289] versus UK Biobank array [n=58,640]). Table 5-10 summaries the numbers of sentinel SNPs (A. 127 SNPs associated with PEF and B. 215 SNPs associated with FEF_{25-75}) that were associated with genotyping array with $P<10^{-5}$. For common variants, reassuringly the majority of identified SNPs do not show significant associations with genotyping array. For identified SNPs with MAF<5% however, a large proportion was found to be associated with genotyping array. Some of the SNPs identified as showing associations with PEF or FEF_{25-75} in these analyses are therefore likely to be spurious associations due to an array effect. For example, rs1426311472 was identified in the GWAS of PEF ($P=1.14x10^{-27}$). The MAFs for this SNP amongst those samples genotyped on each array were markedly different (UK BiLEVE array MAF=0.02%; UK Biobank array MAF=5.9%); this large difference in MAF suggests that the observed association with PEF was a result of a bias due to genotype array effect.

Table 5-10: Summary of associations with genotyping array.
Results shown for A. 127 SNPs associated with PEF and B. 215 SNPs associated with FEF_{25-75}

	A. SNPs associated with PEF		**B. SNPs associated with FEF_{25-75}**	
	Associated with Chip ($P<10^{-5}$)		Associated with Chip ($P<10^{-5}$)	
	Yes	No	Yes	No
MAF≥5%	7	45	12	80
5%≥MAF>1%	60	4	91	5
MAF<1%	11	0	27	0

It may not be the case that all SNPs showing association with genotyping array are resulting in spurious associations with lung function however. The samples selected to be genotyped using the UK BiLEVE array were chosen as their lung function measures (percent predicted FEV_1) were at the extremes or the middle of the lung function distribution in UK Biobank. Consequently, there is a very strong association between genotyping array and lung function (association between genotyping array and the transformed PEF phenotype $P=2.76x10^{-220}$). Figure 5-18 shows the distribution of PEF and FEF_{25-75} phenotypes in never smoker and ever smokers, stratified by genotyping array, and it can be seen that individuals whose lung function is towards the extremes of the distribution are more likely to be genotyped on the UK BiLEVE array. In particular, if there exist low frequency variants with large effects, they are likely to be enriched in individuals at the extremes of the lung function distribution, therefore significant associations with genotyping array would be expected. rs7711789 was associated with PEF ($P=3.82x10^{-9}$), and also showed a significant association genotyping array ($P=1.61x10^{-78}$). There were differences in MAFs between those samples genotyped on the UK BiLEVE array (MAF=1.7%) and the UK Biobank array (MAF=1.0%); however it is plausible that this difference in MAFs is reflective of this SNP having a true effect on lung function.

Figure 5-18: Distribution of PEF and FEF_{25-75} phenotypes in never smoker and ever smokers, stratified by genotyping array.



For the purposes of this chapter, no exclusions have been made based on the inspection of clusterplots, nor due to array associations. The most convincing way to eliminate false positive findings is through following up all identified signals in an independent sample, to identify which SNP associations are replicated. At the time of writing, there was no available resource with sufficient sample size and phenotype data that would be suitable for the replication of the findings described in this chapter. However, the remaining samples in UK Biobank (n=201,415 with PEF and FEF_{25-75} phenotype data available) are currently being genotyped, and these samples would provide a suitable replication resource for future follow-up. Since the UK BiLEVE samples were selected from heavy smokers and never smokers in UK Biobank, a large proportion of individuals with extreme lung function values will have been selected for this analysis and will have been genotyped on the UK BiLEVE array. Individuals with extreme lung function values, but who did not meet either the never or heavy smoking

criteria will be amongst the remaining UK Biobank participants, currently being genotyped on the UK Biobank array. This subset of individuals in particular will be useful for following up potential associations with low frequency SNPs with large effects, and for distinguishing whether these signals represent true associations, or are due to a genotyping array effect.

5.5.6 Effect of identified SNPs on volumetric lung function traits

For all SNPs identified as showing association with PEF and/or FEF_{25-75}, analyses of three volumetric lung function traits (FEV_1, FVC and FEV_1/FVC) were undertaken. As was seen in Figure 5-6 and Figure 5-7, PEF and FEF_{25-75} and the three volumetric lung function traits were all moderately to strongly correlated. A comparison of effect sizes for all SNPs associated ($P<5x10^{-8}$) with A. PEF and B.FEF_{25-75} and their estimated effects on FEV_1, FVC, FEV_1/FVC can be seen in Figure 5-19. The effect estimates for both PEF and FEF_{25-75} and the three volumetric lung function traits are overall highly correlated, with the majority of identified SNPs also reaching genome-wide significance ($P<5x10^{-8}$) for one or more of the volumetric traits (Figure 5-19).

Figure 5-19: Comparison of effect estimates and P-values for SNPs associated ($P<5x10^{-8}$) with A. PEF and B. FEF_{25-75} and other lung function traits.

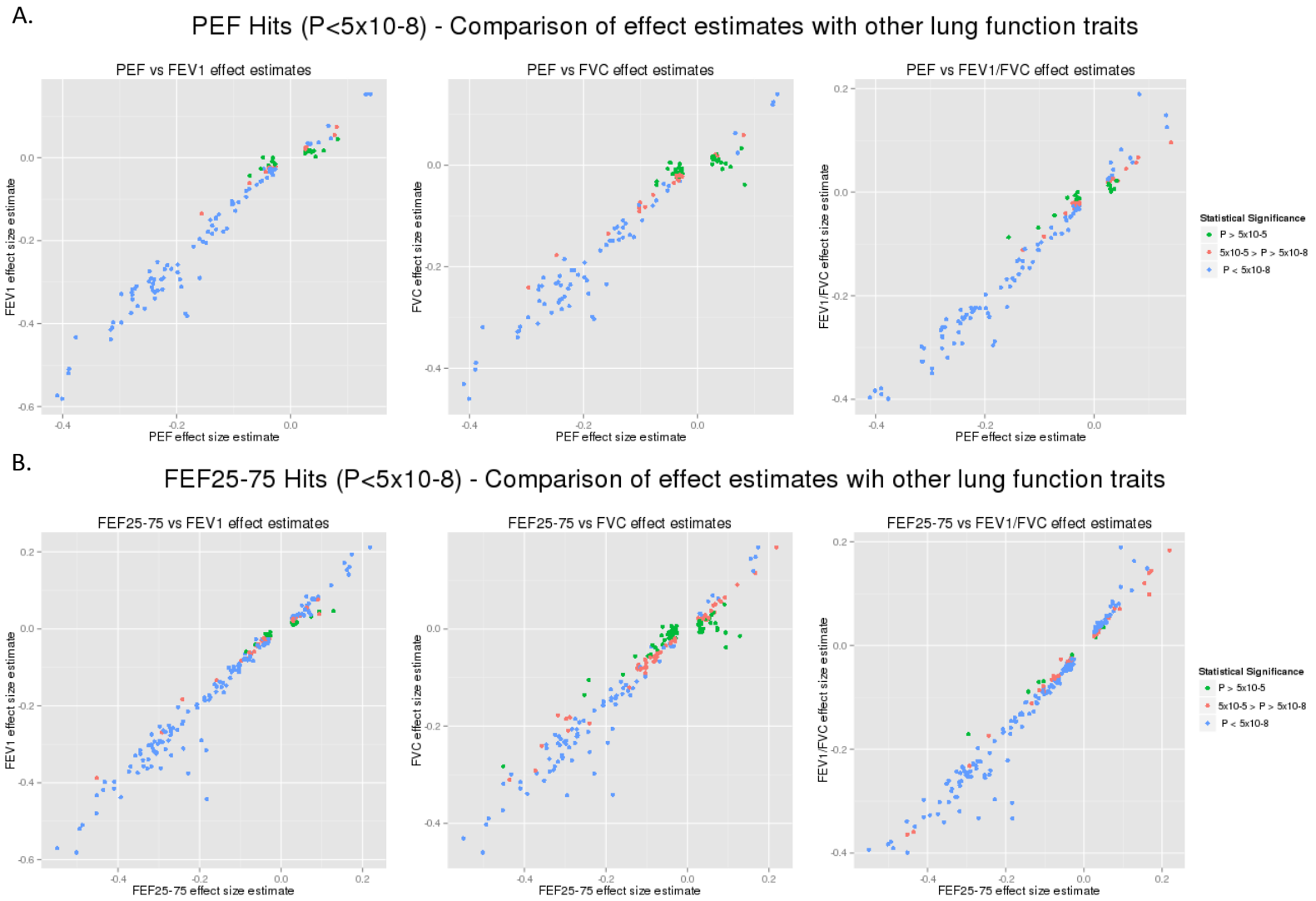
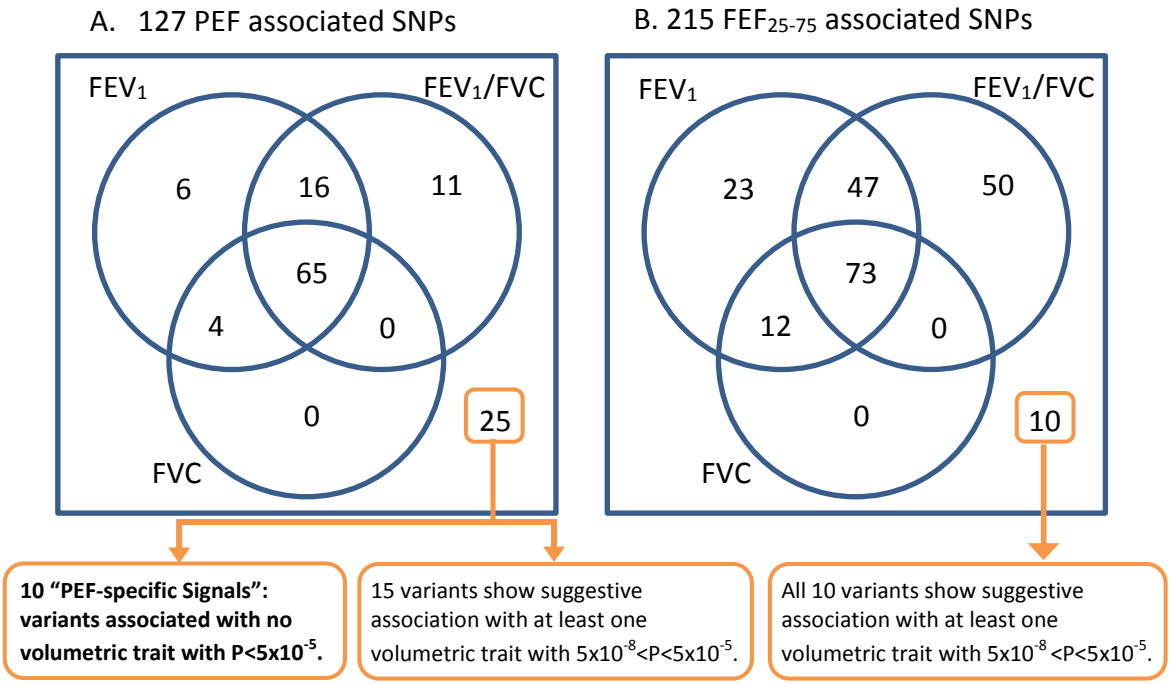


Figure 5-20 further shows the overlap of SNPs showing genome-wide significant associations with the three volumetric traits, for all SNPs identified as showing association with A. PEF and B. FEF_{25-75}.

Figure 5-20: Summary of overlap of traits for which SNPs show genome-wide significant associations ($P<5x10^{-8}$). Shown are the number of variants showing genome-wide significant association ($P<5x10^{-8}$) with each volumetric trait (FEV_1, FVC and FEV_1/FVC) for A. 127 SNPs associated with PEF and B. 215 SNPs associated with FEF_{25-75}.



Consistent with the patterns of correlation of the traits (Figure 5-7), the majority of SNPs associated with FEF_{25-75}, are also significantly associated with one or both of FEV_1 and FEV_1/FVC ($P<5x10^{-8}$). Furthermore, there are no SNPs which show association with FEF_{25-75} and FVC, that are not also associated with one of the other two traits. Ten FEF_{25-75} associated SNPs did not show genome-wide significant associations with any of the volumetric lung function traits; however all of those SNPs did show moderate association at a lower level of significance ($P<5x10^{-5}$) with at least one other trait.

For the PEF associated SNPs, again the majority showed genome-wide significant associations with at least one volumetric trait, in particular FEV_1 and FEV_1/FVC. There are 25 SNPs which were identified through the analysis of PEF which do not show genome-wide significant association with any of FEV_1, FVC or FEV_1/FVC; of those, 15

did show moderate association ($P<5x10^{-5}$), while for 10 SNPs, no associations reached this intermediate significance level for any other lung function trait.

The associations of these ten SNPs and PEF are summarised in Table 5-11 and region plots are shown in Figure 5-21. The most significant PEF specific association was with rs4466136, an intronic SNP in *HAPLN1* (MAF=21.9%, Beta=-0.0485, $P=2.05x10^{-18}$). This showed no statistically significant association with any of the volumetric lung function traits, all with P≥0.045. Other SNPs which showed a significant association with PEF and only very weak associations with other lung function traits (P≥0.01) were rs59538733 (MAF=30.7%, Beta=-0.0311, $P=3.98x10^{-8}$), an intergenic SNP on chromosome 1, upstream of *LAPTM5* and *MATN1*, and rs11111272 (MAF=28.5%, Beta=-0.0303, $P=2.17x10^{-9}$), an intronic SNP in *IGF1*. Other PEF-specific SNPs were located in, or near to *ARHGAP15, CYTL1, LOC105378963, MIR8056, UST, LYL1* and *TASP1*. All of these SNPs did show suggestive association with either FEV_1 or FEV_1/FVC however ($5x10^{-3}>P>5x10^{-5}$).

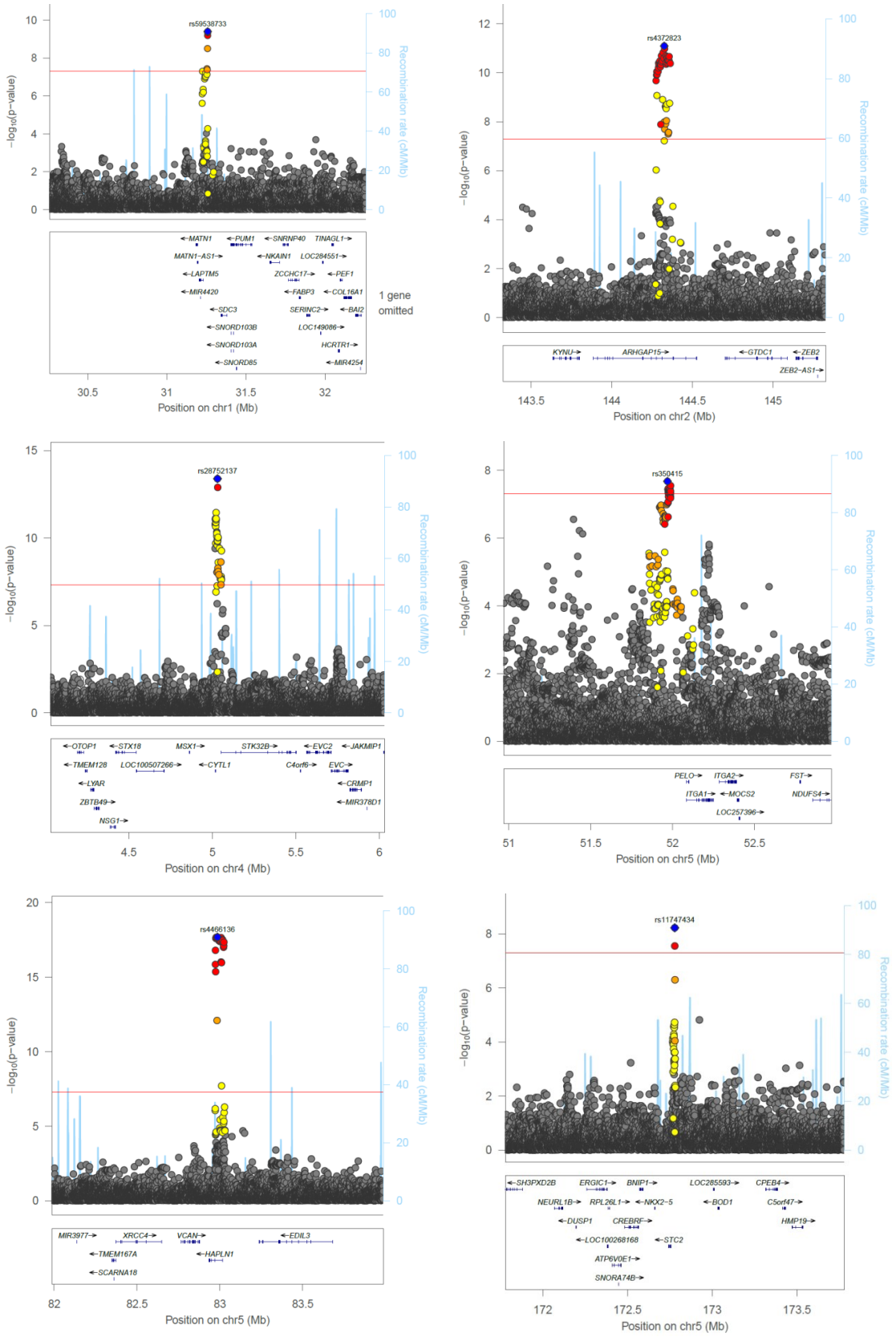
Table 5-11: Summary of SNPs associated with PEF and no other lung function trait ($P<5x10^{-5}$).

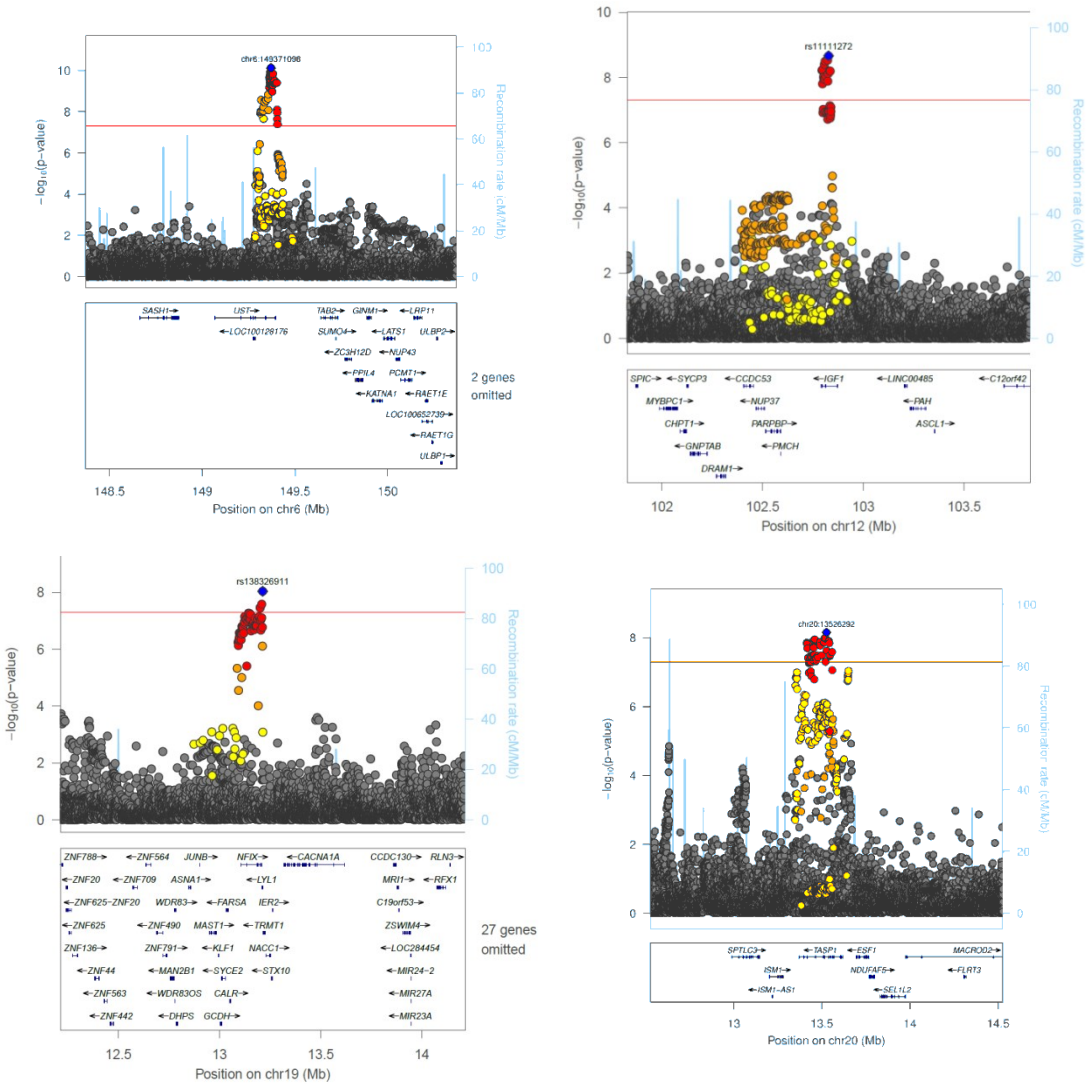
Chromosome (Chr) and position (Pos) in build 37 are given for each SNP. Effect estimates are on an inverse-normal transformed scale after adjustments for age, age^2, sex, height and ancestry principal components, and stratified by ever smoking status.

SNP (Chr:Pos)	Effect/ noneffect allele	INFO	MAF	Effect estimate	Standard error	P-value	Gene(s)	Most significant association with another lung function trait.	
								Trait	P-value
rs59538733 (chr1:31258698)	GC/G	0.9978	30.69%	-0.0311	0.0050	$3.98x10^{-10}$	near *LAPTM5, MATN1, SDC3*, intergenic	FEV_1/FVC	0.0619
rs4372823 (chr2:144325926)	A/G	0.9931	17.59%	0.0413	0.0060	$8.02x10^{-12}$	*ARHGAP15*, intron_variant	FEV_1/FVC	$1.88 x10^{-4}$
rs28752137 (chr4:5030854)	A/G	0.9895	32.57%	0.0370	0.0049	$4.17x10^{-14}$	near *CYTL1, STK32B*, intergenic	FEV_1	$4.39 x10^{-3}$
rs350415 (chr5:51968428)	A/G	0.9948	26.02%	0.0293	0.0052	$2.13x10^{-8}$	*LOC105378963,* downstream_gene_variant	FEV_1	$7.52x10^{-5}$
rs4466136 (chr5:82985576)	T/G	0.9962	21.94%	-0.0485	0.0055	$2.05x10^{-18}$	*HAPLN1*, intron_variant	FEV_1/FVC	0.0451
rs11747434 (chr5:172779211)	T/C	0.9778	27.68%	0.0301	0.0052	$5.93x10^{-9}$	*MIR8056*, downstream_gene_variant	FEV_1	$2.70 x10^{-3}$
rs574284527 (chr6:149371098)	C/CACAG	0.9497	33.67%	-0.0324	0.0050	$7.52x10^{-11}$	*UST*, intron_variant	FEV_1/FVC	$4.40x10^{-3}$
rs11111272 (chr12:102827441)	G/C	0.9961	28.53%	-0.0303	0.0051	$2.17x10^{-9}$	*IGF1*, intron_variant	FVC	0.0105
rs138326911 (chr19:13213277)	T/C	0.9380	3.74%	-0.0715	0.0124	$9.01x10^{-9}$	*LYL1,* intron_variant	FEV_1	$2.26x10^{-4}$
rs34590652 (chr20:13526292)	A/AT	0.9862	43.53%	0.0269	0.0046	$7.05x10^{-9}$	*TASP1*, intron_variant	FEV_1	$1.33x10^{-4}$

Figure 5-21: Region plots for 10 PEF-specific signals.

The sentinel SNP in each region is highlighted in blue with the LD (r^2) of nearby SNPs to the sentinel indicated by colour (red: $r^2>0.8$, orange: $0.8\geq r^2>0.5$, yellow: $0.5\geq r^2>0.2$, grey: $r2<0.2$). The fine scale recombination rate is shown in light blue.





All of these 10 PEF-specific signals appear to be well imputed (all have INFO>0.9) and the sentinel SNP for all but one signal was common. The genotyped proxy SNP in strongest LD with the sentinel was identified for each signal, and cluster plots inspected. For 6 sentinels, there was a genotyped SNP in strong or moderate LD (r^2>0.5) for which clusterplots could be generated. Two sentinel SNPs had weaker proxies ($0.2 \leq r^2 < 0.5$) and only very weak genotyped proxy SNPs ($r^2 \leq 0.01$) were available for the remaining 2 sentinels. The clusterplots for all of these genotyped proxies (Appendix F) showed good clustering and accurate genotype calling. Furthermore, none of these 10 SNPs showed significant associations with genotyping array (all P≥0.02). To fully verify these associations, replication analyses should be carried out using an independent set of samples.

5.6 Discussion

This chapter describes the largest GWAS of flow measures of lung function, PEF and FEF_{25-75}, carried out to date. These analyses were intended to determine the utility of studying these flow measures in addition to FEV_1, FVC and FEV_1/FVC, and to establish whether such analyses could potentially point to novel biological mechanisms underlying lung function. These analyses identified a large number of signals for each trait: 127 SNPs in 93 regions for PEF and 215 SNPs in 153 regions for FEF_{25-75}. Whilst all SNPs associated with FEF_{25-75} were also found to be associated with one or more of FEV_1, FVC and FEV_1/FVC, there were 10 SNPs which were associated with PEF and no other lung function trait ($P<5x10^{-5}$). Future replication in independent samples is required to fully verify all results.

Amongst the top signals identified for each trait were a number of SNPs in regions which had previously been implicated in the volumetric lung function traits, FEV_1 and FEV_1/FVC, or COPD and related traits; these included associations in, or near to *HHIP*, *NPNT* and *KANSL1.* The two most statistically significant associations for both PEF and FEF_{25-75} that were in regions not previously implicated in other lung function traits, were with SNPs in *CPNE8* and *CC2D2A. CPNE8* (Copine VII) is in a family of calcium-dependent membrane binding proteins, characterised as having two N-terminal C2 domains and an A domain at the C-terminus, which may be involved in membrane trafficking and cell signaling pathways (204). *CC2D2A* (coiled-coil and C2 domain containing 2A) encodes a coiled-coil and calcium binding domain protein which plays a critical role in cilia formation, with mutations within this gene known to cause Joubert and Meckel syndromes (205).

The sentinel SNPs in *CPNE8* and *CC2D2A,* were amongst 59% and 57% of the sentinel SNPs for PEF and FEF_{25-75} respectively, which were low frequency or rare (MAF<5%). The estimated effect sizes of low frequency and rare SNPs on PEF and FEF_{25-75} were generally seen to be greater than for SNPs with higher frequencies. Low frequency SNPs have been found to exhibit large effects on a number of other traits, including bone mineral density (206), insulin processing and secretion (57) haematological traits (59), as well as in other lung function traits (43). These low frequency and rare SNPs are likely to be imputed with less accuracy than common SNPs, so replication of the

low frequency signals in the present analysis is of particular importance to eliminate false positive findings. Many of the low frequency and rare SNPs in these analyses were additionally found to be significantly associated with genotyping array ($P<10^{-5}$). Whilst this suggests that some of the identified SNPs might spurious associations resulting from a genotyping array effect, this may not be the case for all SNPs as samples that were genotyped on the UK BiLEVE array were selected on the basis of their lung function. For example, a SNP that is associated with low values of PEF or FEF_{25-75} might also be significantly associated with genotyping array, as samples with poor lung function are more likely to genotyped on the UK BiLEVE array. In some instances, it might be possible to determine whether a signal is likely to be a result of a genotyping array effect, by examining each SNP individually (for example if the differences in MAFs between individuals genotyped on the two arrays was too extreme to be a result of differences in lung function, section 5.5.5), however replication will be the most convincing way to eliminate false positive associations overall.

In terms of common SNPs, the most statistically significant associations were generally within regions previously implicated in lung function. Associations in novel loci included common SNPs within, or close to *SLC26A9*, *HAPLN1*, *BIN3* and *BIRC6*. Of note, the solute carrier 26A9 (*SLC26A9*) gene functions as a chloride ion (Cl^-) channel and is highly expressed in the lung. Missense mutations within *SLC26A9* have been identified in individuals with diffuse bronchiectasis (207), and mice deficient in *SLC26A9* showed airway obstruction after IL-13 induced mucus overproduction, suggesting the *SLC26A9* Cl^- channel is activated in airway inflammation (208).

The present analyses also identified a number of regions where there were two or more independent signals. Recent studies of a number of traits, including body mass index (29), height (33), lipids and coronary artery disease (209) have also identified regions where there are multiple signals. These results highlight the utility of undertaking conditional analyses to identify secondary and potentially further signals in each identified genomic region. Although gene-based tests were not undertaken as part of this chapter, these too could provide an additional means for identifying regions with multiple SNPs influencing a trait.

Overall there was a large overlap in the regions that were significantly associated with both PEF and FEF_{25-75}. The majority of SNPs associated with these flow measures, were also found to be associated with at least one of the three volumetric lung function traits, in particular with FEV_1 and FEV_1/FVC. This is largely unsuprising, given that all of these traits are moderately to strongly correlated. One of the main aims of this chapter was to determine whether GWAS of PEF and FEF_{25-75} might identify signals in regions not associated with the three volumetric lung function measures, usually studied in GWAS of lung function. While all SNPs identified in the analysis of FEF_{25-75} showed at least modest association with one or more of FEV_1, FVC and FEV_1/FVC, there were ten SNPs which were identified as being associated with PEF, but not with any volumetric lung function trait with $P<5x10^{-5}$. In particular, associations with PEF were identified with SNPs in *HAPLN1*, *IGF1* and an intergenic SNP near to *LAPTM5* and *MATN1*; for these SNPs, all associations with other lung function traits had P≥0.01.

The Hyaluronan and proteoglycan link protein 1 (*HAPLN1*) gene assembles and stabilises cartilage proteoglycan aggregate in the cartilage extracellular matrix (210). Mice deficient in *HAPLN1* have been shown to have defective cartilage development and delayed bone formation (211) and the gene is also thought to play a role in cardiac development (212). The matrilin 1, cartilage matrix protein (*MATN1*) is another component of the extracellular matrix, containing two Willebrand Factor A domains. *MATN1* is associated with cartilage proteoglycans and binds to both collagen fibrils and noncollagenous proteins (213, 214).

Insulin-like growth factor 1 (*IGF1*) regulates several cell functions including proliferation, growth and apoptosis (215) and has been implicated in a wide range of traits and diseases, including anthropometric traits (216, 217), bone mineral density (218), breast cancer (219) and glycaemic traits (215, 220). *IGF1* has also been found to play a crucial role in lung development. One study showed that embryonic mice deficient in *IGF1* exhibited delayed distal lung organogenesis (221). Furthermore, *IGF1* has been found to be highly expressed in human lung endothelial cells in early gestation, suggesting a role in vascular development in the foetal lung (222).

Lysosomal protein transmembrane 5 (*LAPTM5*) encodes a transmembrane receptor that has a role in T and B cell activation and the regulation of inflammatory responses by macrophages (223). *LAPTM5* has been shown to display differential expression in foetal and adult lung and may be partially regulated by DNA methylation (224).

In summary, the analyses described in this chapter highlighted a modest number of associations with PEF and SNPs that were not seen with any of the volumetric traits, usually studied in GWAS of lung function. Although the associations identified in this chapter require replication in independent samples, they suggest there may be some utility in studying flow measures, particularly PEF, alongside volumetric measures in future studies. The benefit of studying these additional lung function phenotypes is the potential to discover loci which would not be identified through studying volumetric traits alone and which might reveal the influence of different biological pathways on lung health and disease. The disadvantage of studying multiple related traits is the increased burden of multiple testing, which would need to be taken into account. One way in which these flow measures could be incorporated into prospective analyses, without increasing the burden of multiple testing, is through the use of multivariate methods. A number of multivariate GWAS methods have been proposed for studying the genetic determinants of multiple correlated traits (225-229), which aim to maximise power, by making use of extra information from the between-trait covariance. Given the correlations between PEF, FEF_{25-75} and volumetric lung function traits, these multivariate methods might provide a promising approach for further unravelling the genetic architecture of lung function in future studies.

Chapter 6 Conclusions

The primary aim of the work presented in this thesis was to investigate the genetic basis of lung function and COPD, with a particular focus of exploring the effect of low frequency and rare genetic variants, so far largely overlooked in GWAS of these traits. To this end I have utilised recently developed genotyping chips, imputation panels and statistical methods throughout the thesis. This final chapter summarises the main findings of the thesis (Section 6.1), describes some of the challenges faced whilst undertaking this work (Section 6.2) and discusses ongoing and future work in the field (Section 6.3).

6.1 Summary of work

In Chapter 2, I describe an exome array analysis of COPD risk, and of airflow limitation in COPD cases. These analyses drew 3226 COPD cases from a number of sample collections, as part of the UK COPD exome chip Consortium, and 4784 controls from the wider UK exome chip Consortium, with follow-up using a further 13,210 samples from the UK BiLEVE study (120). I verified associations at a number of previously reported COPD loci, and identified novel associations between COPD risk and low frequency SNPs in *MOCS3* and *IFIT3*. Furthermore, a rare SNP in *SERPINA12* was identified as showing association with %predicted FEV_1 in COPD cases. None of these associations met a predefined exome-wide significance level ($P<3.7x10^{-7}$) however and further follow-up would be required to fully verify these associations. The results of these analyses were published in Thorax in 2016 (148).

In Chapter 3, I undertook an evaluation of methods developed for the meta-analysis of gene-based tests. This evaluation aimed to examine the concordance of the meta-analysis methods with equivalent analyses using individual level data in a dataset with real genome-wide genetic data and phenotypes, far larger than has been utilised for evaluating these methods to date. Using the RAREMETAL (74) package and data from the UK BiLEVE study (120), I found the meta-analysis methods to be approximately equivalent to a mega-analysis using individual level data, for a quantitative trait, and a binary trait with a balanced ratio of cases and controls. I then applied these methods in

a meta-analysis of exome array data and three quantitative lung function measures: FEV_1, FVC and FEV_1/FVC, as described in Chapter 4. These analyses were undertaken in 23,398 individuals from 11 studies from the SpiroMeta Consortium, with follow-up in up to 93,390 independent samples. Through these analyses, I identified an association with a SNP near *LY86* and FEV_1/FVC and with a SNP near *FGF10* and FVC, in ever smokers; both of these associations were replicated in the follow-up samples. *LY86* interacts with the Toll-like receptor signalling pathway and may be involved in immunity and inflammation whilst *FGF10* plays a role embryonic development, growth and repair.

Finally, in Chapter 5 I carried out the largest GWAS to date of two flow lung function measures (PEF and FEF_{25-75}) and genotype data imputed to a combined 1000 Genomes (16) and UK10K (40) imputation panel in 102,929 samples from UK Biobank. These analyses identified a substantial number of SNPs and short indels associated with both PEF and FEF_{25-75}, a large proportion of which were low frequency or rare. Most of the SNPs associated with one or both of PEF and FEF_{25-75} were also found to be associated with at least one volumetric measure of lung function (FEV_1, FVC and FEV_1/FVC). However there were 10 SNPs which were identified as showing association with PEF ($P<5x10^{-8}$), but no other lung function trait with $P<5x10^{-5}$, suggesting there may be some utility in studying these flow measures, in addition to the three volumetric traits usually considered in studies of lung function. All of the associations identified in this chapter require replication in independent samples.

6.2 Challenges and limitations

One of the main challenges in GWA studies, particularly where low frequency variants are a focus, is obtaining large enough samples to have the statistical power to detect associations. In order to achieve such large samples, the analyses described in two of the chapters in this thesis utilised data from various sample collections, either by pooling individual level data (as was done in Chapter 2), or through carrying out a meta-analysis of summary level data (as in Chapter 4). Combining data in this way posed several challenges.

In Chapter 2, I used individual level data from several sample collections. Before undertaking the analyses, I carried out extensive quality control of both phenotype and genotype data. Firstly I had to combine the phenotype data from each sample collection, checking that all phenotype data were consistent. For example, I ensured that all samples had their lung function and height measurements recorded in the same units. I also had to ensure that all samples met the appropriate case or control definition. Where there were seemingly erroneous or missing data, I had to contact the affected study to attempt to resolve these issues. There were also a number of issues regarding the genotype data (Section 2.2.2). The cases and controls were genotyped independently and as a result, there were several SNPs which were subject to genotype calling errors in cases, but which were called correctly in controls. Another issue arose for some very rare SNPs which were subject to batch effects by case collection. A major challenge of the work described in Chapter 2 was identifying and then rectifying each of these issues, in order to eliminate false positive findings.

For the analyses described in Chapter 4, I did not have access to individual level data. Instead I utilised summary data from a number of studies, which posed a different set of difficulties. To ensure comparable analyses were undertaken in each study, I developed detailed analysis plans for study analysts to follow; however there were still a number of issues identified, either by the study analysts, or by myself when undertaking QC of the summary level data (Section 4.2.7). Any heterogeneity in the study level analyses, or programming errors could lead to erroneous results, so great effort was made to detect such errors or inconsistencies. For one study (1958BC), I did have the individual level data which allowed me to pilot the analysis plans, prior to circulation. Having access to the individual level data for this study also assisted with resolving queries and issues faced by other study analysts.

For the analyses undertaken in Chapter 5, I did have available the individual level data for all samples, which were all from a single large sample collection. As a result, there were fewer problems in terms of inconsistent data, batch effects and heterogeneity, as there were in the analyses described in Chapters 2 and 4. These data were not without issue however, for example I identified that a number of samples had erroneous PEF measurements recorded (Section 5.2.2), a problem I was able to resolve using the blow

data points. This issue has also been relayed to UK Biobank for the benefit of future scientific projects. This demonstrates the importance of thorough QC of all genotype and phenotype data, which has made up a substantial part of this thesis.

The analyses described in Chapter 5 utilised imputed genotype data requiring further QC considerations, which were not applicable to the preceding chapters. Since there were no available replication samples for these analyses, I adopted strict filters for imputation quality (Section 5.3.1). These filters are fairly conservative, particularly for the lower frequency SNPs, in comparison to other GWAS using UK10K imputed data (120, 230, 231) and it is likely that these strict filters will have removed some true associations. Limiting false positive associations in these analyses was not a straightforward process, with post-association QC of identified SNPs being more complicated than for data where SNPs are directly genotyped. For example, I described how it was not always possible to generate clusterplots to check genotype calling for the identified SNP, nor for a good proxy which was directly genotyped. In particular a large proportion of the low frequency and rare SNPs did not have a good genotyped proxy and so for these SNPs, the assessment of genotype calling is especially challenging and the likelihood of them representing false positive associations is higher. Another issue specific to these analyses was that of genotyping array effects. Usually where multiple arrays are used in a study, this can be a source of spurious associations; however in the present analyses, the genotype array and lung function were not independent, therefore it might be expected that SNPs showing association with PEF or FEF_{25-75} might also show association with genotyping array. Indeed many of the SNPs identified through these analyses, did show association with genotyping array and this was particularly true for low frequency and rare SNPs. Each SNP would need to be individually examined in attempt to determine whether the signal is a true association or is a result of an array effect. For all associations identified in these analyses, replication in independent samples would be required to fully eliminate false positive findings. For the purposes of the chapter I generated clusterplots for all identified SNPs where this was possible, however due to time constraints, I did not identify SNPs for exclusion based on poor clusterplots, nor were any exclusions made due to array effects. If these associations are to be followed up in the remainder of UK

Biobank, these exclusions should first be made where possible, to reduce the number of SNPs included in the replication analyses, thereby limiting the burden of multiple testing.

The analyses described in Chapters 2 and 4 utilise genotype data generated using the exome array. This array was primarily designed to enable the identification of trait associations with low frequency and rare variants, largely in coding regions, in large sample sizes. Overall the analyses in these chapters had limited success in identifying such associations. In Chapter 2, associations with COPD risk and two low frequency variants were identified, along with an association between a rare variant and percent predicted FEV_1 in COPD cases. None of these associations met "exome-wide" significance however and so require additional verification. The analyses in Chapter 4 identified genome-wide significant associations with two common intergenic SNPs only. There were several low frequency and rare SNPs identified in the discovery stage of these analyses, however none of these were replicated in the follow-up analyses, and indeed for some SNPs replication was not possible as SNPs were unavailable, or monomorphic in the follow-up samples. Studies using exome array data in other traits have also had mixed success in identifying associations with low frequency variants; whilst associations with low frequency variants have been identified for a number of traits including asthma (54), lipids (58), glycaemic (57) and haematological traits (59) and coronary heart disease (232), other exome array studies identified only common, or even no novel statistically significant associations (61, 233-235).

The analyses described in Chapter 5 utilised custom arrays which included: (i) content similar to that of the exome array; (ii) a GWAS grid which allowed for imputation of low frequency variants using the combined 1000 Genomes and UK10K reference panels. This analysis was the largest GWAS of lung function flow measures, and indeed of any measure of lung function, carried out to date and a large number of SNPs with low frequency (MAF<5%) were identified, with larger estimated effect sizes than identified common variants. These results suggest there may be low frequency variants influencing lung function traits, which are revealed when analyses are undertaken with very large numbers of samples (over 100,000 samples in these analyses). However low frequency variants are more likely than common variants to be

false positives due to imputation error, so it is likely that the results in this chapter provide an overestimation of the contribution of low frequency variants with large effects on lung function. Indeed without undertaking replication, potentially preceded by further QC, it is not possible to determine how many of the low frequency signals represent true associations.

As there has been a greater focus on investigating variation at the lower end of the frequency spectrum, there have been a number of developments in statistical methodology for the study of rare variation, in particular with regards to gene-based tests of association. I have utilised some of the most widely used tests (SKAT, WST and SKAT-O) in this thesis. These gene-based tests were intended to identify associations with regions of the genome which would not be detected through single variant association. Many of the genes identified through the gene based tests carried out as part of Chapters 2 and 4 in this thesis were also identified through the single variant association analyses. In each case, the gene-based association appeared to be driven by the single SNP identified in the single variant association analyses. In Chapter 4, there were several genes identified through the gene-based analyses, which did not appear to be driven by a single SNP, however replication of these gene-based signals was found to be difficult. In exome array studies of other traits, there have been some successes of gene-based tests identifying associations with genes that were apparently driven by several rare variants (55, 56, 59).

6.3 Ongoing developments in respiratory genetics and future work

Running concurrently to the work in this thesis were a number of other studies investigating the genetic basis of lung function and lung disease. As mentioned in Section 1.3.4, the SpiroMeta 1000 Genomes (43) and UK BiLEVE (120) efforts utilised genotype data imputed to either the 1000 Genomes only, or 1000 Genomes and UK10K combined reference panels. These analyses had discovery sample sizes of 38,199 (SpiroMeta 1000 Genomes) 48,943 (UK BiLEVE) individuals and between them identified 22 novel lung function signals (associated with FEV_1 FVC or FEV_1/FVC), of which two were low frequency (1%<MAF<5%) and one was rare (MAF<1%). The

CHARGE Consortium has also undertaken a large meta-analysis of exome array data with 44,719 samples. In Chapter 4, a look-up of the results in this analysis was undertaken as part of the replication stage analyses. The SpiroMeta and CHARGE exome array analyses are now being amalgamated into a larger combined analysis, which I am leading and will shortly be writing up for publication.

A further analysis running concurrently with the work in this thesis was a recently published meta-analysis of exome array data and COPD (cases defined as having airflow limitation consistent with GOLD2 or worse) which identified a novel genome-wide significant association between COPD and a common SNP in *IL27* (236). Additionally, a GWAS of post-bronchodilator spirometry in COPD cases identified an association between FEV_1 and a common SNP in *DBH* (237), a region not previously implicated in lung function, nor COPD; however this gene has previously been associated with smoking cessation (132).

Many of the GWAS undertaken for COPD to date are based on airflow limitation, measured by spirometry, largely as this provides an easily measured and objective criterion. The GOLD guidelines however recommend that the impact of COPD on an individual should take into account symptoms and risk of exacerbation, alongside the severity of airflow limitation (94). More recently, studies have been undertaken with more detailed clinical measures and imaging which aim to provide some insight into other aspects of COPD. GWAS studies of emphysema related traits, assessed using computed tomography (CT) identified several genes which were within or near to established lung function loci (*HHIP, AGER, CCDC38, TGFB2, MMP12*) and the 15q25 region associated with smoking, as well as genome-wide significant associations in loci not previously implicated in COPD or related traits (*SERPINA10*, *DLC1, MAN2B1*, *DHX15*, *MGAT5B*, *MAN1C1, VWA8, MYO1D*) (202, 238, 239). These studies all utilised samples from multiple ancestries and the associations in novel regions reached genome-wide significance in the discovery samples alone; however these signals have not yet been replicated in independent samples. A further analysis of COPD cases with chronic bronchitis versus smoking controls with normal spirometry, identified associations with the known lung function gene *FAM13A* and a genome-wide significant association in a novel COPD region, near to *EFCAB4A* and *CHID1* (201).

As described in Chapter 1 of this thesis, genome-wide significant associated variants identified to date collectively only explain small amounts of the expected heritability of the majority of complex traits, including lung function and other respiratory traits. A main aim of this thesis was to investigate whether low frequency and rare variants might explain some of the missing heritability of lung function and COPD. Through the analyses described in this thesis, a number of such variants have been identified as showing associations with COPD (Chapter 2) or quantitative lung function traits (Chapters 4 and 5). In Chapters 4 and 5 in particular, the majority of the variants identified in the discovery analyses had MAF<5%. Where follow-up analyses were undertaken in Chapter 4, none of these low frequency SNP associations were convincingly replicated. However, these findings could still be consistent with rare variants explaining some of the missing heritability of complex traits. Larger sample sizes than those described here are likely required in order to have enough statistical power to detect associations with variants at the lower end of the frequency spectrum. Genotyping of large population based biobanks, such as the UK Biobank and the Kadoori Biobank in China are currently underway and these could provide valuable data resources in which trait associations with low frequency variants can be successfully identified. Other study designs have also been proposed to increase power to detect low frequency associations, such as studies in population isolates (240, 241), family studies with multiple affected members (240), and extreme trait sampling (162, 240). In terms of lung function, the last of these study designs has been adopted in the UK BiLEVE study; as mentioned previously, this analysis identified an association with one rare variant and the extremes of FEV_1, alongside associations with 5 novel common variants (120).

Recent studies of complex traits including height (35), body mass index (BMI) (29, 35) and schizophrenia (31) have estimated that much of the heritability of these traits is a result of the polygenic effect of many common SNPs with small effects, which have not been identified as genome-wide significant associations; however there still remains a small proportion of the heritability that is unexplained. Other potential contributors to heritability include structural variation, epigenetics, epistatic effects, and gene-environment interactions (36). There have been a limited number of gene-

environment interactions influencing lung function identified to date. A genome-wide interaction study (GWIS) of FEV_1 and FEV_1/FVC and smoking identified signals in 3 regions which had not at that time been implicated in lung function (*DNER, HLA-DQ, KCNJ2*); however these associations were largely driven by the SNP main effects (242). More recently, another GWIS of occupational exposure to dust, gases and fumes identified associations with FEV_1 in novel lung function regions (*ZMAT4, PDE4D, ODZ2*) (243). I am currently involved in an analysis of gene-smoking interactions and lung function, focussing on low frequency variants, using gene-based tests, which is being undertaken within the CHARGE and SpiroMeta consortia.

GWAS are continuing to have great success in identifying regions of the genome which are associated with respiratory traits; however the mechanisms through which these loci influence lung function and susceptibility to disease remain largely unknown. It is often the case that the SNP identified through GWAS showing the most statistically significant association does not have a clear biological function on the trait itself, rather it is in LD with a causal variant (244). The identification of a causal variant can highlight molecular mechanisms and provide insight into the pathophysiology of a trait (245). Consequently, there has recently been a greater focus in GWAS on inferring putative causal variants and genes through fine-mapping. A number of programs such as CAVIAR (246), PAINTOR (247) and PICs (248) have been recently developed for identifying credible sets of causal variants using Bayesian methods for statistical fine-mapping. Trans-ethnic fine-mapping methods such as MANTRA (249) have also been proposed. These methods utilise differences in LD structures between different ancestral populations to further restrict credible sets of SNPs to those which are in LD with the causal variant across all populations. Functional annotations can also be used to prioritise variants, in particular genome annotation projects such as ENCODE (250), NIH Roadmap Epigenomics Mapping Consortium (251), and FANTOM5 (252) can help to refine signals in non-coding regions which might be affecting regulation of genes and influencing expression levels. Combinations of statistical and functional methods of fine-mapping have been used in recent studies to refine loci for a number of traits (248, 253-255). The utilisation of these methods for fine-mapping and functional

characterisation of loci is likely to be a focus in ongoing studies of lung function and COPD.

The translation of genetic findings has been a slow process; however examples of GWAS discoveries providing insight into the biological mechanisms of complex traits are now emerging. There are a number of examples of loci identified through GWAS, which themselves exhibit small effects, but for which the genes are already proven to be effective drug targets, such as *HMGCR* (256) and *PCSK9* for LDL cholesterol, *PPARG* and *KCNJ11* for type 2 diabetes (257) and several rheumatoid arthritis genes (253). Recently it has been proposed that the success rate in drug development could potentially be doubled by selecting targets with genetic supporting evidence (256). For most loci discovered through GWAS, the functional mechanism is less apparent; however there are instances of some identified loci leading to novel biological insight. Recent examples include the obesity gene *FTO* which has highlighted a pathway of adipocyte thermogenesis regulation in obesity (258), and an association with schizophrenia and the MHC locus, which has led to the suggestion that excessive complement component 4 (C4) activity could lead to an increased risk of schizophrenia (259). These examples highlight the potential value of GWAS findings and demonstrate the utility of continuing to investigate the genetic determinants of disease traits.

Uncovering the role of low frequency and rare genetic variants on lung function and COPD has proved challenging, and similarly to other complex traits (61, 233-235), very few associations with low frequency SNPs have been identified to date. It is apparent that very large sample sizes are required to evaluate rare variation; such sample sizes are now becoming attainable, for example through the current genotyping efforts of large biobanks. It is likely that a very large number of both common and rare variants, in combination with environmental factors are contributing to these traits and future studies should aim to interrogate genetic variation across the full frequency spectrum. Continuing to uncover the genetic architecture of lung function and COPD has the potential to give insight into the biological mechanisms underlying lung health and disease, and could lead to the future development of preventative and therapeutic interventions.

Appendices

A. Publication resulting from the analyses described in the association of rare variants with COPD risk and airflow limitation (Chapter 2).

Thorax Online First, published on February 25, 2016 as 10.1136/thoraxjnl-2015-207876

OPEN ACCESS

ORIGINAL ARTICLE

Exome-wide analysis of rare coding variation identifies novel associations with COPD and airflow limitation in *MOCS3*, *IFIT3* and *SERPINA12*

Victoria E Jackson,[1] Ioanna Ntalla,[1,2] Ian Sayers,[3] Richard Morris,[4,5] Peter Whincup,[6] Juan-Pablo Casas,[7,8] Antoinette Amuzu,[9] Minkyoung Choi,[9] Caroline Dale,[9] Meena Kumari,[10,11] Jorgen Engmann,[12] Noor Kalsheker,[13] Sally Chappell,[13] Tamar Guetta-Baranes,[13] Tricia M McKeever,[14] Colin N A Palmer,[15] Roger Tavendale,[15] John W Holloway,[16,17] Avan A Sayer,[18,19] Elaine M Dennison,[18,20] Cyrus Cooper,[18,19] Mona Bafadhel,[21] Bethan Barker,[22,23] Chris Brightling,[22,23] Charlotte E Bolton,[24] Michelle E John,[24] Stuart G Parker,[25] Miriam F Moffat,[26] Andrew J Wardlaw,[22,23] Martin J Connolly,[27] David J Porteous,[28] Blair H Smith,[29] Sandosh Padmanabhan,[30] Lynne Hocking,[31] Kathleen E Stirrups,[2,32] Panos Deloukas,[2,33] David P Strachan,[6] Ian P Hall,[3] Martin D Tobin,[1,23] Louise V Wain[1]

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/thoraxjnl-2015-207876).

For numbered affiliations see end of article.

Correspondence to
Victoria E Jackson, Department of Health Sciences, University of Leicester, University Road, Leicester LE1 7RH, UK; vej3@le.ac.uk

Received 24 September 2015
Revised 5 January 2016
Accepted 29 January 2016

To cite: Jackson VE, Ntalla I, Sayers I, *et al*. *Thorax* Published Online First: [*please include* Day Month Year] doi:10.1136/thoraxjnl-2015-207876

ABSTRACT

Background Several regions of the genome have shown to be associated with COPD in genome-wide association studies of common variants.

Objective To determine rare and potentially functional single nucleotide polymorphisms (SNPs) associated with the risk of COPD and severity of airflow limitation.

Methods 3226 current or former smokers of European ancestry with lung function measures indicative of Global Initiative for Chronic Obstructive Lung Disease (GOLD) 2 COPD or worse were genotyped using an exome array. An analysis of risk of COPD was carried out using ever smoking controls (n=4784). Associations with % predicted FEV_1 were tested in cases. We followed-up signals of interest ($p<10^{-5}$) in independent samples from a subset of the UK Biobank population and also undertook a more powerful discovery study by meta-analysing the exome array data and UK Biobank data for variants represented on both arrays.

Results Among the associated variants were two in regions previously unreported for COPD; a low frequency non-synonymous SNP in *MOCS3* (rs7269297, $p_{discovery}=3.08\times10^{-6}$, $p_{replication}=0.019$) and a rare SNP in *IFIT3*, which emerged in the meta-analysis (rs140549288, $p_{meta}=8.56\times10^{-6}$). In the meta-analysis of % predicted FEV_1 in cases, the strongest association was shown for a splice variant in a previously unreported region, *SERPINA12* (rs140198372, $p_{meta}=5.72\times10^{-6}$). We also confirmed previously reported associations with COPD risk at *MMP12, HHIP, GPR126* and *CHRNA5*. No associations in novel regions reached a stringent exome-wide significance threshold ($p<3.7\times10^{-7}$).

Conclusions This study identified several associations with the risk of COPD and severity of airflow limitation, including novel regions *MOCS3*, *IFIT3* and *SERPINA12*, which warrant further study.

Key messages

What is the key question?
- Do low frequency exonic variants influence susceptibility to COPD, and severity of airflow limitation?

What is the bottom line?
- Low frequency single nucleotide polymorphisms (SNPs) in *MOCS3 and IFIT3* were associated with risk of COPD and a rare splice variant in *SERPINA12* was associated with severity of airflow limitation.

Why read on?
- These genomic regions have not previously been implicated in lung function or COPD and these findings could therefore provide further insight into COPD susceptibility and severity.

INTRODUCTION

COPD is a major public health concern, being a leading cause of morbidity and mortality worldwide.[1] The Global Initiative for Chronic Obstructive Lung Disease (GOLD) recommends that the impact of COPD on an individual patient should assessed by considering breathlessness, symptoms and exacerbation risk, in combination with the severity of airflow limitation, which can be graded using %predicted FEV_1.[2] Approximately 1%–2% of COPD cases can be attributed to α1-antitrypsin (AAT) deficiency, a rare inherited disorder, caused by mutations within the *SERPINA1* gene.[3 4] For the remainder of COPD cases, cigarette smoking is recognised as the most

significant risk factor[5]; however, there is also a genetic component, with several genomic regions showing association with COPD risk or airflow limitation to date, including *CHRNA3/5*, *HHIP*,[3] *HTR4, GSTCD, TNS1*,[6] *MMP12*[7 8] and *FAM13A*.[9] COPD diagnosis is confirmed using measures of lung function, so it is likely that the genetic determinants of COPD and lung function will overlap. Indeed, many loci identified in large genome-wide association studies (GWAS) of FEV_1 and the ratio of FEV_1 to forced vital capacity (FEV_1/FVC) in general population samples[10–13] have subsequently being shown to be associated with COPD or airflow limitation.[6 9 14 15]

Despite the successes in identifying genes associated with lung function and COPD, these known loci only explain a small proportion of the expected heritability.[13] Large GWAS undertaken to date have generally focused on common variants (typically >5% minor allele frequency (MAF))[3 9–14]; one hypothesis is that some of the so-called 'missing heritability' might be accounted for by variants of lower frequencies. In this study, we set out to investigate the role of low frequency, functional variants in COPD, and to confirm the role of single nucleotide polymorphisms (SNPs) previously showing association with lung function. It is hypothesised that rare variants are more likely than common variants to have deleterious effects; identifying such SNPs could lead to greater understanding of the pathways and biological mechanisms underlying airflow obstruction and COPD, and could translate to novel targets for treatment.

We genotyped cases with a history of smoking and airflow limitation, indicative of GOLD 2 COPD or worse, and control samples using an exome chip array to which we had added custom content comprising 2585 SNPs tagging regions which had shown suggestive association ($p<2.21\times10^{-3}$) with lung function in a previous large genome-wide HapMap-imputed study.[13] The exome chip genotyping array design contains mostly non-synonymous, splice or stop codon altering variants that are likely to affect protein structure and function, with the majority of variants being low frequency (MAF 1%–5%) or rare (MAF<1%).

In this study, we carried out discovery case–control analyses (COPD cases vs controls) and analyses of %predicted FEV_1 in cases, as a measure of severity of airflow limitation. Replication was undertaken using a subset of the UK Biobank Lung Exome Variant Evaluation (BiLEVE) study, a collection of 48 931 individuals from UK Biobank with high-quality lung function and smoking data who were genotyped on an array that includes substantial overlap with the exome chip.[16] We also adopted a more powerful discovery strategy for COPD risk and severity of airflow limitation, by meta-analysing data for the subset of exome chip variants that were measured in both the COPD exome chip consortium and the UK BiLEVE study.

METHODS

Study participants and phenotypes

A total of 3487 ever smokers with airflow limitation indicative of GOLD 2[2] COPD or worse were identified from 12 UK collections as cases (case collections described in online supplementary table S1). Individuals met case criteria if they had FEV_1/FVC ≤0.7 and %predicted FEV_1 ≤80% (according to the National Health and Nutrition Examination Survey (NHANES) III spirometric reference equations[17]), did not have a doctor diagnosis of asthma and had reported current, or former smoking. Five of the sample collections (n=1398 samples, table 1) were COPD cohorts, with all individuals having irreversible airflow limitation, and meeting GOLD 2 criteria based on post-bronchodilator spirometry. The remaining cases were taken from general population cohorts; for these samples, only pre-bronchodilator spirometry measures were available. We used general population controls with exome chip data, from Generation Scotland: Scottish Family Health Study (GS:SFHS), British 1958 Birth Cohort (1958BC), Oxford Biobank and GoDARTS (Genetics of Diabetes and Audit Research Tayside Study), listed in table 1 with clinical characteristics. All controls were current or former smokers and were free of lung disease, according to available spirometry and phenotype information.

We used a subset of the UK BiLEVE study[16] for replication of novel signals, and for a larger discovery meta-analysis. A total of 24 457 heavy smokers (mean 35 pack-years) were genotyped as part of the UK BiLEVE study, selected such that 9748 individuals formed a low FEV_1 group (based on %predicted FEV_1), 4906 individuals formed a high FEV_1 group and 9803 had average FEV_1. We selected 4231 samples from the low FEV_1 group, with airflow limitation consistent with GOLD 2 or worse as cases and 8979 samples from the high and average FEV_1 groups with FEV_1/FVC >0.7, %predicted FEV_1 >80% and no doctor diagnosis of COPD for use as controls. All spirometry measures were prebronchodilator, all samples were heavy smokers and individuals with a doctor diagnosis of asthma or other lung diseases were excluded. The %predicted FEV_1 was estimated using NHANES III spirometric reference equations.[17]

An overview of the full study design is shown in figure 1.

Genotyping

All 3487 cases and 1032 GS:SFHS controls were genotyped together using the Illumina Human Exome BeadChip with additional custom content for regions which have previously shown modest association with lung function (description of custom content design in online supplementary methods). The remaining discovery analyses control samples were genotyped separately using the Illumina Human Exome BeadChip.

The UK BiLEVE samples were genotyped using the Affymetrix UK BiLEVE array, which includes rare variants selected from the same sequencing project as the Illumina Human Exome BeadChip alongside additional content.[16] Of the 807 411 SNPs included on the Affymetrix UK BiLEVE array, 74 891 were also present on the Illumina Human Exome BeadChip; this subset of SNPs, which were directly genotyped on both arrays, was selected for the discovery meta-analysis.

Quality control of genotype data

Discovery exome analysis

Genotypes were called using Illumina's Gencall algorithm in Genomestudio[18] with refinement of rare variants with missing calls undertaken using zCall.[19] Standard quality control (QC) filters were applied, in accordance with the Exome-chip Quality Control SOP V.5, as developed within the UK exome chip consortium[20] and are fully described in online supplementary methods. In brief, SNPs were excluded if they had low call rate (<99%) or deviated from Hardy Weinberg Equilibrium ($p<10^{-4}$) and samples were excluded if they were duplicates, sex mismatches, heterozygosity outliers (>3 SD from mean), had an excess of singleton SNPs, or were ancestral outliers. Clusterplots for all SNPs of interest were inspected, to ensure accuracy of genotype calling.

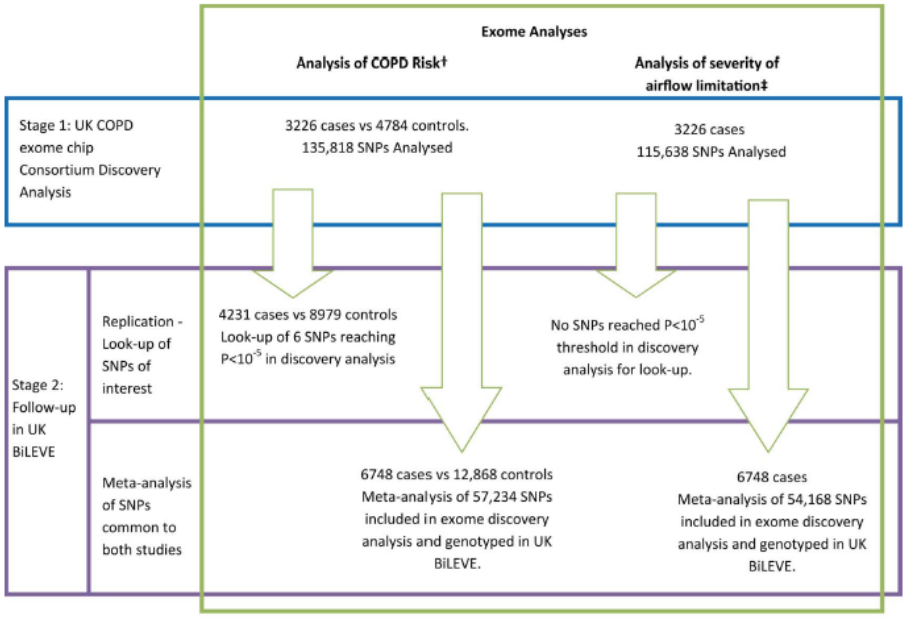
UK BiLEVE data

The QC procedure of the UK BiLEVE genotype data is described elsewhere.[16]

Table 1 Clinical characteristics of samples passing genotype QC

Sample collection	n	Sex	Age	%Predicted FEV_1	FEV_1/FVC	Pack-years	
		Male, n (%)	Mean (SD)	Mean (SD)	Mean (SD)	Samples with data (n)	Mean (SD)
Discovery analyses airflow limitation cases (total n=3226, with pack-years n=2517)							
GS:SFHS	508	224 (44.1%)	58.9 (8.94)	64.84 (12.64)	0.580 (0.108)	482	29.32 (24.96)
British Regional Heart Study	425	425 (100%)	70.1 (5.46)	59.41 (14.66)	0.597 (0.084)	0	–
British Women's Heart and Health Study	254	0 (0%)	69.3 (5.46)	64.26 (12.40)	0.603 (0.074)	203	28.1 (18.36)
UK COPD cohort*	209	129 (61.7%)	68.7 (8.11)	37.94 (15.29)	0.447 (0.119)	199	50.07 (27.79
Hertfordshire Cohort Study	317	203 (64.0%)	66.1 (2.79)	62.89 (13.57)	0.589 (0.101)	312	32.25 (23.37)
COPDBEAT*	87	62 (71.3%)	67.6 (8.77)	45.19 (16.24)	0.480 (0.115)	86	38.69 (21.24)
Nottingham COPD study*	76	48 (63.2%)	67.2 (8.97)	50.29 (15.04)	0.482 (0.111)	74	49.02 (26.86)
Nottingham smokers	125	78 (62.4%)	63.1 (8.60)	46.27 (17.65)	0.503 (0.125)	124	41.75 (20.61)
Gedling study	33	26 (78.8%)	69.0 (8.23)	59.67 (16.81)	0.593 (0.103)	31	45.47 (33.40)
English Longitudinal Study of Aging	166	75 (45.2%)	66.0 (8.17)	54.84 (17.24)	0.526 (0.149)	0	–
EU COPD Gene Scan*	277	155 (56.0%)	67.0 (8.68)	38.51 (14.74)	0.467 (0.120)	277	46.43 (20.56)
GoTARDIS Study*	749	412 (55.0%)	68.8 (8.97)	52.16 (14.14)	0.509 (0.110)	729	43.26 (21.59)
Discovery analyses controls (total n=4784, with pack-years n=3889)							
GS:SFHS	961	552 (57.4%)	54.5 (8.41)	98.18 (10.92)	0.783 (0.051)	961	28.92 (16.86)
British 1958 Birth Cohort	1429	888 (62.1%)	44 (0)	100.90 (13.46)	0.809 (0.060)	1046	14.74 (10.07)
Oxford Biobank	1770	832 (47.0%)	41.6 (5.77)	–	–	1682	9.09 (9.34)
GoDARTS	624	402 (64.4%)	59.0 (10.75)	–	–	200	35.46 (25.89)
UK Biobank Lung Exome Variant Evaluation samples (meta-analysis and replication)							
Airflow limitation cases	4231	2379 (56.2%)	59.54 (6.86)	61.76 (11.8)	0.607 (0.076)	4231	42.41 (21.10)
Controls	8979	4260 (47.4%)	56.19 (7.92)	101.40 (8.1)	0.773 (0.038)	8979	30.43 (14.41)

*Sample collection is COPD case cohort.
GS:SFHS, Generation Scotland: Scottish Family Health Study; GoTARDIS, Tayside Allergy and Respiratory Disease Information System; QC, quality control.



† Case-control analysis of COPD cases with FEV_1/FVC≤0.70 and %predicted FEV_1≤80% vs general population controls
‡Analysis of percent predicted FEV_1 in COPD cases

Figure 1 Two-stage study design. Stage 1: exome discovery analyses. Stage 2: Follow-up in UK BiLEVE: A. Replication of signals; B. meta-analysis of UK COPD exome chip consortium and UK BiLEVE.

Statistical analyses

SNP associations with COPD risk were carried out using a logistic regression model, adjusting for age, sex and pack-years and assuming an additive genetic model. Associations with untransformed %predicted FEV_1 in cases were tested, using a linear regression model, with adjustment for pack-years (analysis of severity of airflow limitation). Since not all samples had pack-years data available, secondary analyses were carried out without adjustment for pack-years, for both the COPD risk and severity of airflow limitation analyses, allowing the inclusion of all samples. Single variant analyses were carried out using PLINK V.1.07.[21] Using a Bonferroni correction for the number of tests undertaken, a significance level of $p<3.7\times10^{-7}$ would be required in the exome single variant analysis to retain a type 1 error of 5%. We defined SNPs of interest as those with $p<10^{-5}$ in the discovery exome analysis; for these SNPs, we undertook replication analyses in the UK BiLEVE study to corroborate findings (see online supplementary methods). We set a Bonferroni corrected significance level for replication, for the number of SNPs in novel loci taken forward to replication ($p<0.017$ for analysis of COPD risk). Gene-based analyses using SKAT-O were additionally undertaken; the methods and results of these analyses are described in the online supplementary information.

Custom content single variant analyses

Custom content comprising 2585 SNPs tagging regions which had shown suggestive association ($p<2.21\times10^{-3}$) with lung function in a previous large genome-wide HapMap-imputed study[13] were also included on the array for cases and GS:SFHS controls. Additional controls from 1958BC and Busselton Health Study (BHS) with genome-wide data were also used; full methods and results of this analysis are given in the supplementary information.

Meta-analysis with UK BiLEVE data

Single variant associations with COPD risk and severity of airflow limitation in the UK BiLEVE samples were carried out using PLINK v1.07,[21] identically to the corresponding discovery analysis with pack-years adjustment. We carried out an inverse-variance–weighted meta-analysis of the union of SNPs included in the discovery exome and UK BiLEVE analyses (described in online supplementary methods).

RESULTS

Discovery exome analysis

3226 cases and 4784 controls passed all sample and SNP genotype QC and were used in the exome analysis (exclusions in online supplementary table S1). Clinical characteristics of these samples are summarised in table 1. Of the SNPs which passed all QC criteria in both cases and controls, 135 818 were polymorphic, of which 101 308 (74.6%) had a MAF<1%.

Analyses of COPD risk

We carried out pack-years adjusted analysis of COPD risk, including 2517 cases and 3889 controls, in addition to an unadjusted analysis, using all 3226 cases and 4784 controls (quantile–quantile plots shown in online supplementary figure S1). A total of four SNPs in three regions met the $p<10^{-5}$ significance threshold in the pack-years adjusted analysis, with five SNPs in four regions showing $p<10^{-5}$ in the unadjusted analysis (figure 2).

In the pack-years adjusted analysis (table 2A and figure 2A), the most significant association was for the previously reported COPD/smoking region 15q25 (sentinel SNP rs8034191 OR: 1.38, MAF=34.8%, $p=2.42\times10^{-7}$). This signal was replicated in the UK BiLEVE study. Two novel signals of association with COPD risk ($p<10^{-5}$) were rs3813803 within *SMPDL3B* (OR: 1.37, MAF=29.2%, $p=1.04\times10^{-6}$) and low frequency SNP rs7269297 within *MOCS3* (OR: 0.25, MAF=1.1%, $p=3.08\times10^{-6}$). There was evidence of replication, just above the Bonferroni corrected level of significance ($p<0.017$) for rs7269297 in the UK BiLEVE study ($p=7.27\times10^{-5}$ for meta-analysis of discovery and UK BiLEVE results, table 2A).

A further two loci were associated with COPD risk in the analysis unadjusted for pack-years: rs3827522 within *PRICKLE1* (OR: 0.12, MAF=0.4%, $p=1.03\times10^{-7}$) and rs17368582 within *MMP12* (OR: 0.712, MAF=12.2% $p=5.01\times10^{-6}$, table 2A and figure 2B); however, there was no evidence of replication of these associations with COPD risk in UK BiLEVE. rs2276109, another SNP within *MMP12*, (MAF=5.6%) which is strongly correlated with rs17368582 ($r^2=0.84$), has previously been associated with COPD risk in smokers.[7] Overall, no associations in novel regions met exome-wide significance ($p<3.7\times10^{-7}$).

Analyses of severity of airflow limitation

Although no SNPs reached the $p<10^{-5}$ significance level in either the pack-years adjusted, or the unadjusted analysis (see online supplementary figures S2 and S3), six SNPs showed some evidence of association ($p<10^{-4}$) in one or both analyses (see online supplementary table S2). Of note, rs28929474, the z-allele within the *SERPINA1* gene, showed modest association in the unadjusted analysis ($\beta=-6.17\%$, MAF=2.0%, $p=2.83\times10^{-5}$).

UK BiLEVE meta-analysis results

Analyses of COPD risk

For the 57 234 polymorphic SNPs common to both the COPD exome chip consortium samples and the UK BiLEVE study, a meta-analysis of discovery and UK BiLEVE study results was undertaken in which three regions showed association with risk of COPD ($p<10^{-5}$, figure 3, online supplementary figure S4 and table 2B). The *GYPA/HHIP* and *GPR126* regions have previously been reported as showing association with lung function and COPD or airflow limitation risk.[3 10 14] The *IFIT3* region signal (rs140549288 p.Val352Leu in *IFIT3*, OR: 1.92, MAF=0.7%, $p=7.49\times10^{-6}$) represents a novel rare variant signal of association with COPD.

Analyses of severity of airflow limitation

A total of 54 168 SNPs were included in the meta-analysis of severity of airflow limitation (see online supplementary figures S5 and S6). One SNP showed association with $p<10^{-5}$: rs140198372, a variant which alters the sequence at a site where the splicing of an intron takes place (splice site) in *SERPINA12* ($\beta=-33.51\%$, MAF=0.03%, $p=5.72\times10^{-6}$, table 3).

Sensitivity analyses to assess COPD case criteria

Of our 3226 COPD cases defined as described above, 1398 also had a GOLD 2 or worse COPD based on postbronchodilator spirometry. We carried out a sensitivity analysis for all SNPs identified in our discovery or meta-analyses of COPD risk, by repeating the discovery analyses including only those 1398 COPD cases which underwent reversibility testing. This analysis

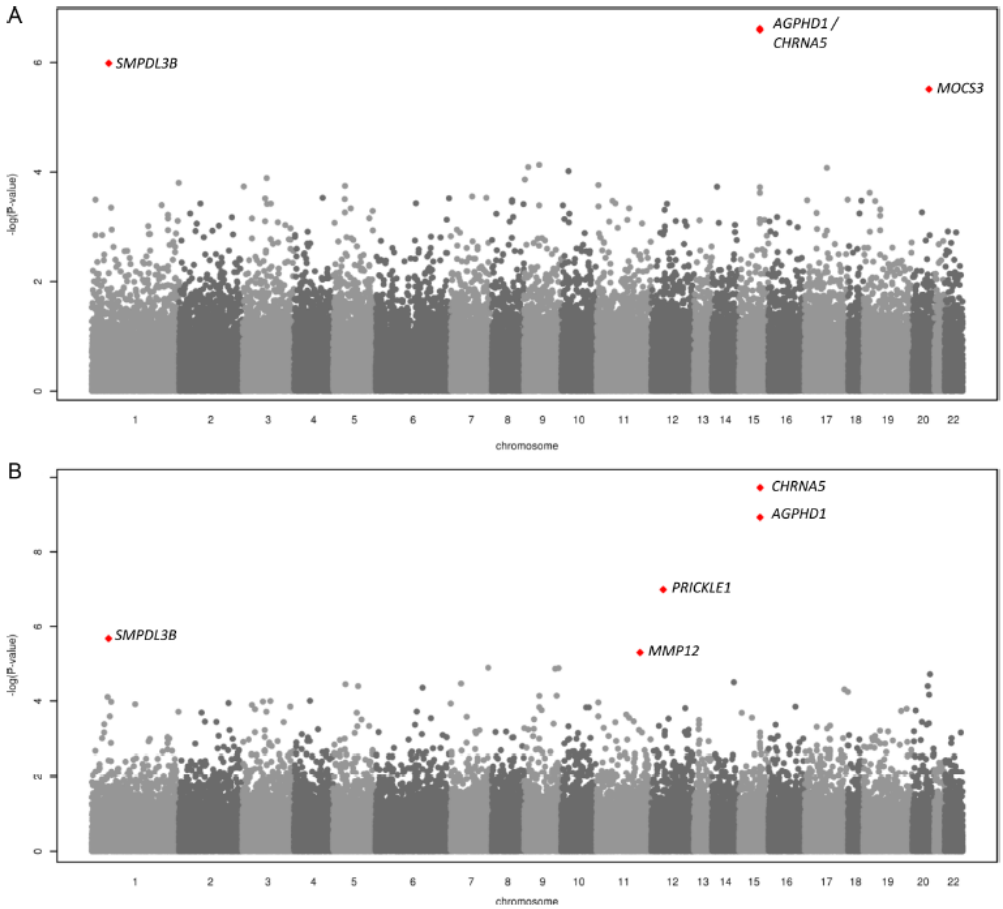


Figure 2 (A) Analysis of COPD risk, with pack-years adjustment (single nucleotide polymorphisms (SNPs) with minor allele frequency (MAF) >0.05% only; SNPs with $p<10^{-5}$ highlighted). (B) Analysis of COPD risk, without pack-years adjustment (SNPs with MAF >0.05% only; SNPs with $p<10^{-5}$ highlighted).

showed consistent estimated effect sizes (see online supplementary table S3 and figure S7), and in particular, the ORs were not substantially attenuated for rs7269297 in *MOCS3* (sensitivity analysis OR: 0.276; original discovery OR: 0.251), nor rs140549288 in *IFIT3* (sensitivity analysis OR: 2.554; original discovery OR: 2.156).

Association of novel loci with smoking behaviour

Given the disparity of smoking behaviour in our cases and control samples (table 1), we further investigated whether either of the two novel COPD risk loci were associated with smoking behaviour, to ascertain whether the associations with COPD may be explained by differences in smoking. Neither of the sentinel SNPs showed significant association with heavy versus never smoking within UK BiLEVE (p=0.956 for rs7269297 and p=0.945 for rs140549288) study. We further undertook a look-up in the publically available results of a GWAS from the Tobacco and Genetics consortium[22] for associations with rs7269297 in *MOCS3* (rs140549288 was not available in data) and a number of smoking traits; however, no evidence for association with smoking behaviour was found (cigarettes per day p=0.610; ever vs never smoking p=0.172; current vs former smoking p=0.699).

DISCUSSION

We carried out analyses of exome chip variants with COPD risk and %predicted FEV_1 among cases, through which we identified a number of SNPs in both known COPD regions and at novel loci that showed suggestive association ($p<10^{-5}$) with risk of COPD. These novel regions (region plots: online supplementary figure S8) warrant further investigation as they may provide insight into the underlying biological mechanisms of COPD and airflow limitation in smokers and could provide novel therapeutic targets. The most significant associations in both the discovery exome analysis and the meta-analysis were with SNPs in the 15q25 region, previously identified through GWAS as being associated with smoking behaviour,[22–24] lung cancer,[25] COPD[3] and airflow obstruction.[14] In addition, we independently replicated previously reported associations of *HHIP*,[3 10] *GPR126*[14] and *MMP12*[7 8] with COPD risk.

We identified novel associations between COPD risk and low frequency or rare coding SNPs in two genes: *MOCS3*

Table 2 Top associations in exome discovery analyses and meta-analysis of COPD risk

(A) SNPs with $p<10^{-5}$ in either the pack-years adjusted or unadjusted discovery analyses

					Discovery pack-years adjusted analysis (2517 cases, 3889 controls)				Discovery unadjusted analysis (3226 cases, 4784 controls)				UK BiLEVE pack-years adjusted analysis (4231 cases, 8979 controls)				Meta-analysis of discovery and UK BiLEVE pack-year adjusted analyses	
					MAF (MAC)		Association result		MAF (MAC)		Association result		MAF (MAC)		Association result		Association result	
rs no.	CHR	Position	Coded allele	Gene	Cases	Controls	OR (95% CI)	p Value*	Cases	Controls	OR (95% CI)	p Value*	Cases	Controls	OR (95% CI)	p Value*	OR (95% CI)	p Value*
rs3813803	1	28282292	C	*SMPDL3B* (non-synonymous)	30.6% (1541)	28.3% (2203)	1.370 (1.207 to 1.554)	**2.41×10^{-6}**	30.3% (1956)	28.5% (2722)	1.288 (1.160 to 1.430)	**2.11×10^{-6}**	28.7% (2418)	29.4% (5269)	0.968 (0.911 to 1.029)	0.298	1.033 (0.978 to 1.092)	0.241
rs17368582	11	102738075	C	*MMP12* (synonymous)	11.1% (561)	12.9% (1001)	0.767 (0.642 to 0.915)	3.22×10^{-3}	11.1% (719)	12.8% (1229)	0.712 (0.615 to 0.824)	**5.01×10^{-6}**	12.0% (1015)	12.2% (2198)	0.982 (0.902 to 1.069)	0.676	0.938 (0.868 to 1.013)	0.101
rs3827522	12	42853871	A	*PRICKLE1* (non-synonymous)	0.2% (11)	0.4% (27)	0.184 (0.065 to 0.519)	1.39×10^{-3}	0.2% (14)	0.5% (46)	0.123 (0.057 to 0.266)	**1.03×10^{-7}**	0.3% (21)	0.3% (45)	0.907 (0.518 to 1.585)	0.731	0.633 (0.386 to 1.039)	0.071
rs8034191	15	78806023	C	near *AGPHD1* (intergenic)	38.0% (1912)	32.7% (2546)	1.374 (1.218 to 1.550)	**2.42×10^{-7}**	37.7% (2432)	32.9% (3144)	1.364 (1.234 to 1.507)	**1.18×10^{-9}**	39.2% (3315)	35.2% (6320)	1.156 (1.092 to 1.224)	**6.85×10^{-7}**	1.193 (1.133 to 1.257)	**2.79×10^{-11}**
rs7269297	20	49576664	G	*MOCS3* (non-synonymous)	0.7% (37)	1.4% (110)	0.251 (0.140 to 0.448)	**3.08×10^{-6}**	0.8% (54)	1.5% (139)	0.423 (0.262 to 0.680)	3.98×10^{-4}	1.2% (98)	1.4% (252)	0.742 (0.578 to 0.953)	0.019	0.626 (0.497 to 0.789)	7.27×10^{-5}

(B) SNPs with $p<10^{-5}$ in the meta-analysis (only most statically significant SNP in each region shown)

					Discovery pack-years adjusted analysis (2517 cases, 3889 controls)				UK BiLEVE pack-years adjusted analysis (4231 cases, 8979 controls)				Meta-analysis of discovery and UK BiLEVE pack-year adjusted analyses	
					MAF (MAC)		Association result		MAF (MAC)		Association result		Association result	
rs no.	CHR	Position	Coded allele	Gene	Cases	Controls	OR (95% CI)	p Value*	Cases	Controls	OR (95% CI)	p Value*	OR (95% CI)	p Value*
rs1828591	4	145480780	A	GYPA/HHIP (intergenic)	35.6% (1794)	39.1% (3042)	0.9167 (0.814, 1.032)	0.153	36.6% (3088)	40.0% (771)	0.867 (0.819, 0.918)	**9.88×10^{-7}**	0.876 (0.832, 0.922)	**5.75×10^{-7}**
rs4896582	6	142703877	A	GPR126 (intronic)	29.3% (1473)	31.7% (2468)	0.8594 (0.757, 0.974)	0.018	28.0% (2349)	30.2% (5344)	0.879 (0.826, 0.934)	3.87×10^{-5}	0.875 (0.827, 0.925)	**2.53×10^{-6}**
rs140549288	10	91099466	C	IFIT3 (exonic), LIPA (intronic)	0.8% (38)	0.6% (44)	2.156 (1.046, 4.445)	0.037	0.9% (79)	0.6% (100)	1.880 (1.378, 2.565)	6.87×10^{-5}	1.924 (1.441, 2.560)	**8.56×10^{-6}**

*p Values in bold significant at $p<10^{-5}$ level.
BiLEVE, Biobank Lung Exome Variant Evaluation; MAC, minor allele count; MAF, minor allele frequency; SNPs, single nucleotide polymorphisms.

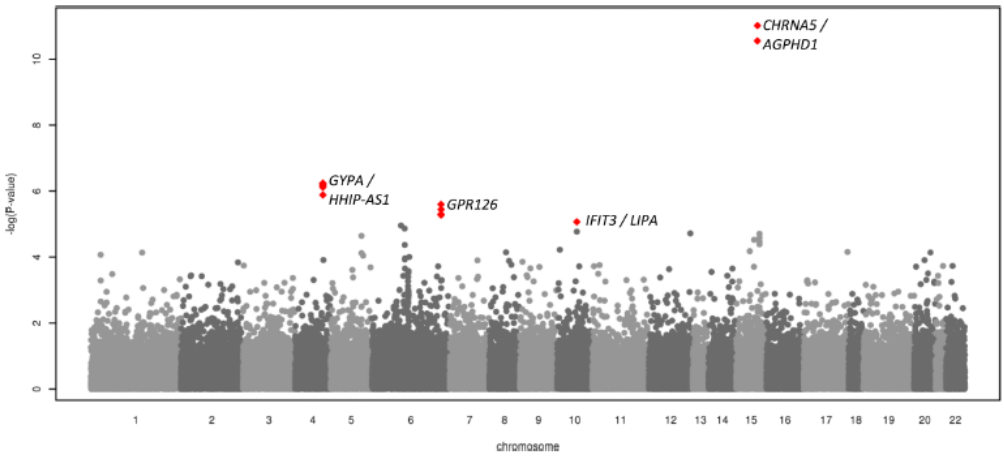


Figure 3 Meta-analysis of COPD risk in discovery exome analysis and UK Biobank Lung Exome Variant Evaluation samples.

(rs7269297, serine to alanine, MAF=1.3%, $p_{discovery-meta}=3.08\times10^{-6}$, PolyPhen prediction: benign) and *IFIT3* (rs140549288, valine to leucine, MAF=0.7%, $p_{meta-meta}=8.56\times10^{-6}$, PolyPhen prediction: benign). The protein encoded by *MOCS3* adenylates and activates molybdopterin synthase, an enzyme required to synthesise molybdenum cofactor[26] and is expressed in bronchial epithelium and smooth muscle layer of the bronchus.[27] *IFIT3* is associated with interferon-α antiviral activity and has been found to be up-regulated in respiratory syncytial virus infection[28] and in human lung epithelial cells infected with dengue virus.[29] The SNP rs140549288 is also located within in an intron of *LIPA*; the product of this gene is involved in the hydrolysis of cholesteryl esters and triglycerides and other SNPs within this gene have previously been associated with coronary artery disease.[30]

The z-allele within the *SERPINA1* gene was associated with a lower %predicted FEV_1 in cases (unadjusted analysis: $p_{discovery-meta}=2.83\times10^{-5}$); as well as being a well-established cause of AAT deficiency,[3 4] this SNP has also previously been associated with an increased annual decline in FEV_1 in a general population sample[31] and increased airflow limitation in COPD cases.[32] In the present study, the z-allele was associated with an increased risk of COPD, although this was not statistically significant (OR: 1.27, p=0.252). The likely reason for the lack of a significant association with this known COPD locus is that some of the case collections excluded individuals with AAT deficiency, resulting in selection bias. In the meta-analysis of severity of airflow limitation, we identified a very rare SNP within another serine protease inhibitor gene, *SERPINA12*, not previously associated with COPD (rs140198372, MAF=0.03%, $p_{meta-meta}=5.72\times10^{-6}$). SERPINA12 and SERPINA1 lie 96.6 kb apart on chromosome 14 (rs140198372 and the z-allele in SERPINA1 are not in linkage disequilibrium ($r2=9.0\times10^{-6}$)). *SERPINA12* has been associated with cardiovascular diseases, being implicated in obesity and type 2 diabetes.[33]

One of the primary challenges associated with identifying low frequency variants associated with disease is limited statistical power, and this could explain our lack of strong statistically significant findings. Indeed, none of the reported associations in novel regions met a stringent exome-wide significance level ($p<3.8\times10^{-7}$) overall. In the present study, we would have just 54% power to detect an association with an SNP associated with COPD risk with a MAF of 1% and an OR of 2, at the $p<3.8\times10^{-7}$ level. Furthermore, recent analyses undertaken by the UK10K Consortium found no evidence of low frequency SNPs having large effects, upon a series of traits.[34] Due to the limited power to detect single variant associations of rare variants with modest effect sizes, we additionally adopted gene-based analyses using SKAT-O, a method which combines information from several rare variants (see online supplementary information). In these analyses, we only identified one gene meeting our elected significance level ($p<10^{-5}$); this gene-based signal in *PRICKLE1* was found however, to be driven by a single SNP, which was identified as being associated with COPD risk in

Table 3 Top associations ($p<10^{-5}$) in meta-analysis of severity of airflow limitation

					Severity of airflow limitation, adjusted for pack-years (n=2517)			UK BiLEVE pack-years adjusted analysis (n=4231)			Meta-analysis of discovery and UK BiLEVE pack-year adjusted analyses	
rs no.	CHR	Position	Coded allele	Gene	MAF (MAC)	Beta (95% CI)	p Value	MAF (MAC)	Beta (95% CI)	p Value	Beta (95% CI)	p Value
rs140198372	14	94953832	A	*SERPINA12* (splice site)	0.059% (3)	−29.23 (−49.50 to −8.96)	2.59×10^{-5}	0.012% (1)	−38.35 (−59.88 to −16.82)	4.11×10^{-4}	−33.51 (−48.27 to −18.76)	**5.72×10^{-6}**

*p Values in bold significant at $p<10^{-5}$ level.
BiLEVE, Biobank Lung Exome Variant Evaluation; MAC, minor allele count; MAF, minor allele frequency.

the single variant discovery analysis, but which was not replicated in the UK BiLEVE data.

Another limitation of this study is that a number of our cases had only prebronchodilator spirometry; for these samples, it could not be determined whether their airflow limitation was reversible, and so a proportion of these cases may not have met the clinical definition of COPD. We undertook case–control sensitivity analyses using our discovery samples, restricting cases to the subset of 1398 individuals taken from COPD cohorts and who had known irreversible airflow limitation. The effect estimates of our top hits did not substantially change in this sensitivity analysis, suggesting that our broader case definition, including samples that did not undergo reversibility testing, did not result in substantial misclassification bias.

A further potential source of bias in this study was the heavier smoking history in our cases compared with the control samples. For the two SNPs identified through the analyses of COPD risk, we found no evidence of association with smoking in data from the UK BiLEVE study, suggesting that the associations with COPD risk were not driven by the imbalances in smoking behaviour.

Finally, it was not possible to validate the findings of this study through additional genotyping; however for the three reported loci, consistent results were observed in both the discovery and the UK BiLEVE samples. It would not be expected to see the same false positive result in these two independent samples, therefore, strengthening the evidence for these being true associations.

In summary, we have identified potentially interesting associations with low frequency and rare SNPs and COPD risk in two regions not previously implicated in COPD or lung function. We further identified an association of %predicted FEV_1 in individuals with COPD with a very rare SNP in *SERPINA12*. Further confirmation of these associations in larger independent collections of COPD cases and controls is needed. This study also provides further evidence that the z-allele within *SERPINA1* may be related to severity of airflow limitation in COPD. While large sample sizes may be required to definitively identify novel loci, we present evidence to support the notion that the genetic contribution to COPD risk comprises polygenic contributions of rare, low frequency and common genetic variants. Future studies, alone or in combination, should aim to target the full allele frequency range to unravel the genetic architecture of COPD.

Author affiliations
[1]Department of Health Sciences, University of Leicester, Leicester, UK
[2]William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK
[3]Division of Respiratory Medicine, University of Nottingham, Queen's Medical Centre, Nottingham, UK
[4]School of Social & Community Medicine, University of Bristol, Bristol, UK
[5]Department of Primary Care & Population Health, UCL, London, UK
[6]Population Health Research Institute, St George's, University of London, London, UK
[7]University College London, Farr Institute of Health Informatics, London, UK
[8]Cochrane Heart Group, London, UK
[9]Department of Non-communicable Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London, UK
[10]ISER, University of Essex, Colchester, Essex, UK
[11]Department of Epidemiology and Public Health, UCL, London, UK
[12]Institute of Cardiovascular Science, UCL, London, UK
[13]School of Life Sciences, University of Nottingham, Nottingham, UK
[14]Division of Epidemiology and Public Health, Nottingham City Hospital, University of Nottingham, Nottingham, UK
[15]Cardiovascular and Diabetes Medicine, School of Medicine, University of Dundee, Dundee, UK.
[16]Human Development & Health, Faculty of Medicine, University of Southampton, Southampton General Hospital, Southampton, UK
[17]NIHR Southampton Respiratory Biomedical Research Unit, University of Southampton and University Hospital Southampton NHS Foundation Trust, Southampton General Hospital, Southampton, UK
[18]MRC Lifecourse Epidemiology Unit, University of Southampton, Southampton General Hospital, Southampton, UK
[19]NIHR Southampton Biomedical Research Centre, University of Southampton and University Hospital Southampton NHS Foundation Trust, Southampton General Hospital, Southampton, UK
[20]Victoria University, Wellington, New Zealand
[21]Respiratory Medicine Unit, Nuffield Department of Medicine, University of Oxford, Oxford, UK
[22]Institute for Lung Health, Department of Infection, Immunity and Inflammation, University of Leicester, Leicester, UK
[23]National Institute for Health Research Respiratory Biomedical Research Unit, Glenfield Hospital, Leicester, UK
[24]Nottingham Respiratory Research Unit, University of Nottingham, City Hospital Campus, Nottingham, UK
[25]Institute for Ageing and Health, Newcastle University, Campus for Ageing and Vitality, Newcastle upon Tyne, UK
[26]Department of Molecular Genetics and Genomics, National Heart and Lung Institute, Imperial College London, London, UK
[27]Freemasons' Department of Geriatric Medicine, University of Auckland, New Zealand
[28]Generation Scotland, Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK
[29]Division of Population Health Sciences, University of Dundee, Dundee, UK
[30]Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow, UK
[31]Institute of Medical Sciences, University of Aberdeen, Aberdeen, UK
[32]Department of Haematology, University of Cambridge, Cambridge, UK
[33]Princess Al-Jawhara Al-Brahim Centre of Excellence in Research of Hereditary Disorders, King Abdulaziz University, Jeddah, Saudi Arabia

Acknowledgements This research used the ALICE and SPECTRE High Performance Computing Facilities at the University of Leicester and was supported by the National Institute for Health Research (NIHR) Leicester Respiratory Biomedical Research Unit. This article/paper/report presents independent research funded partially by the NIHR. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. This research has been conducted using the UK Biobank Resource.

Contributors Case collection study concept, or data acquisition and quality control: IS, IPH, DPS, RM, PW, JPC, AA, MC, CD, MK, JE, NK, SC, TGB, TMM, CNAP, RT, JWH, AAS, EMD, CC, MB, BB, CB, CEB, MEJ, SGP, MFM, AJW, MJC, BJP, BHS, SP and LH. Genotype data acquisition and QC: KES and PD. Central study design, analysis and writing of manuscript: VEJ, IN, LVW, MDT, IPH and IS.

Funding British Women's Heart and Health Study is funded by the Department of Health grant no. 90049 and the British Heart Foundation grant no. PG/09/022. British Regional Heart Study is supported by the British Heart Foundation (grant RG/13/16/30528). CB (COPDBEAT) received funding from the Medical Research Council UK (grant no. G0601369), CB (COPDBEAT) and AJW (UKCOPD) were supported by the National Institute for Health Research (NIHR Leicester Biomedical Research Unit). MB (COPDBEAT) received funding from the NIHR (grant no. PDF-2013-06-052). Hertfordshire Cohort Study received support from the Medical Research Council, Arthritis Research UK, the International Osteoporosis Foundation and the British Heart Foundation; NIHR Biomedical Research Centre in Nutrition, University of Southampton; NIHR Musculoskeletal Biomedical Research Unit, University of Oxford. Generation Scotland: Scottish Family Health Study is funded by the Chief Scientist Office, Scottish Government Health Directorates, grant number CZD/16/6 and the Scottish Funding Council grant HR03006. EU COPD Gene Scan is funded by the European Union, grant no. QLG1-CT-2001-01012. English Longitudinal Study of Aging is funded by the Institute of Aging, NIH grant No. AG1764406S1. GoDARTs is funded by the Wellcome Trust grants 072960, 084726 and 104970. MDT has been supported by MRC fellowship G0902313. UK Biobank Lung Exome Variant Evaluation study was funded by a Medical Research Council strategic award to MDT, IPH, DPS and LVW (MC_PC_12010).

Competing interests None declared.

Ethics approval Several (meta-analysis design).

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Rabe KF, Hurd S, Anzueto A, *et al*. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *Am J Respir Crit Care Med* 2007;176:532–55.
2 *Global Strategy for the Diagnosis, Management and Prevention of COPD*. 2015. http://www.goldcopd.org/uploads/users/files/GOLD_Report_2015_Apr2.pdf
3 Pillai SG, Ge D, Zhu G, *et al*. A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genet* 2009;5:e1000421.
4 Thun GA, Imboden M, Ferrarotti I, *et al*. Causal and synthetic associations of variants in the *SERPINA* gene cluster with alpha1-antitrypsin serum levels. *PLoS Genet* 2013;9:e1003585.
5 Mayer AS, Newman LS. Genetic and environmental modulation of chronic obstructive pulmonary disease. *Respir Physiol* 2001;128:3–11.
6 Soler Artigas M, Wain LV, Repapi E, *et al*. Effect of five genetic variants associated with lung function on the risk of chronic obstructive lung disease, and their joint effects on lung function. *Am J Respir Crit Care Med* 2011;184:786–95.
7 Hunninghake GM, Cho MH, Tesfaigzi Y, *et al*. MMP12, lung function, and COPD in high-risk populations. *N Engl J Med* 2009;361:2599–608.
8 Haq I, Chappell S, Johnson SR, *et al*. Association of MMP-12 polymorphisms with severe and very severe COPD: a case control study of MMPs-1, 9 and 12 in a European population. *BMC Med Genet* 2010;11:7.
9 Cho MH, Boutaoui N, Klanderman BJ, *et al*. Variants in FAM13A are associated with chronic obstructive pulmonary disease. *Nat Genet* 2010;42:200–2.
10 Wilk JB, Chen TH, Gottlieb DJ, *et al*. A Genome-wide association study of pulmonary function measures in the Framingham Heart Study. *PLoS Genet* 2009;5: e1000429.
11 Repapi E, Sayers I, Wain LV, *et al*. Genome-wide association study identifies five loci associated with lung function. *Nat Genet* 2010;42:36–44.
12 Hancock DB, Eijgelsheim M, Wilk JB, *et al*. Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nat Genet* 2010;42:45–52.
13 Soler Artigas M, Loth DW, Wain LV, *et al*. Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. *Nat Genet* 2011;43:1082–90.
14 Wilk JB, Shrine NRG, Loehr LR, *et al*. Genome-wide association studies identify CHRNA5/3 and HTR4 in the development of airflow obstruction. *Am J Respir Crit Care Med* 2012;186:622–32.
15 Castaldi PJ, Cho MH, Litonjua AA, *et al*. The association of genome-wide significant spirometric loci with chronic obstructive pulmonary disease susceptibility. *Am J Respir Cell Mol Biol* 2011;45:1147–53.
16 Wain L, Shrine N, Miller S, *et al*. Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *Lancet Respir Med* 2015;3:769–81.
17 Hankinson JL, Odencrantz JR, Fedan KB. Spirometric reference values from a sample of the general U.S. population. *Am J Respir Crit Care Med* 1999;159:179–87.
18 Illumina Inc. Illumina GenCall Data Analysis Software. 2005.
19 Goldstein JI, Crenshaw A, Carey J, *et al*. zCall: a rare variant caller for array-based genotyping: Genetics and population analysis. *Bioinformatics* 2012;28:2543–5.
20 Mahajan A, Robertson N, Rayner W. Exome-Chip Quality Control SOP. Version 5, 2012-11-20. 2012.
21 Purcell S, Neale B, Todd-Brown K, *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;81:559–75.
22 The Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet* 2010;42:441–7.
23 Thorgeirsson TE, Gudbjartsson DF, Surakka I, *et al*. Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. *Nat Genet* 2010;42:448–53.
24 Liu JZ, Tozzi F, Waterworth DM, *et al*. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet* 2010;42:436–40.
25 McKay JD, Hung RJ, Gaborieau V, *et al*. Lung cancer susceptibility locus at 5p15.33. *Nat Genet* 2008;40:1404–6.
26 Matthies A, Nimtz M, Leimkühler S. Molybdenum cofactor biosynthesis in humans: identification of a persulfide group in the rhodanese-like domain of MOCS3 by mass spectrometry. *Biochemistry* 2005;44:7912–20.
27 Uhlen M, Oksvold P, Fagerberg L, *et al*. Towards a knowledge-based Human Protein Atlas. *Nat Biotech* 2010;28:1248–50.
28 Ternette N, Wright C, Kramer HB, *et al*. Label-free quantitative proteomics reveals regulation of interferon-induced protein with tetratricopeptide repeats 3 (IFIT3) and 5′-3′-exoribonuclease 2 (XRN2) during respiratory syncytial virus infection. *Virol J* 2011;8:442.
29 Hsu YL, Shi SF, Wu WL, *et al*. Protective roles of interferon-induced protein with tetratricopeptide repeats 3 (IFIT3) in dengue virus infection of human lung epithelial cells. *PLoS ONE* 2013;8:e79518.
30 Wild PS, Zeller T, Schillert A, *et al*. A genome-wide association study identifies LIPA as a susceptibility gene for coronary artery disease. *Circ Cardiovasc Genet* 2011;4:403–12.
31 Dahl M, Tybjærg-Hansen A, Lange P, *et al*. Change in lung function and morbidity from chronic obstructive pulmonary disease in α1-antitrypsin MZ heterozygotes: a longitudinal study of the general population. *Ann Intern Med* 2002;136:270–9.
32 Molloy K, Hersh CP, Morris VB, *et al*. Clarification of the risk of chronic obstructive pulmonary disease in α1-antitrypsin deficiency PiMZ heterozygotes. *Am J Respir Crit Care Med* 2014;189:419–27.
33 Kim DS, Burt AA, Crosslin DR, *et al*. Novel common and rare genetic determinants of paraoxonase activity: FTO, SERPINA12, and ITGAL. *J Lipid Res* 2013;54:552–60.
34 Walter K, Min JL, Huang J, *et al*., UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature* 2015;526:82–90.

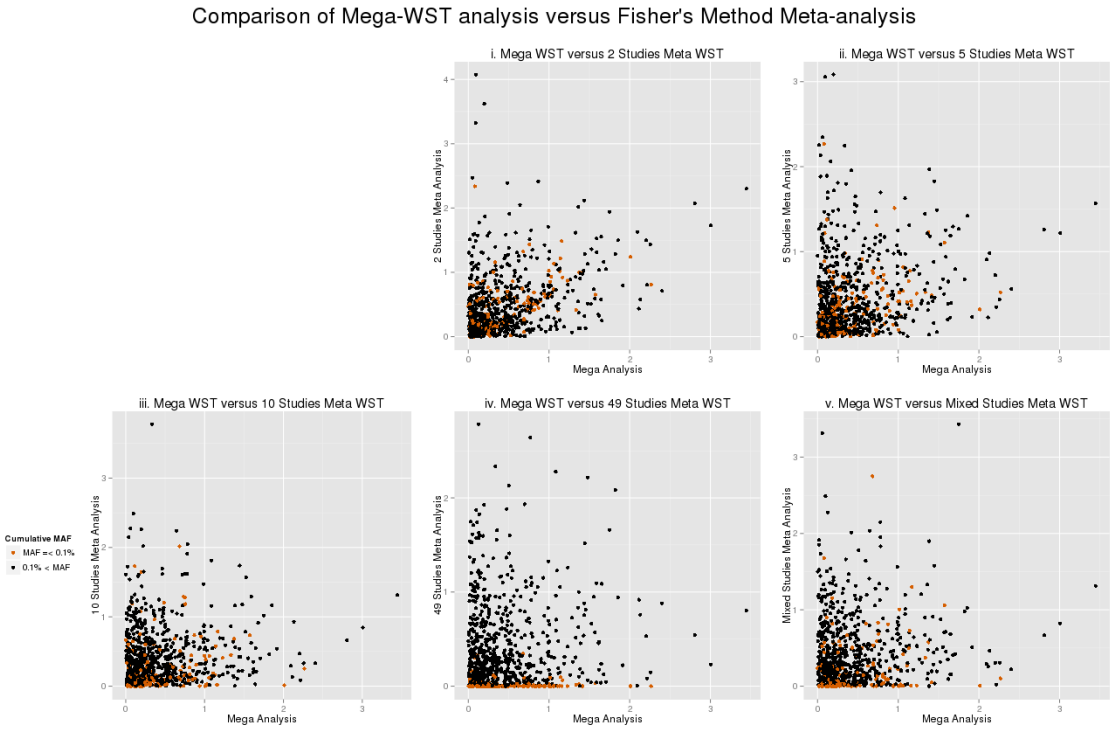
B. Additional Results for the Investigation of RAREMETAL in UK BiLEVE (Chapter 3)

P-value comparisons from the mega-analysis, and each meta-analysis scenario where P-values were combined using both Fisher's Method and the Z-score meta-analysis method, for the following:

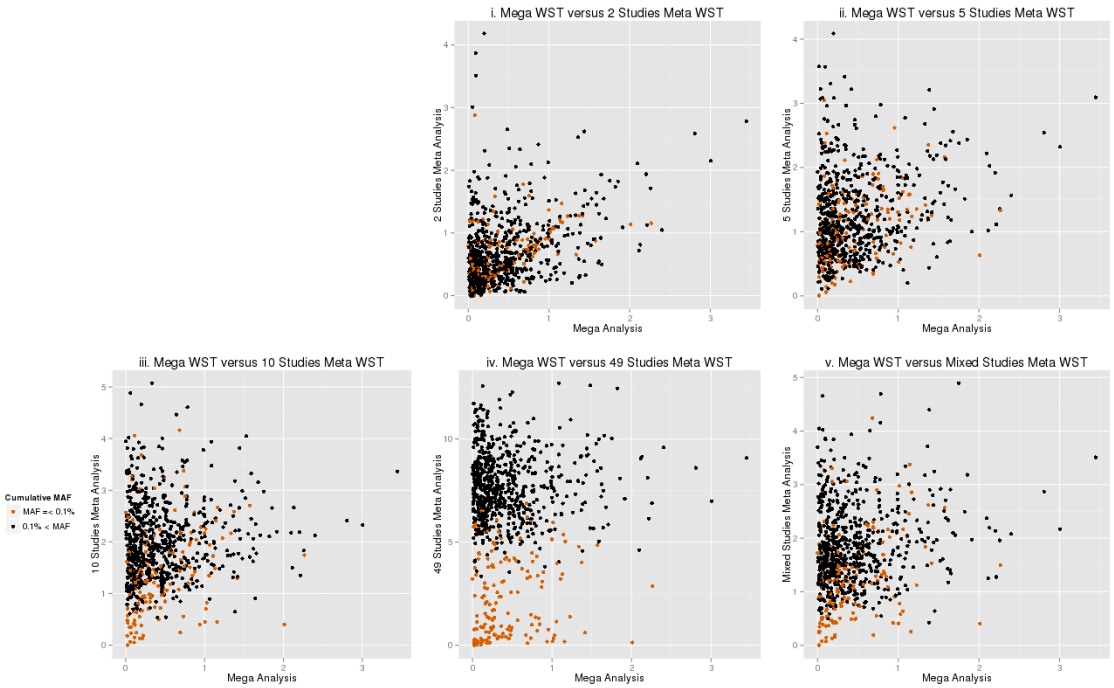
1. WST Analysis of FEV_1
2. SKAT Analysis of Smoking:
3. WST Analysis of Smoking:

1. WST Analysis of FEV_1

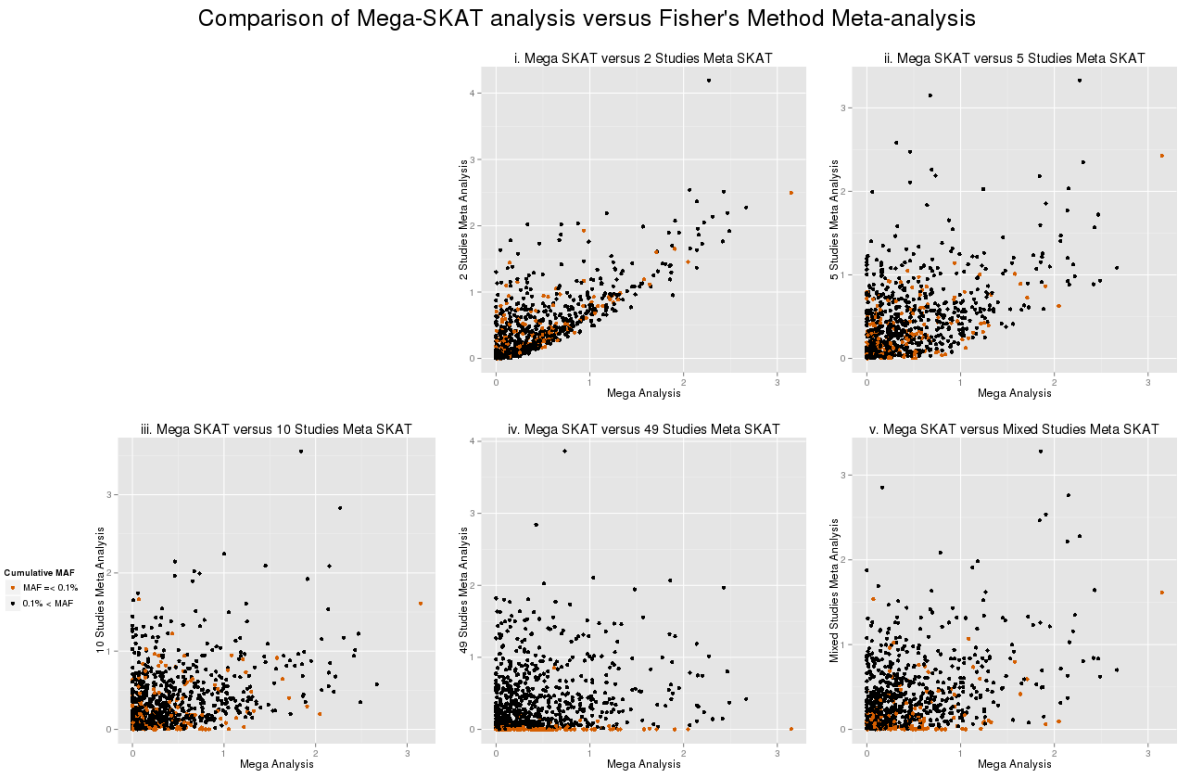
Appendix Figure B-1: Comparison of P-values from the mega-analysis, and each Fisher's Method meta-analysis scenario for the WST analysis of FEV_1. Genes with Cumulative MAF<0.1% highlighted.



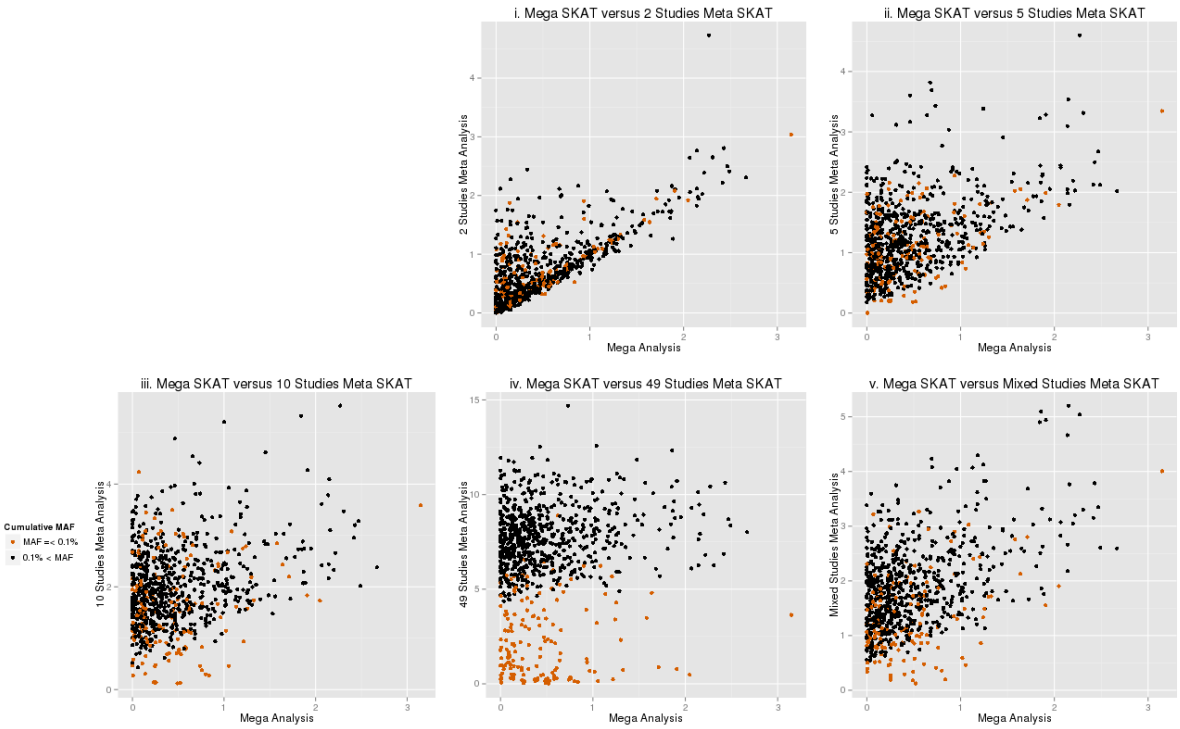
Appendix Figure B-2: Comparison of P-values from the mega-analysis, and each Z-score meta-analysis scenario for the WST analysis of FEV_1.Genes with Cumulative MAF<0.1% highlighted.



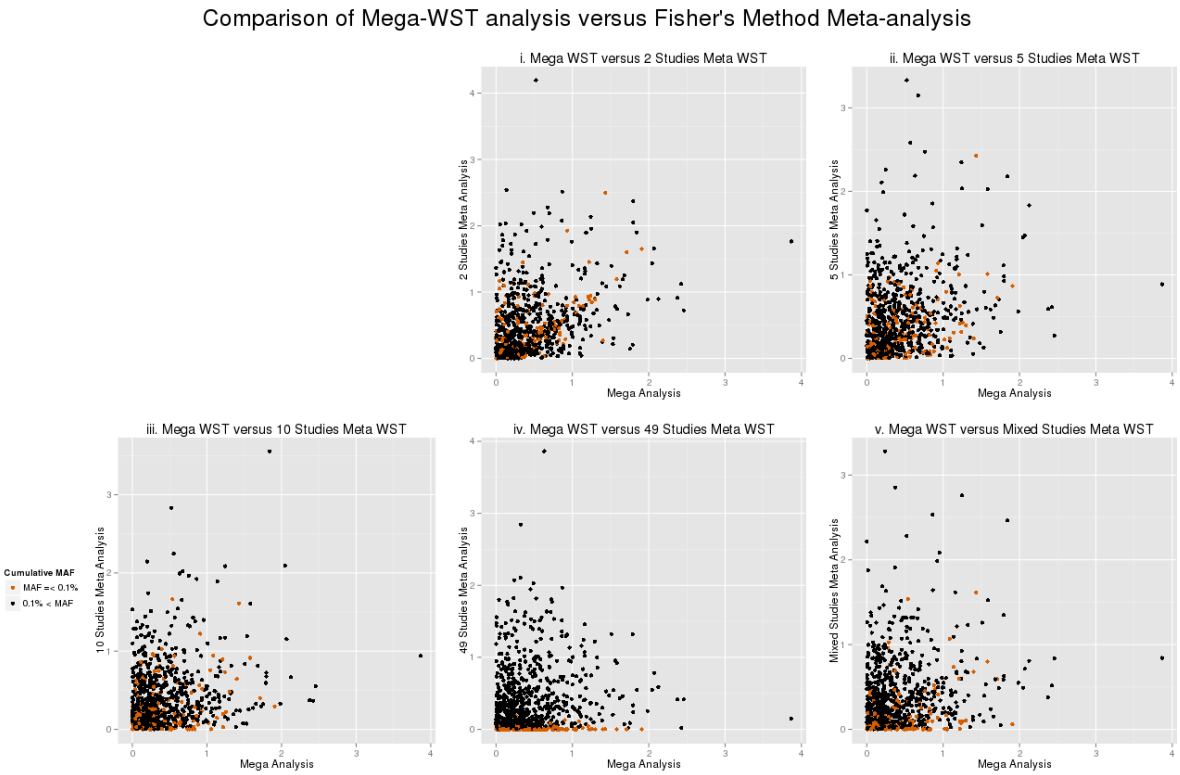
Appendix Figure B-3: Comparison of P-values from the mega-analysis, and each Fisher's Method meta-analysis scenario for the SKAT analysis of Smoking. Genes with Cumulative MAF<0.1% highlighted.



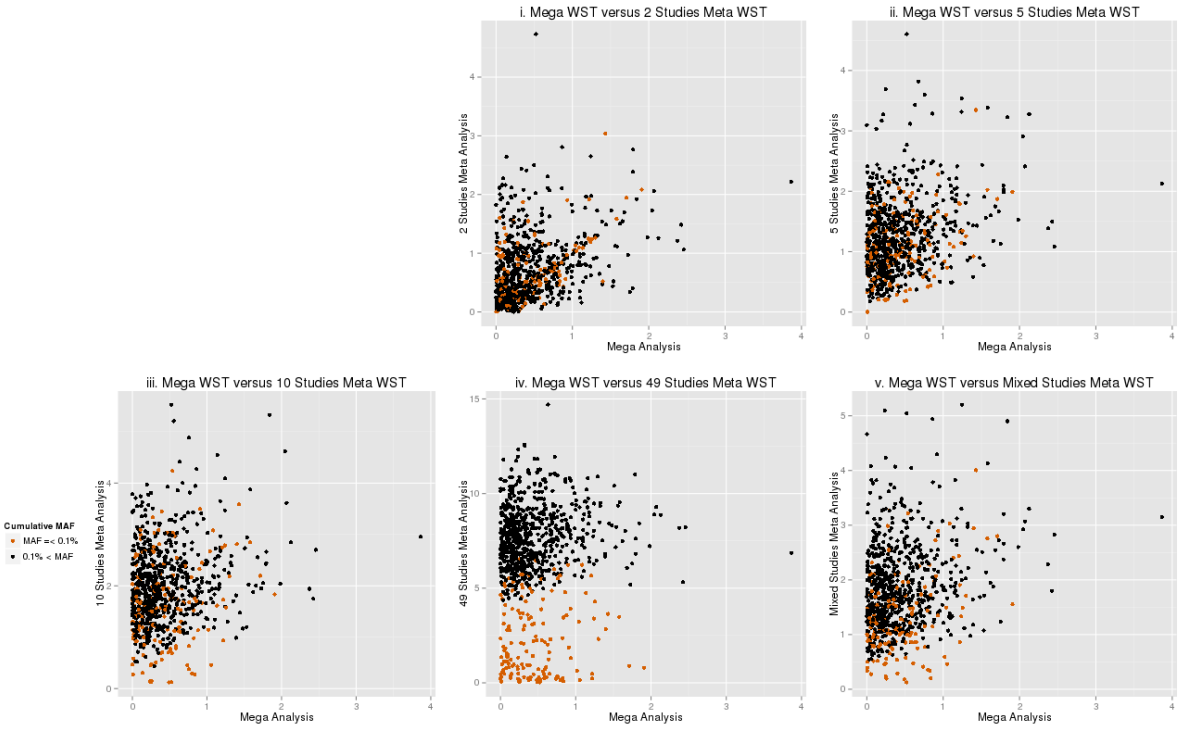
Appendix Figure B-4: Comparison of P-values from the mega-analysis, and each Z-score meta-analysis scenario for the SKAT analysis of smoking. Genes with Cumulative MAF<0.1% highlighted.



Appendix Figure B-5: Comparison of P-values from the mega-analysis, and each Fisher's Method meta-analysis scenario for the WST analysis of Smoking. Genes with Cumulative MAF<0.1% highlighted.



Appendix Figure B-6: Comparison of P-values from the mega-analysis, and each Z-score meta-analysis scenario for the WST analysis of smoking. Genes with Cumulative MAF<0.1% highlighted.



C. Analysis plans for the meta-analysis of exome array data and quantitative lung function traits (Chapter 4)

1. Analysis plan for study-level analysis 1: “SpiroMeta Exome Chip Analysis Plan”
2. Analysis plan for study-level analysis 2: “SpiroMeta Exome Chip Gene-based Analysis Plan”

SpiroMeta Exome Chip Analysis Plan

Draft: 8th May 2013

This document has been adapted from the generic quantitative traits exome chip analysis plan[1] which describes the cohort-specific analysis strategy for quantitative traits where genotype data is available from the exome chip. The focus is on SINGLE-VARIANT analyses as a "fast-track" plan, whilst gene-based burden tests are being finalised.

SPIROMETA SPECIFIC - GENOTYPE CALLING & QC

Gencall has been found to perform poorly when calling very rare genotypes, in particular by calling singleton heterozygotes as missing. Consequentially, cohorts will be asked to run the genotyping algorithm zCall (Goldstein,J.I. et al. 2012) to assign genotypes called as missing by Gencall. The ***Exome-chip Quality Control SOP (Version 5)*** *outlines the full genotype calling and QC procedure, which in brief includes:*

- *Running of Gencall*
- *Initial QC on Gencall data*
- *Callibration of zCall*
- *Running of zCall*
- *Secondary QC on zCall data*

In terms of post-calling QC, we ask each cohort to carry out the following QC checks:

- *Identification of SNPs with an excess of Mendelian inconsistencies (this only applies to data with parent-offspring pairs available)*
- *Identification of monomorphic SNPs.*

We ask that cohorts submit a SNP QC report, for ***all*** *SNPS, as a tab-delimited txt format file, with fields named and formatted as follows:*

Markername: *as produced by the software used.*
Min_all*: minor allele (a single character: "A" "C" "G" "T" for SNPs and "R", "I" or "D" for INDELS and CNVs)*
Maj_all*: major allele (a single character: "A" "C" "G" "T" for SNPs and "R", "I" or "D" for INDELS and CNVs)*
Freq*: allele frequency for minor allele (numeric data, 4 decimal places)*
MAC: *minor allele count(integer)*
Call_rate*: SNP call rate (numeric data, 4 decimal places)*
HWE_P*: Hardy Weinberg equilibrium P-value (scientific notation, coefficient to 3 decimal places)*

[1] **Generic Quantitative Traits Exome Chip Analysis Plan Draft: 25th October 2012**. Initial document drafted by Andrew Morris, Cecilia Lindgren, Anuhba Majahan (Oxford), and subsequently adapted after discussion by wider UK exome chip analysts and PIs.

Cluster_sep: *Cluster separation score (numeric data, 2 decimal places)*
Mono: *Indicator variable for monomorphic SNPs (1=monomorphic SNP; 0=otherwise)*
Men_Incon: *For cohorts with parent-offspring pairs only -Indicator variable for Mendelian Inconsistencies (1=excess of Mendelian inconsistencies; 0=otherwise)*

- *Indicator variables and the MAC field should be a single integer with no decimal point.*
- *Please code missing values as a single dot character (".").*
- *Note that no quotes should be used around any data cells or headers.*

Please use the following name for the SNP QC report file:
*SNP_QC_**COHORTNAME_VERSION**.txt*
where:
***COHORTNAME** will be an identifier for the specific cohort*
***VERSION** will be the date of the day of the uploading (ddmmyy)*

Cohort analysts should perform analyses of traits as described in **Steps 1-4** of the analysis plan below. If you have any questions or problems with implementing these cohort-specific steps in the analysis plan, please contact the central analyses email for generic analyses/QC issues and the trait specific groups for more trait-specific queries. Trait specific analysis teams should then perform meta-analyses of the submitted association summary statistics for each trait, but feedback to central analysts group if they spot issues of generic interest. Guidelines in **Steps 5-6** of the analysis plan.

1. **Cohort-specific "kinship" analyses in non-family based studies**

SPIROMETA SPECIFIC – KINSHIP ANALYSIS

This step of the analysis should be carried out in studies with unrelated individuals only.

The generic QT analysis plan recommends the use of PLINK for kinship analysis, as outlined below. Cohorts may carry out an equivalent analysis using ancestry principal components, estimated using eigenstrat or equivalent software, but are encouraged to liaise with the analysis working group to confirm that comparable methods are being used. Please ensure this analysis is carried out with the subset of LD pruned SNPs, with MAF>0.01.

If the threshold PI_HAT>0.2, suggested below results in the exclusion of a large number of samples, cohorts should liase with the central analysis team regarding alternative approaches to dealing with cryptic relatedness.

Calculate identity by state (IBS) between each pair of samples on the basis of an LD pruned (r2<0.2) set of markers at MAF >0.01 passing QC. From these data, calculate the proportion of the genome shared IBD (PI_HAT). Within PLINK, this can be achieved using the `--genome`

option. For each pair of samples with PI_HAT>0.2 remove the sample with lowest call rate on the basis of all variants passing QC. Repeat this process until only "unrelated" samples remain (i.e. PI_HAT≤0.2 for all pairs).

For all "unrelated samples", perform multidimensional scaling (MDS) on the basis of the IBS calculated from the LD pruned MAF >0.01 markers. Record the projection of each sample onto the first ten principal components (PCs) for use in downstream analyses to adjust for population structure. Final decisions on whether to exclude samples is likely to be cohort specific, but clear European outliers should be excluded from European based studies at a minimum. Within PLINK, this can be achieved using the `--cluster --mds-plot 10` options.

2. Cohort-specific trait transformations

Inverse rank normalisation of the residuals are recommended to reduce the impact of deviations from normality on trait associations (most often observed for rare variants). The pipeline for generation of transformed traits is thus as follows:

Raw values -> Exclusions -> Residuals after covariate adjustment -> Inverse rank normalisation

SPIROMETA SPECIFIC – TRAIT TRANSFORMATIONS

ADJUSTMENTS – No adjustments to the raw trait values need to be made, prior to exclusions.

ADULT COHORTS: Undertake exclusions & trait transformation for FEV1, FEV1/FVC and FVC, for the following subsets:

1. *Never-smokers:*
   - *Exclusions – Restrict dataset to never-smokers only with complete data on both FEV1 and FVC.*
   - *Covariate adjustments – Undertake linear regression of age, age2, sex & height*
2. *Ever-smokers:*
   - *Exclusions – Restrict dataset to ever-smokers only with complete data on both FEV1 and FVC.*
   - *Covariate adjustments – Undertake linear regression of age, age2, sex & height*
3. *Ever-smokers with pack-years:*
   - *Exclusions – Restrict dataset to ever-smokers only with complete data on FEV1,FVC and pack-years smoking.*
   - *Covariate adjustments – Undertake linear regression of age, age2, sex, height & pack-years*

CHILDREN'S COHORTS: Undertake exclusions & trait transformation for FEV1, FEV1/FVC and FVC, for all individuals:

1. *All samples:*

- *Exclusions – Restrict dataset to samples with complete data on both FEV1 and FVC.*
- *Covariate adjustments – Undertake linear regression of age, age2, sex & height*

INVERSE RANK NORMALISATION: For all subsets, transform residuals after covariate adjustment to ranks and then to normally distributed z-scores. These inverse-normal transformed residuals are then used as the phenotype for SNP association testing under an additive genetic model.

We recommend that traits are NOT adjusted for principal components to adjust for population structure because: (i) they can be included in downstream single-variant and burden test analyses; and (ii) they are not necessary for kinship-based association analyses (for example EMMAX & GEMMA).

SPIROMETA SPECIFIC - DESCRIPTIVE STATISTICS

Cohorts will be asked to provide information on the distribution (range, mean, sd) of FEV1, FEV1/FVC, FVC, age, sex, height, smoking status, pack-years of smoking, numbers diagnosed with asthma, COPD (across all individuals, not for subset analyses at this stage).

Cohorts will also be asked to provide mean, sd and histograms of trait residuals (FEV1, FEV1/FVC and FVC) from linear regression after adjusting for covariates age, age2, sex & height .Please provide these for males and females separately and for all individuals combined.

We have asked for copies of the questionnaires used to collect smoking data, and additional information where needed so that we can assess the consistency of approaches used.

3. Cohort-specific single-variant analyses in non-family based studies

SPIROMETA SPECIFIC – ASSOCIATION ANALYSIS SOFTWARE

The generic QT analysis plan recommends the use of EPACTS or PLINK for single variant association analysis, as outlined below. Cohorts wishing to use different packages for association testing are encouraged to liaise with the analysis working group to confirm that consistent approaches are employed and that consistent output is available.
After the association test has been undertaken, GWAtoolbox (www.eurac.edu/GWAtoolbox.html) can be used to quickly check the quality of the results before uploading.

SPIROMETA SPECIFIC – PRIORITISATION OF TRAITS

The analysis should ideally be undertaken on all three lung function traits, however if the deadline is not manageable for all traits, please prioritize in the order FEV1, FEV1/FVC and FVC. In addition, the analysis of the autosomal chromosomes should take priority over that of the X-chromosome.

Test for association of each inverse rank normalised trait with each **autosomal** variant passing QC (irrespective of minor allele count). For each trait, fit a linear regression model assuming an additive effect of the minor allele. Within EPACTS, this can be achieved with the `--test q.linear` option. Within PLINK, this can be achieved using the --linear option. Remeber to adjust for PCs as covariates to account for population structure, as necessary.

Test for association of each inverse rank normalised trait with each **X-chromosome** variant passing QC (irrespective of minor allele count) **in males and females separately**. For each trait, fit a linear regression model assuming an additive effect of the minor allele, in the same way as for autosomal analyses. Code males as 0/2 not 0/1, as females inactivate one X chromosome making homozygous females the equivalent dosage as hemizygous males.

SPIROMETA SPECIFIC – FILE NAMING SCHEME

For each trait and cohort subset, please prepare three files containing unformatted output produced by the software analysis: (i) for all autosomal variants passing QC (sex-combined analysis); (ii) for all X chromosome variants passing QC (male-specific analysis); and (iii) for all X chromosome variants passing QC (female-specific analysis). Please use the following naming convention for files:

***STUDY_ANALYSIS_TRAIT**_EXOMECHIP_SINGLE_AUTOSOMES_**ANALYST_DATE.**txt*
***STUDY_ ANALYSIS _TRAIT** _EXOMECHIP_SINGLE_XMALES_ **ANALYST_DATE**.txt*
***STUDY_ ANALYSIS _TRAIT** _EXOMECHIP_SINGLE_XFEMALES_ **ANALYST_DATE**.txt*

In these filenames:

***STUDY** is replaced with a short name or acronym for the study.*
***ANALYSIS** is replaced with "smk", "smkPY" (for the pack-years adjustment), "nonsmk", or "all" (for childrens cohorts where no stratification for smoking is carried out)* .
***TRAIT** is replaced wiht "FEV1","FVC" or "FF" (for the ratio FEV1/FVC).*
***ANALYST** is replaced with the initials of the analyst.*
***DATE** is replaced with the date of the analysis in the form DDMMYY.*

For example:

***ILFGC_smk_FEV1** _EXOMECHIP_SINGLE_AUTOSOMES_**APM_051012**.txt*
***ILFGC_smkPY _FVC** _EXOMECHIP_SINGLE_XMALES_**APM_051012**.txt*
***ILFGC_nonsmk _FF** _EXOMECHIP_SINGLE_XFEMALES_**APM_051012**.txt*

SPIROMETA SPECIFIC – POST-ASSOCIATION TESTING QC

To ensure that appropriate quality controls have been applied, after the single-variant association tests have been undertaken on the genotype data, QQplots and lambdas should be

generated seperately for SNPs with MAF≥0.05, 0.05>MAF≥0.01 and MAF<0.01 and checked for any gross deviations.
In future, cohorts may be required to inspect the cluster plots of any significant results to ensure those SNPs have not been incorrectly called by Gencall (and therefore not called by z-call).

4. Cohort specific single variant analyses in family based studies.

SPIROMETA SPECIFIC – SINGLE VARIANT ANALYSES IN RELATED INDIVIDUALS

Please test for association with the inverse rank normalised trait (i) for all autosomal variants passing QC (sex-combined analysis); (ii) for all X chromosome variants passing QC (male-specific analysis); and (iii) for all X chromosome variants passing QC (female-specific analysis).

*Given that the analysis should be carried out on the transformed trait and that GEMMA takes into account population structure in its analysis, no adjustments should be needed with the –c flag. Please also use the **–maf 0** flag, so that no MAF filter is applied.*

Please refer to section 3 for details of trait prioritisation, file naming scheme and post-association testing QC.

We recommend the use of GEMMA for single-variant analyses in cohorts with related individuals. Full details of the GEMMA software can be found at:

http://home.uchicago.edu/xz7/software.html

GEMMA accepts plink format files as input. Initially, calculate a centered relatedness matrix by running "gemma –bfile <inputfilename> –gk 1 –o <matrixfilename>". By default, only polymorphic SNPs that have missingness below 5% and minor allele frequency above 1% will be used to estimate the relatedness matrix.

Then perform association of each inverse rank normalised trait with each variant by running "gemma –bfile <inputfilename> –n 1 –k <matrixfilename .cXX.txt> -maf 0 –fa 4 –o <outputfilename>".

One can specify a different column in the .fam file as phenotype column by using "-n [num]", where "-n 1" uses the original sixth column and "-n 2" uses the seventh column as phenotypes etc.

GEMMA can produce Wald, score and LRT test results. By adjusting the number after flag "-fa", you can specify your choice. Above we request all three tests to be run. Asymptotically, these tests are equivalent. It may be worth running only the LRT for small sample sizes and/or low frequency variants.

By default, only polymorphic SNPs that have missingness below 5% and minor allele frequency above 0.01 will be tested. These threshold can be changed using the "-miss" and "-maf" flags.

Both missingness and minor allele frequency of a given SNP are calculated based on analysed individuals (i.e. individuals with no missing phenotypes and no missing covariates).

SpiroMeta Exome Chip Gene-based Analysis Plan

Draft: 4th June 2014

Single variant association analysis has proved successful in identifying common SNPs which influence complex disease; for rare variants however, it is somewhat underpowered. A number of gene-based, or region-based methods have been developed as a more powerful tool to detect rare variants associated with a trait. Several burden tests have been proposed, which combine information from several variants within a specified genomic region into a single quantity, which is then used for association testing with the trait; these methods perform well where variants within a gene have similar effects on the trait, in terms of direction and magnitude. Other proposed methods such as the sequence kernel association test (SKAT) utilise variance-components models and are powerful where a region has a combination of protective, deleterious and neutral variants.
Several packages have been developed to facilitate the meta-analysis of these gene-based tests; within SpiroMeta, we shall be using RAREMETAL developed by Liu et al. (Nature Genetics 46, 200–204 (2014)). RAREMETAL allows for a number of burden tests and SKAT to be carried out, as well as conditional analyses. Further to this, it allows for different genomic regions to be specified centrally, without further analyses at study level.

To run RAREMETAL, we require from each study, score statistics for each variant, along with covariance matrices. This analysis plan outlines how these statistics can be generated using either RAREMETALWORKER, or rvtests. Studies are free to choose whichever of these two softwares to use.

PHENOTYPES

Phenotypes for the gene-based analyses should be identical to those used in the single variant analyses. Required phenotypes are fully described in the single variant analysis plan, but in brief:

Undertake trait transformation for FEV_1, FEV_1/FVC and FVC, for the following subsets:

1. *Never-smokers, adjusted for age, age^2, sex & height*
4. *Ever-smokers, adjusted for age, age^2, sex & height*
5. *Ever-smokers with pack-years, adjusted for age, age^2, sex, height & pack-years*

For all subsets and traits, transform residuals after covariate adjustment to ranks and then to normally distributed z-scores. These inverse-normal transformed residuals are then used as the phenotype. The first ten principal components should be used as covariates when running analyses using RAREMETALWORKER / rvtests.

RAREMETALWORKER and rvtests require the following input files:

RAREMETALWORKER:
- VCF
- PED
- DAT

RVTESTS:
- VCF
- phenotype file
- covariate file

Note- RAREMETALWORKER may be run without the vcf file, with genotypes in the PED file instead; we strongly recommend running with the vcf file as the checkvcf script helps to ensure consistency of alleles/positions etc. across studies.

1. VCF file (Both RAREMETALWORKER and RVTESTS)

To create vcf files from plink files, we recommend using PLINK/SEQ. The vcf file should include genotypes for all autosomal and X chromosome SNPs.
Please note:

- There are a number of SNPs on the exome chip which share chromosomal positions; this has been found to cause some problems when creating the vcf using PLINK/SEQ. The file "duplicate_sites.txt" (circulated with this analysis plan), provides a list of 833 sites for which there are two SNPs for the Illumina Human Exome Beadchip v.1. For each site, the file gives the SNP id and alleles for the SNP to preserve (identified as was present in annotatedList.txt file from exome chip FTP site ftp://share.sph.umich.edu/exomeChip/IlluminaDesigns/) and the corresponding SNP to exclude. The file "SNPs_to_exclude.txt" may be used with the --remove flag in Plink to remove these 833 SNPs from the plink files, prior to the conversion to vcf. If studies are using an alternative version of the exome chip, we ask that duplicate SNPs be identified and removed before converting data to vcf (in this case, **please upload details of duplicate sites and excluded SNP ids with results**).
- To ensure the chromosome X genotypes are coded correctly in the vcf file, all individuals must have sex correctly coded in the Plink fam file.

Full PLINK/SEQ installation instructions and documentation can be found here:
https://atgu.mgh.harvard.edu/plinkseq/start-pseq.shtml

First, PLINK/SEQ may be downloaded and installed using:

```
wget http://atgu.mgh.harvard.edu/plinkseq/dist/version-0.08/plinkseq-0.08-x86_64.tar.gz
tar -xvzf plinkseq-0.08-x86_64.tar.gz
```

The PLINK/SEQ files must be in command path to run; this may be done by modifying the PATH environment variable:

```
export PATH=$PATH:/path/to/plinkseq-0.08-x86_64
```

Create new PLINK/SEQ project:

```
pseq projectname new-project
```

Load Plink files:

```
pseq projectname load-plink --file Plinkfilesname --id iid
```

Create vcf:

```
pseq projectname write-vcf > vcfname.vcf
```

The vcf file must then be edited: "chr" must be removed from the first from first field in vcf file; chromosome X must be labelled X rather than 23, as in Plink and the file compressed using bgzip:

```
sed 's/^chr//' vcfname.vcf | sed 's/^23/X/' | bgzip -c >
vcfname.vcf.gz
```

Use checkVCF python script to check VCF (Download script and reference genome from: http://genome.sph.umich.edu/wiki/CheckVCF.py)

```
python checkVCF.py -r hs37d5.fa -o outputname vcfname.vcf.gz
```

This script will create a number of files, listing SNPs which are monomorphic (*outputname.check.mono*), SNPs where the expected minor allele has a frequency >0.5 (*outputname.check.af*), a list of non-SNP sites, largely indels (*outputname.check.nonSnp*) and genotypes which are not found or are incorrectly formatted (*outputname.check.geno*). It is worth checking these files to identify any issues from the conversion of the data to vcf, but listed variants need not necessarily be excluded or changed. A final file (*outputname.check.ref)* lists SNPs whose reference allele does not match with the required reference allele. Sites will be listed as:

```
MismatchRefBase   1:564766:T-C/T
```

where T is the expected reference allele and C/T are the reference/non-reference allele in the vcf file. For these SNPs, the reference allele should be changed using Plink, by forcing the reference allele (using the --reference-allele command) and /or by strand flipping (--flip command). After changing the reference alles, a new vcf file should be created and again checked using the checkVCF.py. This process should be repeated until the vcf file is consistent with the reference genome, and the checkVCF script outputs the message "*No error found by checkVCF.py, thank you for cleanning VCF*".

2. PED & DAT files (RAREMETALWORKER only)

For each of the three subsets (never smokers, ever smokers and ever smokers with pack-years), ped files containing the phenotypes and covariates must be created. The ped file should include the first five columns of the plink ped file (FamID, IndID, PatID, MatID, Sex), followed by three phenotype columns with the **inverse transformed residuals of each trait** (FEV_1, FVC, FEV_1/FVC), followed by ten columns of covariates (PC1-PC10). The ped file should have no column headers and individuals must be listed in the same order as they appear in the vcf file. Missing values should be coded "NA".

Example ped:

```
FAM1 AB1  0   0   1    2.564     1.434     1.657    -0.01209   0.01324   …     0.008585
FAM2 CD2  0   0   2    1.456     1.336     0.267    -0.009965 -0.00645   …    -0.005685
FAM3 FW5  0   0   2    3.211     2.876     3.022     0.00015  -0.02151   …    -0.000365
FAM4 JK7  0   0   1   -1.654    -0.5164   -0.9634   -0.00454   0.00654   …     0.006354
```

The corresponding dat file describes the ped file; the second column should list trait and covariate names, with the first column coded T for trait and C for covariate. The dat file must have traits and covariates listed in the order they appear as columns in the ped file, have no column headers, and it should not describe the first five rows, as these are always the same.

Corresponding example dat:

```
T      FEV1
T      FVC
T      FF
C      PC1
C      PC2
⋮      ⋮
C      PC10
```

3. Phenotype & Covariate files (RVTESTS only)

For each of the three subsets (never smokers, ever smokers and ever smokers with pack-years), phenotype and covariate files must be created. These files both include the first five columns of the plink ped file. In the phenotype file, these are followed by three phenotype columns with the **inverse transformed residuals of each trait** (FEV_1, FVC, FEV_1/FVC). In the covariates file, the five ped file fields are followed by ten columns of covariates (PC1-PC10). We advise the use of headers in these files, with the first two columns to be labelled "FID IID". Missing values should be coded "NA".

Example phenotype file:

FID	IID	PATID	MATID	SEX	FEV1	FVC	FF
FAM1	AB1	0	0	1	2.564	1.434	1.657
FAM2	CD2	0	0	2	1.456	1.336	0.267
FAM3	FW5	0	0	2	3.211	2.876	3.022
FAM4	JK7	0	0	1	-1.654	-0.5164	-0.9634

Example covariate file:

FID	IID	PATID	MATID	SEX	PC1	PC2	…	PC10
FAM1	AB1	0	0	1	-0.01209	0.01324	…	0.008585
FAM2	CD2	0	0	2	-0.009965	-0.00645	…	-0.005685
FAM3	FW5	0	0	2	0.00015	-0.02151	…	-0.000365
FAM4	JK7	0	0	1	-0.00454	0.00654	…	0.006354

RUNNING RAREMETALWORKER

RAREMETALWORKER may be downloaded from here: http://genome.sph.umich.edu/wiki/RAREMETALWORKER . This wiki page also provides full instructions for the installation and running of RAREMETALWORKER.

Prior to running RAREMETALWORKER, the cleaned vcf file will need to be tabix indexed using:

```
tabix -p vcf vcfname.vcf.gz
```

1. Studies with unrelated individuals

RAREMETALWORKER may be ran with the ped, dat and vcf files as input, using the following command:

```
raremetalworker --ped pedname.ped --dat datname.dat --vcf
vcfname.vcf.gz --makeResiduals --prefix outputname
```

This will carry out the RAREMETAL analysis for all autosomal and chromosome X SNPs, with each of the three traits in turn; the *--makeResiduals* flag specifies that the trait is to be adjusted using the 10PCs, before the linear models are fitted using the resulting residuals.

2. Studies with family data

RAREMETALWORKER has several ways of dealing with family data: either familial relationships may be described in the ped file; alternatively a kinship matrix may be calculated from genotype data, or be read from an existing file.

We recommend that studies run analysis of related individuals, using an empirical kinship matrix, as follows:

```
raremetalworker --ped pedname.ped --dat datname.dat --vcf
vcfname.vcf.gz --makeResiduals --kinGeno --kinMAF 0.01 --kinMiss
0.05 --vcX --prefix outputname
```

This will carry out the RAREMETAL analysis for all autosomal and chromosome X SNPs, with each of the three traits in turn. The `--makeResiduals` flag specifies that the 10PCs are to be used as covariates. To account for relatedness, the `--kinGeno` command creates a kinship matrix using variants with MAF>0.01 (`--kinMAF`) and genotype missing rate <0.05 (`--kinMiss`), with the `--vcX` flag indicating an additional kinship matrix be generated for the non-pseudoautosomal region of chromosome X.

RUNNING RVTESTS

Rvtests may be downloaded from here: http://genome.sph.umich.edu/wiki/Rvtests.

1. Studies with unrelated individuals

Rvtests should be ran with the vcf, phenotype and covariate files, using the `--meta score,cov` command, as follows:

```
rvtest --inVcf vcfname.vcf.gz --pheno phenofilename.txt --pheno-
name FEV1 --covar covarfilename.txt --covar-name
PC1,PC2,PC3,PC4,PC5,PC6,PC7,PC8,PC9,PC10 --meta score,cov --out
outputname
```

The above command would generate a score vector and covariance matrix for all autosomal and chromosome X SNPs, for the FEV_1 phenotype, with adjustment for the first 10 PCs. The `--pheno-name` and `--covar-name` flags specify which of the phenotypes / covariates are to be used from the files specified by `--pheno` and `--covar,` respectively.

2. Studies with family data

Before running rvtests on related individuals, a kinship matrix must first be created using the vcf2kinship command. This kinship matrix may be created either from known pedigree information in a ped file, or by creating an empirical kinship matrix from the vcf file.

We recommend creating an empirical kinship matrix from genotype data as follows:

```
vcf2kinship --inVcf vcfname.vcf.gz --ped phenofilename.txt --bn
--xHemi --minMAF 0.01 --maxMISS 0.05 --out kinshipname
```

The *--bn* flag indicates the kinship will be calculated using the Balding-Nicols method, using variants with MAF>0.01 (*--minMAF*) and genotype missing rate <0.05 (*--maxMiss*). The *--xHemi* flag is required to create an additional kinship matrix for the hemizygote region of chromosome X. To utilise the *-xHemi* flag, vcf2kinship requires the phenotype file to be specified using *--ped*, which should have sex listed in the fifth column.

Once the kinship matrix has been created, rvtests may be run as follows, using the *--meta score,cov* command:

```
rvtest --inVcf vcfname.vcf.gz --pheno phenofilename.txt --pheno-
name FEV1 --covar covarfilename.txt --covar-name
PC1,PC2,PC3,PC4,PC5,PC6,PC7,PC8,PC9,PC10 --kinship
kinshipname.kinship --meta score,cov --out outputname
```

This will generate a score vector and covariance matrix for all autosomal and chromosome X SNPs, for the FEV_1 phenotype, with adjustment for the first 10 PCs. The *--kinship* flag should be used to specify the name of the main kinship file created using the vcf2kinship command. For analysis of chromosome X SNPs, it is not necessary to additionally specify the name of the X hemizygote kinship matrix; if ***kinshipname.****kinship* is specified using the *--kinship* flag, rvtests will automatically attempt to use ***kinshipname.****xHemi.kinship* for the chromosome X analysis.

DATA FOR UPLOAD

1. **RAREMETALWORKER output**

After running RAREMETALWORKER, the following files will have been generated:

outputname.traitname.*singlevar.score.txt*
outputname.traitname.*singlevar.cov.txt*
outputname.*singlevar.log*
outputname.traitname.plots.pdf

For each subgroup and trait, we ask for all unformatted output files and plots to be uploaded, using the following naming scheme:

STUDY_ANALYSIS_TRAIT*_EXOMECHIP_GENE-BASED_****ANALYST_DATE.****score.txt*
STUDY_ANALYSIS_TRAIT*_EXOMECHIP_GENE-BASED_****ANALYST_DATE.****cov.txt*
STUDY_ANALYSIS_TRAIT*_EXOMECHIP_GENE-BASED_****ANALYST_DATE_****RMW.****log*
STUDY_ANALYSIS_TRAIT*_EXOMECHIP_GENE-BASED_****ANALYST_DATE.****plots.pdf*

In these filenames:

STUDY *is replaced with a short name or acronym for the study.*
ANALYSIS *is replaced with "smk", "smkPY" (for the pack-years adjustment), "nonsmk".*
TRAIT *is replaced with "FEV1","FVC" or "FF" (for the ratio FEV1/FVC).*
ANALYST *is replaced with the initials of the analyst.*
DATE *is replaced with the date of the analysis in the form DDMMYY.*

For example:

B58C_smk_FEV1 *_ EXOMECHIP_GENE-BASED _****VEJ_080414****.score.txt*
B58C_smk_FEV1 *_ EXOMECHIP_GENE-BASED _****VEJ_080414****.cov.txt*
B58C_smk_FEV1 *_ EXOMECHIP_GENE-BASED _****VEJ_080414_****RMW.log*
B58C_smk_FEV1 *_ EXOMECHIP_GENE-BASED _****VEJ_080414****.plots.pdf*

2. Rvtests output

After running rvtests, the following files will have been generated:

outputname`.MetaScore.assoc`
outputname`.MetaCov.assoc.gz`
outputname`.log`

For each subgroup and trait, we ask for the unformatted .MetaScore.assoc and the .MetaCov.assoc.gz files to be uploaded, using the following naming scheme:

STUDY_ANALYSIS_TRAIT_EXOMECHIP_GENE-BASED_***ANALYST_DATE.***MetaScore.assoc
STUDY_ANALYSIS_TRAIT_EXOMECHIP_GENE-BASED_***ANALYST_DATE.***MetaCov.assoc.gz
STUDY_ANALYSIS_TRAIT_EXOMECHIP_GENE-BASED_***ANALYST_DATE_***RVTESTS.log

In these filenames:

STUDY *is replaced with a short name or acronym for the study.*
ANALYSIS *is replaced with "smk", "smkPY" (for the pack-years adjustment), "nonsmk".*
TRAIT *is replaced with "FEV1","FVC" or "FF" (for the ratio FEV1/FVC).*
ANALYST *is replaced with the initials of the analyst.*
DATE *is replaced with the date of the analysis in the form DDMMYY.*

For example:

B58C_smk_FEV1 _ EXOMECHIP_GENE-BASED _***VEJ_080414***.MetaScore.assoc
B58C_smk_FEV1 _ EXOMECHIP_GENE-BASED _***VEJ_080414***.MetaCov.assoc.gz
B58C_smk_FEV1 _ EXOMECHIP_GENE-BASED _***VEJ_080414_***RVTESTS.log

3. QC files

Please upload the output files generated by the checkvcf script for the **final** vcf file:

STUDY.check.log
STUDY.check.dup
STUDY.check.noSnp
STUDY.check.ref
STUDY.check.geno
STUDY.check.af
STUDY.check.mono

If studies are not using the Illumina Human Exome BeadChip v.1 version of the exome chip, please upload a list of duplicate sites, with the ids of SNPs which were removed prior to converting the genotype data to vcf (see creating VCF file section of plan for details).

D. Additional Results for the meta-analysis of exome array data and quantitative lung function traits (Chapter 4)

Appendix Table D-1: Replication results for all SNPs identified in single variant association discovery analyses (P<10-4). Only variants in novel loci shown, and variants were only followed up for the trait for which they were most significantly associated. Chromosome (CHR) and position (POS) in build 37 are given for each SNP. For the discovery analysis, UK BiLEVE (replication stage 1) and UKHLS (replication stage 2), beta values reflect effect-size estimates on an inverse-normal transformed scale after adjustments for age, age^2, sex, height and ancestry principal components, and stratified by ever smoking status. For the CHARGE Consortium (replication Stage 2), beta values represent untransformed trait effect estimates, after adjustment for former smoking, current smoking and pack-years of smoking, age, age^2, sex, height , height^2, centre/cohort, principle components and weight (FVC only).

					Discovery Analysis				Stage 1 Replication				Discovery + Stage 1 Replication meta	Stage 2 Replication								Discovery + Stage 1 + 2 Replication meta
					SpiroMeta Consortium				UK BiLEVE					UKHLS				CHARGE Consortium				
rs id	CHR	POS	Effect / other allele	Trait	N	Effect Allele Frequency (EAF)	Beta	P-value	N	EAF	Beta	P-value	P-value	N	EAF	Beta	P-value	N	EAF	Beta	P-value	P-value
rs201163722	1	28661302	T/C	FEV1	23386	0.01%	-2.347	5.29E-05	NA	NA	NA	NA	**5.29E-05**	7446	0.00%	NA	NA	36998	0.02%	-109.569	0.3850	2.57E-03
rs200090958	1	86951099	A/G	FEV1/FVC	23397	0.00%	2.783	9.67E-05	NA	NA	NA	NA	**9.67E-05**	7449	0.00%	NA	NA	36985	0.01%	-2.157	0.4780	5.75E-03
rs141436979	1	89844088	T/C	FEV1/FVC	23396	0.02%	-1.486	2.44E-05	NA	NA	NA	NA	**2.44E-05**	7449	0.01%	-0.7122	0.6911	36985	0.02%	-3.089	0.0903	7.32E-05
rs1192415	1	92077097	A/G	FVC	23376	81.24%	-0.059	6.11E-07	48943	81.02%	-0.0138	0.0872	**5.21E-05**	7448	81.75%	-0.0104	0.6252	36996	82.22%	-8.589	0.0925	2.52E-05
rs11204697	1	150658971	T/C	FEV1	23393	0.03%	1.060	7.62E-05	48943	0.01%	0.2483	0.5775	2.35E-04	-	-	-	-	-	-	-	-	-
rs35608243	2	174131392	C/T	FVC	23395	8.17%	0.081	1.60E-06	48943	8.43%	0.0041	0.7191	0.0048	-	-	-	-	-	-	-	-	-
rs148627602	2	209309610	A/G	FEV1	23397	0.03%	1.009	5.75E-05	48943	0.03%	-0.3466	0.0906	0.2177	-	-	-	-	-	-	-	-	-
rs144052038	3	49720010	G/A	FVC	23392	0.07%	0.755	2.59E-05	NA	NA	NA	NA	**5.46E-05**	7436	0.00%	NA	NA	33029	0.09%	-65.951	0.3484	7.71E-02
rs141921900	3	74334458	A/G	FEV1/FVC	23384	1.32%	-0.178	1.21E-05	48943	1.10%	0.0015	0.9606	0.0098	-	-	-	-	-	-	-	-	-
rs62290268	3	194790799	G/C	FEV1	23394	0.08%	0.681	5.47E-05	48943	1.29%	-0.0491	0.1551	0.5588	-	-	-	-	-	-	-	-	-
rs3733250	4	77192868	A/G	FVC	23392	41.02%	-0.044	2.71E-06	48943	40.41%	-0.0087	0.1830	3.29E-04	-	-	-	-	-	-	-	-	-
rs142127543	4	90833153	A/G	FEV1/FVC	23395	0.14%	0.522	2.09E-05	48943	0.19%	-0.0053	0.9470	0.0244	-	-	-	-	-	-	-	-	-
rs17037102	4	107845794	T/C	FEV1/FVC	23394	10.34%	-0.062	4.78E-05	48943	10.56%	0.0003	0.9742	0.0238	-	-	-	-	-	-	-	-	-
rs79300690	4	122250654	A/G	FEV1/FVC	23385	1.77%	0.143	4.44E-05	48943	1.13%	-0.0350	0.2464	0.0734	-	-	-	-	-	-	-	-	-
rs147517729	4	147561147	A/C	FVC	18294	1.46%	-0.202	4.56E-06	48943	1.40%	0.0141	0.6131	0.0647	-	-	-	-	-	-	-	-	-
rs772835	5	944298	G/A	FVC	23395	9.95%	-0.067	2.03E-05	48943	10.11%	-0.0232	0.0293	**4.89E-05**	7449	10.42%	-0.0355	0.1834	36996	10.20%	-3.347	0.6032	1.26E-03
rs17648108	5	177831556	C/T	FVC	12820	28.38%	0.046	6.93E-06	NA	NA	NA	NA	**1.62E-05**	7449	27.63%	0.0112	0.5393	36996	29.58%	-1.122	0.7941	1.10E-02
rs1294421	6	6743149	G/T	FEV1/FVC	23395	61.47%	0.037	9.14E-05	48943	60.78%	0.0314	1.59E-06	**6.78E-10**	7449	60.35%	0.0284	0.0892	36985	61.94%	0.203	1.46E-04	**1.12E-13**
rs3749903	6	42992825	G/C	FVC	23389	12.38%	-0.072	2.95E-05	48943	13.50%	-0.0088	0.3451	0.0073	-	-	-	-	-	-	-	-	-
rs9784763	6	109624937	A/G	FEV1	15224	39.24%	0.039	4.07E-05	48943	37.87%	0.0063	0.3401	0.0018	-	-	-	-	-	-	-	-	-
rs143974258	6	136552493	A/G	FEV1/FVC	22559	0.06%	0.740	7.35E-05	48943	0.13%	0.0029	0.9775	0.0518	-	-	-	-	-	-	-	-	-
rs57658073	8	24775940	A/G	FEV1	23297	0.28%	0.344	9.04E-05	48943	0.26%	-0.0400	0.5416	0.0642	-	-	-	-	-	-	-	-	-
rs146520900	8	145667730	A/G	FEV1	22392	0.38%	0.386	7.00E-07	48943	0.37%	-0.0913	0.1237	0.0743	-	-	-	-	-	-	-	-	-
rs141834891	9	12694063	T/C	FEV1/FVC	23397	0.08%	0.755	6.45E-06	48943	0.11%	-0.0697	0.5241	0.0532	-	-	-	-	-	-	-	-	-
rs2773347	9	100388197	T/C	FEV1	23391	67.49%	0.043	1.16E-05	48943	67.05%	0.0118	0.0835	**9.05E-05**	7449	67.08%	0.0278	0.1110	36998	66.51%	2.334	0.5176	1.18E-04

					Discovery Analysis				Stage 1 Replication				Discovery + Stage 1 Replication meta	Stage 2 Replication								Discovery + Stage 1 + 2 Replication meta
					SpiroMeta Consortium				UK BiLEVE					UKHLS				CHARGE Consortium				
rs id	CHR	POS	Effect / other allele	Trait	N	Effect Allele Frequency (EAF)	Beta	P-value	N	EAF	Beta	P-value	P-value	N	EAF	Beta	P-value	N	EAF	Beta	P-value	P-value
rs143386455	9	107533244	C/G	FEV1/FVC	23139	0.08%	0.648	8.81E-05	48943	0.11%	-0.0631	0.5924	0.0657	-	-	-	-	-	-	-	-	-
rs41278437	9	113170060	A/G	FEV1/FVC	23397	0.07%	-0.769	2.24E-05	48943	0.11%	0.0191	0.8501	0.0572	-	-	-	-	-	-	-	-	-
rs17578859	9	139879170	A/G	FEV1	23387	25.90%	0.048	5.95E-06	48943	25.81%	0.0122	0.0954	**8.41E-05**	7447	25.49%	0.0235	0.2052	36998	25.56%	-8.100	0.0366	2.45E-02
rs141541697	10	92635830	T/C	FEV1	23361	0.01%	1.897	3.30E-06	NA	NA	NA	NA	**3.30E-06**	-	-	NA	NA	36998	0.02%	-50.711	0.6490	1.67E-02
rs61736639	11	14891141	C/G	FEV1	23396	0.58%	0.238	8.58E-05	48943	0.66%	0.0599	0.1376	6.50E-04	-	-	-	-	-	-	-	-	-
rs188851356	11	125647897	A/G	FEV1/FVC	23397	0.09%	-0.718	1.00E-05	NA	NA	NA	NA	**1.00E-05**	7449	0.00%	NA	NA	36985	0.04%	1.497	0.2898	6.99E-02
rs187124232	11	126144859	G/C	FEV1/FVC	23392	0.08%	-0.780	7.61E-06	NA	NA	NA	NA	**7.61E-06**	7449	0.00%	NA	NA	36985	0.00%	1.232	0.7922	1.49E-02
rs35639297	12	56142553	T/G	FEV1/FVC	23396	0.48%	-0.278	3.26E-05	48943	0.48%	-0.0005	0.9915	0.0139	-	-	-	-	-	-	-	-	-
rs142653430	12	121469271	A/G	FEV1	23395	0.01%	-2.858	1.12E-06	NA	NA	NA	NA	**1.12E-06**	7449	0.00%	NA	NA	36998	0.00%	-139.237	0.5698	1.04E-03
rs201930455	12	129360559	A/G	FEV1	23396	0.00%	-3.011	3.35E-05	NA	NA	NA	NA	**3.35E-05**	7449	0.00%	NA	NA	-	-	NA	NA	3.35E-05
rs7984952	13	31231806	C/T	FVC	23394	40.72%	-0.039	4.14E-05	48943	40.55%	0.0016	0.8019	5.01E-02	-	-	-	-	-	-	-	-	-
rs3742302	13	31233063	A/G	FVC	23358	40.72%	-0.039	3.64E-05	48943	40.49%	0.0009	0.8931	3.85E-02	-	-	-	-	-	-	-	-	-
rs149470963	13	67477723	T/G	FEV1	12633	0.15%	-0.690	2.90E-05	48943	0.17%	-0.0307	0.7128	2.69E-02	-	-	-	-	-	-	-	-	-
rs11558436	14	32257065	C/A	FEV1	23397	0.61%	-0.247	3.40E-05	48943	0.50%	-0.0925	0.0641	**4.49E-05**	7448	0.61%	0.1047	0.3202	36998	0.55%	23.453	0.3063	6.35E-03
rs1952153	14	87775721	C/A	FVC	23390	57.67%	-0.039	3.29E-05	48943	57.43%	-0.0044	0.4966	0.0058	-	-	-	-	-	-	-	-	-
rs61991737	14	93712290	A/C	FEV1	23397	0.15%	0.469	5.85E-05	48943	0.17%	0.1019	0.2087	8.84E-04	-	-	-	-	-	-	-	-	-
rs118125046	15	79586782	G/C	FEV1/FVC	23376	0.75%	0.219	4.38E-05	48943	0.61%	-0.0283	0.5331	0.0304	-	-	-	-	-	-	-	-	-
rs3751093	17	25958304	A/G	FVC	12633	20.90%	-0.070	6.62E-06	48943	21.16%	-0.0155	0.0482	2.55E-04	-	-	-	-	-	-	-	-	-
rs144042976	19	37975803	A/G	FEV1	23386	0.02%	-1.439	5.06E-06	NA	NA	NA	NA	**5.06E-06**	7431	0.02%	0.3427	0.5524	36998	0.01%	235.498	0.1349	1.68E-01
rs149178822	19	40540724	C/A	FEV1/FVC	12631	1.26%	-0.233	5.05E-05	48943	1.49%	-0.0271	0.3041	0.0086	-	-	-	-	-	-	-	-	-
rs146608853	20	49225233	A/G	FEV1/FVC	23397	0.04%	-1.085	7.90E-06	48943	0.03%	0.1487	0.6516	8.75E-04	-	-	-	-	-	-	-	-	-
rs200373931	20	62193999	T/C	FVC	23381	0.02%	1.237	4.46E-05	NA	NA	NA	NA	**9.02E-05**	7449	0.03%	-1.0725	0.0164	33029	0.01%	-45.311	0.8051	1.70E-01
rs140025782	21	28216862	A/C	FEV1	23378	0.30%	-0.338	7.20E-05	48943	0.01%	-0.6916	0.1648	**3.37E-05**	7395	0.10%	0.1079	0.6762	36998	0.20%	2.965	0.9365	1.09E-02
rs35946782	21	40763754	A/G	FEV1/FVC	23397	0.02%	1.458	4.37E-06	48943	0.01%	-0.3965	0.2864	0.00509	-	-	-	-	-	-	-	-	-
rs77543787	22	33264982	T/C	FVC	23396	0.02%	1.547	4.07E-05	48943	0.38%	0.0296	0.6538	0.2742	-	-	-	-	-	-	-	-	-

Appendix Table D-2: Replication results for all SNPs identified in single variant association discovery analyses in ever smokers and never smokers separately ($P<10^{-4}$). Only variants in novel loci and that were not identified in the analyses of ever and never smokers combined are shown. Variants were only followed up for the trait for which they were most significantly associated. Chromosome (CHR) and position (POS) in build 37 are given for each SNP. Beta values reflect effect-size estimates on an inverse-normal transformed scale after adjustments for age, age^2, sex, height and ancestry principal components, and stratified by ever smoking status.

						Discovery Analysis				Stage 1 Replication				Discovery + Stage 1 Replication meta	Stage 2 Replication				Discovery + Stage 1 + 2 Replication meta
						SpiroMeta Consortium				UK BiLEVE					UKHLS				
rs id	CHR	POS	Effect / other allele	Trait	Smoking subset (Ever / never smokers)	N	Effect Allele Frequency (EAF)	Beta	P-value	N	EAF	Beta	P-value	P-value	N	EAF	Beta	P-value	P-value
rs201163722	1	28661302	T/C	FEV1/FVC	ever	11558	0.01%	-2.9956	2.35E-05	NA	NA	NA	NA	**2.35E-05**	4509	0.00%	NA	NA	2.35E-05
rs1414896	1	95692310	A/G	FVC	ever	11562	60.84%	-0.0561	4.14E-05	24460	60.36%	0.0125	0.1770	0.2262	-	-	-	-	-
rs2764504	1	119234198	C/T	FVC	ever	11560	5.89%	0.1133	5.78E-05	24460	5.99%	0.0407	0.0322	**5.26E-05**	4508	5.48%	1.3342	0.1821	2.07E-05
rs199581193	1	201175658	G/A	FVC	never	11418	0.01%	-2.8726	7.45E-07	24483	0.08%	0.4149	0.0203	0.3578	-	-	-	-	-
rs76611705	1	223177974	A/G	FEV1	never	11834	1.55%	-0.2065	8.09E-05	24483	1.61%	-0.0607	0.0907	2.74E-04	-	-	-	-	-
rs200091857	2	97008504	T/C	FEV1/FVC	never	11832	0.07%	-1.0461	1.91E-05	24483	0.09%	-0.0749	0.6736	0.0053	-	-	-	-	-
rs202022630	2	234750666	T/C	FEV1/FVC	ever	11562	0.01%	2.7461	9.20E-05	NA	NA	NA	NA	**9.20E-05**	4509	0.01%	-1.8633	0.0624	0.0198
rs147184138	3	25833094	G/C	FVC	ever	11563	0.09%	0.9031	7.47E-05	24460	0.08%	0.0261	0.8937	0.0186	-	-	-	-	-
rs61747991	3	56653424	T/A	FEV1/FVC	ever	11557	4.90%	0.1249	4.52E-05	24460	4.87%	0.0255	0.2256	9.35E-04	-	-	-	-	-
rs201934751	3	150404106	T/A	FEV1/FVC	ever	11563	0.01%	2.7604	6.75E-05	NA	NA	NA	NA	**6.75E-05**	4507	0.00%	NA	NA	6.75E-05
rs7652177	3	171969077	G/C	FEV1	ever	11558	46.38%	-0.0549	6.90E-05	24460	49.18%	-0.0153	0.0921	3.20E-04	-	-	-	-	-
rs7627615	3	183818416	A/G	FEV1	ever	11555	58.72%	-0.0561	3.54E-05	24460	59.19%	0.0175	0.0579	0.4357	-	-	-	-	-
rs144473454	5	9136617	A/G	FVC	ever	10440	0.27%	0.5816	8.92E-05	24460	0.08%	0.1050	0.5369	0.0065	-	-	-	-	-
rs4634319	5	27418887	G/A	FVC	ever	11564	6.89%	0.1058	5.43E-05	24460	6.81%	0.0011	0.9517	0.0195	-	-	-	-	-
rs1448044	5	44296986	A/G	FVC	ever	11550	31.53%	0.0563	8.41E-05	24460	32.34%	0.0402	3.31E-05	**1.62E-08**	4437	33.55%	0.8785	0.3796	**1.90E-08**
rs255888	5	111103258	T/C	FEV1	ever	11562	55.70%	0.0550	3.56E-05	24460	56.21%	0.0124	0.1776	5.54E-04	-	-	-	-	-
rs147752980	5	130791507	C/T	FEV1	never	11834	0.12%	0.8385	6.44E-06	NA	NA	NA	NA	**6.44E-06**	2940	0.15%	-0.0790	0.9371	6.27E-05
rs148279287	5	140768844	T/C	FVC	ever	10165	0.55%	-0.3905	4.16E-05	24460	0.48%	0.0111	0.8649	0.0291	-	-	-	-	-
rs6941356	6	87967636	G/A	FEV1/FVC	never	11833	10.97%	0.0833	7.48E-05	24483	9.99%	-0.0088	0.5600	0.0747	-	-	-	-	-
rs144830879	6	129649451	A/G	FEV1/FVC	never	11832	0.02%	1.9884	8.17E-05	24483	0.04%	0.0070	0.9777	0.0231	-	-	-	-	-
rs41298397	6	132891977	C/T	FEV1	ever	11562	0.40%	-0.4721	6.28E-06	24460	0.43%	0.1317	0.1032	0.2240	-	-	-	-	-
rs35839363	6	132909838	A/G	FVC	ever	11561	0.03%	1.4025	7.42E-05	24460	0.03%	-0.2007	0.4376	0.1085	-	-	-	-	-
rs13286541	9	113251951	C/T	FEV1/FVC	ever	11559	9.87%	0.0870	8.11E-05	24460	9.84%	0.0216	0.1550	0.0007	-	-	-	-	-
rs5030723	9	120476694	A/G	FEV1/FVC	ever	11553	0.31%	0.4750	7.12E-05	24460	0.36%	-0.1399	0.1161	0.3396	-	-	-	-	-
rs2296957	9	134401335	T/C	FEV1/FVC	never	10069	95.42%	-0.1336	8.65E-05	24483	95.21%	0.0044	0.8343	0.0386	-	-	-	-	-
rs7871194	9	139544437	C/A	FEV1/FVC	never	11830	57.04%	-0.0525	6.44E-05	24483	58.84%	0.0026	0.7765	0.0406	-	-	-	-	-
rs141660796	10	72360577	A/G	FEV1/FVC	ever	11563	0.06%	1.1473	3.65E-05	24460	0.02%	-0.8962	0.0084	0.8683	-	-	-	-	-
rs821205	10	107727810	C/A	FVC	ever	11564	52.01%	0.0522	7.93E-05	24460	51.37%	-0.0056	0.5360	0.0843	-	-	-	-	-
rs5006889	11	5373104	G/A	FEV1	never	11827	26.30%	-0.0696	3.51E-06	24460	27.04%	0.0034	0.7405	0.0175	-	-	-	-	-

						Discovery Analysis				Stage 1 Replication				Discovery + Stage 1 Replication meta	Stage 2 Replication				Discovery + Stage 1 + 2 Replication meta
						SpiroMeta Consortium				UK BiLEVE					UKHLS				
rs id	**CHR**	**POS**	**Effect / other allele**	**Trait**	**Smoking subset (Ever / never smokers)**	**N**	**Effect Allele Frequency (EAF)**	**Beta**	**P-value**	**N**	**EAF**	**Beta**	**P-value**	**P-value**	**N**	**EAF**	**Beta**	**P-value**	**P-value**
rs142159415	11	5776626	T/C	FVC	ever	11564	0.95%	0.2677	7.98E-05	24460	1.01%	0.0540	0.2297	0.0013	-	-	-	-	-
rs199618034	11	114182882	C/G	FEV1	ever	11559	0.02%	1.8491	3.59E-05	24460	0.06%	-0.1230	0.5644	0.0049	-	-	-	-	-
rs1982528	12	132237848	C/T	FVC	ever	10147	98.35%	-0.2244	5.25E-05	24460	98.61%	-0.0372	0.3538	0.0023	-	-	-	-	-
rs140930007	13	51854595	A/G	FVC	never	11419	0.03%	-1.6837	8.63E-06	NA	NA	NA	NA	**8.63E-06**	2940	0.03%	-0.1255	0.9001	5.40E-05
rs144854034	13	98829388	G/T	FEV1/FVC	ever	11560	0.11%	-0.7922	8.43E-05	24460	0.14%	-0.0512	0.7120	0.0113	-	-	-	-	-
rs140501662	14	20711665	T/C	FEV1/FVC	never	11831	0.15%	0.6755	7.66E-05	24483	0.04%	0.1679	0.4809	0.0046	-	-	-	-	-
rs200081065	14	91755506	T/C	FEV1	never	11834	0.08%	0.9679	2.47E-05	NA	NA	NA	NA	**2.47E-05**	2490	0.07%	0.5251	0.5995	6.10E-05
rs200614333	15	42143077	T/C	FVC	never	11832	0.02%	1.7770	6.41E-05	24483	0.02%	0.5671	0.0753	1.83E-04	-	-	-	-	-
rs138439412	15	52017135	T/C	FVC	ever	11561	0.92%	-0.2737	7.75E-05	24460	0.89%	0.0483	0.3377	0.1474	-	-	-	-	-
rs79030022	15	75941897	T/C	FEV1/FVC	ever	11553	0.05%	1.1613	5.71E-05	24460	0.02%	0.5275	0.2122	9.41E-04	-	-	-	-	-
rs144617499	16	21073933	A/G	FEV1	ever	11561	2.15%	-0.1813	8.23E-05	24460	2.32%	-0.0120	0.6885	0.0104	-	-	-	-	-
rs77439178	16	31091757	A/G	FEV1/FVC	ever	11479	0.01%	2.9418	3.61E-05	NA	NA	NA	NA	**3.61E-05**	4509	0.00%	NA	NA	3.61E-05
rs141225776	16	66547713	A/T	FVC	never	11833	0.02%	1.7422	9.80E-05	NA	NA	NA	NA	**9.80E-05**	2936	0.02%	0.3146	0.7530	2.87E-04
rs146239773	17	1387496	A/G	FEV1	ever	11563	0.02%	2.0179	1.11E-05	24460	0.01%	-0.5164	0.4243	0.3357	-	-	-	-	-
rs7207403	17	47210506	A/C	FEV1	ever	11562	56.42%	0.0591	6.02E-05	24460	55.81%	-0.0185	0.0369	0.4004	-	-	-	-	-
rs143270448	17	74274071	A/G	FEV1	never	11834	0.15%	0.7305	1.20E-05	24460	0.24%	0.0277	0.7754	0.0063	-	-	-	-	-
rs201979657	18	77171089	A/G	FEV1/FVC	ever	11557	0.02%	-1.9564	9.00E-05	NA	NA	NA	NA	**9.00E-05**	4509	0.01%	0.4703	0.6382	0.0021
rs200123506	19	38375738	C/T	FEV1	ever	11563	0.18%	0.6251	7.17E-05	24460	0.09%	0.0699	0.6795	0.0563	-	-	-	-	-
rs61737337	19	40197267	A/G	FEV1/FVC	ever	11137	0.01%	2.9111	4.18E-05	24460	0.02%	-0.2268	0.5628	0.0651	-	-	-	-	-
rs201361713	19	48737800	T/C	FVC	ever	11564	0.09%	0.8992	6.58E-05	24460	0.01%	-0.2270	0.6666	0.0566	-	-	-	-	-
rs143501994	19	51870712	T/C	FEV1/FVC	ever	11563	0.37%	0.4191	9.25E-05	24460	0.22%	-0.0107	0.932	0.0319	-	-	-	-	-
rs200402559	21	44838346	A/G	FVC	never	11833	0.01%	2.9643	2.54E-05	NA	NA	NA	NA	**2.54E-05**	2940	0.00%	NA	NA	2.54E-05
rs201423754	22	37893171	T/C	FEV1/FVC	ever	11563	0.01%	-2.5360	1.26E-05	NA	NA	NA	NA	**1.26E-05**	4509	0.00%	NA	NA	1.26E-05
rs12841259	X	118893390	G/A	FEV1/FVC	ever	8744	0.48%	0.3479	4.25E-05	NA	NA	NA	NA	**4.25E-05**	4506	0.03%	-0.6750	0.4997	0.0020

Appendix Table D-3: Replication results for all genes identified discovery SKAT analyses in ever smokers and never smokers separately ($P<10^{-4}$). Discovery, replication and combined meta results calculated using RAREMETAL.

			Discovery Analysis - SpiroMeta Consortium		Replication Analysis - UK BiLEVE & UKHLS		Combined Meta-analysis
Gene Name	Trait	Smoking subset (Ever / never smokers)	No. variants included in test (n_{snp})	P-value (P_{disc})	No. variants included in test (n_{snp})	P-value (P_{rep})	P-value (P_{meta})
C12orf77	FEV1/FVC	Ever	4	5.33E-05	4	0.840701	0.048557
NFATC1	FEV1/FVC	Never	10	8.41E-05	7	0.45282	0.431509

Appendix Table D-4: Replication results for all genes identified discovery WST analyses in ever smokers and never smokers separately ($P<10^{-4}$). Discovery, replication and combined meta results calculated using RAREMETAL. Beta values for Discovery and Stage 1 replication results reflect effect-size estimates on an inverse-normal transformed scale.

			Discovery Analysis - SpiroMeta Consortium			Replication Analysis - UK BiLEVE & UKHLS			Combined Meta-analysis	
Gene Name	Trait	Smoking subset (Ever / never smokers)	No. variants included in test (n_{snp})	Effect size(β_{disc})	P-value (P_{disc})	No. variants included in test (n_{snp})	Effect size(β_{rep})	P-value (P_{rep})	Effect size(β_{meta})	P-value (P_{meta})
FAM45A	FEV_1	Ever	6	-0.01309	5.66E-05	2	0.003047	0.296944	-0.00584	7.79E-04
NPEPL1	FEV_1	Never	2	0.018006	8.46E-05	2	-0.00082	0.742565	-0.00277	0.281712
PGBD1	FEV_1	Never	13	0.006683	2.63E-05	7	-4.9E-05	0.957415	0.001683	0.015263
FAM45A	FVC	Ever	6	-0.01337	3.78E-05	2	0.000384	0.895643	-0.00653	1.65E-04
GPR123	FVC	Ever	4	0.016998	8.40E-05	4	-0.00168	0.343451	0.004207	0.04262
WRB	FVC	Never	2	-0.01768	9.73E-05	1	-0.00221	0.603703	-0.01017	4.37E-05
C12orf77	FEV_1/FVC	Ever	4	0.018913	3.29E-06	3	0.000154	0.933714	0.007473	0.030133
LRPPRC	FEV_1/FVC	Ever	16	-0.01023	8.03E-05	7	-8.95E-04	0.003491	-0.00104	0.003646

E. Full results for the analysis of flow lung function measures PEF and FEF_{25-75} in UK Biobank (Chapter 5).

Appendix Table E-1: All sentinel SNPs identified in the analyses of PEF ($P<5x10^{-8}$).

Chromosome (Chr) and position (Pos) in build 37 are given for each SNP. Effect estimates are on an inverse-normal transformed scale after adjustments for age, age^2, sex, height and ancestry principal components, and stratified by ever smoking status.

SNP (Chr:Pos)	Effect/ noneffect allele	INFO	MAF	Effect Estimate	Standard error	P-value	Conditioned SNP(s)	Gene(s)	Consequence
rs3790760(chr1:19746527)	C/G	0.9887	22.58%	-0.0390	0.0055	1.51E-12	NA	*CAPZB*	intron_variant
rs59538733(chr1:31258698)	GC/G	0.9978	30.69%	-0.0311	0.0050	3.98E-10	NA	*LOC105378621*	upstream_gene_variant
rs111253797(chr1:51216645)	C/CT	0.9720	39.94%	0.0294	0.0047	5.23E-10	NA	*FAF1*	intron_variant
rs72706228(chr1:149177435)	T/A	1.0000	2.41%	-0.1307	0.0151	4.51E-18	NA	-	intergenic_variant
rs186278988(chr1:149681794)	T/C	0.9205	1.01%	0.1405	0.0239	4.20E-09	rs72706228	-	intergenic_variant
rs187954997(chr1:205741433)	T/A	0.7575	1.18%	-0.1981	0.0256	1.01E-14	rs1342062	*RAB29*	intron_variant
rs1342062(chr1:205912786)	G/T	0.9514	32.70%	0.0437	0.0050	2.33E-18	NA	*SLC26A9*	upstream_gene_variant
rs6688548(chr1:239850638)	C/A	0.9840	48.87%	-0.0339	0.0046	1.83E-13	NA	*CHRM3*	intron_variant
rs71389215(chr2:9293365)	CCT/C	0.9767	37.59%	-0.0268	0.0048	2.34E-08	NA	-	intergenic_variant
rs2544531(chr2:15904041)	A/G	1.0000	48.47%	0.0273	0.0046	2.26E-09	NA	-	intergenic_variant
rs143880252(chr2:31878429)	A/T	0.9239	3.48%	-0.1598	0.0131	4.64E-34	rs143268195	*SRD5A2*	intron_variant
rs143268195(chr2:32021240)	C/T	0.6868	1.48%	-0.3114	0.0241	3.47E-38	NA	*LOC105374449*	intron_variant, non_coding_transcript_variant
rs7583334(chr2:32800602)	T/C	0.8608	4.89%	-0.1117	0.0118	2.23E-21	rs143880252, rs143268195	*BIRC6*	intron_variant
rs77972916(chr2:43762112)	G/A	0.9855	7.63%	0.0583	0.0087	1.77E-11	NA	*THADA*	intron_variant
rs702901(chr2:65763552)	A/C	1.0000	3.79%	0.0662	0.0119	2.78E-08	NA	*LOC105369166*	intron_variant, non_coding_transcript_variant
rs188717678(chr2:135593014)	T/G	0.6325	1.08%	-0.2384	0.0294	4.58E-16	NA	*ACMSD*	upstream_gene_variant
rs559940908(chr2:135607570)	G/A	0.6294	1.03%	-0.3155	0.0305	4.18E-25	rs188717678	*ACMSD*	intron_variant
rs148474091(chr2:135825162)	A/G	0.8608	1.29%	-0.1537	0.0224	6.50E-12	rs188717678, rs559940908, rs75840321	*RAB3GAP1*	intron_variant
rs75840321(chr2:136650513)	G/C	0.6878	1.17%	-0.2776	0.0271	1.53E-24	NA	-	intergenic_variant
rs4372823(chr2:144325926)	A/G	0.9931	17.59%	0.0413	0.0060	8.01E-12	NA	*ARHGAP15*	intron_variant

SNP (Chr:Pos)	Effect/ noneffect allele	INFO	MAF	Effect Estimate	Standard error	P-value	Conditioned SNP(s)	Gene(s)	Consequence
rs72855705(chr2:151018807)	T/C	0.9601	4.45%	0.0771	0.0114	1.13E-11	NA	*LOC105373682*	intron_variant, non_coding_transcript_variant
rs115723803(chr2:156842429)	G/A	1.0000	6.58%	-0.0522	0.0092	1.33E-08	NA	*LOC105373705*	upstream_gene_variant
rs150421147(chr2:184213901)	C/T	1.0000	1.92%	-0.1025	0.0166	6.21E-10	NA	-	intergenic_variant
rs558858909(chr2:204693871)	C/G	0.6432	1.05%	-0.3160	0.0301	1.02E-25	rs56102377	-	intergenic_variant
rs56102377(chr2:204737635)	G/A	0.8085	1.28%	-0.2415	0.0234	5.42E-25	NA	*CTLA4*	3_prime_UTR_variant
rs13069216(chr3:55149475)	A/G	0.9820	48.01%	-0.0276	0.0046	2.63E-09	NA	-	intergenic_variant
rs6445925(chr3:57709777)	T/C	0.9890	48.73%	0.0258	0.0046	2.04E-08	NA	-	intergenic_variant
rs79105080(chr3:160779029)	T/C	0.8474	2.28%	-0.2397	0.0173	1.61E-43	NA	*PPM1L*	intron_variant
rs76580162(chr3:160989377)	C/T	0.7294	1.17%	-0.1846	0.0263	2.31E-12	rs79105080	*LOC105374187*	non_coding_transcript_exon_variant, non_coding_transcript_variant
rs17515933(chr3:164971278)	G/A	1.0000	1.76%	-0.1010	0.0173	5.26E-09	NA	*LINC01322*	intron_variant, non_coding_transcript_variant
rs6794830(chr3:168811226)	T/C	0.9959	36.19%	-0.0319	0.0048	2.01E-11	NA	*MECOM*	intron_variant
rs2592831(chr4:1711404)	T/C	0.9972	33.63%	-0.0265	0.0048	4.26E-08	NA	*SLBP*	intron_variant
rs28752137(chr4:5030854)	A/G	0.9895	32.57%	0.0370	0.0049	4.16E-14	NA	*LOC105374361*	intron_variant, non_coding_transcript_variant
rs191050570(chr4:15582642)	T/C	0.6451	1.88%	-0.3890	0.0227	1.42E-65	NA	*CC2D2A*	intron_variant
rs76364661(chr4:15743240)	G/T	0.8421	1.08%	-0.2732	0.0249	4.29E-28	rs191050570	*BST1*	intron_variant
rs73238348(chr4:56656153)	G/T	1.0000	11.15%	-0.0417	0.0072	8.61E-09	NA	-	intergenic_variant
rs192751765(chr4:77236153)	G/T	0.8150	0.92%	-0.2211	0.0275	7.98E-16	rs74936215	*CCDC158*	intron_variant
rs74936215(chr4:77349465)	G/A	0.7627	1.51%	-0.3133	0.0225	3.07E-44	NA	*LOC105377287*	intron_variant, non_coding_transcript_variant
rs113192062(chr4:90832304)	G/T	0.8506	0.56%	-0.2686	0.0340	2.79E-15	rs558009692	*MMRN1*	intron_variant
rs558009692(chr4:90840728)	A/T	0.6600	1.78%	-0.1982	0.0223	6.41E-19	NA	*MMRN1*	intron_variant
rs181375239(chr4:91081880)	T/C	0.7627	1.29%	-0.2802	0.0241	3.03E-31	NA	*CCSER1*	intron_variant
rs34712979(chr4:106819053)	G/A	1.0000	25.85%	-0.0592	0.0052	1.04E-29	NA	*NPNT*	splice_region_variant, intron_variant
rs541066384(chr4:133944260)	C/T	0.8901	1.25%	-0.1409	0.0220	1.57E-10	NA	*LOC105377431*	intron_variant, non_coding_transcript_variant
rs7658614(chr4:145445694)	T/A	0.9987	46.81%	0.0707	0.0046	7.55E-54	NA	*LOC105377462*	intron_variant ,non_coding_transcript_variant

SNP (Chr:Pos)	Effect/ noneffect allele	INFO	MAF	Effect Estimate	Standard error	P-value	Conditioned SNP(s)	Gene(s)	Consequence
rs116391341(chr4:154393826)	G/C	1.0000	3.15%	-0.0727	0.0130	2.20E-08	NA	*KIAA0922*	intron_variant
rs350415(chr5:51968428)	A/G	0.9948	26.02%	0.0293	0.0052	2.12E-08	NA	*LOC105378963*	downstream_gene_variant
rs76494599(chr5:60464399)	T/C	0.9699	4.94%	-0.0784	0.0108	3.27E-13	NA	*CTC-436P18.1*	intron_variant, non_coding_transcript_variant
rs4466136(chr5:82985576)	T/G	0.9962	21.94%	-0.0485	0.0055	2.05E-18	NA	*HAPLN1*	intron_variant
rs55971857(chr5:92441594)	A/C	0.9975	36.77%	0.0315	0.0047	3.34E-11	NA	-	intergenic_variant
rs376542571(chr5:102433408)	C/T	0.8180	0.36%	-0.2438	0.0433	1.87E-08	NA	*GIN1*	intron_variant
rs140877435(chr5:137245396)	G/A	0.9381	1.25%	-0.1375	0.0213	1.07E-10	NA	*PKD2L2*	intron_variant
rs549379352(chr5:137440650)	T/A	0.6525	1.66%	-0.1454	0.0231	3.12E-10	rs140877435	-	intergenic_variant
rs6580550(chr5:147856232)	T/C	1.0000	44.51%	0.0302	0.0046	5.59E-11	NA	*HTR4*	downstream_gene_variant
rs6872356(chr5:156962658)	C/T	0.9800	14.43%	-0.0437	0.0066	2.93E-11	NA	*ADAM19*	intron_variant
rs11747434(chr5:172779211)	T/C	0.9778	27.68%	0.0301	0.0052	5.93E-09	NA	*MIR8056*	downstream_gene_variant
.(chr6:17602870)	A/G	1.0000	1.78%	0.1321	0.0175	4.15E-14	NA	*FAM8A1*	missense_variant
rs3130568(chr6:31102884)	T/C	0.9962	49.30%	-0.0370	0.0046	5.82E-16	rs532524051, rs538489083, rs560438058	*PSORS1C1*	intron_variant
rs532524051(chr6:31194673)	G/A	0.6055	1.83%	-0.2781	0.0232	3.81E-33	rs538489083	-	intergenic_variant
rs538489083(chr6:32095727)	C/T	0.5303	1.70%	-0.3768	0.0256	3.83E-49	NA	*ATF6B*	intron_variant
rs9273229(chr6:32613914)	A/C	0.8656	36.45%	-0.0502	0.0051	8.46E-23	rs532524051, rs538489083, rs560438058	*HLA-DQA1,-*	downstream_gene_variant, intergenic_variant
rs560438058(chr6:32670158)	T/G	0.8307	5.48%	-0.1304	0.0113	1.22E-30	rs538489083	-	intergenic_variant
rs138535200(chr6:108633740)	C/T	0.7463	2.54%	-0.2482	0.0176	4.73E-45	NA	*LACE1*	intron_variant
rs7748807(chr6:108814162)	G/C	1.0000	1.39%	-0.1183	0.0196	1.48E-09	rs138535200	*LACE1*	intron_variant
rs75176386(chr6:112203464)	T/C	0.8008	0.62%	-0.1941	0.0334	6.07E-09	NA	-	intergenic_variant
rs17280293(chr6:142688969)	A/G	1.0000	2.71%	0.0825	0.0141	4.90E-09	rs190516	*ADGRG6*	missense_variant
rs190516(chr6:142813761)	T/C	0.9947	31.09%	0.0494	0.0050	2.29E-23	NA	-	intergenic_variant
rs574284527(chr6:149371098)	C/CACAG	0.9497	33.67%	-0.0324	0.0050	7.52E-11	NA	*UST*	intron_variant
rs213522(chr7:26944252)	G/T	0.9934	49.10%	0.0263	0.0046	1.13E-08	NA	-	intergenic_variant
rs140768661(chr7:81185406)	A/G	1.0000	2.64%	-0.1029	0.0142	3.83E-13	NA	*LOC105369146*	intron_variant
.(chr7:128496289)	T/TCA	1.0000	3.37%	-0.1376	0.0126	1.14E-27	NA	*FLNC*	intron_variant
rs12698403(chr7:156127246)	G/A	0.9936	44.14%	-0.0283	0.0046	9.51E-10	NA	-	intergenic_variant

SNP (Chr:Pos)	Effect/ noneffect allele	INFO	MAF	Effect Estimate	Standard error	P-value	Conditioned SNP(s)	Gene(s)	Consequence
rs2272026(chr8:9757600)	C/T	0.9839	30.52%	0.0278	0.0050	3.15E-08	NA	*LINC00599*	non_coding_transcript_exon_variant, non_coding_transcript_variant
rs117929207(chr8:13901118)	T/A	1.0000	1.55%	-0.1022	0.0186	3.66E-08	NA	-	intergenic_variant
rs6981627(chr8:22530061)	G/C	0.9132	2.56%	-0.2091	0.0156	3.53E-41	NA	*BIN3*	upstream_gene_variant
rs34249114(chr8:22535398)	G/A	0.9461	6.01%	-0.1264	0.0100	1.01E-36	rs6981627	-	intergenic_variant
rs551029716(chr8:118071198)	C/T	0.8578	0.52%	-0.2478	0.0347	9.39E-13	NA	*SLC30A8*	intron_variant
rs190161317(chr8:130919581)	C/T	0.9103	0.74%	-0.2249	0.0281	1.33E-15	NA	*FAM49B*	intron_variant
rs150078012(chr8:145860483)	G/A	0.7090	1.17%	-0.1488	0.0265	1.83E-08	NA	*ARHGAP39,-*	intron_variant, intergenic_variant
rs4925815(chr8:145894656)	C/G	0.6338	47.59%	-0.0328	0.0058	1.24E-08	rs150078012	*ARHGAP39,-*	intron_variant, intergenic_variant
rs116341416(chr9:1924499)	G/C	1.0000	2.60%	-0.0782	0.0143	4.92E-08	NA	*LOC105375952*	downstream_gene_variant
rs17209774(chr9:4145163)	G/C	0.9894	36.77%	-0.0267	0.0048	2.14E-08	NA	*GLIS3*	intron_variant
rs116940183(chr9:27480138)	G/A	0.7998	2.50%	-0.0918	0.0167	3.86E-08	rs548356952	*MOB3B*	intron_variant
rs548356952(chr9:27500658)	G/C	0.8039	1.22%	-0.1707	0.0235	4.16E-13	NA	*MOB3B*	intron_variant
rs558415(chr9:33953770)	T/C	0.8220	1.84%	-0.2231	0.0195	2.61E-30	NA	*UBAP2*	intron_variant
rs117434123(chr9:38996308)	C/G	1.0000	3.33%	-0.0719	0.0129	2.27E-08	NA	*LOC101927042*	intron_variant, non_coding_transcript_variant
rs811689(chr9:119410756)	C/T	0.9978	44.89%	-0.0276	0.0046	1.85E-09	NA	*ASTN2*	intron_variant
rs2271804(chr10:12252217)	G/A	0.9956	47.03%	0.0355	0.0046	1.10E-14	NA	*CDC123*	intron_variant
rs185638441(chr10:15559576)	G/C	0.8150	0.27%	-0.2967	0.0491	1.48E-09	NA	*ITGA8*	intron_variant
.(chr10:97879525)	A/AT	0.5054	1.11%	-0.2705	0.0323	5.06E-17	NA	*CRTAC1*	intron_variant
rs117443545(chr11:83190398)	G/A	0.7085	1.08%	-0.2414	0.0268	2.08E-19	NA	DLG2	intron_variant
rs138857603(chr11:83275030)	A/AT	0.8690	0.63%	-0.2337	0.0315	1.18E-13	rs117443545	DLG2	intron_variant
rs112066330(chr11:83299083)	G/T	0.8023	0.36%	-0.2781	0.0441	2.92E-10	rs117443545, rs138857603	DLG2	intron_variant
rs114902756(chr11:85820597)	G/C	0.8024	0.42%	-0.2968	0.0406	2.62E-13	NA	-	intergenic_variant
rs7105597(chr11:86432083)	G/A	0.9866	15.25%	-0.0360	0.0064	1.94E-08	rs114902756	-	intergenic_variant
rs111400016(chr11:91208353)	G/A	1.0000	1.76%	-0.0966	0.0172	2.16E-08	NA	-	intergenic_variant
rs503441(chr11:126010797)	A/G	0.9904	18.38%	0.0329	0.0059	2.88E-08	NA	LOC105369591	upstream_gene_variant
rs150950471(chr12:39134817)	C/G	0.6038	2.41%	-0.4098	0.0213	3.65E-82	NA	CPNE8	intron_variant
rs558946982(chr12:40462632)	G/GT	0.6928	1.62%	0.1334	0.0217	7.63E-10	rs17519950	SLC2A13	intron_variant

SNP (Chr:Pos)	Effect/ noneffect allele	INFO	MAF	Effect Estimate	Standard error	P-value	Conditioned SNP(s)	Gene(s)	Consequence
rs200767456(chr12:40558001)	A/T	0.6709	3.81%	0.0806	0.0145	2.52E-08	rs150950471, rs558946982, rs17519950	-	intergenic_variant
rs17519950(chr12:40695188)	A/T	0.8380	1.15%	-0.2785	0.0240	4.00E-31	NA	LRRK2	intron_variant
rs4760619(chr12:48499931)	A/T	0.9899	16.64%	-0.0349	0.0062	1.63E-08	NA	PFKM	intron_variant
rs11111272(chr12:102827441)	G/C	0.9961	28.53%	-0.0303	0.0051	2.17E-09	NA	IGF1	intron_variant
rs12427728(chr13:98174730)	C/T	0.5654	1.05%	-0.4013	0.0338	1.36E-32	NA	-	intergenic_variant
rs55962908(chr15:49768429)	G/T	0.8900	47.02%	-0.0270	0.0049	2.58E-08	NA	FGF7	intron_variant
rs1441358(chr15:71612514)	T/G	1.0000	33.60%	-0.0468	0.0049	6.39E-22	NA	THSD4	intron_variant
rs77111785(chr15:74930527)	C/T	0.5803	1.29%	-0.1563	0.0265	3.82E-09	NA	EDC3	intron_variant
rs12594577(chr15:76828793)	A/C	0.9995	47.86%	0.0255	0.0046	2.71E-08	NA	SCAPER	intron_variant
rs140330585(chr15:78866445)	G/A	0.9996	33.46%	-0.0308	0.0049	2.46E-10	NA	CHRNA5	intron_variant
rs74930371(chr16:19273328)	G/T	0.9152	1.95%	-0.2447	0.0176	7.15E-44	NA	SYT17	intron_variant
rs544559569(chr16:30880829)	TC/T	0.5605	1.99%	-0.1307	0.0225	6.46E-09	NA	BCL7C	intron_variant
rs3104770(chr16:52627368)	A/T	0.7563	2.77%	-0.2379	0.0168	1.98E-45	NA	CASC16	intron_variant, non_coding_transcript_variant
rs11149827(chr16:75435143)	A/G	0.9878	40.82%	-0.0380	0.0047	4.54E-16	NA	CFDP1	intron_variant
rs561060101(chr17:17776177)	G/C	0.7202	1.18%	-0.2338	0.0258	1.34E-19	rs201183826	TOM1L2	intron_variant
rs76926781(chr17:17805129)	C/T	0.8439	1.26%	-0.1916	0.0230	9.04E-17	rs561060101, rs201183826	TOM1L2	intron_variant
rs201183826(chr17:18107778)	AT/A	0.8990	1.26%	-0.2268	0.0221	1.06E-24	NA	ALKBH5	intron_variant
rs561223711(chr17:36861636)	T/TG	0.9847	19.69%	-0.0319	0.0058	3.55E-08	NA	MLLT6	upstream_gene_variant
rs73314997(chr17:44061123)	C/T	0.6376	2.31%	-0.3898	0.0210	3.00E-77	NA	MAPT,MAPT	missense_variant, missense_variant
rs139935845(chr17:44307598)	G/A	0.7489	1.55%	-0.1817	0.0224	5.41E-16	rs73314997, rs56332949	LOC105371799	intron_variant, non_coding_transcript_variant
rs56332949(chr17:44316480)	G/C	0.7377	1.69%	-0.1595	0.0210	3.00E-14	rs73314997	LOC105371799	intron_variant, non_coding_transcript_variant
rs227726(chr17:54777585)	C/T	0.9884	33.61%	-0.0296	0.0049	1.21E-09	NA	-	intergenic_variant
rs9898150(chr17:69185195)	T/G	0.9976	48.59%	-0.0355	0.0046	8.59E-15	NA	CASC17	intron_variant, non_coding_transcript_variant
rs112990608(chr17:69384158)	G/C	0.9772	7.36%	-0.0522	0.0088	3.41E-09	rs9898150	-	intergenic_variant
rs8089099(chr18:10078071)	G/A	0.9802	27.21%	0.0348	0.0052	2.12E-11	NA	-	intergenic_variant

SNP (Chr:Pos)	Effect/ noneffect allele	INFO	MAF	Effect Estimate	Standard error	P-value	Conditioned SNP(s)	Gene(s)	Consequence
rs138326911(chr19:13213277)	T/C	0.9380	3.74%	-0.0715	0.0124	9.01E-09	NA	LYL1	intron_variant
rs564454164(chr19:54957229)	AAAG/A	0.9289	1.43%	-0.2168	0.0204	2.06E-26	NA	LENG8,-	upstream_gene_variant, intergenic_variant
rs143792972(chr19:55834177)	G/C	0.9281	0.62%	-0.2561	0.0307	6.86E-17	rs564454164	TMEM150B	intron_variant
rs34590652(chr20:13526292)	A/AT	0.9862	43.53%	0.0269	0.0046	7.05E-09	NA	TASP1	intron_variant

Appendix Table E-2: All sentinel SNPs identified in the analyses of FEF_{25-75} ($P<5x10^{-8}$).

Chromosome (Chr) and position (Pos) in build 37 are given for each SNP. Effect estimates are on an inverse-normal transformed scale after adjustments for age, age^2, sex, height and ancestry principal components, and stratified by ever smoking status.

SNP (Chr:Pos)	Effect/ noneffect allele	INFO	MAF	Effect Estimate	Standard error	P-value	Conditioned SNP(s)	Gene(s)	Consequence
rs11588995 (chr1:17417958)	C/T	0.9325	37.72%	-0.0295	0.0049	2.19E-09	NA	*PADI2*	intron_variant
rs2986163 (chr1:25622288)	T/A	0.8132	41.01%	-0.0333	0.0052	1.24E-10	NA	*RHD*	intron_variant
rs67452844 (chr1:39616282)	A/G	1.0000	25.09%	-0.0442	0.0053	8.36E-17	NA	*MACF1*	intron_variant
rs72673454 (chr1:60962424)	T/C	0.9868	5.02%	-0.0633	0.0106	2.16E-09	rs146229204	-	intergenic_variant
rs146229204 (chr1:61789514)	A/G	1.0000	1.60%	-0.1223	0.0183	2.38E-11	NA	*NFIA*	intron_variant
rs79745125 (chr1:68617354)	G/A	1.0000	3.12%	-0.0776	0.0132	3.92E-09	NA	*WLS*	intron_variant
rs6693314 (chr1:92058290)	C/T	0.9616	13.84%	-0.0390	0.0068	9.03E-09	NA	-	intergenic_variant
rs2282248 (chr1:111736594)	C/T	0.9903	32.69%	0.0290	0.0049	4.27E-09	NA	*DENND2D*	intron_variant
rs72706228 (chr1:149177435)	T/A	1.0000	2.41%	-0.1763	0.0152	3.45E-31	NA	-	intergenic_variant
rs587733913 (chr1:149698458)	G/A	0.9325	0.97%	0.1671	0.0242	5.49E-12	rs77421422	*RP11-353N4.5*	non_coding_transcript_exon_variant,non_coding_transcript_variant
rs77421422 (chr1:150604958)	A/G	1.0000	1.58%	-0.1160	0.0185	3.21E-10	rs72706228, rs587733913	*ENSA*	upstream_gene_variant
rs187954997 (chr1:205741433)	T/A	0.7575	1.18%	-0.2647	0.0258	9.32E-25	NA	*RAB29*	intron_variant
rs142495088 (chr1:219929662)	CAA/C	0.9475	45.08%	0.0296	0.0048	4.89E-10	NA	*SLC30A10*	intron_variant,non_coding_transcript_variant
rs6694220 (chr1:239883616)	A/G	0.9989	49.05%	-0.0359	0.0046	6.65E-15	NA	*CHRM3*	intron_variant
. (chr2:15913628)	C/CA	1.0000	49.03%	0.0300	0.0046	6.95E-11	NA	-	intergenic_variant
rs55884799 (chr2:18287623)	T/C	0.9949	17.42%	0.0541	0.0061	7.45E-19	NA	*KCNS3*	intron_variant,non_coding_transcript_variant
rs143880252 (chr2:31878429)	A/T	0.9239	3.48%	-0.2065	0.0132	6.11E-55	rs139999372	*AL133247.3*	downstream_gene_variant
rs139999372 (chr2:31941464)	C/A	0.6270	1.23%	-0.4530	0.0286	1.40E-56	NA	-	intergenic_variant
rs73922761 (chr2:32816432)	T/C	0.8892	5.06%	-0.1403	0.0115	1.62E-34	rs143880252, rs139999372	*BIRC6*	intron_variant
rs702901 (chr2:65763552)	A/C	1.0000	3.79%	0.0753	0.0120	3.38E-10	NA	*AC074391.1*	intron_variant,non_coding_transcript_variant

SNP (Chr:Pos)	Effect/ noneffect allele	INFO	MAF	Effect Estimate	Standard error	P-value	Conditioned SNP(s)	Gene(s)	Consequence
rs13018626 (chr2:67087652)	C/T	1.0000	6.67%	-0.0560	0.0092	1.36E-09	NA	-	intergenic_variant
rs77524493 (chr2:128618594)	A/T	1.0000	4.18%	0.0631	0.0114	3.51E-08	NA	*POLR2D*	upstream_gene_variant
rs188717678 (chr2:135593014)	T/G	0.6325	1.08%	-0.3259	0.0298	7.09E-28	NA	*ACMSD*	upstream_gene_variant
rs559940908 (chr2:135607570)	G/A	0.6294	1.03%	-0.3935	0.0309	3.55E-37	rs188717678	*ACMSD*	intron_variant
rs149870752 (chr2:135707918)	CTG/C	0.5716	2.46%	-0.1866	0.0209	4.16E-19	rs188717678, rs559940908, rs75840321, rs17467077	*CCNT2*	intron_variant
rs75840321 (chr2:136650513)	G/C	0.6878	1.17%	-0.3388	0.0276	1.33E-34	NA	-	intergenic_variant
rs186885313 (chr2:136736556)	T/A	0.6100	1.48%	0.1551	0.0237	6.23E-11	rs188717678, rs559940908, rs75840321, rs17467077	*DARS*	intron_variant
rs17467077 (chr2:136976914)	G/A	0.7473	41.12%	-0.0407	0.0055	1.15E-13	rs75840321	-	intergenic_variant
rs72855705 (chr2:151018807)	T/C	0.9601	4.45%	0.0634	0.0114	2.97E-08	NA	-	intergenic_variant
rs16840048 (chr2:156992261)	A/G	0.9789	13.92%	-0.0417	0.0067	5.59E-10	NA	-	intergenic_variant
rs16858920 (chr2:171556798)	T/C	1.0000	2.74%	-0.0827	0.0141	4.24E-09	NA	*AC007277.3*	upstream_gene_variant
rs150421147 (chr2:184213901)	C/T	1.0000	1.92%	-0.1217	0.0167	2.93E-13	NA	-	intergenic_variant
rs533421015 (chr2:204645768)	A/C	0.8762	0.10%	-0.4362	0.0785	2.81E-08	rs558858909, rs56102377	*RNU6-474P*	upstream_gene_variant
rs558858909 (chr2:204693871)	C/G	0.6432	1.05%	-0.4102	0.0304	1.85E-41	rs56102377	-	intergenic_variant
rs56102377 (chr2:204737635)	G/A	0.8085	1.28%	-0.3251	0.0238	1.87E-42	NA	*CTLA4*	3_prime_UTR_variant
rs7602943 (chr2:217632085)	A/G	0.9583	14.58%	-0.0391	0.0067	4.55E-09	NA	*AC007563.5*	intron_variant,non_coding_transcript_variant
rs2571445 (chr2:218683154)	A/G	1.0000	39.70%	0.0373	0.0047	2.41E-15	NA	*TNS1*	missense_variant
rs56398110 (chr2:221353688)	C/T	1.0000	3.33%	-0.0734	0.0128	9.35E-09	NA	*AC067956.1*	intron_variant,non_coding_transcript_variant
rs16825267 (chr2:229569919)	C/G	0.9900	8.06%	0.0734	0.0085	6.45E-18	NA	-	intergenic_variant
rs61332075 (chr2:239316560)	G/C	0.9824	12.12%	0.0406	0.0071	9.99E-09	rs11124197	*RNU6-234P*	upstream_gene_variant
rs11124197 (chr2:239882327)	T/C	0.9947	19.84%	0.0518	0.0058	4.43E-19	NA	-	intergenic_variant
rs1286664 (chr3:25529280)	C/T	1.0000	17.61%	0.0385	0.0061	1.98E-10	NA	*RARB*	intron_variant

SNP (Chr:Pos)	Effect/ noneffect allele	INFO	MAF	Effect Estimate	Standard error	P-value	Conditioned SNP(s)	Gene(s)	Consequence
rs553112184 (chr3:53692219)	C/CT	0.9371	22.89%	-0.0362	0.0057	1.48E-10	NA	*CACNA1D*	intron_variant
rs62256243 (chr3:55163321)	C/T	0.9903	33.64%	-0.0373	0.0049	3.19E-14	NA	-	intergenic_variant
rs72874879 (chr3:57713464)	C/T	0.9959	23.55%	0.0399	0.0054	2.10E-13	NA	-	intergenic_variant
rs114153976 (chr3:120798364)	G/C	1.0000	2.79%	0.0795	0.0140	1.22E-08	NA	*STXBP5L*	intron_variant
rs2999089 (chr3:127935159)	C/G	0.9985	12.00%	-0.0480	0.0071	1.43E-11	NA	*EEFSEC*	intron_variant
rs372836737 (chr3:134261782)	C/CAA	0.9152	26.14%	-0.0302	0.0055	4.06E-08	NA	*CEP63*	intron_variant
rs76904649 (chr3:143776664)	C/T	1.0000	3.10%	-0.0756	0.0132	1.10E-08	NA	-	intergenic_variant
rs79105080 (chr3:160779029)	T/C	0.8474	2.28%	-0.2985	0.0175	1.67E-65	NA	*PPM1L*	intron_variant
rs185502807 (chr3:160977292)	C/T	0.6696	1.45%	-0.1423	0.0244	5.86E-09	rs79105080, rs112769734	-	intergenic_variant
rs112769734 (chr3:160984308)	C/A	0.7292	1.17%	-0.2282	0.0266	1.08E-17	rs79105080	-	intergenic_variant
rs17515933 (chr3:164971278)	G/A	1.0000	1.76%	-0.1201	0.0174	5.45E-12	NA	*LINC01322*	intron_variant,non_coding_transcript_variant
rs6794830 (chr3:168811226)	T/C	0.9959	36.19%	-0.0357	0.0048	9.96E-14	NA	*MECOM*	intron_variant
rs191050570 (chr4:15582642)	T/C	0.6451	1.88%	-0.4881	0.0232	2.40E-98	NA	*CC2D2A*	intron_variant
rs76364661 (chr4:15743240)	G/T	0.8421	1.08%	-0.3461	0.0251	3.02E-43	rs191050570	*RP11-442P12.2*	upstream_gene_variant
rs73238348 (chr4:56656153)	G/T	1.0000	11.15%	-0.0529	0.0073	3.79E-13	NA	-	intergenic_variant
rs62316308 (chr4:75676337)	C/A	0.9979	26.27%	0.0329	0.0052	3.34E-10	NA	*BTC*	intron_variant
rs1530294 (chr4:77202186)	G/A	0.8191	0.22%	-0.3734	0.0554	1.58E-11	rs62316308, rs192751765, rs116291420	*FAM47E-STBD1*	intron_variant
rs192751765 (chr4:77236153)	G/T	0.8150	0.92%	-0.2837	0.0276	7.56E-25	rs116291420	*STBD1*	downstream_gene_variant
rs116291420 (chr4:77346196)	G/C	0.7825	1.52%	-0.4094	0.0222	5.10E-76	NA	*CCDC158*	upstream_gene_variant
rs2013701 (chr4:89885086)	G/T	0.9955	49.23%	-0.0363	0.0046	4.04E-15	NA	*FAM13A*	intron_variant
rs113192062 (chr4:90832304)	G/T	0.8506	0.56%	-0.3184	0.0343	1.52E-20	rs2013701, rs558009692, rs181375239	*MMRN1*	intron_variant
rs558009692 (chr4:90840728)	A/T	0.6600	1.78%	-0.2713	0.0225	1.50E-33	rs2013701	*MMRN1*	intron_variant
rs181375239 (chr4:91081880)	T/C	0.7627	1.29%	-0.3515	0.0243	2.20E-47	NA	*CCSER1*	intron_variant
rs6533183 (chr4:106133184)	C/T	0.9984	34.06%	-0.0352	0.0048	4.02E-13	rs34712979	*TET2*	intron_variant

SNP (Chr:Pos)	Effect/ noneffect allele	INFO	MAF	Effect Estimate	Standard error	P-value	Conditioned SNP(s)	Gene(s)	Consequence
rs145501437 (chr4:106815984)	TAGAGAC/ T	0.9830	5.92%	0.0616	0.0098	3.98E-10	rs6533183, rs34712979	*NPNT*	upstream_gene_variant
rs34712979 (chr4:106819053)	G/A	1.0000	25.85%	-0.0879	0.0053	1.61E-62	NA	*NPNT*	splice_region_variant,intron_variant
rs73844242 (chr4:122572542)	A/T	1.0000	2.15%	-0.0973	0.0158	7.69E-10	NA	-	intergenic_variant
rs541066384 (chr4:133944260)	C/T	0.8901	1.25%	-0.1573	0.0221	1.23E-12	rs146919752	-	intergenic_variant
rs146919752 (chr4:134018576)	C/A	0.5872	1.31%	-0.2070	0.0275	4.88E-14	NA	*RP11-9G1.3*	intron_variant,non_coding_transcript_variant
rs6817273 (chr4:145492003)	T/C	0.9980	39.56%	0.0686	0.0047	3.16E-48	NA	*KRT18P51*	upstream_gene_variant
rs2251639 (chr4:154656410)	T/C	0.9693	24.07%	-0.0311	0.0055	1.30E-08	NA	*RNF175*	intron_variant
rs115884234 (chr4:177128291)	T/A	1.0000	4.52%	0.0679	0.0111	9.17E-10	NA	-	intergenic_variant
rs17064246 (chr4:178003392)	A/G	1.0000	3.27%	-0.0750	0.0129	5.45E-09	rs115884234	-	intergenic_variant
rs111310362 (chr5:3008207)	T/A	1.0000	3.83%	-0.0659	0.0120	4.17E-08	NA	-	intergenic_variant
rs12520489 (chr5:43515280)	A/C	0.9825	19.98%	-0.0322	0.0058	3.05E-08	NA	*C5orf34*	upstream_gene_variant
rs1551943 (chr5:52195033)	G/A	1.0000	22.81%	-0.0424	0.0055	1.16E-14	NA	*ITGA1*	intron_variant
rs12186544 (chr5:52255140)	C/A	0.9818	20.52%	0.0369	0.0057	1.42E-10	rs1551943	*ITGA1*	downstream_gene_variant
rs17659497 (chr5:55903639)	A/T	1.0000	2.14%	0.0917	0.0159	7.35E-09	NA	*C5orf67*	upstream_gene_variant
rs115007883 (chr5:60076358)	C/T	0.8889	0.40%	-0.2434	0.0395	7.36E-10	rs75848589	*ELOVL7*	intron_variant
rs75848589 (chr5:60469357)	C/T	0.8560	2.14%	-0.1466	0.0176	8.97E-17	NA	*CTC-436P18.1*	intron_variant,non_coding_transcript_variant
rs71626454 (chr5:66180965)	G/A	1.0000	1.86%	-0.1030	0.0170	1.29E-09	NA	*MAST4*	intron_variant
rs1501911 (chr5:98342868)	T/A	0.9831	38.53%	0.0296	0.0048	5.28E-10	NA	-	intergenic_variant
rs376542571 (chr5:102433408)	C/T	0.8180	0.36%	-0.2838	0.0441	1.27E-10	NA	*GIN1*	synonymous_variant
rs17163397 (chr5:128767384)	A/G	0.9904	12.55%	0.0445	0.0070	1.87E-10	NA	-	intergenic_variant
rs10060626 (chr5:131803967)	T/G	0.9674	23.93%	-0.0299	0.0055	4.69E-08	NA	*C5orf56*	downstream_gene_variant
rs140877435 (chr5:137245396)	G/A	0.9381	1.25%	-0.1595	0.0214	9.95E-14	rs549379352	*PKD2L2*	intron_variant
rs549379352 (chr5:137440650)	T/A	0.6525	1.66%	-0.1881	0.0234	8.19E-16	NA	-	intergenic_variant
rs7733410 (chr5:147856522)	G/A	1.0000	44.04%	0.0571	0.0046	8.49E-35	NA	*HTR4*	downstream_gene_variant

SNP (Chr:Pos)	Effect/ noneffect allele	INFO	MAF	Effect Estimate	Standard error	P-value	Conditioned SNP(s)	Gene(s)	Consequence
rs1800888 (chr5:148206885)	C/T	1.0000	1.51%	-0.1283	0.0188	9.41E-12	rs7733410	*ADRB2*	missense_variant
rs13361953 (chr5:156926442)	T/C	0.9958	33.69%	-0.0457	0.0049	8.06E-21	NA	*ADAM19*	intron_variant
rs6896218 (chr5:179598009)	G/C	0.9808	49.32%	0.0272	0.0047	5.41E-09	NA	*RASGEF1C*	intron_variant
rs1294417 (chr6:6741932)	T/C	0.9822	46.04%	0.0296	0.0047	2.15E-10	NA	-	intergenic_variant
. (chr6:17602870)	A/G	1.0000	1.78%	0.1618	0.0176	3.58E-20	NA	*FAM8A1*	missense_variant
rs6905736 (chr6:19843767)	C/A	0.9321	15.22%	-0.0386	0.0066	5.45E-09	NA	*ID4*	downstream_gene_variant
rs2517611 (chr6:30169327)	A/G	1.0000	22.96%	-0.0321	0.0055	4.05E-09	rs149405105, rs532524051, rs538489083, rs560438058	*TRIM26*	intron_variant
rs149405105 (chr6:30976349)	G/A	0.9678	6.16%	0.0936	0.0097	7.53E-22	NA	*MUC22*	upstream_gene_variant
rs532524051 (chr6:31194673)	G/A	0.6055	1.83%	-0.3128	0.0232	2.24E-41	rs538489083	*XXbac-BPG299F13.16*	upstream_gene_variant
rs2442724 (chr6:31319907)	C/T	0.9857	14.89%	-0.0586	0.0065	1.26E-19	rs149405105, rs532524051, rs538489083, rs560438058	*HLA-B*	downstream_gene_variant
rs538489083 (chr6:32095727)	C/T	0.5303	1.70%	-0.4520	0.0260	6.70E-68	NA	*FKBPL*	downstream_gene_variant
rs9270377 (chr6:32558260)	G/T	0.7291	44.95%	-0.0685	0.0054	5.77E-37	rs149405105, rs532524051, rs538489083, rs560438058	*HLA-DRB1*	upstream_gene_variant
rs560438058 (chr6:32670158)	T/G	0.8307	5.48%	-0.1633	0.0114	2.58E-46	rs538489083	*MTCO3P1*	downstream_gene_variant
rs115830429 (chr6:56239955)	G/A	1.0000	2.45%	0.0901	0.0149	1.47E-09	NA	*COL21A1*	intron_variant
rs13206617 (chr6:73663745)	G/T	0.9942	19.90%	0.0351	0.0058	1.22E-09	NA	*KCNQ5*	intron_variant
rs138535200 (chr6:108633740)	C/T	0.7463	2.54%	-0.3021	0.0178	2.88E-64	NA	*LACE1*	intron_variant
rs112818441 (chr6:108709124)	T/A	0.6688	1.18%	-0.2692	0.0278	3.06E-22	rs138535200	*LACE1*	intron_variant
rs7748807 (chr6:108814162)	G/C	1.0000	1.39%	-0.1741	0.0197	8.77E-19	rs138535200, rs112818441	*LACE1*	intron_variant
rs75176386 (chr6:112203464)	T/C	0.8008	0.62%	-0.3110	0.0339	4.54E-20	NA	-	intergenic_variant

SNP (Chr:Pos)	Effect/ noneffect allele	INFO	MAF	Effect Estimate	Standard error	P-value	Conditioned SNP(s)	Gene(s)	Consequence
rs117677851 (chr6:113532326)	T/G	1.0000	2.22%	-0.1049	0.0155	1.40E-11	NA	-	intergenic_variant
rs7765914 (chr6:142560307)	T/C	1.0000	27.50%	0.0384	0.0052	1.05E-13	rs3748069	-	intergenic_variant
rs17280293 (chr6:142688969)	A/G	1.0000	2.71%	0.0938	0.0142	3.64E-11	rs7765914, rs3748069	*ADGRG6*	missense_variant
rs3748069 (chr6:142767633)	A/G	1.0000	28.91%	0.0729	0.0051	1.02E-46	NA	*ADGRG6*	downstream_gene_variant
rs117606901 (chr6:143970809)	C/G	1.0000	3.51%	0.0823	0.0125	4.29E-11	NA	*PHACTR2*	intron_variant
rs140768661 (chr7:81185406)	A/G	1.0000	2.64%	-0.1153	0.0143	6.10E-16	NA	*AC008163.4*	intron_variant,non_coding_transcript_variant
. (chr7:128496289)	T/TCA	1.0000	3.37%	-0.1874	0.0127	3.18E-49	NA	*FLNC*	intron_variant
rs12698403 (chr7:156127246)	G/A	0.9936	44.14%	-0.0307	0.0047	4.52E-11	NA	-	intergenic_variant
rs372477046 (chr8:7275432)	C/T	0.5825	4.90%	-0.0903	0.0141	1.68E-10	NA	*DEFB4B*	upstream_gene_variant
rs2272026 (chr8:9757600)	C/T	0.9839	30.52%	0.0296	0.0051	5.02E-09	NA	*LINC00599*	non_coding_transcript_exon_variant,non_coding_transcript_variant
rs4128298 (chr8:11823332)	T/C	0.9866	28.45%	0.0310	0.0052	1.73E-09	NA	-	intergenic_variant
rs117929207 (chr8:13901118)	T/A	1.0000	1.55%	-0.1031	0.0187	3.32E-08	NA	-	intergenic_variant
rs75621048 (chr8:13928632)	T/A	1.0000	2.21%	-0.0863	0.0156	3.33E-08	rs117929207	-	intergenic_variant
rs372505725 (chr8:22452602)	AG/A	0.6410	1.48%	-0.1828	0.0260	2.19E-12	rs6981627, rs34249114	*PDLIM2*	downstream_gene_variant
rs6981627 (chr8:22530061)	G/C	0.9132	2.56%	-0.2699	0.0157	2.31E-66	NA	*BIN3*	upstream_gene_variant
rs34249114 (chr8:22535398)	G/A	0.9461	6.01%	-0.1503	0.0100	1.25E-50	rs6981627	*CTD-3247F14.2*	downstream_gene_variant
rs659398 (chr8:103131300)	T/C	0.9797	26.92%	0.0295	0.0052	1.90E-08	NA	*NCALD*	intron_variant
rs551029716 (chr8:118071198)	C/T	0.8578	0.52%	-0.3177	0.0351	1.52E-19	NA	*RP11-1059L18.1*	downstream_gene_variant
. (chr8:130893236)	TTC/T	0.9003	0.59%	-0.3379	0.0322	9.58E-26	NA		
rs150078012 (chr8:145860483)	G/A	0.7090	1.17%	-0.2056	0.0266	1.05E-14	NA	-	intergenic_variant
rs10108089 (chr8:145958174)	G/A	0.6506	2.20%	0.1224	0.0191	1.51E-10	rs150078012, rs190738032	*ZNF251*	intron_variant
rs190738032 (chr8:146126686)	G/C	0.8596	0.50%	-0.2528	0.0357	1.50E-12	rs150078012	*ZNF250*	intron_variant
rs116341416 (chr9:1924499)	G/C	1.0000	2.60%	-0.0918	0.0144	2.06E-10	NA	-	intergenic_variant
rs7872188 (chr9:4124377)	C/T	0.9738	40.06%	-0.0315	0.0048	3.53E-11	NA	*GLIS3*	intron_variant

SNP (Chr:Pos)	Effect/ noneffect allele	INFO	MAF	Effect Estimate	Standard error	P-value	Conditioned SNP(s)	Gene(s)	Consequence
rs113374461 (chr9:23583610)	A/ATATAT	0.9911	48.48%	0.0295	0.0046	1.52E-10	NA	-	intergenic_variant
rs116940183 (chr9:27480138)	G/A	0.7998	2.50%	-0.1139	0.0169	1.46E-11	rs548356952	*MOB3B*	intron_variant
rs548356952 (chr9:27500658)	G/C	0.8039	1.22%	-0.2286	0.0236	3.36E-22	NA	*MOB3B*	intron_variant
rs150361628 (chr9:33941516)	C/T	0.8057	0.37%	-0.3185	0.0438	3.57E-13	rs558415	*UBAP2*	intron_variant
rs558415 (chr9:33953770)	T/C	0.8220	1.84%	-0.2763	0.0197	7.03E-45	NA	*UBAP2*	intron_variant
rs12156614 (chr9:34040152)	C/A	0.5561	1.11%	0.1730	0.0283	1.04E-09	rs150361628, rs558415	*UBAP2*	intron_variant
rs117434123 (chr9:38996308)	C/G	1.0000	3.33%	-0.0783	0.0129	1.45E-09	NA	-	intergenic_variant
rs2493637 (chr9:109521257)	C/T	0.9878	21.16%	-0.0326	0.0057	9.16E-09	NA	-	intergenic_variant
rs73558998 (chr9:117954724)	T/A	0.8647	0.37%	-0.2422	0.0418	6.84E-09	NA	*37226*	intron_variant
rs803909 (chr9:119413447)	T/G	0.9988	46.18%	-0.0254	0.0046	3.76E-08	NA	*ASTN2*	intron_variant
rs2271804 (chr10:12252217)	G/A	0.9956	47.03%	0.0462	0.0046	1.73E-23	NA	*CDC123*	intron_variant
rs185638441 (chr10:15559576)	G/C	0.8150	0.27%	-0.3575	0.0500	8.93E-13	NA	*ITGA8*	intron_variant
rs3847402 (chr10:30267810)	G/A	0.9829	40.34%	-0.0279	0.0047	3.98E-09	NA	-	intergenic_variant
rs2579762 (chr10:78318879)	A/C	1.0000	47.15%	-0.0398	0.0046	6.63E-18	NA	*C10orf11*	downstream_gene_variant
rs4933356 (chr10:82113885)	A/T	0.9837	49.95%	0.0257	0.0047	3.25E-08	NA	*DYDC1*	intron_variant
rs116411520 (chr10:91533071)	G/A	0.8433	2.55%	-0.1025	0.0163	2.99E-10	NA	*KIF20B*	intron_variant
. (chr10:97879525)	A/AT	0.5054	1.11%	-0.3158	0.0330	1.02E-21	NA		
rs140192357 (chr10:121338937)	C/A	0.8629	0.24%	-0.2953	0.0525	1.87E-08	NA	*TIAL1*	intron_variant
rs736962 (chr10:124257996)	A/G	1.0000	2.80%	-0.0832	0.0139	1.97E-09	NA	*HTRA1*	intron_variant
rs11231161 (chr11:62378221)	A/G	1.0000	37.17%	0.0300	0.0048	2.73E-10	NA	*B3GAT3*	downstream_gene_variant
rs2027761 (chr11:73036179)	C/T	0.9992	11.19%	0.0456	0.0073	4.23E-10	NA	*ARHGEF17*	intron_variant
rs117443545 (chr11:83190398)	G/A	0.7085	1.08%	-0.3064	0.0272	2.06E-29	NA	*DLG2*	intron_variant
rs138857603 (chr11:83275030)	A/AT	0.8690	0.63%	-0.3078	0.0318	3.23E-22	rs117443545	*DLG2*	intron_variant
rs17146129 (chr11:83525694)	A/T	0.9009	0.78%	-0.2543	0.0280	1.02E-19	rs117443545, rs138857603	*DLG2*	intron_variant
rs114902756 (chr11:85820597)	G/C	0.8024	0.42%	-0.4324	0.0410	4.89E-26	NA	-	intergenic_variant

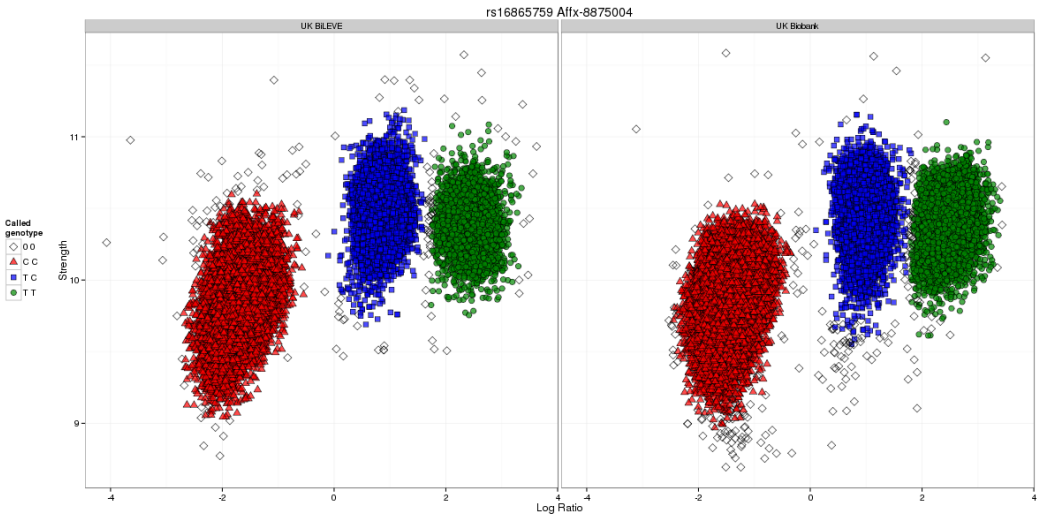
SNP (Chr:Pos)	Effect/ noneffect allele	INFO	MAF	Effect Estimate	Standard error	P-value	Conditioned SNP(s)	Gene(s)	Consequence
rs117261012 (chr11:86444761)	A/G	0.9557	15.38%	-0.0399	0.0065	1.09E-09	rs114902756	*CTD-2005H7.2*	intron_variant,non_coding_transcript_variant
rs111400016 (chr11:91208353)	G/A	1.0000	1.76%	-0.1246	0.0174	7.20E-13	NA	-	intergenic_variant
rs503441 (chr11:126010797)	A/G	0.9904	18.38%	0.0369	0.0060	6.57E-10	NA	-	intergenic_variant
rs7112357 (chr11:131981029)	G/A	0.9871	31.48%	-0.0280	0.0050	1.98E-08	NA	*NTM*	intron_variant
rs569915721 (chr12:34781058)	A/C	0.9896	0.36%	0.2186	0.0384	1.23E-08	NA	-	intergenic_variant
rs539643279 (chr12:38396219)	G/A	0.8244	0.34%	-0.2974	0.0445	2.38E-11	rs115903505	-	intergenic_variant
rs148892628 (chr12:38501518)	T/C	0.5120	1.31%	-0.1836	0.0310	3.15E-09	rs539643279, rs115903505, rs17519950	-	intergenic_variant
rs115903505 (chr12:39146668)	T/A	0.6039	2.48%	-0.5497	0.0215	5.86E-144	NA	*CPNE8*	intron_variant
rs17519950 (chr12:40695188)	A/T	0.8380	1.15%	-0.3457	0.0243	9.33E-46	NA	*LRRK2*	intron_variant
rs11107915 (chr12:95549025)	G/A	0.9960	21.46%	-0.0325	0.0056	7.10E-09	NA	*FGD6*	intron_variant
rs12427728 (chr13:98174730)	C/T	0.5654	1.05%	-0.5018	0.0336	1.74E-50	NA	-	intergenic_variant
rs976224 (chr14:24244518)	T/A	1.0000	4.45%	-0.0607	0.0111	4.91E-08	NA	-	intergenic_variant
rs147261823 (chr14:36578852)	T/C	1.0000	2.71%	-0.0772	0.0141	4.74E-08	NA	*LINC00609*	intron_variant,non_coding_transcript_variant
rs74810641 (chr14:82590256)	G/T	1.0000	3.85%	-0.0730	0.0119	9.68E-10	NA	-	intergenic_variant
rs10137684 (chr14:93494500)	G/A	0.9723	11.81%	-0.0447	0.0072	6.37E-10	NA	*ITPK1*	intron_variant
rs72731149 (chr15:49984710)	G/C	0.9905	6.69%	0.0588	0.0093	2.14E-10	NA	-	intergenic_variant
rs72750950 (chr15:63841893)	C/G	1.0000	6.60%	0.0506	0.0093	4.82E-08	NA	*USP3*	intron_variant
rs140396483 (chr15:66381494)	C/T	1.0000	1.82%	-0.1017	0.0171	2.78E-09	NA	*MEGF11*	intron_variant
rs1441358 (chr15:71612514)	T/G	1.0000	33.60%	-0.0614	0.0049	3.89E-36	NA	*THSD4*	intron_variant
rs72531998 (chr15:71797532)	CCAT/C	0.9952	17.75%	-0.0364	0.0061	1.94E-09	rs1441358	*THSD4*	intron_variant
rs116015662 (chr15:74866502)	T/C	0.8588	0.25%	-0.3692	0.0505	2.61E-13	NA	*ARID3B*	intron_variant
rs11852372 (chr15:78801394)	A/C	0.9882	33.62%	-0.0435	0.0049	1.06E-18	NA	*HYKK*	upstream_gene_variant
. (chr15:84541379)	C/CACACACACAG	0.9170	23.58%	0.0348	0.0057	8.36E-10	NA		
rs74930371 (chr16:19273328)	G/T	0.9152	1.95%	-0.2974	0.0178	7.04E-63	NA	*SYT17*	intron_variant

SNP (Chr:Pos)	Effect/ noneffect allele	INFO	MAF	Effect Estimate	Standard error	P-value	Conditioned SNP(s)	Gene(s)	Consequence
rs544559569 (chr16:30880829)	TC/T	0.5605	1.99%	-0.1342	0.0227	3.15E-09	rs371168229	*BCL7C*	intron_variant
rs371168229 (chr16:30990454)	C/T	0.9032	0.40%	-0.2408	0.0391	7.03E-10	NA	*SETD1A*	intron_variant
rs569251473 (chr16:31081057)	C/A	0.8314	0.25%	-0.2926	0.0516	1.38E-08	rs544559569, rs371168229	*ZNF646*	upstream_gene_variant
rs3104770 (chr16:52627368)	A/T	0.7563	2.77%	-0.3065	0.0171	3.84E-72	NA	*CASC16*	intron_variant,non_coding_transcript_variant
rs2060573 (chr16:58057627)	A/G	0.9906	44.77%	0.0303	0.0047	8.45E-11	NA	*MMP15*	upstream_gene_variant
rs11149827 (chr16:75435143)	A/G	0.9878	40.82%	-0.0393	0.0047	6.54E-17	NA	*CFDP1*	intron_variant
rs112426952 (chr17:9391686)	T/A	1.0000	0.78%	-0.1585	0.0262	1.35E-09	NA	*STX8*	intron_variant
rs561060101 (chr17:17776177)	G/C	0.7202	1.18%	-0.2751	0.0259	2.74E-26	rs76926781, rs201183826	*TOM1L2*	intron_variant
rs76926781 (chr17:17805129)	C/T	0.8439	1.26%	-0.2395	0.0231	3.89E-25	rs201183826	*TOM1L2*	intron_variant
rs201183826 (chr17:18107778)	AT/A	0.8990	1.26%	-0.2905	0.0222	5.13E-39	NA	*ALKBH5*	intron_variant
rs35491131 (chr17:28267098)	CT/C	0.9815	44.40%	-0.0266	0.0047	1.30E-08	NA	*EFCAB5*	upstream_gene_variant
rs35246838 (chr17:36915540)	T/C	0.9654	13.36%	-0.0452	0.0069	5.82E-11	NA	*PSMB3*	intron_variant
rs540802774 (chr17:43881141)	T/C	0.6421	24.74%	0.0661	0.0066	1.76E-23	rs73314997	*CRHR1*	intron_variant
rs73314997 (chr17:44061123)	C/T	0.6376	2.31%	-0.4944	0.0213	1.24E-119	NA	*MAPT*	missense_variant
rs56332949 (chr17:44316480)	G/C	0.7377	1.69%	-0.1954	0.0212	3.16E-20	rs540802774, rs73314997, rs556156663	*RP11-259G18.2*	upstream_gene_variant
rs556156663 (chr17:45029053)	A/G	0.8041	0.50%	-0.2889	0.0378	2.05E-14	rs73314997	*RP11-156P1.2*	intron_variant,NMD_transcript_variant
rs1878688 (chr17:64052034)	G/A	0.6725	1.26%	-0.1768	0.0257	6.39E-12	NA	*CEP112*	intron_variant
rs17178530 (chr17:69236112)	G/C	0.9921	49.16%	-0.0306	0.0046	3.38E-11	NA	-	intergenic_variant
rs8066839 (chr17:77239697)	T/A	1.0000	2.75%	-0.1043	0.0141	1.19E-13	NA	*RBFOX3*	intron_variant
rs633286 (chr18:8809273)	C/T	0.9832	27.17%	-0.0295	0.0052	1.65E-08	NA	*MTCL1*	intron_variant
rs12607689 (chr18:20016299)	T/G	0.9933	40.55%	0.0259	0.0047	4.05E-08	NA	-	intergenic_variant
rs10513996 (chr18:68744370)	T/C	1.0000	1.67%	-0.1037	0.0179	7.37E-09	NA	-	intergenic_variant
rs1995745 (chr18:74301015)	T/G	1.0000	1.71%	-0.1010	0.0177	1.21E-08	NA	*LINC00908*	intron_variant,non_coding_transcript_variant

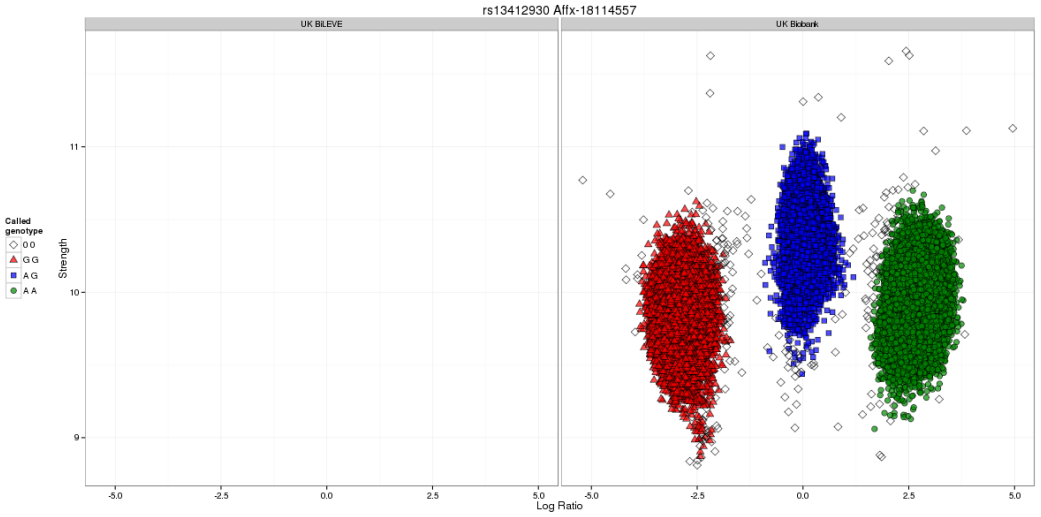
SNP (Chr:Pos)	Effect/ noneffect allele	INFO	MAF	Effect Estimate	Standard error	P-value	Conditioned SNP(s)	Gene(s)	Consequence
rs9636166 (chr19:31829613)	A/C	0.9786	12.61%	-0.0436	0.0070	5.90E-10	NA	*TSHZ3*	intron_variant
rs34093919 (chr19:41117300)	G/A	1.0000	1.20%	0.1290	0.0211	1.06E-09	NA	*LTBP4*	missense_variant,splice_region_variant
rs564454164 (chr19:54957229)	AAAG/A	0.9289	1.43%	-0.2715	0.0205	6.25E-40	NA	*LENG8*	upstream_gene_variant
rs550705585 (chr19:55832694)	T/C	0.8179	0.11%	-0.4517	0.0801	1.70E-08	rs564454164, rs143792972	*TMEM150B*	intron_variant
rs143792972 (chr19:55834177)	G/C	0.9281	0.62%	-0.3286	0.0310	2.55E-26	rs564454164	*TMEM150B*	intron_variant
rs182804848 (chr20:26290295)	T/C	0.6653	1.06%	0.1662	0.0269	6.09E-10	NA	-	intergenic_variant
rs3833318 (chr20:31031590)	TA/T	0.9908	16.89%	-0.0340	0.0062	3.70E-08	NA	*NOL4L*	3_prime_UTR_variant
rs1110660 (chr22:18444693)	A/C	0.9885	23.63%	0.0322	0.0055	3.37E-09	NA	*MICAL3*	intron_variant
rs2283847 (chr22:28181399)	C/T	0.9289	44.57%	-0.0333	0.0048	4.15E-12	NA	*MN1*	intron_variant
rs73883355 (chr22:30401775)	C/A	0.9975	8.87%	0.0580	0.0081	9.41E-13	NA	*MTMR3*	intron_variant

F. Cluster plots for PEF-specific SNPs (Chapter 5).

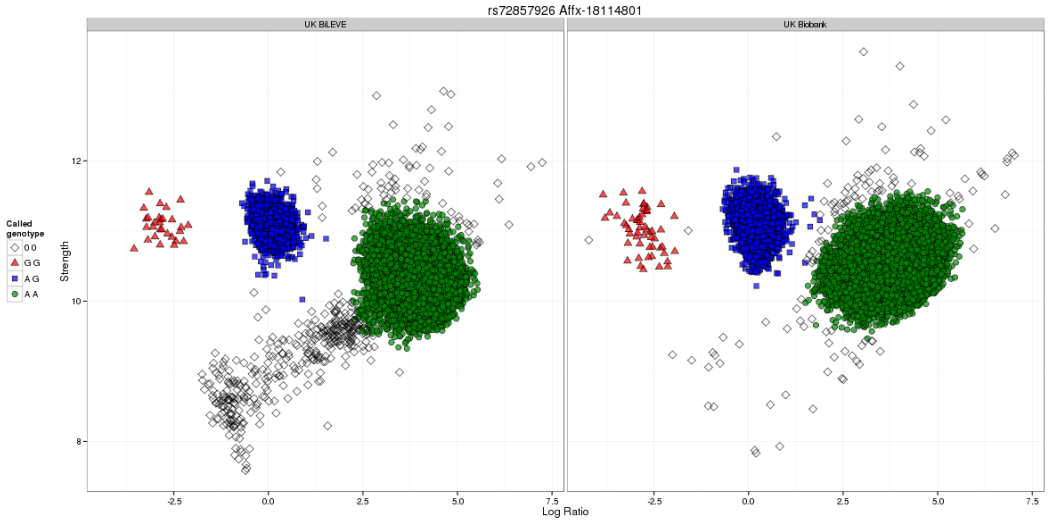
Appendix Figure F-1: Clusterplot for rs16865759 (chr1:31258724) r2=0.9997 with rs34590652.



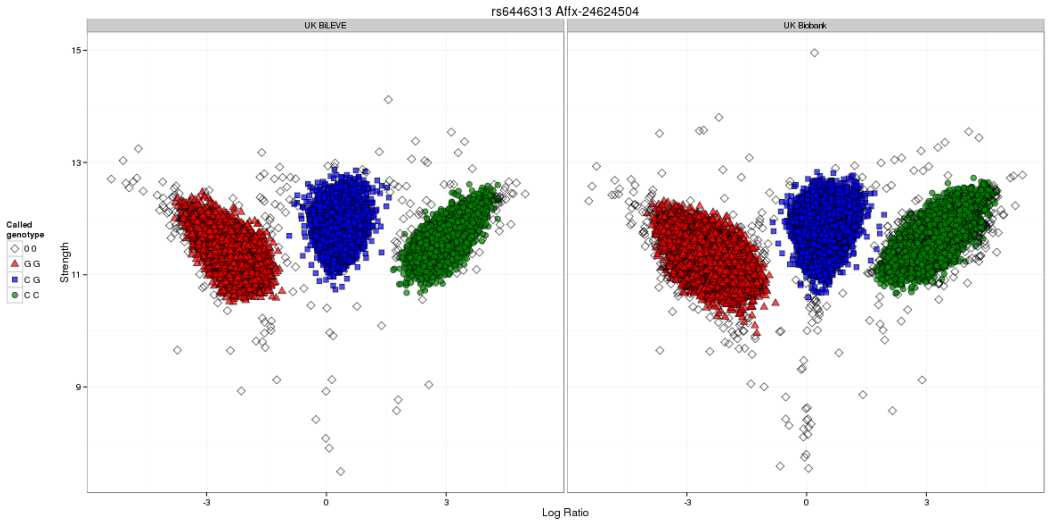
Appendix Figure F-2: Clusterplot for rs16865759, (chr1:31258724) r2=0.9997 with rs34590652.
rs16865759 only included on UK Biobank array. Strongest proxy genotyped on UK BiLEVE array was rs72857926 (chr2: 144327971); see Appendix Figure F-3.



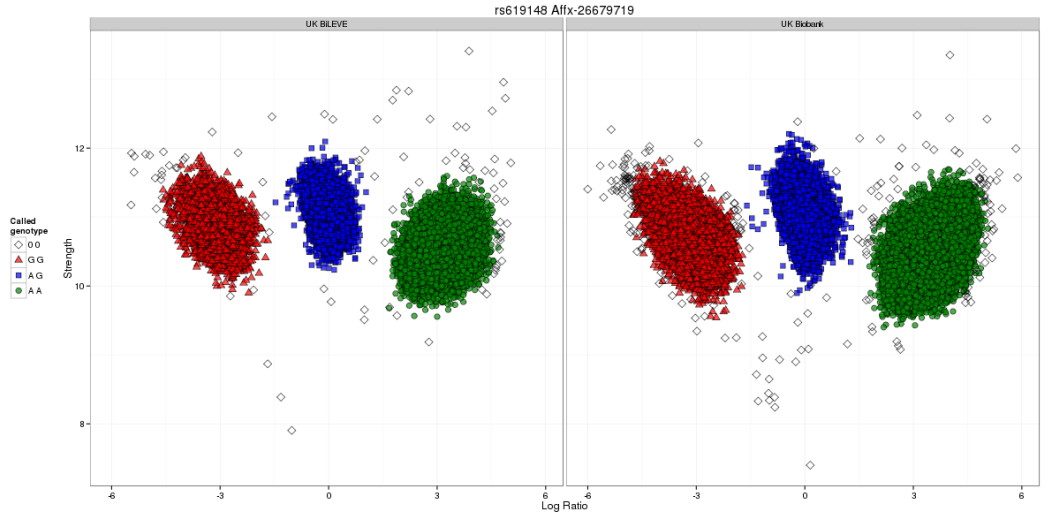
Appendix Figure F-3: Clusterplot for rs72857926 (chr2: 144327971), r2=0.1277 with rs34590652.



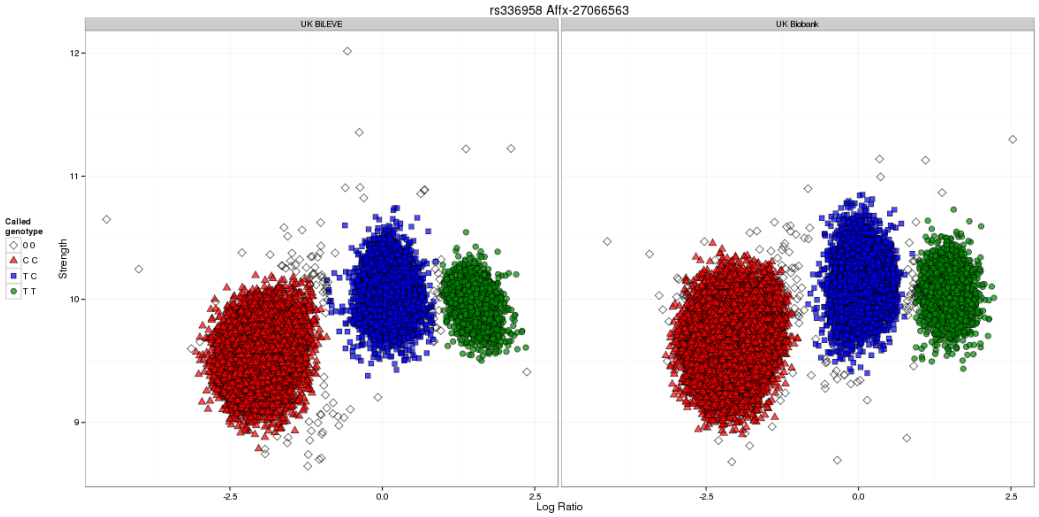
Appendix Figure F-4: Clusterplot for rs6446313 (chr4:5032174) r2=0.9459 with rs28752137.



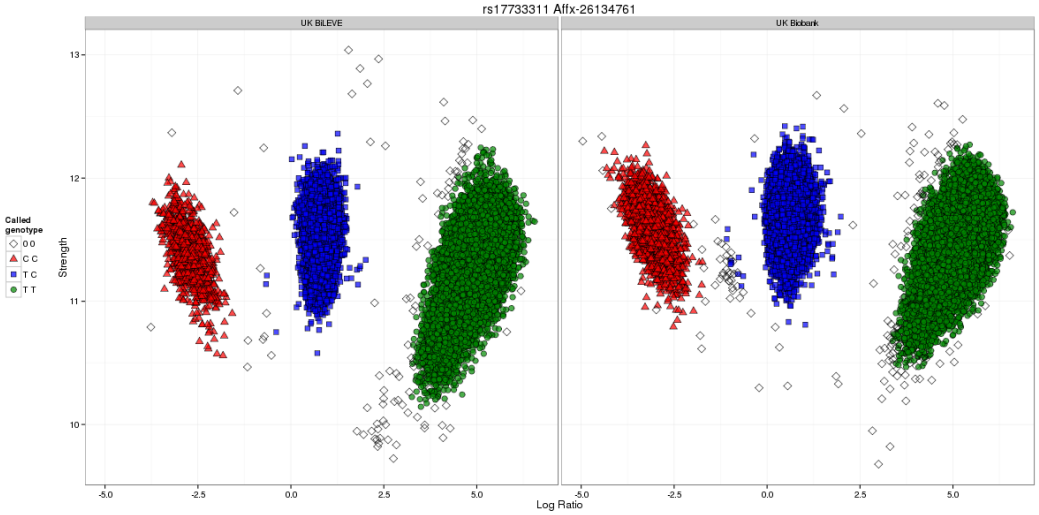
Appendix Figure F-5: Clusterplot for rs619148 (chr5:51972476) r2=0.9911 with rs350415.



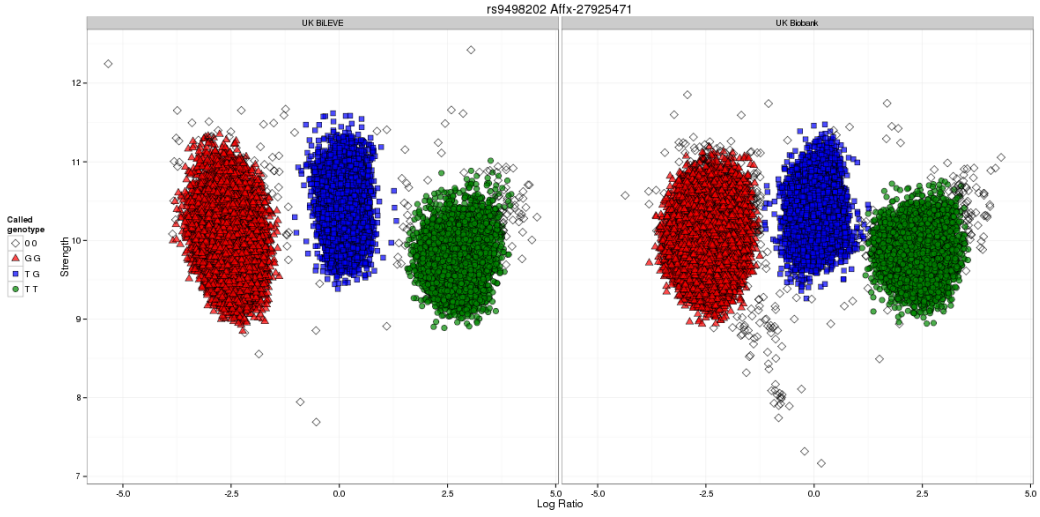
Appendix Figure F-6: Clusterplot for rs336958 (chr5:82973396) r2=0.9815 with rs4466136.



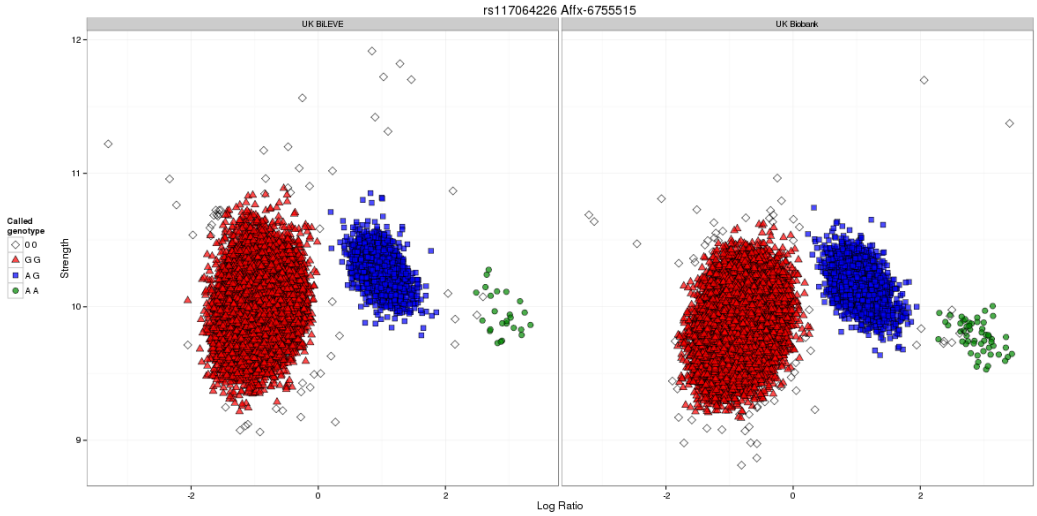
Appendix Figure F-7: Clusterplot for rs17733311 (chr5:172780104) r2=0.5066 with rs11747434.



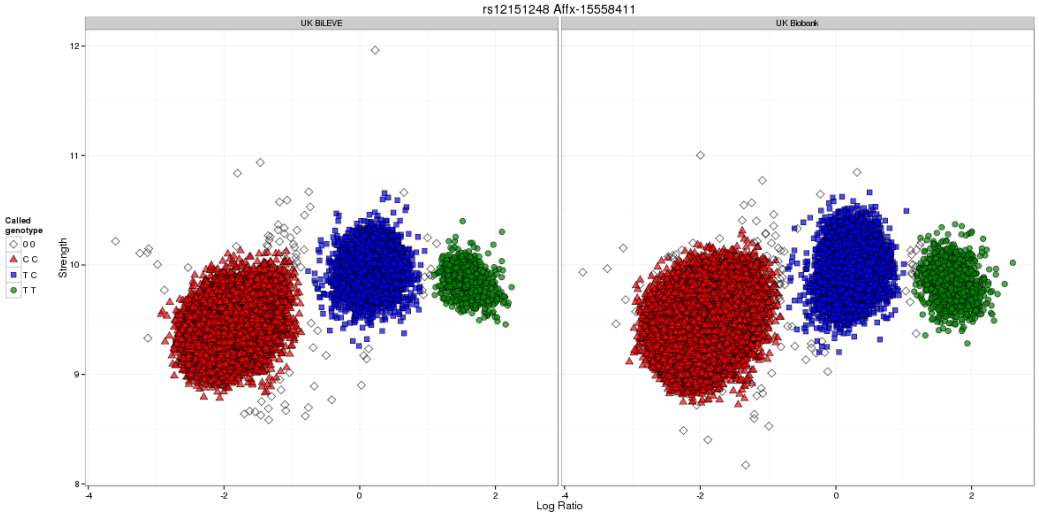
Appendix Figure F-8: Clusterplot for rs9498202 (chr6:149383257) r2=0.9336 with r s574284527.



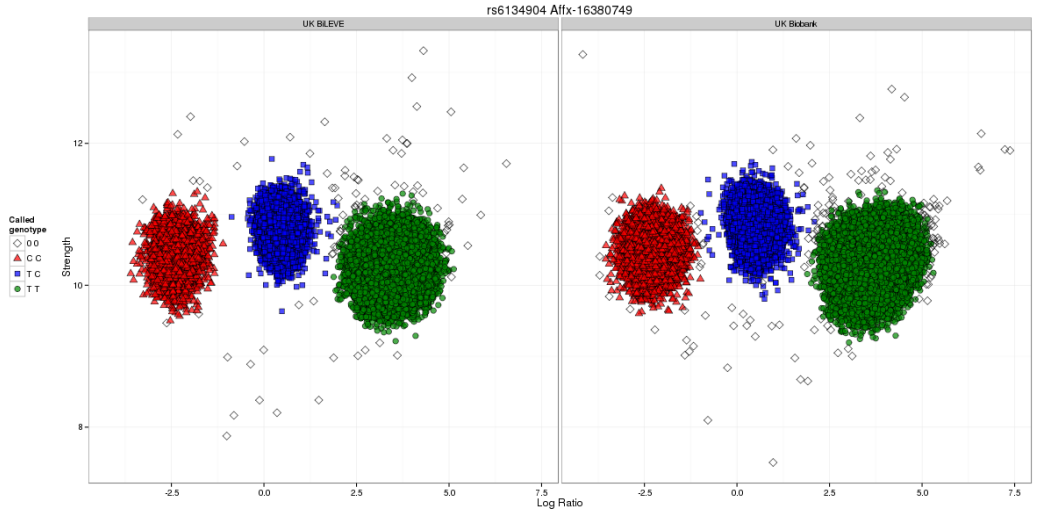
Appendix Figure F-9: Clusterplot for rs117064226 (chr12:102845393) r2=0.0102 with rs11111272.



Appendix Figure F-10: Clusterplot for rs12151248 (chr19:13212025) r2=0.0047 with rs138326911.



Appendix Figure F-11: Clusterplot for rs6134904 (chr20:13516026) r2=0.2449 with rs34590652.



References

1. Palmer LJ, Burton PR, Davey Smith G, editors. An Introduction to Genetic Epidemiology. Bristol, UK: The Policy Press; 2011.

2. Hartl DL, Jones EW. Genetics: Principles and Analysis. 4th ed. Sudbury, MA. USA: Jones and Bartlett Publishers; 1998.

3. Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. Nature Reviews Genetics. 2009;10(4):241-51.

4. Slatkin M. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. Nat Rev Genet. 2008 Jun;9(6):477-85.

5. Padmanabhan S. Handbook of Pharmacogenomics and Stratified Medicine. Academic Press; 2014.

6. Shah S, Arnett DK. Cardiovascular Genetics and Genomics in Clinical Practice. Springer Publishing Company; 2014.

7. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era—concepts and misconceptions. Nat Rev Genet. 2008 Apr;9(4):255-66.

8. Thomas DC. Statistical methods in genetic epidemiology. Oxford University Press; 2004.

9. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet. 2003;33:228-37.

10. Tabor HK, Risch NJ, Myers RM. Candidate-gene approaches for studying complex genetic traits: practical considerations. Nat Rev Genet. 2002 May;3(5):391-7.

11. Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, et al. The international HapMap project. Nature. 2003;426(6968):789-96.

12. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nature reviews. 2005;6(2):95-108.

13. Illumina Inc. Illumina GenCall Data Analysis Software. 2005.

14. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet. 2008 May;9(5):356-69.

15. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. Nature protocols. 2010;5(9):1564-73.

16. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012 Nov 1;491(7422):56-65.

17. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38(8):904-9.

18. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007 Sep;81(3):559-75.

19. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet. 2007;39(7):906-13.

20. Eu-ahsunthornwattana J, Miller EN, Fakiola M, Jeronimo SMB, Blackwell JM, Cordell HJ. Comparison of Methods to Account for Relatedness in Genome-Wide Association Studies with Family-Based Data. PLoS Genetics. 2014;10(7):e1004445.

21. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999;55(4):997-1004.

22. Lin D, Zeng D. Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. Genet Epidemiol. 2010;34(1):60-6.

23. Nyholt DR. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet. 2004 Apr;74(4):765-9.

24. Pe'er I, Yelensky R, Altshuler D, Daly MJ. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genet Epidemiol. 2008;32(4):381-5.

25. Dudbridge F, Koeleman BP. Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. The American Journal of Human Genetics. 2004;75(3):424-35.

26. Seaman S, Müller-Myhsok B. Rapid simulation of P values for product methods and multiple-testing adjustment in association studies. Am J Hum Genet. 2005 Mar;76(3):399-408.

27. Stephens M, Balding DJ. Bayesian statistical methods for genetic association studies. Nat Rev Genet. 2009 Oct;10(10):681-90.

28. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Optimal designs for two-stage genome-wide association studies. Genet Epidemiol. 2007;31(7):776-88.

29. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015;518(7538):197-206.

30. Hedman ÅK, Lindgren CM, McCarthy MI. Genome-wide association studies of obesity. In: The Genetics of Obesity. Springer; 2014. p. 33-53.

31. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014;511(7510):421-7.

32. Light G, Greenwood TA, Swerdlow NR, Calkins ME, Freedman R, Green MF, et al. Comparison of the heritability of schizophrenia and endophenotypes in the COGS-1 family study. Schizophr Bull. 2014 Nov;40(6):1404-11.

33. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet. 2014;46(11):1173-86.

34. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010;42(7):565-9.

35. Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AA, Lee SH, et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat Genet. 2015 Oct;47(10):1114-20.

36. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010;11(6):446-50.

37. Gibson G. Rare and common variants: twenty arguments. Nature Reviews Genetics. 2012;13(2):135-45.

38. Kryukov GV, Pennacchio LA, Sunyaev SR. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am J Hum Genet. 2007 Apr;80(4):727-39.

39. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nature Reviews Genetics. 2010;11(6):415-25.

40. Huang J, Howie B, McCarthy S, Memari Y, Walter K, Min J, et al. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. Nat Commun. 2015 Sep 14;6:8111.

41. McCarthy S, Das S, Kretzschmar W, Durbin R, Abecasis G, Marchini J. A reference panel of 64,976 haplotypes for genotype imputation. bioRxiv. 2015:doi: http://dx.doi.org/10.1101/035170.

42. Michailidou K, Beesley J, Lindstrom S, Canisius S, Dennis J, Lush MJ, et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. Nat Genet. 2015;47(4):373-80.

43. Soler Artigas M, Wain LV, Miller S, Kheirallah AK, Huffman JE, Ntalla I, et al. Sixteen new lung function signals identified through 1000 Genomes Project reference panel imputation. Nat Commun. 2015;6:8658.

44. Surakka I, Horikoshi M, Mägi R, Sarin A, Mahajan A, Lagou V, et al. The impact of low-frequency and rare variants on lipid levels. Nat Genet. 2015;47(6):589-97.

45. de Vries PS, Chasman DI, Sabater-Lleal M, Chen MH, Huffman JE, Steri M, et al. A meta-analysis of 120 246 individuals identifies 18 new loci for fibrinogen concentration. Hum Mol Genet. 2015 Nov 10;25(2):358-70.

46. Howie BN, Donnelly P, Marchini J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLoS Genet. 2009 Jun 19;5(6):e1000529.

47. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010;34(8):816-34.

48. Fuchsberger C, Abecasis GR, Hinds DA. Minimac2: Faster Genotype Imputation. Bioinformatics. 2015 Mar 1;31(5):782-4.

49. Browning B, Browning S. Genotype Imputation with Millions of Reference Samples. The American Journal of Human Genetics. 2016 Jan 7;98(1):116-26.

50. Exome Chip Design Wiki [Internet].; 2013 [cited 2013 Aug 30]. Available from: http://genome.sph.umich.edu/wiki/Exome_Chip_Design.

51. Goldstein JI, Crenshaw A, Carey J, Grant GB, Maguire J, Fromer M, et al. zCall: a rare variant caller for array-based genotyping: Genetics and population analysis. Bioinformatics. 2012 Oct 01;28(19):2543-5.

52. Zhou J, Tantoso E, Wong LP, Ong RT, Bei JX, Li Y, et al. iCall: a genotype-calling algorithm for rare, low-frequency and common variants on the Illumina exome array. Bioinformatics. 2014 Jun 15;30(12):1714-20.

53. Grove ML, Yu B, Cochran BJ, Haritunians T, Bis JC, Taylor KD, et al. Best practices and joint calling of the HumanExome BeadChip: the CHARGE Consortium. PLoS One. 2013;8(7):e68095.

54. Igartua C, Myers RA, Mathias RA, Pino-Yanes M, Eng C, Graves PE, et al. Ethnic-specific associations of rare and low-frequency DNA sequence variants with asthma. Nat Commun. 2015;6:5965.

55. Wessel J, Chu AY, Willems SM, Wang S, Yaghootkar H, Brody JA, et al. Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. Nat Commun. 2015 Jan 29;6:5897.

56. Mahajan A, Sim X, Ng HJ, Manning A, Rivas MA, Highland HM, et al. Identification and functional characterization of G6PC2 coding variants influencing glycemic traits define an effector transcript at the G6PC2-ABCB11 locus. PLoS Genet. 2015 Jan 27;11(1):e1004876.

57. Huyghe JR, Jackson AU, Fogarty MP, Buchkovich ML, Stančáková A, Stringham HM, et al. Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. Nat Genet. 2013;45(2):197-201.

58. Peloso GM, Auer PL, Bis JC, Voorman A, Morrison AC, Stitziel NO, et al. Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. Am J Hum Genet. 2014 Feb;94(2):223-32.

59. Auer PL, Teumer A, Schick U, O'Shaughnessy A, Lo KS, Chami N, et al. Rare and low-frequency coding variants in CXCR2 and other genes are associated with hematological traits. Nat Genet. 2014;46(6):629-34.

60. Li Z, Allingham RR, Nakano M, Jia L, Chen Y, Ikeda Y, et al. A common variant near TGFBR3 is associated with primary open angle glaucoma. Hum Mol Genet. 2015 Jul 1;24(13):3880-92.

61. Zuo X, Sun L, Yin X, Gao J, Sheng Y, Xu J, et al. Whole-exome SNP array identifies 15 new susceptibility loci for psoriasis. Nat Commun. 2015;6:6793.

62. Li B, Leal SM. Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. Am J Hum Genet. 2008 Sep;83(3):311-21.

63. Panoutsopoulou K, Tachmazidou I, Zeggini E. In search of low frequency and rare variants affecting complex traits. Hum Mol Genet. 2013 Aug 6;22(R1):R16-21.

64. Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). Mutat Res. 2007 Feb 3;615(1–2):28-56.

65. Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, Wei L, et al. Pooled Association Tests for Rare Variants in Exon-Resequencing Studies. Am J Hum Genet. 2010 Jun 11;86(6):832-8.

66. Madsen BE, Browning SR. A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. PLoS Genet. 2009 Feb 13;5(2):e1000384.

67. Mukhopadhyay I, Feingold E, Weeks DE, Thalamuthu A. Association tests using kernel-based measures of multi-locus genotype similarity between individuals. Genet Epidemiol. 2010;34(3):213-21.

68. Wu M, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. Am J Hum Genet. 2011 Jul 15;89(1):82-93.

69. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al. Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies. Am J Hum Genet. 2012 Aug 10;91(2):224-37.

70. Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. Nature Reviews Genetics. 2010;11(11):773-85.

71. Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. Nat Genet. 2012;44(3):243-6.

72. O'Connor TD, Kiezun A, Bamshad M, Rich SS, Smith JD, Turner E, et al. Fine-Scale Patterns of Population Stratification Confound Rare Variant Association Tests. PLoS ONE. 2013 Jul 4;8(7):e65834.

73. RAREMETAL Documentation [Internet].; 2013 [cited 2013 Sep]. Available from: http://genome.sph.umich.edu/wiki/RAREMETAL_Documentation.

74. Liu DJ, Peloso GM, Zhan X, Holmen O, Zawistowski M, Feng S, et al. Meta-Analysis of Gene Level Association Tests. Nat Genet. 2014;46(2):200-4.

75. Lee S, Teslovich T, Boehnke M, Lin X. General Framework for Meta-analysis of Rare Variants in Sequencing Association Studies. Am J Hum Genet. 2013 Jul 11;93(1):42-53.

76. Tang ZZ, Lin DY. MASS: meta-analysis of score statistics for sequencing studies. Bioinformatics. 2013 Jul 15;29(14):1803-5.

77. Tang Z, Lin D. Meta-Analysis of Sequencing Studies With Heterogeneous Genetic Associations. Genet Epidemiol. 2014;38(5):389-401.

78. Meta-analysis of a rare-variant association test [Internet].; 2012 [cited 2013 Sep]. Available from: http://stattech.wordpress.fos.auckland.ac.nz/files/2012/11/skat-meta-paper.pdf.

79. Miller MR, Hankinson J, Brusasco V, Burgos F, Casaburi R, Coates A, et al. Standardisation of spirometry. Eur Respir J. 2005 Aug;26(2):319-38.

80. Hayes D,Jr, Kraman SS. The physiologic basis of spirometry. Respir Care. 2009 Dec;54(12):1717-26.

81. Cotes JE, Chinn DJ, Miller MR. Lung function: physiology, measurement and application in medicine. John Wiley & Sons; 2009.

82. Elliott EA, Dawson SV. Test of wave-speed theory of flow limitation in elastic tubes. J Appl Physiol. 1977;43(3):516-22.

83. Mead J, Turner JM, Macklem PT, Little JB. Significance of the relationship between lung recoil and maximum expiratory flow. J Appl Physiol. 1967 Jan;22(1):95-108.

84. Pride NB, Permutt S, Riley RL, Bromberger-Barnea B. Determinants of maximal expiratory flow from the lungs. J Appl Physiol. 1967 Nov;23(5):646-62.

85. Hankinson J, Odencrantz J, Fedan K. Spirometric Reference Values from a Sample of the General U.S. Population. Am J Respir Crit Care Med. 1999 Jan;159(1):179-87.

86. Quanjer PH, Tammeling GJ, Cotes JE, Pedersen OF, Peslin R, Yernault JC. Lung volumes and forced ventilatory flows. Eur Respir J. 1993 Mar;6 Suppl 16:5-40.

87. Quanjer PH, Stanojevic S, Cole TJ, Baur X, Hall GL, Culver BH, et al. Multi-ethnic reference values for spirometry for the 3-95-yr age range: the global lung function 2012 equations. Eur Respir J. 2012 Dec;40(6):1324-43.

88. Brusasco V, Crapo R, Viegi G, American Thoracic Society, European Respiratory Society. Coming together: the ATS/ERS consensus on clinical pulmonary function testing. Eur Respir J. 2005 Jul;26(1):1-2.

89. Rabe KF, Hurd S, Anzueto A, Barnes PJ, Buist SA, Calverley P, et al. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease. Am J Respir Crit Care Med. 2007 Sep 15;176(6):532-55.

90. Mathers C, Fat DM, Boerma JT. The global burden of disease: 2004 update. World Health Organization; 2008.

91. Chronic obstructive pulmonary disease (COPD) [Internet].: World Health Organisation; 2015 [cited 2016 Jan 12]. Available from: http://www.who.int/mediacentre/factsheets/fs315/en/.

92. Burgel PR. The role of small airways in obstructive airway diseases. Eur Respir Rev. 2011 Mar;20(119):23-33.

93. Mannino DM, Buist AS. Global burden of COPD: risk factors, prevalence, and future trends. Lancet. 2007 Sep 1;370(9589):765-73.

94. Global Strategy for the Diagnosis, Management and Prevention of Chronic Obstructive Pulmonary Disease [Internet].; December 2015 [cited 2016 Jan 7]. Available from: http://www.goldcopd.org/uploads/users/files/GOLD_Report%202016.pdf.

95. Rennard SI, Drummond MB. Early chronic obstructive pulmonary disease: definition, assessment, and prevention. Lancet. 2015 May 2;385(9979):1778-88.

96. Postma DS, Bush A, van den Berge M. Risk factors and early origins of chronic obstructive pulmonary disease. Lancet. 2015 Mar 7;385(9971):899-909.

97. Lange P, Celli B, Agustí A, Boje Jensen G, Divo M, Faner R, et al. Lung-function trajectories leading to chronic obstructive pulmonary disease. N Engl J Med. 2015;373(2):111-22.

98. Salvi SS, Barnes PJ. Chronic obstructive pulmonary disease in non-smokers. Lancet. 2009 Aug 29;374(9691):733-43.

99. Silverman E, Speizer F. Risk factors for the development of chronic obstructive pulmonary disease. Med Clin North Am. 1996 May;80(3):501-22.

100. Mayer AS, Newman LS. Genetic and environmental modulation of chronic obstructive pulmonary disease. Respir Physiol. 2001 Oct;128(1):3-11.

101. Barnes PJ, Shapiro SD, Pauwels RA. Chronic obstructive pulmonary disease: molecular and cellular mechanisms. Eur Respir J. 2003 Oct;22(4):672-88.

102. Hogg JC, Chu F, Utokaparch S, Woods R, Elliott WM, Buzatu L, et al. The nature of small-airway obstruction in chronic obstructive pulmonary disease. N Engl J Med. 2004;350(26):2645-53.

103. Larson RK, Barman ML. The Familial Occurrence of Chronic Obstructive Pulmonary Disease. Ann Intern Med. 1965 Dec 1;63(6):1001-8.

104. Tager I, Tishler PV, Rosner B, Speizer FE, Litt M. Studies of the Familial Aggregation of Chronic Bronchitis and Obstructive Airways Disease. Int J Epidemiol. 1978 Mar 1;7(1):55-62.

105. Palmer LJ, Knuiman MW, Divitini ML, Burton PR, James AL, Bartholomew HC, et al. Familial aggregation and heritability of adult lung function: results from the Busselton Health Study. Eur Respir J. 2001 Apr 1;17(4):696-702.

106. Gottlieb D, Wilk J, Harmon M, Evans J, Joost O, Levy D, et al. Heritability of Longitudinal Change in Lung Function. Am J Respir Crit Care Med. 2001 Nov 01;164(9):1655-9.

107. Hukkinen M, Kaprio J, Broms U, Viljanen A, Kotz D, Rantanen T, et al. Heritability of Lung Function: A Twin Study Among Never-Smoking Elderly Women. Twin Research and Human Genetics. 2011;14(05):401.

108. Klimentidis YC, Vazquez AI, de lC, Allison DB, Dransfield MT, Thannickal VJ. Heritability of pulmonary function estimated from pedigree and whole-genome markers. Front Genet. 2013 Sep 19;4:174.

109. Lewiiter FI, Tager IB, Mcgue M, Tishler PV, Speizer FE. Genetic and environmental determinants of level of pulmonary function. American Journal of Epidemiology. 1984 Oct 01;120(4):518-30.

110. Wilk JB, DeStefano AL, Joost O, Myers RH, Cupples LA, Slater K, et al. Linkage and association with pulmonary function measures on chromosome 6q27 in the Framingham Heart Study. Human Molecular Genetics. 2003 Nov 01;12(21):2745-51.

111. Joost O, Wilk JB, Cupples LA, Harmon M, Shearman AM, Baldwin CT, et al. Genetic loci influencing lung function: a genome-wide scan in the Framingham Study. Am J Respir Crit Care Med. 2002 Mar 15;165(6):795-9.

112. Whitfield KE, Wiggins SA, Belue R, Brandon DT. Genetic and environmental influences on forced expiratory volume in African Americans: the Carolina African-American Twin Study of Aging. Ethn Dis. 2004;14(2):206-11.

113. Ingebrigtsen TS, Thomsen SF, van der Sluis S, Miller M, Christensen K, Sigsgaard T, et al. Genetic influences on pulmonary function: a large sample twin study. Lung. 2011;189(4):323-30.

114. DeMeo DL, Celedón JC, Lange C, Reilly JJ, Chapman HA, Sylvia JS, et al. Genome-wide linkage of forced mid-expiratory flow in chronic obstructive pulmonary disease. American journal of respiratory and critical care medicine. 2004;170(12):1294-301.

115. DeMeo DL, Carey VJ, Chapman HA, Reilly JJ, Ginns LC, Speizer FE, et al. Familial aggregation of FEF(25-75) and FEF(25-75)/FVC in families with severe, early onset COPD. Thorax. 2004 May;59(5):396-400.

116. Wilk JB, Chen T, Gottlieb DJ, Walter RE, Nagle MW, Brandler BJ, et al. A Genome-Wide Association Study of Pulmonary Function Measures in the Framingham Heart Study. PLoS Genet. 2009 Mar 20;5(3):e1000429.

117. Repapi E, Sayers I, Wain LV, Burton PR, Johnson T, Obeidat M, et al. Genome-wide association study identifies five loci associated with lung function. Nat Genet. 2010;42(1):36-44.

118. Hancock DB, Eijgelsheim M, Wilk JB, Gharib SA, Loehr LR, Marciante KD, et al. Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. Nat Genet. 2010;42(1):45-52.

119. Soler Artigas M, Loth DW, Wain LV, Gharib SA, Obeidat M, Tang W, et al. Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. Nat Genet. 2011;43(11):1082-90.

120. Wain LV, Shrine N, Miller S, Jackson VE, Ntalla I, Soler Artigas M, et al. Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. Lancet Respir Med. 2015 Oct;3(10):769-81.

121. Imboden M, Kumar A, Curjuric I, Adam M, Thun GA, Haun M, et al. Modification of the association between PM10 and lung function decline by cadherin 13 polymorphisms in the SAPALDIA cohort: a genome-wide interaction analysis. Environ Health Perspect. 2015 Jan;123(1):72-9.

122. Wilk JB, Walter RE, Laramie JM, Gottlieb DJ, O'Connor GT. Framingham Heart Study genome-wide association: results for pulmonary function measures. BMC Med Genet. 2007;8(Suppl 1):S8.

123. Freidin M, Bragina EY, Fedorova O, Deev I, Kulikov E, Ogorodova L, et al. Genome-wide association study of allergic diseases in Russians of West Siberia. Mol Biol (N Y). 2011;45(3):421-9.

124. Dominiczak AF, Connell JMC. Genetics of Hypertension, Volume 24. Elsevier Health Sciences; 2007.

125. Castaldi PJ, Cho MH, Litonjua AA, Bakke P, Gulsvik A, Lomas DA, et al. The Association of Genome-Wide Significant Spirometric Loci with Chronic Obstructive Pulmonary Disease Susceptibility. Am J Respir Cell Mol Biol. 2011 Dec;45(6):1147-53.

126. Cho MH, Boutaoui N, Klanderman BJ, Sylvia JS, Ziniti JP, Hersh CP, et al. Variants in FAM13A are associated with chronic obstructive pulmonary disease. Nat Genet. 2010;42(3):200-2.

127. Soler Artigas M, Wain LV, Repapi E, Obeidat M, Sayers I, Burton PR, et al. Effect of Five Genetic Variants Associated with Lung Function on the Risk of Chronic Obstructive Lung Disease, and Their Joint Effects on Lung Function. Am J Respir Crit Care Med. 2011 Oct 1;184(7):786-95.

128. Cho MH, McDonald MN, Zhou X, Mattheisen M, Castaldi PJ, Hersh CP, et al. Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis. The Lancet Respiratory Medicine. 2014 Mar;2(3):214-25.

129. Pillai SG, Ge D, Zhu G, Kong X, Shianna KV, Need AC, et al. A Genome-Wide Association Study in Chronic Obstructive Pulmonary Disease (COPD): Identification of Two Major Susceptibility Loci. PLoS Genet. 2009 Mar 20;5(3):e1000421.

130. Wilk JB, Shrine NRG, Loehr LR, Zhao JH, Manichaikul A, Lopez LM, et al. Genome-Wide Association Studies Identify CHRNA5/3 and HTR4 in the Development of Airflow Obstruction. Am J Respir Crit Care Med. 2012 Oct 1;186(7):622-32.

131. Munafò MR, Timofeeva MN, Morris RW, Prieto-Merino D, Sattar N, Brennan P, et al. Association Between Genetic Variants on Chromosome 15q25 Locus and Objective Measures of Tobacco Exposure. Journal of the National Cancer Institute. 2012 May 16;104(10):740-8.

132. The Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nat Genet. 2010;42(5):441-7.

133. Liu JZ, Tozzi F, Waterworth DM, Pillai SG, Muglia P, Middleton L, et al. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. Nat Genet. 2010;42(5):436-40.

134. Thorgeirsson TE, Gudbjartsson DF, Surakka I, Vink JM, Amin N, Geller F, et al. Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. Nat Genet. 2010;42(5):448-53.

135. Thun GA, Imboden M, Ferrarotti I, Kumar A, Obeidat M, Zorzetto M, et al. Causal and Synthetic Associations of Variants in the *SERPINA* Gene Cluster with Alpha1-antitrypsin Serum Levels. PLoS Genet. 2013 Aug 22;9(8):e1003585.

136. Loth DW, Soler Artigas M, Gharib SA, Wain LV, Franceschini N, Koch B, et al. Genome-wide association analysis identifies six new loci associated with forced vital capacity. Nat Genet. 2014;46(7):669-77.

137. Hunninghake GM, Cho MH, Tesfaigzi Y, Soto-Quiros M, Avila L, Lasky-Su J, et al. MMP12, Lung Function, and COPD in High-Risk Populations. N Engl J Med. 2009 Dec 31;361(27):2599-608.

138. Seals DF, Courtneidge SA. The ADAMs family of metalloproteases: multidomain proteins with multiple functions. Genes Dev. 2003 Jan 1;17(1):7-30.

139. Visse R, Nagase H. Matrix metalloproteinases and tissue inhibitors of metalloproteinases: structure, function, and biochemistry. Circ Res. 2003 May 2;92(8):827-39.

140. Khokha R, Murthy A, Weiss A. Metalloproteinases and their natural inhibitors in inflammation and immunity. Nat Rev Immunol. 2013 Sep 13;13(9):649-65.

141. Whitsett JA, Wert SE, Trapnell BC. Genetic disorders influencing lung formation and function at birth. Hum Mol Genet. 2004 Oct 1;13 Spec No 2:R207-15.

142. Warburton D, Bellusci S, De Langhe S, Del Moral P, Fleury V, Mailleux A, et al. Molecular mechanisms of early lung specification and branching morphogenesis. Pediatr Res. 2005;57:26R-37R.

143. Nebert DW, Vasiliou V. Analysis of the glutathione S-transferase (GST) gene family. Hum Genomics. 2004 Nov;1(6):460-4.

144. Buckley ST, Ehrhardt C. The receptor for advanced glycation end products (RAGE) and the lung. J Biomed Biotechnol. 2010;2010:917108.

145. Sterner-Kock A, Thorey IS, Koli K, Wempe F, Otte J, Bangsow T, et al. Disruption of the gene encoding the latent transforming growth factor-beta binding protein 4 (LTBP-4) causes abnormal lung development, cardiomyopathy, and colorectal cancer. Genes Dev. 2002 Sep 1;16(17):2264-73.

146. Dabovic B, Robertson IB, Zilberberg L, Vassallo M, Davis EC, Rifkin DB. Function of latent TGFβ binding protein 4 and fibulin 5 in elastogenesis and lung development. J Cell Physiol. 2015;230(1):226-36.

147. Zhernakova A, van Diemen CC, Wijmenga C. Detecting shared pathogenesis from the shared genetics of immune-related diseases. Nat Rev Genet. 2009;10(1):43-55.

148. Jackson VE, Ntalla I, Sayers I, Morris R, Whincup P, Casas JP, et al. Exome-wide analysis of rare coding variation identifies novel associations with COPD and airflow limitation in MOCS3, IFIT3 and SERPINA12. Thorax. 2016 Feb 25:doi:10.1136/thoraxjnl-2015-207876.

149. Mahajan A, Robertson N, Rayner W. Exome-Chip Quality Control SOP. Version 5, 2012-11-20. 2012.

150. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research. 2010 Sep 1;38(16):e164.

151. McKay JD, Hung RJ, Gaborieau V, Boffetta P, Chabrier A, Byrnes G, et al. Lung cancer susceptibility locus at 5p15.33. Nat Genet. 2008;40(12):1404-6.

152. Landi MT, Chatterjee N, Yu K, Goldin LR, Goldstein AM, Rotunno M, et al. A Genome-wide Association Study of Lung Cancer Identifies a Region of Chromosome 5p15 Associated with Risk for Adenocarcinoma. Am J Hum Genet. 2009 Nov 13;85(5):679-91.

153. Kastury K, Taylor WE, Shen R, Arver S, Gutierrez M, Fisher CE, et al. Complementary Deoxyribonucleic Acid Cloning and Characterization of a Putative Human Axonemal Dynein Light Chain Gene. Journal of Clinical Endocrinology & Metabolism. 1997 Sep 01;82(9):3047-53.

154. Matthies A, Nimtz M, Leimkühler S. Molybdenum cofactor biosynthesis in humans: identification of a persulfide group in the rhodanese-like domain of MOCS3 by mass spectrometry. Biochemistry (N Y). 2005;44(21):7912-20.

155. Dahl M, Tybjærg-Hansen A, Lange P, Vestbo J, Nordestgaard BG. Change in Lung Function and Morbidity from Chronic Obstructive Pulmonary Disease in α1-Antitrypsin MZ Heterozygotes: A Longitudinal Study of the General Population. Annals of Internal Medicine. 2002 Feb 19;136(4):270-9.

156. Molloy K, Hersh CP, Morris VB, Carroll TP, O'Connor CA, Lasky-Su J, et al. Clarification of the Risk of Chronic Obstructive Pulmonary Disease in α1-Antitrypsin Deficiency PiMZ Heterozygotes. Am J Respir Crit Care Med. 2014 Feb 15;189(4):419-27.

157. Ternette N, Wright C, Kramer H, Altun M, Kessler B. Label-free quantitative proteomics reveals regulation of interferon-induced protein with tetratricopeptide repeats 3 (IFIT3) and 5'-3'-exoribonuclease 2 (XRN2) during respiratory syncytial virus infection. Virology Journal. 2011;8(1):442.

158. Hsu Y, Shi S, Wu W, Ho L, Lai J. Protective Roles of Interferon-Induced Protein with Tetratricopeptide Repeats 3 (IFIT3) in Dengue Virus Infection of Human Lung Epithelial Cells. PLoS ONE. 2013 Nov 04;8(11):e79518.

159. Wild PS, Zeller T, Schillert A, Szymczak S, Sinning CR, Deiseroth A, et al. A genome-wide association study identifies LIPA as a susceptibility gene for coronary artery disease. Circ Cardiovasc Genet. 2011 Aug 1;4(4):403-12.

160. Kim DS, Burt AA, Crosslin DR, Robertson PD, Ranchalis JE, Boyko EJ, et al. Novel common and rare genetic determinants of paraoxonase activity: FTO, SERPINA12, and ITGAL. Journal of Lipid Research. 2013 Feb 1;54(2):552-60.

161. Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. Nat Rev Genet. 2014 May;15(5):335-46.

162. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet. 2014 Jul 3;95(1):5-23.

163. Evangelou E, Ioannidis JPA. Meta-analysis methods for genome-wide association studies and beyond. Nat Rev Genet. 2013;14(6):379-89.

164. Fisher RA. Statistical methods for research workers. Genesis Publishing Pvt Ltd; 1925.

165. Stouffer SA, Suchman EA, DeVinney LC, Star SA, Williams RM, Jr. The American soldier. Vol 1: Adjustment during army life. Princeton University Press; 1949.

166. Feng S, Pistis G, Zhang H, Zawistowski M, Mulas A, Zoledziewska M, et al. Methods for Association Analysis and Meta-Analysis of Rare Variants in Families. Genet Epidemiol. 2015;39(4):227-38.

167. Tang Z, Lin D. Meta-analysis for Discovering Rare-Variant Associations: Statistical Methods and Software Programs. Am J Hum Genet. 2015 Jul 2;97(1):35-53.

168. Hu Y, Berndt S, Gustafsson S, Ganna A, Hirschhorn J, North KE, et al. Meta-analysis of Gene-Level Associations for Rare Variants Based on Single-Variant Statistics. The American Journal of Human Genetics. 2013 Aug 8;93(2):236-48.

169. Brown T. Introduction to genetics: a molecular approach. Garland Science; 2011.

170. Ma C, Blackwell T, Boehnke M, Scott LJ, the GoT2D investigators. Recommended Joint and Meta-Analysis Strategies for Case-Control Association Testing of Single Low-Count Variants. Genet Epidemiol. 2013;37(6):539-50.

171. Consortium DS, Consortium DM, Mahajan A, Go MJ, Zhang W, Below JE, et al. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nat Genet. 2014;46(3):234-44.

172. Liu L, Sabo A, Neale BM, Nagaswamy U, Stevens C, Lim E, et al. Analysis of rare, exonic variation amongst subjects with autism spectrum disorders and population controls. PLoS Genet. 2013 Apr;9(4):e1003443.

173. Wilk JB, Djousse L, Arnett DK, Rich SS, Province MA, Hunt SC, et al. Evidence for major genes influencing pulmonary function in the NHLBI Family Heart Study. Genet Epidemiol. 2000;19(1):81-94.

174. Lin D, Zeng D. Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. Genet Epidemiol. 2010;34(1):60-6.

175. EPACTS - Genome Analysis Wiki [Internet].; 2015 [cited 9/25/2015]. Available from: http://genome.sph.umich.edu/wiki/EPACTS.

176. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nature genetics. 2012 Jul;44(7):821-4.

177. Yang J, Ferreira T, Morris AP, Medland SE, Madden PA, Heath AC, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet. 2012;44(4):369-75.

178. Westra H, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013;45(10):1238-43.

179. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45(6):580-5.

180. Obeidat M, Miller S, Probert K, Billington CK, Henry AP, Hodge E, et al. GSTCD and INTS12 Regulation and Expression in the Human Lung. PLoS ONE. 2013 Sep 18;8(9):e74630.

181. Lamontagne M, Couture C, Postma DS, Timens W, Sin DD, Pare PD, et al. Refining susceptibility loci of chronic obstructive pulmonary disease with lung eqtls. PLoS One. 2013;8(7):e70220.

182. Hao K, Bossé Y, Nickle DC, Paré PD, Postma DS, Laviolette M, et al. Lung eQTLs to help reveal the molecular underpinnings of asthma. PLoS Genet. 2012;8(11):e1003029.

183. Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T, et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nat Genet. 2008;40(5):616-22.

184. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, et al. Towards a knowledge-based Human Protein Atlas. Nat Biotech. 2010;28(12):1248-50.

185. Kimoto M, Nagasawa K, Miyake K. Role of TLR4/MD-2 and RP105/MD-1 in innate recognition of lipopolysaccharide. Scand J Infect Dis. 2003;35(9):568-72.

186. Heid IM, Jackson AU, Randall JC, Winkler TW, Qi L, Steinthorsdottir V, et al. Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. Nat Genet. 2010;42(11):949-60.

187. Tan JY, Luo YL, Huang X, Shao JL, Lin L, Yang XX. Association of single nucleotide polymorphisms of MD-1 gene with asthma in adults of Han Nationality in Southern China. Zhonghua Jie He He Hu Xi Za Zhi. 2011 Feb;34(2):104-8.

188. Lee S, Wang J, Hsieh Y, Wu Y, Ting H, Wu L. Association of single nucleotide polymorphisms of MD-1 gene with pediatric and adult asthma in the Taiwanese population. J Microbiol Immunol Infect. 2008;41(6):445-9.

189. Klar J, Blomstrand P, Brunmark C, Badhai J, Håkansson HF, Brange CS, et al. Fibroblast growth factor 10 haploinsufficiency causes chronic obstructive pulmonary disease. Journal of Medical Genetics. 2011 Oct 1;48(10):705-9.

190. Sekine K, Ohuchi H, Fujiwara M, Yamasaki M, Yoshizawa T, Sato T, et al. Fgf10 is essential for limb and lung formation. Nat Genet. 1999;21(1):138-41.

191. Gupte VV, Ramasamy SK, Reddy R, Lee J, Weinreb PH, Violette SM, et al. Overexpression of fibroblast growth factor-10 during both inflammatory and fibrotic phases attenuates bleomycin-induced pulmonary fibrosis in mice. American journal of respiratory and critical care medicine. 2009;180(5):424-36.

192. Quigley DA, Fiorito E, Nord S, Van Loo P, Alnæs GG, Fleischer T, et al. The 5p12 breast cancer susceptibility locus affects MRPS30 expression in estrogen-receptor positive tumors. Molecular oncology. 2014;8(2):273-84.

193. Quanjer P, Lebowitz M, Gregg I, Miller M, Pedersen O. Peak expiratory flow: conclusions and recommendations of a Working Party of the European Respiratory Society. Eur Respir J Suppl. 1997 Feb;10(24):2S-8S.

194. Tepper RS, Wise RS, Covar R, Irvin CG, Kercsmar CM, Kraft M, et al. Asthma outcomes: pulmonary physiology. J Allergy Clin Immunol. 2012;129(3):S65-87.

195. Simon MR, Chinchilli VM, Phillips BR, Sorkness CA, Lemanske RF, Szefler SJ, et al. Forced expiratory flow between 25% and 75% of vital capacity and FEV 1/forced vital capacity ratio in relation to clinical and physiological parameters in asthmatic children with normal FEV 1 values. J Allergy Clin Immunol. 2010;126(3):527,534. e8.

196. Cirillo I, Klersy C, Marseglia GL, Vizzaccaro A, Pallestrini E, Tosca M, et al. Role of FEF 25%–75% as a predictor of bronchial hyperreactivity in allergic patients. Annals of Allergy, Asthma & Immunology. 2006;96(5):692-700.

197. Pellegrino R, Viegi G, Brusasco V, Crapo RO, Burgos F, Casaburi R, et al. Interpretative strategies for lung function tests. Eur Respir J. 2005 Nov;26(5):948-68.

198. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010 Nov 15;26(22):2867-73.

199. Boettger LM, Handsaker RE, Zody MC, McCarroll SA. Structural haplotypes and recent evolution of the human 17q21. 31 region. Nat Genet. 2012;44(8):881-5.

200. 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.

201. Lee JH, Cho MH, Hersh CP, McDonald MN, Crapo JD, Bakke PS, et al. Genetic susceptibility for chronic bronchitis in chronic obstructive pulmonary disease. Respir Res. 2014 Sep 21;15:113.

202. Castaldi PJ, Cho MH, San José Estépar R, McDonald MN, Laird N, Beaty TH, et al. Genome-wide association identifies regulatory Loci associated with distinct local histogram emphysema patterns. American journal of respiratory and critical care medicine. 2014;190(4):399-409.

203. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009 Oct 08;461(7265):747-53.

204. Damer CK, Bayeva M, Hahn ES, Rivera J, Socec CI. Copine A, a calcium-dependent membrane-binding protein, transiently localizes to the plasma membrane and intracellular vacuoles in Dictyostelium. BMC cell biology. 2005;6(1):1.

205. Bachmann-Gagescu R, Dona M, Hetterschijt L, Tonnaer E, Peters T, de Vrieze E, et al. The Ciliopathy Protein CC2D2A Associates with NINL and Functions in RAB8-MICAL3-Regulated Vesicle Trafficking. PLoS Genet. 2015;11(10):e1005575.

206. Zheng H, Forgetta V, Hsu Y, Estrada K, Rosello-Diez A, Leo PJ, et al. Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. Nature. 2015;526(7571):112-7.

207. Bakouh N, Bienvenu T, Thomas A, Ehrenfeld J, Liote H, Roussel D, et al. Characterization of SLC26A9 in Patients with CF-Like Lung Disease. Hum Mutat. 2013;34(10):1404-14.

208. Anagnostopoulou P, Riederer B, Duerr J, Michel S, Binia A, Agrawal R, et al. SLC26A9-mediated chloride secretion prevents mucus obstruction in airway inflammation. J Clin Invest. 2012 Oct;122(10):3629-34.

209. Tada H, Won HH, Melander O, Yang J, Peloso GM, Kathiresan S. Multiple associated variants increase the heritability explained for plasma lipids and coronary artery disease. Circ Cardiovasc Genet. 2014 Oct;7(5):583-7.

210. Spicer AP, Joo A, Bowling RA,Jr. A hyaluronan binding link protein gene family whose members are physically linked adjacent to chondroitin sulfate proteoglycan core protein genes: the missing links. J Biol Chem. 2003 Jun 6;278(23):21083-91.

211. Watanabe H, Yamada Y. Mice lacking link protein develop dwarfism and craniofacial abnormalities. Nat Genet. 1999;21(2):225-9.

212. Wirrig EE, Snarr BS, Chintalapudi MR, O'Neal JL, Phelps AL, Barth JL, et al. Cartilage link protein 1 (Crtl1), an extracellular matrix component playing an important role in heart development. Dev Biol. 2007;310(2):291-303.

213. Klatt AR, Becker AA, Neacsu CD, Paulsson M, Wagener R. The matrilins: modulators of extracellular matrix assembly. Int J Biochem Cell Biol. 2011;43(3):320-30.

214. Deák F, Wagener R, Kiss I, Paulsson M. The matrilins: a novel family of oligomeric extracellular matrix proteins. Matrix Biology. 1999;18(1):55-64.

215. Willems SM, Cornes BK, Brody JA, Morrison AC, Lipovich L, Dauriz M, et al. Association of the IGF1 gene with fasting insulin levels. Eur J Hum Genet. 2016:doi: 10.1038/ejhg.2016.4.

216. Okada Y, Kamatani Y, Takahashi A, Matsuda K, Hosono N, Ohmiya H, et al. A genome-wide association study in 19 633 Japanese subjects identified LHX3-QSOX2 and IGF1 as adult height loci. Hum Mol Genet. 2010 Jun 1;19(11):2303-12.

217. Johnston LB, Dahlgren J, Leger J, Gelander L, Savage MO, Czernichow P, et al. Association between insulin-like growth factor I (IGF-I) polymorphisms, circulating IGF-I, and pre-and postnatal growth in two European small for gestational age populations. J Clin Endocrinol Metab. 2003 Oct;88(10):4805-10.

218. Rivadeneira F, Houwing-Duistermaat JJ, Vaessen N, Vergeer-Drop JM, Hofman A, Pols HA, et al. Association between an insulin-like growth factor I gene promoter polymorphism and bone mineral density in the elderly: the Rotterdam Study. J Clin Endocrinol Metab. 2003;88(8):3878-84.

219. Al-Zahrani A, Sandhu MS, Luben RN, Thompson D, Baynes C, Pooley KA, et al. IGF1 and IGFBP3 tagging polymorphisms are associated with circulating levels of IGF1, IGFBP3 and risk of breast cancer. Hum Mol Genet. 2006 Jan 1;15(1):1-10.

220. Dupuis J, Langenberg C, Prokopenko I, Saxena R, Soranzo N, Jackson AU, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet. 2010;42(2):105-16.

221. Pais RS, Moreno-Barriuso N, Hernández-Porras I, López IP, De Las Rivas J, Pichel JG. Transcriptome analysis in prenatal IGF1-deficient mice identifies molecular

pathways and target genes involved in distal lung differentiation. PLoS ONE. 2013 Dec 31;8(12):e83028.

222. Han RN, Post M, Tanswell AK, Lye SJ. Insulin-like growth factor-I receptor–mediated vasculogenesis/angiogenesis in human lung development. American journal of respiratory cell and molecular biology. 2003;28(2):159-69.

223. Ouchida R, Kurosaki T, Wang JY. A role for lysosomal-associated protein transmembrane 5 in the negative regulation of surface B cell receptor levels and B cell activation. J Immunol. 2010 Jul 1;185(1):294-301.

224. Cortese R, Hartmann O, Berlin K, Eckhardt F. Correlative gene expression and DNA methylation profiling in lung development nominate new biomarkers in lung cancer. Int J Biochem Cell Biol. 2008;40(8):1494-508.

225. Guo X, Li Y, Ding X, He M, Wang X, Zhang H. Association Tests of Multiple Phenotypes: ATeMP. PloS one. 2015;10(10):e0140348.

226. Ray D, Pankow JS, Basu S. USAT: A Unified Score-Based Association Test for Multiple Phenotype-Genotype Analysis. Genet Epidemiol. 2016;40(1):20-34.

227. Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat Methods. 2014 Apr;11(4):407-9.

228. Stephens M. A unified framework for association analysis with multiple related phenotypes. PLoS ONE. 2013 Jul 5;8(7):e65245.

229. O'Reilly PF, Hoggart CJ, Pomyen Y, Calboli FC, Elliott P, Jarvelin M, et al. MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. PLoS ONE. 2012;7(5):e34861.

230. UK10K Consortium. The UK10K project identifies rare variants in health and disease. Nature. 2015;526(7571):82-90.

231. Kinnersley B, Labussière M, Holroyd A, Di Stefano A, Broderick P, Vijayakrishnan J, et al. Genome-wide association study identifies multiple susceptibility loci for glioma. Nat Commun. 2015 Oct 1;6:8559.

232. Myocardial Infarction Genetics and CARDIoGRAM Exome Consortia Investigators. Coding Variation in ANGPTL4, LPL, and SVEP1 and the Risk of Coronary Disease. N Engl J Med. 2016 Mar 24;374(12):1134-44.

233. Holmen OL, Zhang H, Zhou W, Schmidt E, Hovelson DH, Langhammer A, et al. No large-effect low-frequency coding variation found for myocardial infarction. Human Molecular Genetics. 2014 Sep 1;23(17):4721-8.

234. Chung SJ, Kim M, Kim J, Kim YJ, You S, Koh J, et al. Exome array study did not identify novel variants in Alzheimer's disease. Neurobiol Aging. 2014;35(8):1958. e13-14.

235. Richards AL, Leonenko G, Walters JT, Kavanagh DH, Rees EG, Evans A, et al. Exome arrays capture polygenic rare variant contributions to schizophrenia. Hum Mol Genet. 2016 Mar 1;25(5):1001-7.

236. Hobbs BD, Parker MM, Chen H, Lao T, Hardin M, Qiao D, et al. Exome Array Analysis Identifies A Common Variant in IL27 Associated with Chronic Obstructive Pulmonary Disease. Am J Respir Crit Care Med. 2016 Jan 15:doi:10.1164/rccm.201510-2053OC.

237. Lutz SM, Cho MH, Young K, Hersh CP, Castaldi PJ, McDonald M, et al. A genome-wide association study identifies risk loci for spirometric measures among smokers of European and African ancestry. BMC Genetics. 2015 Dec 3;16:138.

238. Manichaikul A, Hoffman EA, Smolonska J, Gao W, Cho MH, Baumhauer H, et al. Genome-wide study of percent emphysema on computed tomography in the general population. The Multi-Ethnic Study of Atherosclerosis Lung/SNP Health Association Resource Study. Am J Respir Crit Care Med. 2014 Feb 15;189(4):408-18.

239. Cho MH, Castaldi PJ, Hersh CP, Hobbs BD, Barr RG, Tal-Singer R, et al. A genome-wide association study of emphysema and airway quantitative imaging phenotypes. American journal of respiratory and critical care medicine. 2015;192(5):559-69.

240. Auer PL, Lettre G. Rare variant association studies: considerations, challenges and opportunities. Genome Med. 2015;7(1):16.

241. Hatzikotoulas K, Gilly A, Zeggini E. Using population isolates in genetic association studies. Brief Funct Genomics. 2014 Sep;13(5):371-7.

242. Hancock DB, Soler Artigas M, Gharib SA, Henry A, Manichaikul A, Ramasamy A, et al. Genome-wide joint meta-analysis of SNP and SNP-by-smoking interaction identifies novel loci for pulmonary function. PLoS Genet. 2012;8(12):e1003098.

243. de Jong K, Vonk JM, Timens W, Bossé Y, Sin DD, Hao K, et al. Genome-wide interaction study of gene-by-occupational exposure and effects on FEV 1 levels. J Allergy Clin Immunol. 2015;136(6):1664,1672. e14.

244. Spain SL, Barrett JC. Strategies for fine-mapping complex traits. Hum Mol Genet. 2015 Oct 15;24(R1):R111-9.

245. van de Bunt M, Cortes A, Brown MA, Morris AP, McCarthy MI, IGAS Consortium. Evaluating the Performance of Fine-Mapping Strategies at Common Variant GWAS Loci. PLoS Genet. 2015 Sep 25;11(9):e1005535.

246. Hormozdiari F, Kichaev G, Yang WY, Pasaniuc B, Eskin E. Identification of causal genes for complex traits. Bioinformatics. 2015 Jun 15;31(12):i206-13.

247. Kichaev G, Yang W, Lindstrom S, Hormozdiari F, Eskin E, Price AL, et al. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genet. 2014 Oct 30;10(10):e1004722.

248. Farh KK, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature. 2015;518(7539):337-43.

249. Morris AP. Transethnic meta-analysis of genomewide association studies. Genet Epidemiol. 2011;35(8):809-22.

250. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57-74.

251. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. The NIH roadmap epigenomics mapping consortium. Nat Biotechnol. 2010;28(10):1045-8.

252. Consortium TF. A promoter-level mammalian expression atlas. Nature. 2014;507(7493):462-70.

253. 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.

254. Orr N, Dudbridge F, Dryden N, Maguire S, Novo D, Perrakis E, et al. Fine-mapping identifies two additional breast cancer susceptibility loci at 9q31.2. Hum Mol Genet. 2015 May 15;24(10):2966-84.

255. DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Asian Genetic Epidemiology Network Type 2 Diabetes (AGEN-T2D) Consortium, South Asian Type 2 Diabetes (SAT2D) Consortium, Mexican American Type 2 Diabetes (MAT2D) Consortium, Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples (T2D-GENES) Consortium. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nat Genet. 2014;46(3):234-44.

256. Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, Shen Y, et al. The support of human genetic evidence for approved drug indications. Nat Genet. 2015;47(8):856-60.

257. Price AL, Spencer CC, Donnelly P. Progress and promise in understanding the genetic basis of common diseases. Proc Biol Sci. 2015 Dec 22;282(1821):20151684.

258. Claussnitzer M, Dankel SN, Kim K, Quon G, Meuleman W, Haugen C, et al. FTO obesity variant circuitry and adipocyte browning in humans. N Engl J Med. 2015;373(10):895-907.

259. Sekar A, Bialas AR, de Rivera H, Davis A, Hammond TR, Kamitaki N, et al. Schizophrenia risk from complex variation of complement component 4. Nature. 2016;530(7589):177-83.